Complete Digital Marketing Guide Book for SEO, Social Media & Brand awareness

Definitive & Hidden Secrets of Digital Marketing to grow your business

By: Team at Publicancy

PUBLICANCY
What's your story?

First Printing, 2019

www.Publicancy.com

Feedback:

If you have any comments regarding the quality of this book, or otherwise alter it to better suit your needs, you can contact us through email at info@publicancy.com

Please make sure to include the book title and ISBN in your message

About Publicancy:

Publicancy® is a publishing and marketing platform that is registered in London, England & Wales, which unleashes the creative genius inside everyone. Publicancy makes it easy for authors to get their books designed, published, promoted, and sell professionally on worldwide scale with eBook + Print distribution. Publicancy was founded in 2018, and includes a team of design; Internet and media veterans who share a passion for helping people bring their stories to life & help them to earn passive income effortlessly

Contents

Search Engine Reputation Management

Reputation management, is the process of tracking an entity's actions and other entities' opinions about those actions; reporting on those actions and opinions; and reacting to that report creating a feedback loop. All entities involved are generally people, but that need not always be the case. Other examples of entities include animals, businesses, or even locations or materials. The tracking and reporting may range from word-of-mouth to statistical analysis of thousands of data points.

Reputation management has come into wide use with the advent of widespread computing. This is evidenced by a front page story in the Washington Post. [1][2] featuring several online reputation management firms. Reputation management systems use various predefined criteria for processing complex data to report reputation. However, these systems only facilitate and automate the process of determining trustworthiness. This process is central to all kinds of human interaction, including interpersonal relationships, international diplomacy, stock markets, communication through marketing and public relations and sports. Reputation management is also a professional communications practice – a specialization within the public relations industry. Reputation management ensures that the information about an individual, business or organization is accessible to the public online as well as through traditional outlets and is accurate, up-to-date and authentic. [3]

Real-world communities

Small town

The classic example of reputation management is the small town. Population is small and interactions between members frequent; most interactions are face-to-face and *positively identified* -- that is, there is no question who said or did what. Reputation accrues not only throughout one's lifetime, but is passed down to one's offspring; one's individual reputation depends both on one's own actions and one's inherited reputation.

There are generally few formal mechanisms to manage this *implicit reputation*. Implicit Reputation is the accumulated reputation one gets in a small town from previous actions. The town diner and barber shop serve as forums for exchange of gossip, in which community members' reputations are discussed (implicit reputation), often in frank terms. Outstanding members may receive small, symbolic awards or titles, but these are mere confirmations of general knowledge.

There is exceedingly little deviation from community norms in a small town. This may be seen as either good or bad; there is little crime, but also little room for dissent or change. The small-town model scales poorly; it depends on each member having enough experience of a large number of other members, and this is only possible up to a point.

Big city

The large metropolitan area is at the other end of the spectrum from the small rural town. Community members come and go daily, and most members are only personally acquainted with a small fraction of the whole. Implicit reputation management continues to work within subcommunities, but for the city as a whole, it *cannot*.

Big cities have developed a large **array** of formal reputation management methods. Some apply only to subcommunities, such as, say, an association of local dentists. There are four methods (among others) which apply quite generally to the entire population: **elections**, **appointments**, the **criminal justice** system, and racial or ethnic **prejudice**.

- The city is governed in part by *elected officials* -- persons who are given special powers by popular vote at regular intervals. Campaigns are often well-financed efforts to *force* a positive image of a candidate's reputation upon the electorate; television is often decisive. Elected officials are primarily concerned with preserving this good reputation, which concern dictates their every public action. Failure to preserve a good reputation, not to mention

failure to avoid a bad one, is often cause for removal from office, sometimes prematurely. Candidates and officials frequently concentrate on damaging the reputations of their opponents.

- *Appointed officials* are not elected; they are granted special powers, usually by elected officials, without public deliberation. Persons wishing to be appointed to office also campaign to increase their perceived reputation, but the audience is much smaller. Effective actions and demonstrated merit are often important factors in gaining a positive reputation, but the definition of this merit is made by the elected, appointing officials, who tend to evaluate merit as it applies to them, personally. Thus persons who work hard to increase an elected official's reputation increase their own, at least in their patron's eyes. Some appointees have no other qualification beyond the fact that they may be depended on at all times to support their patrons.
- The stresses of big city life lead to much *crime*, which demands *punishment*, on several grounds. The severity of this punishment and of the efforts of the system to inflict it upon a community member depends in no small part on that individual's prior experiences within the system. Elaborate records are kept of every infraction, even of the suspicion of infractions, and these records are consulted before any decision is made, no matter how trivial. Great effort is expended to positively identify members—driver's licenses and fingerprints, for example—and any use of an alias is carefully recorded. Some small punishments are meted out informally, but most punishments, especially severe ones, are given only after a long, detailed, and formal process: a trial, which must result in a conviction, or finding of guilt, before a punishment is ordered.

  Although it is sometimes said that serving one's punishment is sufficient penalty for the commission of a crime, in truth the damage to one's reputation may be the greater penalty -- damage both within the system itself and within other systems of urban reputation management, such as that of elections to office. Between the explicit punishment and the damage to one's reputation, the total effect of a conviction in the criminal justice system so damages a person's ability to lead a normal life that the process, at least ostensibly, is meticulous in determining guilt or lack thereof. In case of "reasonable" doubt, a suspected malefactor is freed -- though the mere fact of the trial is recorded, and affects his future reputation.
- The ordinary citizen, meeting a stranger, another citizen unknown to the first, is rarely concerned that the second may be an official, elected or otherwise; even so, he may be aware of the relative failure of reputation management in this regard. He does not have easy access to the database of the criminal justice system, and portions are not publicly available at all. Lacking other means, he often turns to the mock-system of *racial or ethnic prejudice*. This attempts to extend the small-town model to large communities by grouping individuals who look alike, dress alike, or talk alike. One reputation serves for all. Each individual is free to compose his personal measure of a group's reputation, and actions of strangers raise or lower that reputation for all group members.

The high incidence of crime, the proverbial incompetence of officials, and constant wars between rival, self-identified groups speaks poorly of **all** systems of urban reputation management. Together, they do not function as well as that of the small town, with no formal system at all.

Profession

Reputation management is also a professional communications practice - a specialization within the public relations industry. It ensures that the information about an individual, business, or organization is accessible to the public online as well as through traditional outlets and is accurate, up-to-date, and authentic. Reputation strategy is competitive strategy. Reputation initiatives drive stakeholder perceptions, which drive the likelihood of eliciting supportive behaviors and fuel business results. Leveraging reputation allows individuals or businesses to build advantage in the marketplace and reduce risk exposure. You manage a reputation by building a 'reputation platform' and by carrying out 'reputing programs'. Reputing aligns identity (what a company is), communication (what a company says), and action (what a company does). Reputing is designed to build and reinforce trusting relationships between companies and their stakeholders. [4]

Reputation management is a process that integrates business strategy, marketing, branding, organizational engagement, and communications. Through this integration, the organization creates "a formal reputation risk plan" that allows it to identify where and how it creates perceived value with each stakeholder and where there are gaps. The process manages reputation cross-functionally and at all "touch points". The research that is done focuses on key reputation metrics. Activities performed by individual or organization which attempt to maintain or create a certain frame of mind regarding themselves in the public eye. Reputation management is the process of identifying what other people are saying or feeling about a person or a business; and taking steps to ensure that the general consensus is in line with your goals. Many people and organizations use various forms of social media to monitor their reputation.[5]

Online communities

eBay

eBay is an online marketplace, a forum for the exchange of goods. The feedback system on eBay asks each user to post his opinion (positive or negative) on the person with whom he transacted. Every place a user's system handle ("ID") is displayed, his feedback is displayed with it.

Since having primarily positive feedback will improve a user's reputation and therefore make other users more comfortable in dealing with him, users are encouraged to behave in acceptable ways—that is, by dealing squarely with other users, both as buyers and as sellers.

Most users are extremely averse to negative feedback and will go to great lengths to avoid it. There is even such a thing as **feedback blackmail**, in which a party to a transaction threatens negative feedback to gain an unfair concession. The fear of getting negative feedback is so great that many users automatically leave positive feedback, with strongly worded comments, in hopes of getting the same in return. Thus, research has shown that a very large number (greater than 98%) of all transactions result in positive feedback. Academic researchers have called the entire eBay system into question based on these results.

The **main result** of the eBay reputation management system is that buyers and sellers are *generally* honest. There are abuses, but not to the extent that there might be in a completely open or unregulated marketplace.[6]

Everything2

Everything2 is a general knowledge base. E2 manages both user and article reputation strongly; one might say it is central to the project's paradigm. Users submit articles, called "writeups", that are published immediately. For each article, each user may cast one vote, positive or negative. Voting is anonymous and each vote cast is final. The article keeps track of its total of positive and negative votes (and the resulting score), all of which can be seen by the submitting user and any user who has already cast their vote on that particular article. Articles with strong positive scores may also be featured on the site's main page, propelling them to even higher scores. Articles with low or negative scores are deleted, hopefully to make way for better articles.

Users themselves are explicitly ranked, using a complicated "level" system[7] loosely based on number of articles submitted (and not deleted) and the overall average article score. Users of higher levels gain various privileges, the first being the ability to cast votes; any user may submit an article, but only users who have a minimum number of "good" articles may vote.

E2's system has a number of detrimental effects. Many new users leave the site after their first article gets multiple negative votes, and is sometimes then also deleted, all without any explanation required. Even experienced users hesitate to submit less than perfect articles since negative votes cannot be retracted. There are also more direct rewards for users submitting new articles than for revising and improving their existing ones. Finally, many users focus heavily on their position in the hierarchy and pander for positive votes. Fiction and amusing essay-style articles tend dominate over long, difficult, boring, less well-written, or controversial ones. Excellent contributions

are still rewarded, but so are merely decent ones and the difference in reward is not proportional to the additional effort.

Slashdot

Slashdot contains little original content, instead revolving around short reviews of content exterior to the site. "Karma" is Slashdot's name for reputation management. "Moderators" are able to vote on both reviews themselves and comments on those reviews in a system not too dissimilar from E2's. In a novel twist, votes are not merely "+1 point" or "-1 point"; moderators also attach one of a list of predefined labels, such as *Flamebait* or *Informative*. This change was made in June 2002 to help prevent some users from taking karma too seriously.[8]

Score is displayed next to each comment. Additionally, any user may set a personal preference to exclude the display of comments with low scores. Users acquire "karma" based, among other things, on the scores of their comments, and karma affects a user's powers. Almost any user may become a moderator, although this status is temporary; thus the average user is not able to vote on any comment. Once a moderator uses up his votes, he returns to the status of ordinary user.

Slashdot has become extremely popular and well-read; used as a verb, it refers to the fact that a website mentioned in Slashdot is often overwhelmed with visitors. There is frequent criticism of Slashdot, on many grounds; the karma system is intentionally not transparent and trolling is quite common. Anonymous cowards are permitted and range freely, as do sockpuppets.

Nonetheless, Slashdot's karma system may account for at least part of its endurance and popularity.

Meatball Wiki

Meatball is a wiki devoted to discussion of online communities, including wikis themselves. Its membership is not large. Meatball permits anonymous users, but relegates them to an inferior status: "If you choose not to introduce yourself, it's assumed you aren't there to participate in exchanging help, but just to 'hang out.'" [9]

While anonymous posters are tolerated, pseudonymous users are not. Thus online handles are supposed to mirror users' *real* names – their names in the outside world, on their birth certificates. The control on this is not rigorous – users are not required to fax in their passports in order to verify their identities – but the convention is supposed to be generally followed; at least it is not openly mocked.

Thus identified, Meatball's users' reputations are managed much as they are in the small town. That is, there is little formal management, but every user carries in his head his own "score", according to his own rating system, based on his personal evaluation of a given other user's character. This *implicit* reputation system is, of course, a part of *every* online community in which handles or names of any kind are used; but in Meatball, it is the whole.

Despite (or because of?) this lack of formal method, Meatball has discussed the problems of reputation management extensively. We will not attempt to link to every relevant page, but one might begin to explore that discussion here
[10]

Wikipedia

Wikipedia is an encyclopedia-content wiki; it includes a very wide range of topics, and exclusion of almost any topic is disputed. There is a large number of community members. Anonymous users are welcomed, and most users are pseudonymous, though many do use *real* names. As in many online communities, some users are sock puppets, although these are discouraged.

Wikipedia, like Meatball or the small town, has no formal method for managing reputation. Barnstars may be awarded for merit, but any user may make such an award. There is a hierarchy of privileges, such as in Slashdot or Everything2. As in most wikis, there is an elaborate history feature, which may be explored by any user to determine which contributions were made by which users. Any user may examine a list of another user's contributions. Edits may be discussed in a variety of forums, but there is no particular grading or rating system, either for edits or community members.

Search Engine Reputation Management

Search Engine Reputation Management (or SERM) tactics are often employed by companies and increasingly by individuals who seek to proactively shield their brands or reputations from damaging content brought to light through search engine queries. Some use these same tactics reactively, in attempts to minimize damage inflicted by inflammatory (or "flame") websites (and weblogs) launched by consumers and, as some believe, competitors.

Given the increasing popularity and development of search engines, these tactics have become more important than ever. Consumer generated media (like blogs) has amplified the public's voice, making points of view - good or bad - easily expressed.[11] This is further explained in this front page article in the Washington Post.[1]

Search Engine Reputation Management strategies include Search engine optimization (SEO) and Online Content Management. Because search engines are dynamic and in constant states of change and revision, it is essential that results are constantly monitored. This is one of the big differences between SEO and online reputation management. SEO involves making technological and content changes to a website in order to make it more friendly for search engines. Online reputation management is about controlling what information users will see when they search for information about a company or person.[12]

Social networking giant Facebook has been known to practice this form of reputation management. When they released their Polls service in Spring 2007, the popular blog TechCrunch found that it could not use competitors' names in Polls. Due largely to TechCrunch's authority in Google's algorithms, its post ranked for Facebook polls. A Facebook rep joined the comments, explained the situation and that the bugs in the old code had been updated so that it was now possible.[13]

Also until social sites like Facebook allow Google to fully spider their site then they won't really have a massive effect on reputation management results in the search engine. The only way to take advantage of such site is to make sure you make your pages public.

It is suggested that if a company website has a negative result directly below it then up to 70% of surfers will click on the negative result first rather than the company website. It is important for a company to ensure that its website gets close to the top of search results for terms relevant to its business. In one study, a number one search result attracted 50,000 monthly visitors. The number 5 result only attracted 6,000 visitors in the same time period.

References

[1] http://www.washingtonpost.com/wp-dyn/content/article/2007/07/01/AR2007070101355.html?hpid=artslot

[2] Kinzie, Susan; Ellen Nakashima (2007-07-02). "Calling In Pros to Refine Your Google Image" (http://www.washingtonpost.com/wp-dyn/content/article/2007/07/01/AR2007070101355.html?hpid=artslot). Washington Post. . Retrieved 2009-09-25.

[3] http://www.reputation-communications.com/resources/"Reputation-Communications: Resources" accessed February 7th, 2012.]

[4] "Reputation Institute: Advisory Services" accessed January 31st, 2012. (http://www.reputationinstitute.com/advisory-services/)

[5] Schreiber, Elliot. "If PR Wants to Manage Reputation, It Needs to Become More Than Messengers" PRSA Blog, March 1st,accessed 2011 January 31st 2012. (http://prsay.prsa.org/index.php/2011/03/01/pr-role-in-reputation-management)

[6] Paul Resnick, Richard Zeckhause. "Trust among strangers in internet transactions: Empirical analysis of eBay' s reputation system" (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.123.5332&rep=rep1&type=pdf), *Emerald Group Publishing Limited*, May 2, 2001, accessed May 30, 2011.

[7] http://www.everything2.com/index.pl?node=Voting%2FExperience%20System [8]
http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2005.tb00241.x/full#ss5

[9] http://www.usemod.com/cgi-bin/mb.pl?UseRealNames

[10] http://www.usemod.com/cgi-bin/mb.pl?RewardReputation

[11] Leake, William (2008-01-29). "Using SEO for Reputation Management" (http://searchenginewatch.com/showPage.html?page=3628265). *Search Engine Watch*.

[12] "Online Reputation Management (ORM) versus Search Engine Optimization (SEO)" (http://www.reputation.com/how_to/ online-reputation-management-orm-versus-search-engine-optimization-seo/). . Retrieved 30 May 2011.

[13] Riley, Duncan. "Facebook Polls: Don't Mention The Competition" (http://techcrunch.com/2007/06/02/facebook-polls-dont-mention-the-competition/). *Facebook Polls: Don't Mention The Competition*. . Retrieved 30 May 2011.

Further reading

- Monarth, Harrison (2009). *Executive Presence: The Art of Commanding Respect Like a CEO*. McGraw-Hill. ISBN 9780071632874.
- Klewes, Joachim and Wreschniok, Robert (2010). *Building and Maintaining Trust in the 21st Century*. ISBN 978-3-642-01629-5.

Semantic Web

The **Semantic Web** is a collaborative movement led by the World Wide Web Consortium (W3C) [1] that promotes common formats for data on the World Wide Web. By encouraging the inclusion of semantic content in web pages, the Semantic Web aims at converting the current web of unstructured documents into a "web of data". It builds on the W3C's Resource Description Framework (RDF).[2]

According to the W3C, "The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries."[2]

The term was coined by Tim Berners-Lee,[3] the inventor of the World Wide Web and director of the World Wide Web Consortium ("W3C"), which oversees the development of proposed Semantic Web standards. He defines the Semantic Web as "a web of data that can be processed directly and indirectly by machines."

While its critics have questioned its feasibility, proponents argue that applications in industry, biology and human sciences research have already proven the validity of the original concept.[4]

History

The concept of the *Semantic Network Model* was coined in the early sixties by the cognitive scientist Allan M. Collins, linguist M. Ross Quillian and psychologist Elizabeth F. Loftus in various publications,[5][6][7][8][8] as a form to represent semantically structured knowledge. It extends the network of hyperlinked human-readable web pages by inserting machine-readable metadata about pages and how they are related to each other, enabling automated agents to access the Web more intelligently and perform tasks on behalf of users. The term was coined by Tim Berners-Lee,[9] the inventor of the World Wide Web and director of the World Wide Web Consortium ("W3C"), which oversees the development of proposed Semantic Web standards. He defines the Semantic Web as "a web of data that can be processed directly and indirectly by machines."

Many of the technologies proposed by the W3C already existed before they were positioned under the W3C umbrella. These are used in various contexts, particularly those dealing with information that encompasses a limited and defined domain, and where sharing data is a common necessity, such as scientific research or data exchange among businesses. In addition, other technologies with similar goals have emerged, such as microformats.

Purpose

The main purpose of the Semantic Web is driving the evolution of the current Web by enabling users to find, share, and combine information more easily. Humans are capable of using the Web to carry out tasks such as finding the Irish word for "folder", reserving a library book, and searching for the lowest price for a DVD. However, machines cannot accomplish all of these tasks without human direction, because web pages are designed to be read by people, not machines. The semantic web is a vision of information that can be readily interpreted by machines, so machines can perform more of the tedious work involved in finding, combining, and acting upon information on the web.

The Semantic Web, as originally envisioned, is a system that enables machines to "understand" and respond to complex human requests based on their meaning. Such an "understanding" requires that the relevant information sources is semantically structured, a challenging task.

Tim Berners-Lee originally expressed the vision of the Semantic Web as follows:[10]

> I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A 'Semantic Web', which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The 'intelligent agents' people have touted for ages will finally materialize.

The Semantic Web is regarded as an integrator across different content, information applications and systems. It has applications in publishing, blogging, and many other areas.

Often the terms "semantics", "metadata", "ontologies" and "Semantic Web" are used inconsistently. In particular, these terms are used as everyday terminology by researchers and practitioners, spanning a vast landscape of different fields, technologies, concepts and application areas. Furthermore, there is confusion with regard to the current status of the enabling technologies envisioned to realize the Semantic Web. In a paper presented by Gerber, Barnard and Van der Merwe[11] the Semantic Web landscape is charted and a brief summary of related terms and enabling technologies is presented. The architectural model proposed by Tim Berners-Lee is used as basis to present a status model that reflects current and emerging technologies.[12]

Limitations of HTML

Many files on a typical computer can be loosely divided into human readable documents and machine readable data. Documents like mail messages, reports, and brochures are read by humans. Data, like calendars, addressbooks, playlists, and spreadsheets are presented using an application program which lets them be viewed, searched and combined in different ways.

Currently, the World Wide Web is based mainly on documents written in Hypertext Markup Language (HTML), a markup convention that is used for coding a body of text interspersed with multimedia objects such as images and interactive forms. Metadata tags provide a method by which computers can categorise the content of web pages, for example:

With HTML and a tool to render it (perhaps web browser software, perhaps another user agent), one can create and present a page that lists items for

```
<meta name="keywords" content="computing, computer studies, computer">
<meta name="description" content="Cheap widgets for sale">
<meta name="author" content="John Doe">
```

sale. The HTML of this catalog page can make simple, document-level assertions such as "this document's title is 'Widget Superstore'", but there is no capability within the HTML itself to assert unambiguously that, for example, item number X586172 is an Acme Gizmo with a retail price of €199, or that it is a consumer product. Rather, HTML can only say that the span of text "X586172" is something that should be positioned near "Acme Gizmo" and "€199", etc. There is no way to say "this is a catalog" or even to establish that "Acme Gizmo" is a kind of title or that "€199" is a price. There is also no way to express that these pieces of information are bound together in describing a discrete item, distinct from other items perhaps listed on the page.

Semantic HTML refers to the traditional HTML practice of markup following intention, rather than specifying layout details directly. For example, the use of `<em>` denoting "emphasis" rather than `<i>`, which specifies italics. Layout details are left up to the browser, in combination with Cascading Style Sheets. But this practice falls short of specifying the semantics of objects such as items for sale or prices.

Microformats represent unofficial attempts to extend HTML syntax to create machine-readable semantic markup about objects such as retail stores and items for sale.

Semantic Web solutions

The Semantic Web takes the solution further. It involves publishing in languages specifically designed for data: Resource Description Framework (RDF), Web Ontology Language (OWL), and Extensible Markup Language (XML). HTML describes documents and the links between them. RDF, OWL, and XML, by contrast, can describe arbitrary things such as people, meetings, or airplane parts.

These technologies are combined in order to provide descriptions that supplement or replace the content of Web documents. Thus, content may manifest itself as descriptive data stored in Web-accessible databases,[13] or as markup within documents (particularly, in Extensible HTML (XHTML) interspersed with XML, or, more often, purely in XML, with layout or rendering cues stored separately). The machine-readable descriptions enable content managers to add meaning to the content, i.e., to describe the structure of the knowledge we have about that content. In this way, a machine can process knowledge itself, instead of text, using processes similar to human deductive reasoning and inference, thereby obtaining more meaningful results and helping computers to perform automated information gathering and research.

An example of a tag that would be used in a non-semantic web page:

```
<item>cat</item>
```

Encoding similar information in a semantic web page might look like this:

```
<item rdf:about="http://dbpedia.org/resource/Cat">Cat</item>
```

Tim Berners-Lee calls the resulting network of Linked Data the Giant Global Graph, in contrast to the HTML-based World Wide Web. Berners-Lee posits that if the past was document sharing, the future is data sharing. His answer to the question of "how" provides three points of instruction. One, a URL should point to the data. Two, anyone accessing the URL should get data back. Three, relationships in the data should point to additional URLs with data.

Web 3.0

Tim Berners-Lee has described the semantic web as a component of 'Web 3.0'.[14]

> People keep asking what Web 3.0 is. I think maybe when you've got an overlay of scalable vector graphics – everything rippling and folding and looking misty — on Web 2.0 and access to a semantic Web integrated across a huge space of data, you'll have access to an unbelievable data resource..."
>
> — Tim Berners-Lee, 2006

"Semantic Web" is sometimes used as a synonym for "Web 3.0", though each term's definition varies.

Examples

When we talk about the Semantic Web, we speak about many "howto's" which are often incomprehensible because the required notions of linguistics are very often ignored by most people. Thus, we rather imagine how emergence of the Semantic Web looks in the future.

Meta-Wiki

The sites of Wiki type soar. Their administrations and their objectives can be very different. These wikis are more and more specialized. But most of wikis limit the search engines to index them because these search engines decrease the wikis' efficiency and record pages which are obsolete, by definition, outside the wiki (perpetual update). Meta- search-engines are going to aggregate the obtained result by requesting individually at each of these wikis. The wikis become silos of available data for consultation by people and machines through access points (triplestore).

Semantic detectives & Semantic identity

The young bloggers are now on the labour market. The companies do not ask any longer for the judicial file of a new employee. To have information, the companies appeal in a systematic way to engines which are going to interrogate all the sites which reference and index the accessible information on the Web. The differentiation between search engines is going to concern the capacity to respond at requests where the sense is going to take more and more importance (evolution of the requests with keywords towards the semantic requests). There will be three types of person: the unknown, the "without splash" and the others. The others will have to erase in a systematic way the information which could carry disadvantages and which will be more and more accessible. It will be the same engines of semantic search which also charge this service.

Profile Privacy/Consumer/Public

The Web's children became parents. They use tools which can limit the access and the spreading of the information by their children. So, the parents can see at any time the web's logs of their children but they also have a net which is going to filter their "private" identity before it is broadcasted on the network. For example, a third-part trust entity, along with their mobile telephone provider, the post office and the bank, will possess the consumer's identity so as to mask the address of delivery and the payment of this consumer. A public identity also exists to spread a resume (CV), a blog or an avatar for example but the data remain the property of the owner of the server who hosts this data. So, the mobile telephone provider offers a personal server who will contain one public zone who will automatically be copied on the network after every modification. If I want that my resume is not any longer on the network, I just have to erase it of my public zone from my server. So, the mobile telephone provider creates a controllable silo of information for every public profile.

Personal agent

In a few years, the last generation of robot is now mobile and transcribes the human voice. However, it has to transfer the semantic interpretation to more powerful computers. These servers can so interpret the sense of simple sentences and interrogate other servers to calculate the answer to be given. Example: "Arthur returned at him. He ordered a pizza by his personal digital agent. His agent is going to send the information to the home server which will accept or not the purchase. It refuses because it received the order of the Arthur's parents to buy only a well-balanced menu. So, the home server displays on the TV3D the authorized menus to allow Arthur to choose a new meal."

Research assistant

In 20??, the Semantic Web is now a reality. Marc is a researcher. He has a new idea. He is going to clarify it with his digital assistant which is immediately going to show him the incoherence of his demonstration by using the accessible knowledge in silos on the Web. Marc will be able to modify his reasoning or to find the proofs which demonstrate that the existing knowledge is false and so to advance the scientific knowledge within the Semantic Web.

Challenges

Some of the challenges for the Semantic Web include vastness, vagueness, uncertainty, inconsistency, and deceit. Automated reasoning systems will have to deal with all of these issues in order to deliver on the promise of the Semantic Web.

- Vastness: The World Wide Web contains many billions of pages[15]. The SNOMED CT medical terminology ontology alone contains 370,000 class names, and existing technology has not yet been able to eliminate all semantically duplicated terms. Any automated reasoning system will have to deal with truly huge inputs.

- Vagueness: These are imprecise concepts like "young" or "tall". This arises from the vagueness of user queries, of concepts represented by content providers, of matching query terms to provider terms and of trying to combine different knowledge bases with overlapping but subtly different concepts. Fuzzy logic is the most common technique for dealing with vagueness.
- Uncertainty: These are precise concepts with uncertain values. For example, a patient might present a set of symptoms which correspond to a number of different distinct diagnoses each with a different probability. Probabilistic reasoning techniques are generally employed to address uncertainty.
- Inconsistency: These are logical contradictions which will inevitably arise during the development of large ontologies, and when ontologies from separate sources are combined. Deductive reasoning fails catastrophically when faced with inconsistency, because "anything follows from a contradiction". Defeasible reasoning and paraconsistent reasoning are two techniques which can be employed to deal with inconsistency.
- Deceit: This is when the producer of the information is intentionally misleading the consumer of the information. Cryptography techniques are currently utilized to alleviate this threat.

This list of challenges is illustrative rather than exhaustive, and it focuses on the challenges to the "unifying logic" and "proof" layers of the Semantic Web. The World Wide Web Consortium (W3C) Incubator Group for Uncertainty Reasoning for the World Wide Web (URW3-XG) final report [16] lumps these problems together under the single heading of "uncertainty". Many of the techniques mentioned here will require extensions to the Web Ontology Language (OWL) for example to annotate conditional probabilities. This is an area of active research.[17]

Standards

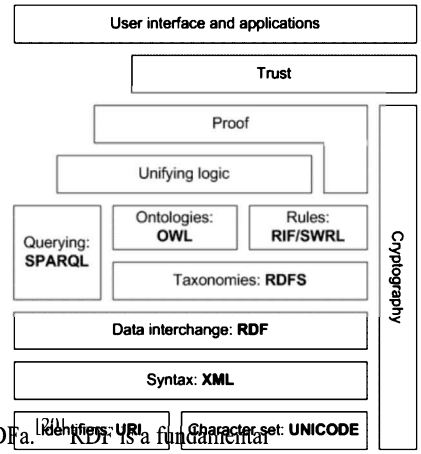
Standardization for Semantic Web in the context of Web 3.0 is under the care of W3C.[18]

Components

The term "Semantic Web" is often used more specifically to refer to the formats and technologies that enable it.[2] The collection, structuring and recovery of linked data are enabled by technologies that provide a formal description of concepts, terms, and relationships within a given knowledge domain. These technologies are specified as W3C standards and include:

- Resource Description Framework (RDF), a general method for describing information
- RDF Schema (RDFS)
- Simple Knowledge Organization System (SKOS)
- SPARQL, an RDF query language
- Notation3 (N3), designed with human-readability in mind
- N-Triples, a format for storing and transmitting data
- Turtle (Terse RDF Triple Language)
- Web Ontology Language (OWL), a family of knowledge representation languages

The Semantic Web Stack illustrates the architecture of the Semantic Web. The functions and relationships of the components can be summarized as follows:[19]



The Semantic Web Stack.

- XML provides an elemental syntax for content structure within documents, yet associates no semantics with the meaning of the content contained within. XML is not at present a necessary component of Semantic Web technologies in most cases, as alternative syntaxes exists, such as Turtle. Turtle is a de facto standard, but has not been through a formal standardization process.
- XML Schema is a language for providing and restricting the structure and content of elements contained within XML documents.
- RDF is a simple language for expressing data models, which refer to objects ("resources") and their relationships. An RDF-based model can be represented in a variety of syntaxes, e.g., RDF/XML, N3, Turtle, and RDFa. RDF is a fundamental standard of the Semantic Web.[21][22][23]
- RDF Schema extends RDF and is a vocabulary for describing properties and classes of RDF-based resources, with semantics for generalized-hierarchies of such properties and classes.
- OWL adds more vocabulary for describing properties and classes: among others, relations between classes (e.g. disjointness), cardinality (e.g. "exactly one"), equality, richer typing of properties, characteristics of properties (e.g. symmetry), and enumerated classes.
- SPARQL is a protocol and query language for semantic web data sources.

Current state of standardization

Current ongoing standardizations include:

- Rule Interchange Format (RIF) as the Rule Layer of the Semantic Web Stack Not yet fully

realized layers include:

- Unifying Logic and Proof layers are undergoing active research.

The intent is to enhance the usability and usefulness of the Web and its interconnected resources through:

- Servers which expose existing data systems using the RDF and SPARQL standards. Many converters to RDF [24] exist from different applications. Relational databases are an important source. The semantic web server attaches to the existing system without affecting its operation.
- Documents "marked up" with semantic information (an extension of the HTML `<meta>` tags used in today's Web pages to supply information for Web search engines using web crawlers). This could be machine-understandable information about the human-understandable content of the document (such as the creator, title, description, etc., of the document) or it could be purely metadata representing a set of facts (such as resources and services elsewhere in the site). (Note that *anything* that can be identified with a *Uniform Resource Identifier* (URI) can be described, so the semantic web can reason about animals, people, places, ideas, etc.) Semantic markup is often generated automatically, rather than manually.
- Common metadata vocabularies (ontologies) and maps between vocabularies that allow document creators to know how to mark up their documents so that agents can use the information in the supplied metadata (so that *Author* in the sense of 'the Author of the page' won't be confused with *Author* in the sense of a book that is the subject of a book review).
- Automated agents to perform tasks for users of the semantic web using this data

- Web-based services (often with agents of their own) to supply information specifically to agents (for example, a Trust service that an agent could ask if some online store has a history of poor service or spamming)

Skeptical reactions

Practical feasibility

Critics (e.g. Which Semantic Web? [25]) question the basic feasibility of a complete or even partial fulfillment of the semantic web. Cory Doctorow's critique ("metacrap") is from the perspective of human behavior and personal preferences. For example, people may include spurious metadata into Web pages in an attempt to mislead Semantic Web engines that naively assume the metadata's veracity. This phenomenon was well-known with metatags that fooled the AltaVista ranking algorithm into elevating the ranking of certain Web pages: the Google indexing engine specifically looks for such attempts at manipulation. Peter Gärdenfors and Timo Honkela point out that logic-based semantic web technologies cover only a fraction of the relevant phenomena related to semantics.[26][27]

Where semantic web technologies have found a greater degree of practical adoption, it has tended to be among core specialized communities and organizations for intra-company projects.[28] The practical constraints toward adoption have appeared less challenging where domain and scope is more limited than that of the general public and the World-Wide Web.[28]

Potential of an idea in fast progress

The original 2001 Scientific American article by Berners-Lee described an expected evolution of the existing Web to a Semantic Web.[29] A complete evolution as described by Berners-Lee has yet to occur. In 2006, Berners-Lee and colleagues stated that: "This simple idea, however, remains largely unrealized."[30] While the idea is still in the making, it seems to evolve quickly and inspire many. Between 2007–2010 several scholars have explored the social potential of the semantic web in the business and health sectors, and for social networking.[31] They have also explored the broader evolution of democracy: how a society forms its common will in a democratic manner through a semantic web.[32]

Censorship and privacy

Enthusiasm about the semantic web could be tempered by concerns regarding censorship and privacy. For instance, text-analyzing techniques can now be easily bypassed by using other words, metaphors for instance, or by using images in place of words. An advanced implementation of the semantic web would make it much easier for governments to control the viewing and creation of online information, as this information would be much easier for an automated content-blocking machine to understand. In addition, the issue has also been raised that, with the use of FOAF files and geo location meta-data, there would be very little anonymity associated with the authorship of articles on things such as a personal blog. Some of these concerns were addressed in the "Policy Aware Web" project[33] and is an active research and development topic.

Doubling output formats

Another criticism of the semantic web is that it would be much more time-consuming to create and publish content because there would need to be two formats for one piece of data: one for human viewing and one for machines. However, many web applications in development are addressing this issue by creating a machine-readable format upon the publishing of data or the request of a machine for such data. The development of microformats has been one reaction to this kind of criticism. Another argument in defense of the feasibility of semantic web is the likely falling price of human intelligence tasks in digital labor markets, such as the Amazon Mechanical Turk.

Specifications such as eRDF and RDFa allow arbitrary RDF data to be embedded in HTML pages. The GRDDL (Gleaning Resource Descriptions from Dialects of Language) mechanism allows existing material (including

microformats) to be automatically interpreted as RDF, so publishers only need to use a single format, such as HTML.

Projects

This section lists some of the many projects and tools that exist to create Semantic Web solutions.[34]

DBpedia

DBpedia is an effort to publish structured data extracted from Wikipedia: the data is published in RDF and made available on the Web for use under the GNU Free Documentation License, thus allowing Semantic Web agents to provide inferencing and advanced querying over the Wikipedia-derived dataset and facilitating interlinking, re-use and extension in otherdata-sources.

FOAF

A popular vocabulary on the semantic web is Friend of a Friend (or FoaF), which uses RDF to describe the relationships people have to other people and the "things" around them. FOAF permits intelligent agents to make sense of the thousands of connections people have with each other, their jobs and the items important to their lives; connections that may or may not be enumerated in searches using traditional web search engines. Because the connections are so vast in number, human interpretation of the information may not be the best way of analyzing them.

FOAF is an example of how the Semantic Web attempts to make use of the relationships within a social context.

SIOC

The Semantically-Interlinked Online Communities project (SIOC, pronounced "shock") provides a vocabulary of terms and relationships that model web data spaces. Examples of such data spaces include, among others: discussion forums, blogs, blogrolls / feed subscriptions, mailing lists, shared bookmarks and image galleries.

NextBio

A database consolidating high-throughput life sciences experimental data tagged and connected via biomedical ontologies. Nextbio is accessible via a search engine interface. Researchers can contribute their findings for incorporation to the database. The database currently supports gene or protein expression data and sequence centric data and is steadily expanding to support other biological data types.

References

[1] "XML and Semantic Web W3C Standards Timeline" (http://www.dblab.ntua.gr/~bikakis/XML and Semantic Web W3C Standards Timeline-History.pdf). 2012-02-04. .

[2] "W3C Semantic Web Activity" (http://www.w3.org/2001/sw/). World Wide Web Consortium (W3C). November 7, 2011. . Retrieved November 26, 2011.

[3] Berners-Lee, Tim; James Hendler and Ora Lassila (May 17, 2001). "The Semantic Web" (http://www.sciam.com/article. cfm?id=the-semantic-web&print=true). *Scientific American Magazine*. . Retrieved March 26, 2008.

[4] Klyne, Graham (February 26, 2004). "Semantic Web Applications" (http://www.ninebynine.net/Papers/SemanticWebApplications.pdf). Nine by Nine. . Retrieved November 26, 2011.

[5] Allan M. Collins, A; M.R. Quillian (1969). "Retrieval time from semantic memory". *Journal of verbal learning and verbal behavior* **8** (2): 240–247. doi:10.1016/S0022-5371(69)80069-1.

[6] Allan M. Collins, A; M. Ross Quillian (1970). "Does category size affect categorization time?". *Journal of verbal learning and verbal behavior* **9** (4): 432–438. doi:10.1016/S0022-5371(70)80084-6.

[7] Allan M. Collins, Allan M.; Elizabeth F. Loftus (1975). "A spreading-activation theory of semantic processing". *Psychological Review* **82** (6): 407–428. doi:10.1037/0033-295X.82.6.407.

[8] Quillian, MR (1967). "Word concepts. A theory and simulation of some basic semantic capabilities". *Behavioral Science* **12** (5): 410–430. doi:10.1002/bs.3830120511. PMID 6059773.

[9] Berners-Lee, Tim; James Hendler and Ora Lassila (May 17, 2001). "The Semantic Web" (http://www.sciam.com/article. cfm?id=the-semantic-web&print=true). *Scientific American Magazine*. . Retrieved March 26, 2008.

[10] Berners-Lee, Tim; Fischetti, Mark (1999). *Weaving the Web*. HarperSanFrancisco. chapter 12. ISBN 978-0-06-251587-2.

[11] Gerber, AJ, Barnard, A & Van der Merwe, Alta (2006), A Semantic Web Status Model, Integrated Design & Process Technology, Special Issue: IDPT 2006

[12] Gerber, Aurona; Van der Merwe, Alta; Barnard, Andries; (2008), A Functional Semantic Web architecture, European Semantic Web Conference 2008, ESWC'08, Tenerife, June 2008.

[13] Artem Chebotko and Shiyong Lu, "Querying the Semantic Web: An Efficient Approach Using Relational Databases", LAP Lambert Academic Publishing, ISBN 978-3-8383-0264-5, 2009.

[14] Victoria Shannon (June 26, 2006). "A 'more revolutionary' Web" (http://www.iht.com/articles/2006/05/23/business/web.php). *International Herald Tribune*. . Retrieved May 24, 2006.

[15] http://www.worldwidewebsize.com/

[16] http://www.w3.org/2005/Incubator/urw3/XGR-urw3-20080331/

[17] Lukasiewicz, Thomas; Umberto Straccia. "Managing uncertainty and vagueness in description logics for the Semantic Web" (http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B758F-4SPSPKW-1&_user=147018&_coverDate=11/30/2008&_rdoc=1&_fmt=&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000012179&_version=1&_urlVersion=0&_userid=147018&md5=8123c273189b1148cadb12f95b87a5ef). .

[18] Semantic Web Standards published by the W3C (http://www.w3.org/2001/sw/wiki/Main_Page)

[19] "OWL Web Ontology Language Overview" (http://www.w3.org/TR/owl-features/). World Wide Web Consortium (W3C). February 10, 2004. . Retrieved November 26, 2011.

[20] "RDF tutorial" (http://www.lesliesikos.com/tutorials/rdf/). Dr. Leslie Sikos. . Retrieved 2011-07-05.

[21] "Resource Description Framework (RDF)" (http://www.w3.org/RDF/). World Wide Web Consortium. .

[22] "Standard websites" (http://www.lesliesikos.com/). Dr. Leslie Sikos. . Retrieved 2011-07-05.

[23] Allemang, D., Hendler, J. (2011). "RDF—The basis of the Semantic Web. In: Semantic Web for the Working Ontologist (2nd Ed.)". Morgan Kaufmann. doi:10.1016/B978-0-12-385965-5.10003-2.

[24] http://esw.w3.org/topic/ConverterToRdf

[25] http://portal.acm.org/citation.cfm?id=900051.900063&coll=ACM&dl=ACM&CFID=29933182&CFTOKEN=24611642

[26] Gärdenfors, Peter (2004). *How to make the Semantic Web more semantic*. IOS Press. pp. 17–34.

[27] Timo Honkela, Ville Könönen, Tiina Lindh-Knuutila and Mari-Sanna Paukkeri (2008). "Simulating processes of concept formation and communication" (http://www.informaworld.com/smpp/content~content=a903999101). *Journal of Economic Methodology*. .

[28] Ivan Herman (2007). "State of the Semantic Web" (http://www.w3.org/2007/Talks/0424-Stavanger-IH/Slides.pdf). *Semantic Days 2007*. . Retrieved July 26, 2007.

[29] Berners-Lee, Tim (May 1, 2001). "The Semantic Web" (http://www.sciam.com/article. cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21). Scientific American. . Retrieved March 13, 2008.

[30] Nigel Shadbolt, Wendy Hall, Tim Berners-Lee (2006). "The Semantic Web Revisited" (http://eprints.ecs.soton.ac.uk/12614/1/Semantic_Web_Revisted.pdf). *IEEE Intelligent Systems*. . Retrieved April 13, 2007.

[31] Lee Feigenbaum (May 1, 2007). "The Semantic Web in Action" (http://www.thefigtrees.net/lee/sw/sciam/semantic-web-in-action). Scientific American. . Retrieved February 24, 2010.

[32] Martin Hilbert (April, 2009). "The Maturing Concept of E-Democracy: From E-Voting and Online Consultations to Democratic Value Out of Jumbled Online Chatter" (http://www.informaworld.com/smpp/content~db=all~content=a911066517). Journal of Information Technology and Politics. . Retrieved February 24, 2010.

[33] http://policyawareweb.org/

[34] See, for instance: Bergman, Michael K. "Sweet Tools" (http://www.mkbergman.com/?page_id=325). *AI3; Adaptive Information, Adaptive Innovation, Adaptive Infrastructure*. . Retrieved January 5, 2009.

- Roger Chaffin: *The concept of a semantic Relation.* In: Adrienne Lehreru. a. (Hrsg.): *Frames, Fields and contrasts. New essays in semantic and lexical organisation*, Erlbaum, Hillsdale, N.J. 1992, ISBN 0-8058-1089-7, S. 253–288.
- Hermann Helbig: *Die semantische Struktur natürlicher Sprache. Wissenspräsentation mit MultiNet*, Springer, Heidelberg 2001, ISBN 3-540-67784-4.
- M. Ross Quillian: *Word concepts. A theory and simulation of some basic semantic capabilities.* In: *Behavioral Science* 12 (1967), S. 410–430.
- M. Ross Quillian: *Semantic memory.* In: Marvin Minsky (Hrsg.): *Semantic information processing*, MIT Press, Cambridge, Mass. 1988.

- KlausReichenberger:*Kompendium semantische Netze: Konzepte, Technologie, Modellierung*,Springer, Heidelberg 2010, ISBN 3-642-04314-3.
- JohnF.Sowa:*Principles of semantic networks. Explorations in the representation of knowledge*,Morgan Kaufmann, San Mateo, Cal. 1991, ISBN 1-55860-088-4.

Further reading

- Liyang Yu (January 6, 2011). *A Developer's Guide to the Semantic Web* (http://www.amazon.com/ Developers-Guide-Semantic-Web/dp/3642159699/ref=sr_1_1?ie=UTF8&qid=1321027111&sr=8-1). Springer. ISBN 978-3642159695.
- GrigorisAntoniou,FrankvanHarmelen(March31,2008).*A Semantic Web Primer, 2nd Edition* (http://www. amazon.com/Semantic-Primer-Cooperative-Information-Systems/dp/0262012421/). The MIT Press. ISBN 0-262-01242-1.
- DeanAllemang,JamesHendler(May9,2008).*Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL* (http://www.amazon.com/Semantic-Web-Working-Ontologist-Effective/dp/0123735564/). Morgan Kaufmann. ISBN 978-0-12-373556-0.
- John Davies (July 11, 2006). *Semantic Web Technologies: Trends and Research in Ontology-based Systems* (http://www.amazon.com/Semantic-Web-Technologies-Research-Ontology-based/dp/0470025964/). Wiley. ISBN 0-470-02596-4.
- Pascal Hitzler, Markus Krötzsch, Sebastian Rudolph (August 25, 2009). *Foundations of Semantic Web Technologies* (http://www.semantic-web-book.org). CRC Press. ISBN 1-4200-9050-X.
- ThomasB.Passin(March1,2004).*Explorer's Guide to the Semantic Web* (http://www.amazon.com/ Explorers-Guide-Semantic-Thomas-Passin/dp/1932394206/). Manning Publications. ISBN 1-932394-20-6.
- LiyangYu(June14,2007).*Introduction to Semantic Web and Semantic Web Services* (http://www.amazon. com/Introduction-Semantic-Web-Services/dp/1584889330/). CRC Press. ISBN 1-58488-933-0.
- Jeffrey T. Pollock (March 23, 2009). *Semantic Web For Dummies* (http://www.amazon.com/gp/product/ 0470396792). For Dummies. ISBN 0-470-39679-2.
- Martin Hilbert (April, 2009). *The Maturing Concept of E-Democracy: From E-Voting and Online Consultations to Democratic Value Out of Jumbled Online Chatter* (http://www.informaworld.com/smpp/ content~db=all~content=a911066517). Journal of Information Technology & Politics. ISBN 1-68080-271-5242.
- Tim Berners-Lee Gives the Web a New Definition (http://computemagazine.com/ man-who-invented-world-wide-web-gives-new-definition/)

External links

- Official website (http://semanticweb.org)
- links collection (http://www.semanticoverflow.com/questions/1/where-can-i-learn-about-the-semantic-web) on Semantic Overflow (http://semanticoverflow.com)
- Semantic Technology and the Enterprise (http://www.semanticarts.com)
- SSWAP: Simple Semantic Web Architecture and Protocol (http://sswap.info)
- How Stuff Works: The Semantic Web (http://www.howstuffworks.com/semantic-web.htm)
- The Semantic Web Journal (http://www.semantic-web-journal.net:)

Microformat

A **microformat** (sometimes abbreviated **µF**) is a web-based approach to semantic markup which seeks to re-use existing HTML/XHTML tags to convey metadata[1] and other attributes in web pages and other contexts that support (X)HTML, such as RSS. This approach allows software to process information intended for end-users (such as contact information, geographic coordinates, calendar events, and the like) automatically.

Although the content of web pages is technically already capable of "automated processing", and has been since the inception of the web, such processing is difficult because the traditional markup tags used to display information on the web do not describe what the information means.[2] Microformats can bridge this gap by attaching semantics, and thereby obviate other, more complicated, methods of automated processing, such as natural language processing or screen scraping. The use, adoption and processing of microformats enables data items to be indexed, searched for, saved or cross-referenced, so that information can be reused or combined.[2]

As of 2010, microformats allow the encoding and extraction of events, contact information, social relationships and so on. While more are still being developed, it appears that other formats such as schema.org have achieved greater industry support.[3]

Background

Microformats emerged as part of a grassroots movement to make recognizable data items (such as events, contact details or geographical locations) capable of automated processing by software, as well as directly readable by end-users.[2][4] Link-based microformats emerged first. These include vote links that express opinions of the linked page, which search engines can tally into instant polls.[5]

As the microformats community grew, CommerceNet, a nonprofit organization that promotes electronic commerce on the Internet, helped sponsor and promote the technology and support the microformats community in various ways.[5] CommerceNet also helped co-found the Microformats.org community site.[5]

Neither CommerceNet nor Microformats.org operates as a standards body. The microformats community functions through an open wiki, a mailing list, and an Internet relay chat (IRC) channel.[5] Most of the existing microformats were created at the Microformats.org wiki and the associated mailing list, by a process of gathering examples of web publishing behaviour, then codifying it. Some other microformats (such as rel=nofollow and unAPI) have been proposed, or developed, elsewhere.

Technical overview

XHTML and HTML standards allow for the embedding and encoding of semantics within the attributes of markup tags. Microformats take advantage of these standards by indicating the presence of metadata using the following attributes:

class

: Classname

rel

: relationship, description of the target address in an anchor-element (`<a href=... rel=...>...</a>`)

rev

: reverse relationship, description of the referenced document (in one case, otherwise deprecated in microformats[6])

For example, in the text "The birds roosted at 52.48,-1.89" is a pair of numbers which may be understood, from their context, to be a set of geographic coordinates. With wrapping in spans (or other HTML elements) with specific class

names (in this case `geo`, `latitude` and `longitude`, all part of the geo microformat specification):

```
The birds roosted at
   <span class="geo">
     <span class="latitude">52.48</span>,
     <span class="longitude">-1.89</span>
   </span>
```

software agents can recognize exactly what each value represents and can then perform a variety of tasks such as indexing, locating it on a map and exporting it to a GPS device.

Example

In this example, the contact information is presented as follows:

```
<ul>
  <li>Joe Doe</li>
  <li>The Example Company</li>
  <li>604-555-1234</li>
  <li><a href="http://example.com/">http://example.com/</a></li>
</ul>
```

With hCard microformat markup, that becomes:

```
<ul class="vcard">
  <li class="fn">Joe Doe</li>
  <li class="org">The Example Company</li>
  <li class="tel">604-555-1234</li>
  <li><a class="url" href="http://example.com/">http://example.com/</a></li>
</ul>
```

Here, the formatted name (*fn*), organisation (*org*), telephone number (*tel*) and web address (*url*) have been identified using specific class names and the whole thing is wrapped in `class="vcard"`, which indicates that the other classes form an hCard (short for "HTML vCard") and are not merely coincidentally named. Other, optional, hCard classes also exist. Software, such as browser plug-ins, can now extract the information, and transfer it to other applications, such as an address book.

In-context examples

For annotated examples of microformats on live pages, see HCard#Live example and Geo (microformat)#Three_classes.

Specific microformats

Several microformats have been developed to enable semantic markup of particular types of information.

- hAtom – for marking up Atom feeds from within standard HTML
- hCalendar – for events
- hCard – for contact information; includes:
  - adr – for postal addresses
  - geo – for geographical coordinates (latitude, longitude)
- hMedia - for audio/video content [7][8]
- hNews - for news content
- hProduct – for products
- hRecipe - for recipes and foodstuffs.
- hResume – for resumes or CVs
- hReview – for reviews
- rel-directory – for distributed directory creation and inclusion[9]
- rel-enclosure – for multimedia attachments to web pages[10]
- rel-license – specification of copyright license[11]
- rel-nofollow, an attempt to discourage third-party content spam (e.g. spam in blogs)
- rel-tag – for decentralized tagging (Folksonomy)[12]
- xFolk – for tagged links
- XHTML Friends Network (XFN) – for social relationships
- XOXO – for lists and outlines

Microformats under development

Among the many proposed microformats,[13] the following are undergoing active development:

- hAudio – for audio files and references to released recordings
- citation – for citing references
- currency – for amounts of money
- figure – for associating captions with images[14]
- geo extensions – for places on Mars, the Moon, and other such bodies; for altitude; and for collections of waypoints marking routes or boundaries
- species – for the names of living things (already used by Wikipedia[15] and the BBC Wildlife Finder)
- measure – for physical quantities, structured data-values[16]

Uses of microformats

Using microformats within HTML code provides additional formatting and semantic data that applications can use. For example, applications such as web crawlers can collect data about on-line resources, or desktop applications such as e-mail clients or scheduling software can compile details. The use of microformats can also facilitate "mash ups" such as exporting all of the geographical locations on a web page into (for example) Google Maps to visualize them spatially.

Several browser extensions, such as Operator for Firefox and Oomph for Internet Explorer, provide the ability to detect microformats within an HTML document. When hCard or hCalendar are involved, such browser extensions allow to export them into formats compatible with contact management and calendar utilities, such as Microsoft Outlook. When dealing with geographical coordinates, they allow to send the location to maps applications such as Google Maps. Yahoo! Query Language can be used to extract microformats from web pages.[17] On 12 May 2009, Google announced that they would be parsing the hCard, hReview and hProduct microformats, and using them to populate search result pages[18]. They have since extended this to use hCalendar for events[19] and hRecipe for cookery recipes[19]. Similarly, microformats are also consumed by Bing[20] and Yahoo![21]. Together, these are the world's top three search engines.[22]

Microsoft expressed a desire to incorporate Microformats into upcoming projects;[23] as have other software companies.

Alex Faaborg summarizes the arguments for putting the responsibility for microformat user interfaces in the web browser rather than making more complicated HTML:[24]

- Only the web browser knows what applications are accessible to the user and what the user's preferences are
- It lowers the barrier to entry for website developers if they only need to do the markup and not handle "appearance" or "action" issues
- Retains backwards compatibility with web browsers that don't support microformats
- The web browser presents a single point of entry from the web to the user's computer, which simplifies security issues

Evaluation of microformats

Various commentators have offered review and discussion on the design principles and practical aspects of microformats. Additionally, microformats have been compared to other approaches that seek to serve the same or similar purpose.[25] From time to time, there is criticism of a single, or all, microformats.[25] Documented efforts to advocate both the spread and use of microformats are known to exist as well.[26][27] Opera Software CTO and CSS creator Håkon Wium Lie said in 2005 "We will also see a bunch of microformats being developed, and that's how the semantic web will be built, I believe."[28] However, as of August 2008, Toby Inkster, author of the "Swignition" (formerly "Cognition") microformat parsing service pointed out that no new microformat specifications had been published for over three years.[29]

Design principles

Computer scientist and entrepreneur, Rohit Khare stated that *reduce, reuse, and recycle* is "shorthand for several design principles" that motivated the development and practices behind microformats.[5]:71-72 These aspects can be summarized as follows:

- Reduce: favor the simplest solutions and focus attention on specific problems;
- Reuse: work from experience and favor examples of current practice;
- Recycle: encourage modularity and the ability to embed, valid XHTML can be reused in blog posts, RSS feeds, and anywhere else you can access the web.[5]

Accessibility

Because some microformats make use of title attribute of HTML's `abbr` element to conceal machine-readable data (particularly date-times and geographical coordinates) in the "abbr design pattern [30]", the plain text content of the element is inaccessible to those screen readers that expand abbreviations.[31] In June 2008, the BBC announced that it would be dropping use of microformats using the `abbr` design pattern because of accessibility concerns.[32]

Comparison with alternative approaches

Microformats are not the only solution for providing "more intelligent data" on the web. Alternative approaches exist and are under development as well. For example, the use of XML markup and standards of the Semantic Web are cited as alternative approaches.[5] Some contrast these with microformats in that they do not necessarily coincide with the design principles of "reduce, reuse, and recycle", at least not to the same extent.[5]

One advocate of microformats, Tantek Çelik, characterized a problem with alternative approaches:

> Here's a new language we want you to learn, and now you need to output these additional files on your server. It's a hassle. (Microformats) lower the barrier to entry.[2]

For some applications the use of other approaches may be valid. If one wishes to use microformat-style embedding but the type of data one wishes to embed does not map to an existing microformat, one can use RDFa to embed arbitrary vocabularies into HTML, for example: embedding domain-specific scientific data on the Web like zoological or chemical data where no microformat for such data exists. Furthermore, standards such as W3C's GRDDL allow microformats to be converted into data compatible with the Semantic Web.[33]

Another advocate of microformats, Ryan King, put the compatibility of microformats with other approaches this way:

> Microformats provide an easy way for many people to contribute semantic data to the web. With GRDDL all of that data is made available for RDF Semantic Web tools. Microformats and GRDDL can work together to build a better web.[33]

Notes

[1] "Class Names Across All Microformats" (http://microformats.org/wiki/existing-classes). *Microformats.org*. 2007-09-23.. Retrieved 2008-09-06.

[2] "What's the Next Big Thing on the Web? It May Be a Small, Simple Thing -- Microformats" (http://knowledge.wharton.upenn.edu/index.cfm?fa=printArticle&ID=1247). *Knowledge@Wharton*. Wharton School of the University of Pennsylvania. 2005-07-27..

[3] Skalbeck, Roger (2012). "Top 10 Law School Home Pages of 2011" (http://ssrn.com/abstract=2001967) (in English) (PDF). *Journal of Law* (Georgetown Public Law and Legal Theory Research Paper) **2** (1): 25-52. Archived from the original (http://www.icebox500.com/ wp-content/uploads/2012/02/Top-10-Law-School-Home-Pages-of-2011.pdf) on 2012-02-10. . Retrieved 2012-02-15. "This is similar to Microformats, but it seems to have more industry backing."

[4] In this context, the definition of "end-user" includes a person reading a web page on a computer screen or mobile device, or an assistive technology software program such as a screen reader.

[5] Khare, Rohit (January/February 2006). "Microformats: The Next (Small) Thing on the Semantic Web?" (http://csdl2.computer.org/persagen/DLAbsToc.jsp?resourcePath=/dl/mags/ic/&toc=comp/mags/ic/2006/01/w1toc.xml&DOI=10.1109/MIC.2006.13). *IEEE Internet Computing* (IEEE Computer Society) **10** (1): 68–75. doi:10.1109/MIC.2006.13. . Retrieved 2008-09-06.

[6] ""rel" attribute frequently asked questions" (http://microformats.org/wiki/rel-faq). *Microformats.org*. 2008-08-06. . Retrieved 2008-09-06.

[7] http://microformats.org/wiki/hmedia

[8] http://sixrevisions.com/web-development/ultimate-guide-to-microformats-reference-and-examples/

[9] http://microformats.org/wiki/rel-directory

[10] http://microformats.org/wiki/rel-enclosure

[11] http://microformats.org/wiki/rel-license

[12] http://microformats.org/wiki/rel-tag

[13] "Exploratory Discussions" (http://microformats.org/wiki/exploratory-discussions). *Microformats.org*. 2008-08-15. . Retrieved 2008-09-06.

[14] http://microformats.org/wiki/figure

[15] http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Microformats/classes#Species

[16] http://microformats.org/wiki/measure

[17] Heilman, Chris (2009-01-19). "Retrieving and displaying data from Wikipedia with YQL" (http://developer.yahoo.net/blog/archives/2009/01/wikipedia_w_yql.html). *Yahoo Developer Network*. Yahoo. . Retrieved 2009-01-19.

[18] Goel, Kavi; Ramanathan V. Guha, Othar Hansson (2009-05-12). "Introducing Rich Snippets" (http://googlewebmastercentral.blogspot. com/2009/05/introducing-rich-snippets.html). *Google Webmaster Central Blog*. Google. . Retrieved 2009-05-25.

[19] Gong, Jun; Kosuke Suzuki, Yu Watanabe (2010-04-13). "Better recipes on the web: Introducing recipe rich snippets" (http://googlewebmastercentral.blogspot.com/2010/04/better-recipes-on-web-introducing.html). Google. . Retrieved 17 March 2011.

[20] "Bing Introducing Schema.org: Bing, Google and Yahoo Unite to Build the Web of Objects - Search Blog - Site Blogs - Bing Community" (http://www.bing.com/community/site_blogs/b/search/archive/2011/06/02/bing-google-and-yahoo-unite-to-build-the-web-of-objects.aspx). *Bing*. 2011-06-02. . Retrieved 2 June 2011.

[21] "Introducing schema.org: A Collaboration on Structured Data" (http://www.ysearchblog.com/2011/06/02/ introducing-schema-org-a-collaboration-on-structured-data). 2011-06-02. . Retrieved 2 June 2011.

[22] "Top 5 Search Engines from Oct to Dec 10 | StatCounter Global Stats" (http://gs.statcounter.com/ #search_engine-ww-monthly-201010-201012). StatCounter. . Retrieved 17 January 2011.

[23] "Bill Gates at Mix06 – "We need microformats"" (http://microformats.org/blog/2006/03/20/bill-gates-at-mix06-we-need-microformats). 2006-03-20. . Retrieved 2008-09-06. "We need microformats and to get people to agree on them. It is going to bootstrap exchanging data on the Web… …we need them for things like contact cards, events, directions…"

[24] http://blog.mozilla.com/faaborg/2007/02/04/microformats-part-4-the-user-interface-of-microformat-detection/

[25] "Criticism" (http://microformats.org/wiki?title=criticism&oldid=18478). *Microformats.org*. 2007-03-24. . Retrieved 2007-08-15.

[26] "Advocacy" (http://microformats.org/wiki/advocacy). *Microformats.org*. 2008-08-27. . Retrieved 2007-08-15.

[27] "Spread Microformats" (http://microformats.org/wiki/spread-microformats). *Microformats.org*. 2008-08-29. . Retrieved 2007-08-15. This includes community resources for marketing microformats such as buttons, banners, wallpaper/desktop screens, logo graphics, etc.

[28] Holzschlag, Molly E. (2005-03-31). "Interview with Håkon Wium Lie" (http://www.molly.com/2005/03/31/ interview-with-hkon-wium-lie/). *Molly.com*. . Retrieved 2007-11-18.

[29] Inkster, Toby A. (2008-04-22). "More than three years" (http://microformats.org/discuss/mail/microformats-discuss/2008-August/ 012402.html). *Microformats.org*. . Retrieved 2008-08-24.

[30] http://microformats.org/wiki/abbr-design-pattern

[31] Craig, James (2007-04-27). "hAccessibility" (http://www.webstandards.org/2007/04/27/haccessibility/). Web Standards Project. . Retrieved 2007-08-16.

[32] Smethurst, Michael (2008-06-23). "Removing Microformats from bbc.co.uk/programmes" (http://www.bbc.co.uk/blogs/radiolabs/2008/06/removing_microformats_from_bbc.shtml). BBC. . Retrieved 2008-08-24.

[33] "W3C GRDDL Recommendation Bridges HTML/Microformats and the Semantic Web" (http://xml.coverpages.org/ni2007-09-13-a. html). *XML Coverpages*. OASIS. 2007-09-13. . Retrieved 2007-11-23.

References

- Allsopp, John (March 2007). *Microformats: Empowering Your Markup for Web 2.0*. Friends of ED. p. 368. ISBN 978-1-59059-814-6.
- Orchard, Leslie M (September 2005). *Hacking RSS and Atom*. John Wiley & Sons. p. 602. ISBN 978-0-7645-9758-9.
- Robbins, Jennifer Niederst; Tantek Çelik, Derek Featherstone, Aaron Gustafson (February 2006). *Web Design In A Nutshell* (Third ed.). O'Reilly Media. p. 826. ISBN 978-0-596-00987-8.

Further reading

- Suda, Brian (September 2006). *Using Microformats*. O'Reilly Media. p. 45. ISBN 978-0-596-528218.
- Ahmet Soylu, Patrick De Causmaecker, Fridolin Wild Ubiquitous Web for Ubiquitous Environments: The Role of Embedded Semantics (http://www.rintonpress.com/journals/jmmonline.html#v6n1), article in Journal of Mobile Multimedia, Vol. 6, No.1, pp. 26–48, (2010). PDF (https://lirias.kuleuven.be/bitstream/123456789/ 243944/2/JMM_soylu_et_al_2010.pdf)

External links

- microformats.org (http://microformats.org/)
- Microformats Primer (http://www.digital-web.com/articles/microformats_primer/)
- Optimus (http://microformatique.com/optimus/) microformats parser and validator
- A four-part discussion of Microformats, UI issues, and possible presentation in Firefox 3 by Alex Faaborg of Mozilla (http://blog.mozilla.com/faaborg/2006/12/11/microformats-part-0-introduction)

Web 2.0

Web 2.0 is a loosely defined intersection of web application features that facilitate participatory information sharing, interoperability, user-centered design,[1] and collaboration on the World Wide Web. A Web 2.0 site allows users to interact and collaborate with each other in a social media dialogue as creators (prosumers) of user-generated content in a virtual community, in contrast to websites where users (consumers) are limited to the passive viewing of content that was created for them. Examples of Web 2.0 include social networking sites, blogs, wikis, video sharing sites, hosted services, web applications, mashups and folksonomies.

A tag cloud typical of Web 2.0 services organized around themes[2][3]

The term is closely associated with Tim O'Reilly because of the O'Reilly Media Web 2.0 conference in late 2004.[2] Although the term suggests a new version of the World Wide Web, it does not refer to an update to any technical specification, but rather to cumulative changes in the ways software developers and end-users use the Web. Whether Web 2.0 is qualitatively different from prior web technologies has been challenged by World Wide Web

inventor Tim Berners-Lee, who called the term a "piece of jargon",[4] precisely because he intended the Web in his vision as "a collaborative medium, a place where we [could] all meet and read and write". He called it the "Read/Write Web".[5]

History

The term "Web 2.0" was first used in January 1999 by Darcy DiNucci, a consultant on electronic information design (information architecture). In her article, "Fragmented Future", DiNucci writes:[6]

> The Web we know now, which loads into a browser window in essentially static screenfuls, is only an embryo of the Web to come. The first glimmerings of Web 2.0 are beginning to appear, and we are just starting to see how that embryo might develop. The Web will be understood not as screenfuls of text and graphics but as a transport mechanism, the ether through which interactivity happens. It will [...] appear on your computer screen, [...] on your TV set [...] your car dashboard [...] your cell phone [...] hand-held game machines [...] maybe even your microwave oven.

Her use of the term deals mainly with Web design, aesthetics, and the interconnection of everyday objects with the Internet; she argues that the Web is "fragmenting" due to the widespread use of portable Web-ready devices. Her article is aimed at designers, reminding them to code for an ever-increasing variety of hardware. As such, her use of the term hints at, but does not directly relate to, the current uses of the term.

The term Web 2.0 did not resurface until 2002.[7][8][9][10] These authors focus on the concepts currently associated with the term where, as Scott Dietzen puts it, "the Web becomes a universal, standards-based integration platform".[9] John Robb wrote: "What is Web 2.0? It is a system that breaks with the old model of centralized Web sites and moves the power of the Web/Internet to the desktop."[10]

In 2003, the term began its rise in popularity when O'Reilly Media and MediaLive hosted the first Web 2.0 conference. In their opening remarks, John Battelle and Tim O'Reilly outlined their definition of the "Web as Platform", where software applications are built upon the Web as opposed to upon the desktop. The unique aspect of this migration, they argued, is that "customers are building your business for you".[11] They argued that the activities of users generating content (in the form of ideas, text, videos, or pictures) could be "harnessed" to create value. O'Reilly and Battelle contrasted Web 2.0 with what they called "Web 1.0". They associated Web 1.0 with the business models of Netscape and the Encyclopædia Britannica Online. For example,

> Netscape framed "the web as platform" in terms of the old software paradigm: their flagship product was the web browser, a desktop application, and their strategy was to use their dominance in the browser market to establish a market for high-priced server products. Control over standards for displaying content and applications in the browser would, in theory, give Netscape the kind of market power enjoyed by Microsoft in the PC market. Much like the "horseless carriage" framed the automobile as an extension of the familiar, Netscape promoted a "webtop" to replace the desktop, and planned to populate that webtop with information updates and applets pushed to the webtop by information providers who would purchase Netscape servers.[12]

In short, Netscape focused on creating software, updating it on occasion, and distributing it to the end users. O'Reilly contrasted this with Google, a company that did not at the time focus on producing software, such as a browser, but instead on providing a service based on data such as the links Web page authors make between sites. Google exploits this user-generated content to offer Web search based on reputation through its "PageRank" algorithm. Unlike software, which undergoes scheduled releases, such services are constantly updated, a process called "the perpetual beta". A similar difference can be seen between the Encyclopædia Britannica Online and Wikipedia: while the Britannica relies upon experts to create articles and releases them periodically in publications, Wikipedia relies on trust in anonymous users to constantly and quickly build content. Wikipedia is not based on expertise but rather an adaptation of the open source software adage "given enough eyeballs, all bugs are shallow", and it produces and updates articles constantly. O'Reilly's Web 2.0 conferences have been held every year since 2003, attracting entrepreneurs, large companies, and technology reporters.

In terms of the lay public, the term Web 2.0 was largely championed by bloggers and by technology journalists, culminating in the 2006 *TIME magazine* Person of The Year (*You*).[13] That is, *TIME* selected the masses of users who were participating in content creation on social networks, blogs, wikis, and media sharing sites. In the cover story, Lev Grossman explains:

> It's a story about community and collaboration on a scale never seen before. It's about the cosmic compendium of knowledge Wikipedia and the million-channel people's network YouTube and the online metropolis MySpace. It's about the many wresting power from the few and helping one another for nothing and how that will not only change the world but also change the way the world changes.

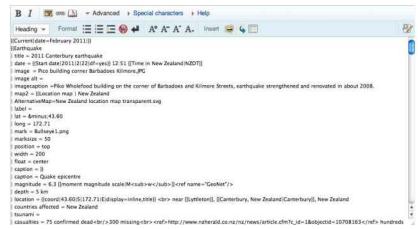
Since that time, Web 2.0 has found a place in the lexicon; in 2009 Global Language Monitor declared it to be the one-millionth English word.[14]

Characteristics

Web 2.0 websites allow users to do more than just retrieve information. By increasing what was already possible in "Web 1.0", they provide the user with more user-interface, software and storage facilities, all through their browser. This has been called "Network as platform" computing.[3] Users can provide the data that is on a Web 2.0 site and exercise some control over that data.[3][15] These sites may have an "Architecture of participation" that encourages users to add value to the application as they use it.[2][3] Some scholars have made the case that cloud computing is a form of Web 2.0 because cloud computing is simply an implication of computing on the Internet.[16]

A list of ways that people can volunteer to improve Mass Effect Wiki, on the main page of that site. Mass Effect Wiki is an example of content generated by users working collaboratively.

The concept of Web-as-participation-platform captures many of these characteristics. Bart Decrem, a founder and former CEO of Flock, calls Web 2.0 the "participatory Web"[17] and regards the Web-as-information-source as Web1.0.



Edit box interface through which anyone could edit a Wikipedia article.

The Web 2.0 offers all users the same freedom to contribute. While this opens the possibility for rational debate and collaboration, it also opens the possibility for "spamming" and "trolling" by less rational users. The impossibility of excluding group members who don't contribute to the provision of goods from sharing profits gives rise to the possibility that rational members will prefer to withhold their contribution of effort and free ride on the contribution of others.[18] This requires what is sometimes called radical trust by the management of the website. According to Best,[19] the characteristics of Web 2.0 are: rich user experience, user participation, dynamic content, metadata, web standards and scalability. Further characteristics, such as openness, freedom[20] and collective intelligence[21] by way of user participation, can also be viewed as essential attributes of Web 2.0.

Technologies

The client-side/web browser technologies used in Web 2.0 development are Asynchronous JavaScript and XML (Ajax), Adobe Flash and the Adobe Flex framework, and JavaScript/Ajax frameworks such as YUI Library, Dojo Toolkit, MooTools, jQuery and Prototype JavaScript Framework. Ajax programming uses JavaScript to upload and download new data from the web server without undergoing a full page reload.

To allow users to continue to interact with the page, communications such as data requests going to the server are separated from data coming back to the page (asynchronously). Otherwise, the user would have to routinely wait for the data to come back before they can do anything else on that page, just as a user has to wait for a page to complete the reload. This also increases overall performance of the site, as the sending of requests can complete quicker independent of blocking and queueing required to send data back to the client....

The data fetched by an Ajax request is typically formatted in XML or JSON (JavaScript Object Notation) format, two widely used structured data formats. Since both of these formats are natively understood by JavaScript, a programmer can easily use them to transmit structured data in their web application. When this data is received via Ajax, the JavaScript program then uses the Document Object Model (DOM) to dynamically update the web page based on the new data, allowing for a rapid and interactive user experience. In short, using these techniques, Web designers can make their pages function like desktop applications. For example, Google Docs uses this technique to create a Web based word processor.

Adobe Flex is another technology often used in Web 2.0 applications. Compared to JavaScript libraries like jQuery, Flex makes it easier for programmers to populate large data grids, charts, and other heavy user interactions.[22] Applications programmed in Flex, are compiled and displayed as Flash within the browser. As a widely available plugin independent of W3C (World Wide Web Consortium, the governing body of web standards and protocols) standards, Flash is capable of doing many things that were not possible pre-HTML5, the language used to construct web pages. Of Flash's many capabilities, the most commonly used in Web 2.0 is its ability to play audio and video files. This has allowed for the creation of Web 2.0 sites where video media is seamlessly integrated with standard HTML.

In addition to Flash and Ajax, JavaScript/Ajax frameworks have recently become a very popular means of creating Web 2.0 sites. At their core, these frameworks do not use technology any different from JavaScript, Ajax, and the DOM. What frameworks do is smooth over inconsistencies between web browsers and extend the functionality available to developers. Many of them also come with customizable, prefabricated 'widgets' that accomplish such common tasks as picking a date from a calendar, displaying a data chart, or making a tabbed panel.

On the server side, Web 2.0 uses many of the same technologies as Web 1.0. New languages such as PHP, Ruby, Perl, Python and JSP are used by developers to output data dynamically using information from files and databases. What has begun to change in Web 2.0 is the way this data is formatted. In the early days of the Internet, there was little need for different websites to communicate with each other and share data. In the new "participatory web", however, sharing data between sites has become an essential capability. To share its data with other sites, a website must be able to generate output in machine-readable formats such as XML (Atom, RSS, etc.) and JSON. When a site's data is available in one of these formats, another website can use it to integrate a portion of that site's functionality into itself, linking the two together. When this design pattern is implemented, it ultimately leads to data that is both easier to find and more thoroughly categorized, a hallmark of the philosophy behind the Web 2.0 movement.

In brief, Ajax is a key technology used to build Web 2.0 because it provides rich user experience and works with any browser whether it is Firefox, Chrome, Internet Explorer or another popular browser. Then, a language with very good web services support should be used to build Web 2.0 applications. In addition, the language used should be iterative meaning that the addition and deployment of features can be easily and quickly achieved.

Concepts

Web 2.0 can be described in 3 parts, which are as follows:

- Rich Internet application (RIA) — defines the experience brought from desktop to browser whether it is from a graphical point of view or usability point of view. Some buzzwords related to RIA are Ajax and Flash.
- Web-oriented architecture (WOA) — is a key piece in Web 2.0, which defines how Web 2.0 applications expose their functionality so that other applications can leverage and integrate the functionality providing a set of much richer applications (Examples are: Feeds, RSS, Web Services, Mash-ups)
- Social Web — defines how Web 2.0 tends to interact much more with the end user and make the end-user an integral part.

As such, Web 2.0 draws together the capabilities of client- and server-side software, content syndication and the use of network protocols. Standards-oriented web browsers may use plug-ins and software extensions to handle the content and the user interactions. Web 2.0 sites provide users with information storage, creation, and dissemination capabilities that were not possible in the environment now known as "Web 1.0".

Web 2.0 websites include the following features and techniques: Andrew McAfee used the acronym SLATES to refer to them:[23]

Search

Finding information through keyword search.

Links

Connects information together into a meaningful information ecosystem using the model of the Web, and provides low-barrier social tools.

Authoring

The ability to create and update content leads to the collaborative work of many rather than just a few web authors. In wikis, users may extend, undo and redo each other's work. In blogs, posts and the comments of individuals build up overtime.

Tags

Categorization of content by users adding "tags"—short, usually one-word descriptions—to facilitate searching, without dependence on pre-made categories. Collections of tags created by many users within a single system may be referred to as "folksonomies" (i.e., folk taxonomies).

Extensions

Software that makes the Web an application platform as well as a document server. These include software like Adobe Reader, Adobe Flash player, Microsoft Silverlight, ActiveX, Oracle Java, Quicktime, Windows Media, etc.

Signals

The use of syndication technology such as RSS to notify users of content changes.

While SLATES forms the basic framework of Enterprise 2.0, it does not contradict all of the higher level Web 2.0 design patterns and business models. In this way, a new Web 2.0 report from O'Reilly is quite effective and diligent in interweaving the story of Web 2.0 with the specific aspects of Enterprise 2.0. It includes discussions of self-service IT, the long tail of enterprise IT demand, and many other consequences of the Web 2.0 era in the enterprise. The report also makes many sensible recommendations around starting small with pilot projects and measuring results, among a fairly long list.[24]

Usage

A third important part of Web 2.0 is the social Web, which is a fundamental shift in the way people communicate. The social web consists of a number of online tools and platforms where people share their perspectives, opinions, thoughts and experiences. Web 2.0 applications tend to interact much more with the end user. As such, the end user is not only a user of the application but also a participant by:

- Podcasting
- Blogging
- Tagging
- Contributing to RSS
- Social bookmarking
- Social networking

The popularity of the term Web 2.0, along with the increasing use of blogs, wikis, and social networking technologies, has led many in academia and business to coin a flurry of 2.0s,[25] including Library 2.0,[26] Social Work 2.0,[27] Enterprise 2.0, PR 2.0,[28] Classroom 2.0,[29] Publishing 2.0,[30] Medicine 2.0,[31] Telco 2.0, Travel 2.0, Government 2.0,[32] and even Porn 2.0.[33] Many of these 2.0s refer to Web 2.0 technologies as the source of the new version in their respective disciplines and areas. For example, in the Talis white paper "Library 2.0: The Challenge of Disruptive Innovation", Paul Miller argues

> Blogs, wikis and RSS are often held up as exemplary manifestations of Web 2.0. A reader of a blog or a wiki is provided with tools to add a comment or even, in the case of the wiki, to edit the content. This is what we call the Read/Write web. Talis believes that Library 2.0 means harnessing this type of participation so that libraries can benefit from increasingly rich collaborative cataloging efforts, such as including contributions from partner libraries as well as adding rich enhancements, such as book jackets or movie files, to records from publishers and others.[34]

Here, Miller links Web 2.0 technologies and the culture of participation that they engender to the field of library science, supporting his claim that there is now a "Library 2.0". Many of the other proponents of new 2.0s mentioned here use similar methods.

The meaning of web 2.0 is role dependent, as Dennis D. McDonalds noted. For example, some use Web 2.0 to establish and maintain relationships through social networks, while some marketing managers might use this promising technology to "end-run traditionally unresponsive I.T. department[s]."[35]

There is a debate over the use of Web 2.0 technologies in mainstream education. Issues under consideration include the understanding of students' different learning modes; the conflicts between ideas entrenched in informal on-line communities and educational establishments' views on the production and authentication of 'formal' knowledge; and questions about privacy, plagiarism, shared authorship and the ownership of knowledge and information produced and/or published on line.[36]

Marketing

> For marketers, Web 2.0 offers an opportunity to engage consumers. A growing number of marketers are using Web 2.0 tools to collaborate with consumers on product development, service enhancement and promotion. Companies can use Web 2.0 tools to improve collaboration with both its business partners and consumers. Among other things, company employees have created wikis—Web sites that allow users to add, delete, and edit content — to list answers to frequently asked questions about each product, and consumers have added significant contributions. Another marketing Web 2.0 lure is to make sure consumers can use the online community to network among themselves on topics of their own choosing.[37]
>
> Mainstream media usage of web 2.0 is increasing. Saturating media hubs—like *The New York Times, PC Magazine* and *Business Week* — with links to popular new web sites and services, is critical to achieving the threshold for mass adoption of those services.[38]

Web 2.0 offers financial institutions abundant opportunities to engage with customers. Networks such as Twitter, Yelp and Facebook are now becoming common elements of multichannel and customer loyalty strategies, and banks are beginning to use these sites proactively to spread their messages. In a recent article for Bank Technology News, Shane Kite describes how Citigroup's Global Transaction Services unit monitors social media outlets to address customer issues and improve products. Furthermore, the FI uses Twitter to release "breaking news" and upcoming events, and YouTube to disseminate videos that feature executives speaking about market news.[39]

Small businesses have become more competitive by using Web 2.0 marketing strategies to compete with larger companies. As new businesses grow and develop, new technology is used to decrease the gap between businesses and customers. Social networks have become more intuitive and user friendly to provide information that is easily reached by the end user. For example, companies use Twitter to offer customers coupons and discounts for products and services.[40]

According to Google Timeline, the term Web 2.0 was discussed and indexed most frequently in 2005, 2007 and 2008. Its average use is continuously declining by 2–4% per quarter since April 2008.

Web 2.0 in education

Web 2.0 technologies provide teachers with new ways to engage students in a meaningful way. "Children raised on new media technologies are less patient with filling out worksheets and listening to lectures"[41] because students already participate on a global level. The lack of participation in a traditional classroom stems more from the fact that students receive better feedback online. Traditional classrooms have students do assignments and when they are completed, they are just that, finished. However, Web 2.0 shows students that education is a constantly evolving entity. Whether it is participating in a class discussion, or participating in a forum discussion, the technologies available to students in a Web 2.0 classroom does increase the amount they participate.

Will Richardson stated in *Blogs, Wikis, Podcasts and other Powerful Web tools for the Classrooms*, 3rd Edition that, "The Web has the potential to radically change what we assume about teaching and learning, and it presents us with important questions to ponder: What needs to change about our curriculum when our students have the ability to reach audiences far beyond our classroom walls?"[42] Web 2.0 tools are needed in the classroom to prepare both students and teachers for the shift in learning that Collins and Halverson describe. According to Collins and Halverson, the self-publishing aspects as well as the speed with which their work becomes available for consumption allows teachers to give students the control they need over their learning. This control is the preparation students will need to be successful as learning expands beyond the classroom."[41]

Some may think that these technologies could hinder the personal interaction of students, however all of the research points to the contrary. "Social networking sites have worried many educators (and parents) because they often bring with them outcomes that are not positive: narcissism, gossip, wasted time, 'friending', hurt feelings, ruined reputations, and sometimes unsavory, even dangerous activities, [on the contrary,] social networking sites promote conversations and interaction that is encouraged by educators."[43] By allowing students to use the technology tools of Web 2.0, teachers are actually giving students the opportunity to learn for themselves and share that learning with their peers. One of the many implications of Web 2.0 technologies on class discussions is the idea that teachers are no longer in control of the discussions. Instead, Russell and Sorge (1999) conclude that integrating technology into instruction tends to move classrooms from teacher-dominated environments to ones that are more student-centered. While it is still important for them to monitor what students are discussing, the actual topics of learning are being guided by the students themselves.

Web 2.0 calls for major shifts in the way education is provided for students. One of the biggest shifts that Will Richardson points out in his book *Blogs, Wikis, Podcasts, and Other Powerful Web Tools for Classrooms*[42] is the fact that education must be not only socially but collaboratively constructed. This means that students, in a Web 2.0 classroom, are expected to collaborate with their peers. By making the shift to a Web 2.0 classroom, teachers are

creating a more open atmosphere where students are expected to stay engaged and participate in the discussions and learning that is taking place around them. In fact, there are many ways for educators to use Web 2.0 technologies in their classrooms.

"Weblogs are not built on static chunks of content. Instead they are comprised of reflections and conversations that in many cases are updated every day [...] They demand interaction."[42] Will Richardson's observation of the essence of weblogs speaks directly to why blogs are so well suited to discussion based classrooms. Weblogs give students a public space to interact with one another and the content of the class. As long as the students are invested in the project, the need to see the blog progress acts as motivation as the blog itself becomes an entity that can demand interaction.

For example, Laura Rochette implemented the use of blogs in her American History class and noted that in addition to an overall improvement in quality, the use of the blogs as an assignment demonstrated synthesis level activity from her students. In her experience, asking students to conduct their learning in the digital world meant asking students "to write, upload images, and articulate the relationship between these images and the broader concepts of the course, [in turn] demonstrating that they can be thoughtful about the world around them."[44] Jennifer Hunt, an 8th grade language arts teacher of pre-Advanced Placement students shares a similar story. She used the WANDA project and asked students to make personal connections to the texts they read and to describe and discuss the issues raised in literature selections through social discourse. They engaged in the discussion via wikis and other Web 2.0 tools, which they used to organize, discuss, and present their responses to the texts and to collaborate with others in their classroom and beyond.

The research shows that students are already using these technological tools, but they still are expected to go to a school where using these tools is frowned upon or even punished. If educators are able to harness the power of the Web 2.0 technologies students are using, it could be expected that the amount of participation and classroom discussion would increase. It may be that how participation and discussion is produced is very different from the traditional classroom, but nevertheless it does increase.

Web 2.0 and philanthropy

The spread of participatory information-sharing over the internet, combined with recent improvements in low-cost internet access in developing countries, has opened up new possibilities for peer-to-peer charities, which allow individuals to contribute small amounts to charitable projects for other individuals. Websites such as Donors Choose and Global Giving now allow small-scale donors to direct funds to individual projects of their choice.

A popular twist on internet-based philanthropy is the use of peer-to-peer lending for charitable purposes. Kiva pioneered this concept in 2005, offering the first web-based service to publish individual loan profiles for funding. Kiva raises funds for local intermediary microfinance organizations which post stories and updates on behalf of the borrowers. Lenders can contribute as little as $25 to loans of their choice, and receive their money back as borrowers repay. Kiva falls short of being a pure peer-to-peer charity, in that loans are disbursed before being funded by lenders and borrowers do not communicate with lenders themselves.[45][46] However, the recent spread of cheap internet access in developing countries has made genuine peer-to-peer connections increasingly feasible. In 2009 the US-based nonprofit Zidisha tapped into this trend to offer the first peer-to-peer microlending platform to link lenders and borrowers across international borders without local intermediaries. Inspired by interactive websites such as Facebook and eBay, Zidisha's microlending platform facilitates direct dialogue between lenders and borrowers and a performance rating system for borrowers. Web users worldwide can fund loans for as little as a dollar.[47]

Web-based applications and desktops

Ajax has prompted the development of websites that mimic desktop applications, such as word processing, the spreadsheet, and slide-show presentation. In 2006 Google, Inc. acquired one of the best-known sites of this broad class, Writely.[48] WYSIWYG wiki and blogging sites replicate many features of PC authoring applications.

Several browser-based "operating systems" have emerged, including EyeOS[49] and YouOS.(No longer active.)[50] Although coined as such, many of these services function less like a traditional operating system and more as an application platform. They mimic the user experience of desktop operating-systems, offering features and applications similar to a PC environment, and are able to run within any modern browser. However, these so-called "operating systems" do not directly control the hardware on the client's computer.

Numerous web-based application services appeared during the dot-com bubble of 1997–2001 and then vanished, having failed to gain a critical mass of customers. In 2005, WebEx acquired one of the better-known of these, Intranets.com, for $45 million.[51]

Distribution of media

XML and RSS

Many regard syndication of site content as a Web 2.0 feature. Syndication uses standardized protocols to permit end-users to make use of a site's data in another context (such as another website, a browser plugin, or a separate desktop application). Protocols permitting syndication include RSS (really simple syndication, also known as web syndication), RDF (as in RSS 1.1), and Atom, all of them XML-based formats. Observers have started to refer to these technologies as webfeeds.

Specialized protocols such as FOAF and XFN (both for social networking) extend the functionality of sites or permit end-users to interact without centralized websites.

Web APIs

Web 2.0 often uses machine-based interactions such as REST and SOAP. Servers often expose proprietary Application programming interfaces (API), but standard APIs (for example, for posting to a blog or notifying a blog update) have also come into use. Most communications through APIs involve XML or JSON payloads.

REST APIs, through their use of self-descriptive messages and hypermedia as the engine of application state, should be self-describing once an entry URI is known. Web Services Description Language (WSDL) is the standard way of publishing a SOAP API and there are a range of web service specifications. EMML, or Enterprise Mashup Markup Language by the Open Mashup Alliance, is an XML markup language for creating enterprise mashups.

Criticism

Critics of the term claim that "Web 2.0" does not represent a new version of the World Wide Web at all, but merely continues to use so-called "Web 1.0" technologies and concepts. First, techniques such as AJAX do not replace underlying protocols like HTTP, but add an additional layer of abstraction on top of them. Second, many of the ideas of Web 2.0 had already been featured in implementations on networked systems well before the term "Web 2.0" emerged. Amazon.com, for instance, has allowed users to write reviews and consumer guides since its launch in 1995, in a form of self-publishing. Amazon also opened its API to outside developers in 2002.[52] Previous developments also came from research in computer-supported collaborative learning and computer supported cooperative work (CSCW) and from established products like Lotus Notes and Lotus Domino, all phenomena that preceded Web 2.0.

But perhaps the most common criticism is that the term is unclear or simply a buzzword. For example, in a podcast interview,[4] Tim Berners-Lee described the term "Web 2.0" as a "piece of jargon":

> "Nobody really knows what it means...If Web 2.0 for you is blogs and wikis, then that is people to people. But that was what the Web was supposed to be all along."[4]

Other critics labeled Web 2.0 "a second bubble" (referring to the Dot-com bubble of circa 1995–2001), suggesting that too many Web 2.0 companies attempt to develop the same product with a lack of business models. For example, *The Economist* has dubbed the mid- to late-2000s focus on Web companies "Bubble 2.0".[53] Venture capitalist Josh Kopelman noted that Web 2.0 had excited only 53,651 people (the number of subscribers at that time to TechCrunch, a Weblog covering Web 2.0 startups and technology news), too few users to make them an economically viable target for consumer applications.[54] Although Bruce Sterling reports he's a fan of Web 2.0, he thinks it is now dead as a rallying concept.[55]

Critics have cited the language used to describe the hype cycle of Web 2.0[56] as an example of Techno-utopianist rhetoric.[57]

In terms of Web 2.0's social impact, critics such as Andrew Keen argue that Web 2.0 has created a cult of digital narcissism and amateurism, which undermines the notion of expertise by allowing anybody, anywhere to share and place undue value upon their own opinions about any subject and post any kind of content, regardless of their particular talents, knowledge, credentials, biases or possible hidden agendas. Keen's 2007 book, Cult of the Amateur, argues that the core assumption of Web 2.0, that all opinions and user-generated content are equally valuable and relevant, is misguided. Additionally, Sunday Times reviewer John Flintoff has characterized Web 2.0 as "creating an endless digital forest of mediocrity: uninformed political commentary, unseemly home videos, embarrassingly amateurish music, unreadable poems, essays and novels", and also asserted that Wikipedia is full of "mistakes, half truths and misunderstandings".[58] Michael Gorman, former president of the American Library Association has been vocal about his opposition to Web 2.0 due to the lack of expertise that it outwardly claims though he believes that there is some hope for the future as "The task before us is to extend into the digital world the virtues of authenticity, expertise, and scholarly apparatus that have evolved over the 500 years of print, virtues often absent in the manuscript age that preceded print".[59]

Trademark

In November 2004, CMP Media applied to the USPTO for a service mark on the use of the term "WEB 2.0" for live events.[60] On the basis of this application, CMP Media sent a cease-and-desist demand to the Irish non-profit organization IT@Cork on May 24, 2006,[61] but retracted it two days later.[62] The "WEB 2.0" service mark registration passed final PTO Examining Attorney review on May 10, 2006, and was registered on June 27, 2006.[60] The European Union application (application number 004972212, which would confer unambiguous status in Ireland) was [63] refused on May 23, 2007.

Web 3.0

Definitions of Web 3.0 vary greatly. Some[64] believe its most important features are the Semantic Web and personalization. Focusing on the computer elements, Conrad Wolfram has argued that Web 3.0 is where "the computer is generating new information", rather than humans.[65]

Andrew Keen, author of *The Cult of the Amateur*, considers the Semantic Web an "unrealisable abstraction" and sees Web 3.0 as the return of experts and authorities to the Web. For example, he points to Bertelsmann's deal with the German Wikipedia to produce an edited print version of that encyclopedia.[66] CNN Money's Jessi Hempel expects Web 3.0 to emerge from new and innovative Web 2.0 services with a profitable business model.[67]

Futurist John Smart, lead author of the Metaverse Roadmap[68] defines Web 3.0 as the first-generation Metaverse (convergence of the virtual and physical world), a web development layer that includes TV-quality open video, 3D simulations, augmented reality, human-constructed semantic standards, and pervasive broadband, wireless, and sensors. Web 3.0's early geosocial (Foursquare, etc.) and augmented reality (Layar, etc.) webs are an extension of

Web 2.0's participatory technologies and social networks (Facebook, etc.) into 3D space. Of all its metaverse-like developments, Smart suggests Web 3.0's most defining characteristic will be the mass diffusion of NTSC-or-better quality open video [69] to TVs, laptops, tablets, and mobile devices, a time when "the internet swallows the television."[70] Smart considers Web 4.0 to be the Semantic Web and in particular, the rise of statistical, machine-constructed semantic tags and algorithms, driven by broad collective use of conversational interfaces, perhaps circa 2020.[71] David Siegel's perspective in *Pull: The Power of the Semantic Web*, 2009, is consonant with this, proposing that the growth of human-constructed semantic standards and data will be a slow, industry-specific incremental process for years to come, perhaps unlikely to tip into broad social utility until after 2020.

According to some Internet experts, Web 3.0 will allow the user to sit back and let the Internet do all of the work for them.[72] Rather than having search engines gear towards your keywords, the search engines will gear towards the user. Keywords will be searched based on your culture, region, and jargon.[73]

References

[1] "Core Characteristics of Web 2.0 Services" (http://www.techpluto.com/web-20-services/). .

[2] Paul Graham (November 2005). "Web 2.0" (http://www.paulgraham.com/web20.html). . Retrieved 2006-08-02. "I first heard the phrase 'Web 2.0' in the name of the Web 2.0 conference in 2004."

[3] Tim O'Reilly (2005-09-30). "What Is Web 2.0" (http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20. html). O'Reilly Network. . Retrieved 2006-08-06.

[4] "DeveloperWorks Interviews: Tim Berners-Lee" (http://www.ibm.com/developerworks/podcast/dwi/cm-int082206txt.html). 2006-07-28. . Retrieved 2007-02-07.

[5] "Berners-Lee on the read/write web" (http://news.bbc.co.uk/2/hi/technology/4132752.stm). BBC News. 2005-08-09. . Retrieved 2011-02-06.

[6] DiNucci, Darcy (1999). "Fragmented Future" (http://darcyd.com/fragmented_future.pdf) (pdf). *Print* **53** (4): 32. .

[7] Idehen, Kingsley. 2003. RSS: INJAN (It's not just about news). Blog. Blog Data Space. August 21 OpenLinksW.com (http://www. openlinksw.com/dataspace/kidehen@openlinksw.com/weblog/kidehen@openlinksw.com's BLOG [127]/241)

[8] Idehen, Kingsley. 2003. Jeff Bezos Comments about Web Services. Blog. Blog Data Space. September 25. OpenLinksW.com (http://www. openlinksw.com/blog/~kidehen/index.vspx?id=373)

[9] Knorr, Eric. 2003. The year of Web services. CIO, December 15.

[10] "John Robb's Weblog" (http://jrobb.mindplex.org/2003/08/16.html). Jrobb.mindplex.org. . Retrieved 2011-02-06.

[11] O'Reilly, Tim, and John Battelle. 2004. Opening Welcome: State of the Internet Industry. In San Francisco, California, October 5.

[12] O'Reilly, T., 2005.

[13] Grossman, Lev. 2006. Person of the Year: You. December 25. Time.com (http://www.time.com/time/covers/0,16641,20061225,00. html)

[14] "'Millionth English Word' declared". NEWS.BBC.co.uk (http://news.bbc.co.uk/1/hi/world/americas/8092549.stm)

[15] Dion Hinchcliffe (2006-04-02). "The State of Web 2.0" (http://web2.wsj2.com/the_state_of_web_20.htm). Web Services. . Retrieved 2006-08-06.

[16] [SSRN: http://ssrn.com/abstract=732483 Wireless Communications and Computing at a Crossroads: New Paradigms and Their Impact on Theories Governing the Public's Right to Spectrum Access], Patrick S. Ryan, Journal on Telecommunications & High Technology Law, Vol. 3, No. 2, p. 239, 2005.

[17] Bart Decrem (2006-06-13). "Introducing Flock Beta 1" (http://www.flock.com/node/4500). Flock official blog. . Retrieved 2007-01-13.

[18] Gerald Marwell and Ruth E. Ames: "Experiments on the Provision of Public Goods. I. Resources, Interest, Group Size, and the Free-Rider Problem". *The American Journal of Sociology*, Vol. 84, No. 6 (May, 1979), pp. 1335–1360

[19] Best, D., 2006. Web 2.0 Next Big Thing or Next Big Internet Bubble? Lecture Web Information Systems. Technische Universiteit Eindhoven.

[20] Greenmeier, Larry and Gaudin, Sharon. "Amid The Rush To Web 2.0, Some Words Of Warning – Web 2.0 – InformationWeek" (http:// www.informationweek.com/news/management/showArticle. jhtml;jsessionid=EWRPGLVJ53OW2QSNDLPCKHSCJUNN2JVN?articleID=199702353&_requestid=494050). www.informationweek.com. . Retrieved 2008-04-04.

[21] O'Reilly, T., 2005. What is Web 2.0. Design Patterns and Business Models for the Next Generation of Software, 30, p. 2005

[22] Marak Squires dot com : JavaScript/jQuery versus Actionscript/Flex : Take 1 (http://maraksquires.com/articles/2009/11/16/ javascript-jquery-versus-actionscript-flex-take-1/)

[23] McAfee, A. (2006). Enterprise 2.0: The Dawn of Emergent Collaboration. MIT Sloan Management review. Vol. 47, No. 3, p. 21–28.

[24] Web 2.0 definition updated and Enterprise 2.0 emerges | ZDNet (http://blogs.zdnet.com/Hinchcliffe/?p=71)

[25] Schick, S., 2005. I second that emotion. IT Business.ca (Canada).

[26] Miller, P., 2008. Library 2.0: The Challenge of Disruptive Innovation. Available at: Google.com (http://www.talis.com/resources/documents/447_Library_2_prf1.pdf)

[27] Singer, Jonathan B. (2009). *The Role and Regulations for Technology in Social Work Practice and E-Therapy: Social Work 2.0. In A. R. Roberts (Ed).* (http://www.us.oup.com/us/catalog/general/subject/SocialWork/~~/ dmlldz11c2EmY2k9OTc4MDE5NTM2OTM3Mw==). New York, U.S.A.: Oxford University Press. ISBN 978-0195369373. .

[28] Breakenridge, D., 2008. PR 2.0: New Media, New Tools, New Audiences 1st ed., FT Press.

[29] "Classroom 2.0" (http://www.classroom20.com/). . Retrieved 2010-09-22

[30] Karp, Scott. "Publishing 2.0" (http://publishing2.com/). Publishing2.com. . Retrieved 2011-02-06.

[31] Medicine 2.0

[32] Eggers, William D. (2005). *Government 2.0: Using Technology to Improve Education, Cut Red Tape, Reduce Gridlock, and Enhance Democracy* (http://www.manhattan-institute.org/government2.0/). Lanham MD, U.S.A.: Rowman & Littlefield Publishers, Inc.. ISBN 978-0742541757. .

[33] Rusak, Sergey (2009). *Web 2.0 Becoming An Outdated Term* (http://www.progressiveadvertiser.com/ web-2-0-becoming-an-outdated-term/). Boston, Massachusetts, U.S.A.: Progressive Advertiser. .

[34] Miller 10–11

[35] "i-Technology Viewpoint: It's Time to Take the Quotation Marks Off "Web 2.0" | Web 2.0 Journal" (http://web2.sys-con.com/node/ 207411). Web2.sys-con.com. . Retrieved 2011-02-06.

[36] Anderson, Paul (2007). "What is Web 2.0? Ideas, technologies and implications for education" (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.108.9995&rep=rep1&type=pdf). *JISC Technology and Standards Watch*. .

[37] Parise, Salvatore (2008). "The Secrets of Marketing in a Web 2.0 World" (http://online.wsj.com/article/SB122884677205091919.html). *The Wall Street Journal*. .

[38] MacManus, Richard (2007). "Mainstream Media Usage of Web 2.0 Services is Increasing" (http://www.readwriteweb.com/archives/mainstream_media_web20.php). *Read Write Web*. .

[39] "Banks use Web 2.0 to increase customer retention" (http://www.pntmarketingservices.com/newsfeed/article/Banks_use_Web_2_0_to_increase_customer_retention-800226524.html). *PNT Marketing Services*. 2010. .

[40] "Small Businesses Need Innovation — New Company May Have Their Solution" (http://www.sfgate.com/cgi-bin/article.cgi?f=/g/a/2010/10/25/prwebprweb4693214.DTL). *San Francisco Chronicle*. 2010. .

[41] Collins, Allan (2009). *Rethinking Education in the Age of Technology*. New York, NY: Teachers College Press. pp. 176. ISBN 978-0-8077-5002-5.

[42] .

[43] Hargadon, Steve. "Educational Networking: The Important Role Web 2.0 Will Play In Education" (http://www.scribd.com/doc/ 24161189/Educational-Networking-The-Important-Role-Web-2-0-Will-Play-in-Education). . Retrieved 19 May 2011.

[44] Rochette, Laura (2007). "What Classroom Technology Has Taught Me about Curriculum, Teaching, and Infinite Possibilities". *English Journal*. 2 **37**: 43–48.

[45] *Kiva Is Not Quite What It Seems* (http://blogs.cgdev.org/open_book/2009/10/kiva-is-not-quite-what-it-seems.php), by David Roodman, Center for Global Development, Oct. 2, 2009, as accessed Jan. 2 & 16, 2010

[46] *Confusion on Where Money Lent via Kiva Goes* (http://www.nytimes.com/2009/11/09/business/global/09kiva.html?_r=1&scp=1& sq=Kiva&st=cse), by Stephanie Strom, in The New York Times, Nov. 8, 2009, as accessed Jan. 2 & 16, 2010

[47] "Zidisha Set to "Expand" in Peer-to-Peer Microfinance", Microfinance Focus, Feb 2010 (http://www.microfinancefocus.com/news/2010/ 02/07/zidisha-set-to-expand-in-peer-to-peer-microfinance-julia-kurnia/)

[48] "Google buys Web word-processing technology" (http://www.news.com/2100-1032_3-6048136.html). www.news.com. . Retrieved 2007-12-12.

[49] "Can eyeOS Succeed Where Desktop.com Failed?" (http://www.techcrunch.com/2006/11/27/eyeos-open-source-webos-for-the-masses/). www.techcrunch.com. . Retrieved 2007-12-12.

[50] "Tech Beat Hey YouOS! – BusinessWeek" (http://www.businessweek.com/the_thread/techbeat/archives/2006/03/hey_youos.html). www.businessweek.com. . Retrieved 2007-12-12.

[51] "PC World — WebEx Snaps Up Intranets.com" (http://www.pcworld.com/article/id,122068-page,1/article.html). www.pcworld.com. . Retrieved 2007-12-12.

[52] Tim O'Reilly (2002-06-18). "Amazon Web Services API" (http://www.oreillynet.com/pub/wlg/1707?wlg=yes). *O'Reilly Network*. . Retrieved 2006-05-27.

[53] "Bubble 2.0" (http://www.economist.com/business/displaystory.cfm?story_id=E1_QQNVDDS). *The Economist*. 2005-12-22. . Retrieved 2006-12-20.

[54] Josh Kopelman (2006-05-11). "53,651" (http://redeye.firstround.com/2006/05/53651.html). *Redeye VC*. . Retrieved 2006-12-21.

[55] "Bruce Sterling presenta il web 2.0" (http://www.lastampa.it/multimedia/multimedia.asp?p=1&IDmsezione=29&IDalbum=8558& tipo=VIDEO#mpos). *"LASTAMPA.it"*. .

[56] "Gartner 2006 Emerging Technologies Hype Cycle" (http://www.gartner.com/it/page.jsp?id=495475). .

[57] ""Critical Perspectives on Web 2.0", Special issue of *First Monday*, 13(3), 2008. UIC.edu (http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/issue/view/263/showToc)".

[58] Flintoff, JohnPaul (2007-06-03). "Thinking is so over" (http://technology.timesonline.co.uk/tol/news/tech_and_web/personal_tech/ article1874668.ece). *The Times* (London). .
[59] Gorman, Michael. "Web 2.0: The Sleep of Reason, Part 1" (http://www.britannica.com/blogs/2007/06/ web-20-the-sleep-of-reason-part-i/). . Retrieved 26th April 2011.
[60] "USPTO serial number 78322306" (http://tarr.uspto.gov/servlet/tarr?regser=serial&entry=78322306). Tarr.uspto.gov.. Retrieved 2011-02-06.
[61] "O'Reilly and CMP Exercise Trademark on 'Web 2.0'" (http://yro.slashdot.org/article.pl?sid=06/05/26/1238245). *Slashdot*. 2006-05-26. . Retrieved 2006-05-27.
[62] Nathan Torkington (2006-05-26). "O'Reilly's coverage of Web 2.0 as a service mark" (http://radar.oreilly.com/archives/2006/05/more_on_our_web_20_service_mar.html). *O'Reilly Radar*. . Retrieved 2006-06-01.
[63] http://oami.europa.eu/CTMOnline/RequestManager/en_Result?transition=ResultsDetailed&ntmark=&application=CTMOnline&bAdvanced=0&language=en&deno=&source=search_basic.jsp&idappli=004972212#
[64] Agarwal, Amit. "Web 3.0 concepts explained in plain English". Labnol.org (http://www.labnol.org/internet/web-3-concepts-explained/ 8908/)
[65] Conrad Wolfram on Communicating with apps in web 3.0 (http://www.itpro.co.uk/621535/ q-a-conrad-wolfram-on-communicating-with-apps-in-web-3-0) IT PRO, 17 Mar 2010
[66] Keen, Andrew. "Web 1.0 + Web 2.0 = Web 3.0." TypePad.com (http://andrewkeen.typepad.com/the_great_seduction/2008/04/ web-10-web-20-w.html)
[67] Hempel, Jessi. "Web 2.0 is soo over. Welcome to Web 3.0." CNNMoney.CNN.com (http://money.cnn.com/2009/01/07/technology/hempel_threepointo.fortune/index.htm)
[68] "Metaverse Roadmap Overview" (http://www.metaverseroadmap.org/MetaverseRoadmapOverview.pdf) (PDF).. Retrieved 2011-02-06.
[69] http://openvideoalliance.org/about/?l=en
[70] Smart, John. 2010. "The Television Will Be Revolutionized: The iPad, Internet TV, and Web 3.0." (http://www.accelerating.org/articles/televisionwillberevolutionized.html)
[71] "Smart, John. 2003. "The Conversational Interface."" (http://www.accelerationwatch.com/lui.html). Accelerationwatch.com. 2008-11-14. . Retrieved 2011-02-06.
[72] HowStuffWorks "Web 3.0 Basics" (http://computer.howstuffworks.com/web-302.htm)
[73] STI International (http://www.sti2.org)

External links

- McKinsey & Company - Global Survey - McKinseyQuarterly.com (http://www.mckinseyquarterly.com/Information_Technology/Applications/ How_businesses_are_using_Web_20_A_McKinsey_Global_Survey_1913?gp=1), How businesses are using Web 2.0, June 2008
- UIC.edu (http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/issue/view/263/showToc), "Critical Perspectives on Web 2.0", Special issue of *First Monday*, 13(3), 2008.
- MacManus, Richard. Porter, Joshua. Digital-Web.com (http://www.digital-web.com/articles/web_2_for_designers/), "Web 2.0 for Designers", *Digital Web Magazine*, May 4, 2005.
- Graham Vickery, Sacha Wunsch-Vincent: OECD.org (http://www.oecd.org/document/40/0,3343,en_2649_201185_39428648_1_1_1_1,00.html), "Participative Web and User-Created Content: Web 2.0, Wikis and Social Networking; OECD, 2007

Web 1.0

Web 1.0, or web, refers to the first stage of the World Wide Web linking webpages with hyperlinks.

History

Hyperlinks between webpages began with the release of the WWW to the public in 1993,[1] and describe the Web before the "bursting of the Dot-com bubble" in 2001.

Since 2004, Web 2.0 has been the term used to describe social web, especially the current business models of sites on the World Wide Web.[2]

Characteristics

Terry Flew, in his 3rd Edition of *New Media* described what he believed to characterize the differences between Web 1.0 and Web 2.0:

> "move from personal websites to blogs and blog site aggregation, from publishing to participation, from web content as the outcome of large up-front investment to an ongoing and interactive process, and from content management systems to links based on tagging (folksonomy)".

Flew believed it to be the above factors that form the basic change in trends that resulted in the onset of the Web 2.0 "craze".[3]

The shift from Web 1.0 to Web 2.0 can be seen as a result of technological refinements, which included such adaptations as "broadband, improved browsers, and AJAX, to the rise of Flash application platforms and the mass development of widgetization, such as Flickr and YouTube badges". As well as such adjustments to the Internet, the shift from Web 1.0 to Web 2.0 is a direct result of the change in the behavior of those who use the World Wide Web. Web 1.0 trends included worries over privacy concerns resulting in a one-way flow of information, through websites which contained "read-only" material. Now, during Web 2.0, the use of the Web can be characterized as the decentralization of website content, which is now generated from the "bottom-up", with many users being contributors and producers of information, as well as the traditional consumers.

To take an example from above, Personal web pages were common in Web 1.0, and these consisted of mainly static pages hosted on free hosting services such as Geocities. Nowadays, dynamically generated blogs and social networking profiles, such as Myspace and Facebook, are more popular, allowing for readers to comment on posts in a way that was not available during Web 1.0.

At the Technet Summit in November 2006, Reed Hastings, founder and CEO of Netflix, stated a simple formula for defining the phases of the Web:

> Web 1.0 was dial-up, 50K average bandwidth, Web 2.0 is an average 1 megabit of bandwidth and Web 3.0 will be 10 megabits of bandwidth all the time, which will be the full video web, and that will feel like Web 3.0.
>
> —Reed Hastings

Web 1.0 design elements

Some design elements of a Web 1.0 site include:

- Static pages instead of dynamic user-generated content.[4]
- The use of framesets.
- The use of tables to position and align elements on a page. These were often used in combination with "spacer" GIFs (1x1 pixel transparent images in the GIF format.)
- Proprietary HTML extensions such as the <blink> and <marquee> tags introduced during the first browser war.
- Online guestbooks.
- GIF buttons, typically 88x31 pixels in size promoting web browsers and other products.[5]
- HTML forms sent via email. A user would fill in a form, and upon clicking submit their email client would attempt to send an email containing the form's details.[6]

References

[1] (Berners-Lee 2000) Tim Berners-Lee invented the world wide web.
[2] http://www.moveo.com/data/White_Papers/GettingThere_Dave_103006.pdf
[3] Flew, Terry (2008). *New Media: An Introduction* (3rd Edition ed.). Melbourne: Oxford University Press. p. 19.
[4] Web 1.0 defined - How stuff works (http://computer.howstuffworks.com/web-101.htm)
[5] Web 1.0 Revisited - Too many stupid buttons (http://www.complexify.com/buttons/)
[6] WEBalley - forms tutorial (http://www.weballey.nl/forms/emailform.html)

Search engine optimization

Search engine optimization (**SEO**) is the process of improving the visibility of a website or a web page in search engines via the "natural," or un-paid ("organic" or "algorithmic"), search results. In general, the earlier (or higher ranked on the search results page), and more frequently a site appears in the search results list, the more visitors it will receive from the search engine's users. SEO may target different kinds of search, including image search, local search, video search, academic search,[1] news search and industry-specific vertical search engines.

As an Internet marketing strategy, SEO considers how search engines work, what people search for, the actual search terms or keywords typed into search engines and which search engines are preferred by their targeted audience. Optimizing a website may involve editing its content and HTML and associated coding to both increase its relevance to specific keywords and to remove barriers to the indexing activities of search engines. Promoting a site to increase the number of backlinks, or inbound links, is another SEO tactic.

The acronym "SEOs" can refer to "search engine optimizers," a term adopted by an industry of consultants who carry out optimization projects on behalf of clients, and by employees who perform SEO services in-house. Search engine optimizers may offer SEO as a stand-alone service or as a part of a broader marketing campaign. Because effective SEO may require changes to the HTML source code of a site and site content, SEO tactics may be incorporated into website development and design. The term "search engine friendly" may be used to describe website designs, menus, content management systems, images, videos, shopping carts, and other elements that have been optimized for the purpose of search engine exposure.

History

Webmasters and content providers began optimizing sites for search engines in the mid-1990s, as the first search engines were cataloging the early Web. Initially, all webmasters needed to do was to submit the address of a page, or URL, to the various engines which would send a "spider" to "crawl" that page, extract links to other pages from it, and return information found on the page to be indexed.[2] The process involves a search engine spider downloading a page and storing it on the search engine's own server, where a second program, known as an indexer, extracts various information about the page, such as the words it contains and where these are located, as well as any weight for specific words, and all links the page contains, which are then placed into a scheduler for crawling at a later date.

Site owners started to recognize the value of having their sites highly ranked and visible in search engine results, creating an opportunity for both white hat and black hat SEO practitioners. According to industry analyst Danny Sullivan, the phrase "search engine optimization" probably came into use in 1997.[3] The first documented use of the term Search Engine Optimization was John Audette and his company Multimedia Marketing Group as documented by a web page from the MMG site from August, 1997.[4]

Early versions of search algorithms relied on webmaster-provided information such as the keyword meta tag, or index files in engines like ALIWEB. Meta tags provide a guide to each page's content. Using meta data to index pages was found to be less than reliable, however, because the webmaster's choice of keywords in the meta tag could potentially be an inaccurate representation of the site's actual content. Inaccurate, incomplete, and inconsistent data in meta tags could and did cause pages to rank for irrelevant searches.[5] Web content providers also manipulated a number of attributes within the HTML source of a page in an attempt to rank well in search engines.[6]

By relying so much on factors such as keyword density which were exclusively within a webmaster's control, early search engines suffered from abuse and ranking manipulation. To provide better results to their users, search engines had to adapt to ensure their results pages showed the most relevant search results, rather than unrelated pages stuffed with numerous keywords by unscrupulous webmasters. Since the success and popularity of a search engine is determined by its ability to produce the most relevant results to any given search, allowing those results to be false would turn users to find other search sources. Search engines responded by developing more complex ranking algorithms, taking into account additional factors that were more difficult for webmasters to manipulate.

Graduate students at Stanford University, Larry Page and Sergey Brin, developed "Backrub," a search engine that relied on a mathematical algorithm to rate the prominence of web pages. The number calculated by the algorithm, PageRank, is a function of the quantity and strength of inbound links.[7] PageRank estimates the likelihood that a given page will be reached by a web user who randomly surfs the web, and follows links from one page to another. In effect, this means that some links are stronger than others, as a higher PageRank page is more likely to be reached by the random surfer.

Page and Brin founded Google in 1998. Google attracted a loyal following among the growing number of Internet users, who liked its simple design.[8] Off-page factors (such as PageRank and hyperlink analysis) were considered as well as on-page factors (such as keyword frequency, meta tags, headings, links and site structure) to enable Google to avoid the kind of manipulation seen in search engines that only considered on-page factors for their rankings. Although PageRank was more difficult to game, webmasters had already developed link building tools and schemes to influence the Inktomi search engine, and these methods proved similarly applicable to gaming PageRank. Many sites focused on exchanging, buying, and selling links, often on a massive scale. Some of these schemes, or link farms, involved the creation of thousands of sites for the sole purpose of link spamming.[9]

By 2004, search engines had incorporated a wide range of undisclosed factors in their ranking algorithms to reduce the impact of link manipulation. Google says it ranks sites using more than 200 different signals.[10] The leading search engines, Google, Bing, and Yahoo, do not disclose the algorithms they use to rank pages. SEO service providers, such as Rand Fishkin, Barry Schwartz, Aaron Wall and Jill Whalen, have studied different approaches to search engine optimization, and have published their opinions in online forums and blogs.[11][12] SEO practitioners may also study patents held by various search engines to gain insight into the algorithms.[13]

In 2005, Google began personalizing search results for each user. Depending on their history of previous searches, Google crafted results for logged in users.[14] In 2008, Bruce Clay said that "ranking is dead" because of personalized search. It would become meaningless to discuss how a website ranked, because its rank would potentially be different for each user and each search.[15]

In 2007, Google announced a campaign against paid links that transfer PageRank.[16] On June 15, 2009, Google disclosed that they had taken measures to mitigate the effects of PageRank sculpting by use of the nofollow attribute on links. Matt Cutts, a well-known software engineer at Google, announced that Google Bot would no longer treat nofollowed links in the same way, in order to prevent SEO service providers from using nofollow for PageRank sculpting.[17] As a result of this change the usage of nofollow leads to evaporation of pagerank. In order to avoid the above, SEO engineers developed alternative techniques that replace nofollowed tags with obfuscated Javascript and thus permit PageRank sculpting. Additionally several solutions have been suggested that include the usage of iframes, Flash and Javascript.[18]

In December 2009, Google announced it would be using the web search history of all its users in order to populate search results.[19]

Google Instant, real-time-search, was introduced in late 2009 in an attempt to make search results more timely and relevant. Historically site administrators have spent months or even years optimizing a website to increase search rankings. With the growth in popularity of social media sites and blogs the leading engines made changes to their algorithms to allow fresh content to rank quickly within the search results.[20]

Relationship with search engines

By 1997, search engines recognized that webmasters were making efforts to rank well in their search engines, and that some webmasters were even manipulating their rankings in search results by stuffing pages with excessive or irrelevant keywords. Early search engines, such as Altavista and Infoseek, adjusted their algorithms in an effort to prevent webmasters from manipulating rankings.[21]


Yahoo and Google offices

Due to the high marketing value of targeted search results, there is potential for an adversarial relationship between search engines and SEO service providers. In 2005, an annual conference, AIRWeb, Adversarial Information Retrieval on the Web,[22] was created to discuss and minimize the damaging effects of aggressive web content providers.

Companies that employ overly aggressive techniques can get their client websites banned from the search results. In 2005, the Wall Street Journal reported on a company, Traffic Power, which allegedly used high-risk techniques and failed to disclose those risks to its clients.[23] Wired magazine reported that the same company sued blogger and SEO Aaron Wall for writing about the ban.[24] Google's Matt Cutts later confirmed that Google did in fact ban Traffic Power and some of its clients.[25]

Some search engines have also reached out to the SEO industry, and are frequent sponsors and guests at SEO conferences, chats, and seminars. Major search engines provide information and guidelines to help with site optimization.[26][27] Google has a Sitemaps program[28] to help webmasters learn if Google is having any problems indexing their website and also provides data on Google traffic to the website. Bing Toolbox provides a way from webmasters to submit a sitemap and web feeds, allowing users to determine the crawl rate, and how many pages have been indexed by their search engine.

Methods

Getting indexed

The leading search engines, such as Google, Bing and Yahoo!, use crawlers to find pages for their algorithmic search results. Pages that are linked from other search engine indexed pages do not need to be submitted because they are found automatically. Some search engines, notably Yahoo!, operate a paid submission service that guarantee crawling for either a set fee or cost per click.[29] Such programs usually guarantee inclusion in the database, but do not guarantee specific ranking within the search results.[30] Two major directories, the Yahoo Directory and the Open Directory Project both require manual submission and human editorial review.[31] Google offers Google Webmaster Tools, for which an XML Sitemap feed can be created and submitted for free to ensure that all pages are found, especially pages that aren't discoverable by automatically following links.[32]

Search engine crawlers may look at a number of different factors when crawling a site. Not every page is indexed by the search engines. Distance of pages from the root directory of a site may also be a factor in whether or not pages get crawled.[33]

Preventing crawling

To avoid undesirable content in the search indexes, webmasters can instruct spiders not to crawl certain files or directories through the standard robots.txt file in the root directory of the domain. Additionally, a page can be explicitly excluded from a search engine's database by using a meta tag specific to robots. When a search engine visits a site, the robots.txt located in the root directory is the first file crawled. The robots.txt file is then parsed, and will instruct the robot as to which pages are not to be crawled. As a search engine crawler may keep a cached copy of this file, it may on occasion crawl pages a webmaster does not wish crawled. Pages typically prevented from being crawled include login specific pages such as shopping carts and user-specific content such as search results from internal searches. In March 2007, Google warned webmasters that they should prevent indexing of internal search results because those pages are considered search spam.[34]

Increasing prominence

A variety of methods can increase the prominence of a webpage within the search results. Cross linking between pages of the same website to provide more links to most important pages may improve its visibility.[35] Writing content that includes frequently searched keyword phrase, so as to be relevant to a wide variety of search queries will tend to increase traffic.[35] Updating content so as to keep search engines crawling back frequently can give additional weight to a site. Adding relevant keywords to a web page's meta data, including the title tag and meta description, will tend to improve the relevancy of a site's search listings, thus increasing traffic. URL normalization of web pages accessible via multiple urls, using the "canonical" meta tag[36] or via 301 redirects can help make sure links to different versions of the url all count towards the page's link popularity score.

Image search optimization

Image search optimization is the process of organizing the content of a webpage to increase relevance to a specific keyword on image search engines. Like search engine optimization, the aim is to achieve a higher organic search listing and thus increasing the volume of traffic from search engines.

Image search optimization techniques can be viewed as a subset of search engine optimization techniques that focuses on gaining high ranks on image search engine results.

Unlike normal SEO process, there is not much to do for ISO. Making high quality images accessible to search engines and providing some description about images is almost all that can be done for ISO.

White hat versus black hat

SEO techniques can be classified into two broad categories: techniques that search engines recommend as part of good design, and those techniques of which search engines do not approve. The search engines attempt to minimize the effect of the latter, among them spamdexing. Industry commentators have classified these methods, and the practitioners who employ them, as either white hat SEO, or black hat SEO.[37] White hats tend to produce results that last a long time, whereas black hats anticipate that their sites may eventually be banned either temporarily or permanently once the search engines discover what they are doing.[38]

An SEO technique is considered white hat if it conforms to the search engines' guidelines and involves no deception. As the search engine guidelines[26][27][39] are not written as a series of rules or commandments, this is an important distinction to note. White hat SEO is not just about following guidelines, but is about ensuring that the content a search engine indexes and subsequently ranks is the same content a user will see. White hat advice is generally summed up as creating content for users, not for search engines, and then making that content easily accessible to the spiders, rather than attempting to trick the algorithm from its intended purpose. White hat SEO is in many ways similar to web development that promotes accessibility,[40] although the two are not identical.

Black hat SEO attempts to improve rankings in ways that are disapproved of by the search engines, or involve deception. One black hat technique uses text that is hidden, either as text colored similar to the background, in an invisible div, or positioned off screen. Another method gives a different page depending on whether the page is being requested by a human visitor or a search engine, a technique known as cloaking.

Search engines may penalize sites they discover using black hat methods, either by reducing their rankings or eliminating their listings from their databases altogether. Such penalties can be applied either automatically by the search engines' algorithms, or by a manual site review. One infamous example was the February 2006 Google removal of both BMW Germany and Ricoh Germany for use of deceptive practices.[41] Both companies, however, quickly apologized, fixed the offending pages, and were restored to Google's list.[42]

As a marketing strategy

SEO is not an appropriate strategy for every website, and other Internet marketing strategies can be more effective, depending on the site operator's goals.[43] A successful Internet marketing campaign may also depend upon building high quality web pages to engage and persuade, setting up analytics programs to enable site owners to measure results, and improving a site's conversion rate.[44]

SEO may generate an adequate return on investment. However, search engines are not paid for organic search traffic, their algorithms change, and there are no guarantees of continued referrals. Due to this lack of guarantees and certainty, a business that relies heavily on search engine traffic can suffer major losses if the search engines stop sending visitors.[45] Search engines can change their algorithms, impacting a website's placement, possibly resulting in a serious loss of traffic. According to Google's CEO, Erick Schmidt, in 2010, Google made over 500 algorithm changes - almost 1.5 per day.[46] It is considered wise business practice for website operators to liberate themselves from dependence on search engine traffic.[47] Seomoz.org has suggested that "search marketers, in a twist of irony, receive a very small share of their traffic from search engines." Instead, their main sources of traffic are links from other websites.[48]

International markets

Optimization techniques are highly tuned to the dominant search engines in the target market. The search engines' market shares vary from market to market, as does competition. In 2003, Danny Sullivan stated that Google represented about 75% of all searches.[49] In markets outside the United States, Google's share is often larger, and Google remains the dominant search engine worldwide as of 2007.[50] As of 2006, Google had an 85-90% market share in Germany.[51] While there were hundreds of SEO firms in the US at that time, there were only about five in

Germany.[51] As of June 2008, the market share of Google in the UK was close to 90% according to Hitwise.[52] That market share is achieved in a number of countries.

As of 2009, there are only a few large markets where Google is not the leading search engine. In most cases, when Google is not leading in a given market, it is lagging behind a local player. The most notable markets where this is the case are China, Japan, South Korea, Russia and the Czech Republic where respectively Baidu, Yahoo! Japan, Naver, Yandex and Seznam are market leaders.

Successful search optimization for international markets may require professional translation of web pages, registration of a domain name with a top level domain in the target market, and web hosting that provides a local IP address. Otherwise, the fundamental elements of search optimization are essentially the same, regardless of language.[51]

Legal precedents

On October 17, 2002, SearchKing filed suit in the United States District Court, Western District of Oklahoma, against the search engine Google. SearchKing's claim was that Google's tactics to prevent spamdexing constituted a tortious interference with contractual relations. On May 27, 2003, the court granted Google's motion to dismiss the complaint because SearchKing "failed to state a claim upon which relief may be granted."[53][54]

In March 2006, KinderStart filed a lawsuit against Google over search engine rankings. Kinderstart's website was removed from Google's index prior to the lawsuit and the amount of traffic to the site dropped by 70%. On March 16, 2007 the United States District Court for the Northern District of California (San Jose Division) dismissed KinderStart's complaint without leave to amend, and partially granted Google's motion for Rule 11 sanctions against KinderStart's attorney, requiring him to pay part of Google's legal expenses.[55][56]

Notes

[1] Beel, Jöran and Gipp, Bela and Wilde, Erik (2010). "Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar and Co." (http://www.sciplore.org/publications/2010-ASEO--preprint.pdf). Journal of Scholarly Publishing. pp. 176–190. . Retrieved 2010-04-18.

[2] Brian Pinkerton. "Finding What People Want: Experiences with the WebCrawler" (http://www.webir.org/resources/phd/pinkerton_2000. pdf) (PDF). The Second International WWW Conference Chicago, USA, October 17–20, 1994. . Retrieved 2007-05-07.

[3] Danny Sullivan (June 14, 2004). "Who Invented the Term "Search Engine Optimization"?" (http://forums.searchenginewatch.com/showpost.php?p=2119&postcount=10). Search Engine Watch. . Retrieved 2007-05-14. See Google groups thread (http://groups.google. com/group/alt.current-events.net-abuse.spam/browse_thread/thread/6fee2777dc17b8ab/3858bff94e56aff3?lnk=st&q="search+engine+ optimization"&rnum=1#3858bff94e56aff3).

[4] (Document Number 19970801004204) "Documentation of Who Invented SEO at the Internet Way Back Machine" (http://web.archive.org/web/19970801004204/www.mmgco.com/campaign.html). Internet Way Back Machine. Archived from (Document Number 19970801004204) the original (http://www.mmgco.com/campaign.html) on 1997-08-01. (Document Number 19970801004204).

[5] Cory Doctorow (August 26, 2001). "Metacrap: Putting the torch to seven straw-men of the meta-utopia" (http://web.archive.org/web/ 20070409062313/http://www.e-learningguru.com/articles/metacrap.htm). e-LearningGuru. Archived from the original (http://www. e-learningguru.com/articles/metacrap.htm) on 2007-04-09. . Retrieved 2007-05-08.

[6] Pringle, G., Allison, L., and Dowe, D. (April 1998). "What is a tall poppy among web pages?" (http://www.csse.monash.edu.au/~lloyd/tilde/InterNet/Search/1998_WWW7.html). Proc. 7th Int. World Wide Web Conference. . Retrieved 2007-05-08.

[7] Brin, Sergey and Page, Larry (1998). "The Anatomy of a Large-Scale Hypertextual Web Search Engine" (http://www-db.stanford.edu/~backrub/google.html). Proceedings of the seventh international conference on World Wide Web. pp. 107–117. . Retrieved 2007-05-08.

[8] Thompson, Bill (December 19, 2003). "Is Google good for you?" (http://news.bbc.co.uk/1/hi/technology/3334531.stm). BBC News. . Retrieved 2007-05-16.

[9] Zoltan Gyongyi and Hector Garcia-Molina (2005). "Link Spam Alliances" (http://infolab.stanford.edu/~zoltan/publications/ gyongyi2005link.pdf) (PDF). Proceedings of the 31st VLDB Conference, Trondheim, Norway. . Retrieved 2007-05-09.

[10] Hansell, Saul (June 3, 2007). "Google Keeps Tweaking Its Search Engine" (http://www.nytimes.com/2007/06/03/business/yourmoney/ 03google.html). New York Times. . Retrieved 2007-06-06.

[11] Danny Sullivan (September 29, 2005). "Rundown On Search Ranking Factors" (http://blog.searchenginewatch.com/blog/ 050929-072711). Search Engine Watch. . Retrieved 2007-05-08.

[12] "Search Engine Ranking Factors V2" (http://www.seomoz.org/article/search-ranking-factors). SEOmoz.org. April 2, 2007. . Retrieved 2007-05-14.
[13] Christine Churchill (November 23, 2005). "Understanding Search Engine Patents" (http://searchenginewatch.com/showPage. html?page=3564261). Search Engine Watch. . Retrieved 2007-05-08.
[14] "Google Personalized Search Leaves Google Labs - Search Engine Watch (SEW)" (http://searchenginewatch.com/3563036). searchenginewatch.com. . Retrieved 2009-09-05.
[15] "Will Personal Search Turn SEO On Its Ear?" (http://www.webpronews.com/topnews/2008/11/17/seo-about-to-get-turned-on-its-ear). www.webpronews.com. . Retrieved 2009-09-05.
[16] "8 Things We Learned About Google PageRank" (http://www.searchenginejournal.com/8-things-we-learned-about-google-pagerank/ 5897/). www.searchenginejournal.com. . Retrieved 2009-08-17.
[17] "PageRank sculpting" (http://www.mattcutts.com/blog/pagerank-sculpting/). Matt Cutts. . Retrieved 2010-01-12.
[18] "Google Loses "Backwards Compatibility" On Paid Link Blocking & PageRank Sculpting" (http://searchengineland.com/google-loses-backwards-compatibility-on-paid-link-blocking-pagerank-sculpting-20408). searchengineland.com. . Retrieved 2009-08-17.
[19] "Personalized Search for everyone" (http://googleblog.blogspot.com/2009/12/personalized-search-for-everyone.html). Google. . Retrieved 2009-12-14.
[20] "Relevance Meets Real Time Web" (http://googleblog.blogspot.com/2009/12/relevance-meets-real-time-web.html). Google Blog. .
[21] Laurie J. Flynn (November 11, 1996). "Desperately Seeking Surfers" (http://query.nytimes.com/gst/fullpage. html?res=940DE0DF123BF932A25752C1A960958260). New York Times. . Retrieved 2007-05-09.
[22] "AIRWeb" (http://airweb.cse.lehigh.edu/). Adversarial Information Retrieval on the Web, annual conference. . Retrieved 2007-05-09.
[23] David Kesmodel (September 22, 2005). "Sites Get Dropped by Search Engines After Trying to 'Optimize' Rankings" (http://online.wsj. com/article/SB112714166978744925.html?apl=y&r=947596). *Wall Street Journal*. . Retrieved 2008-07-30.
[24] Adam L. Penenberg (September 8, 2005). "Legal Showdown in Search Fracas" (http://www.wired.com/news/culture/0,1284,68799,00. html). Wired Magazine. . Retrieved 2007-05-09.
[25] Matt Cutts (February 2, 2006). "Confirming a penalty" (http://www.mattcutts.com/blog/confirming-a-penalty/). mattcutts.com/blog. . Retrieved 2007-05-09.
[26] "Google's Guidelines on Site Design" (http://www.google.com/webmasters/guidelines.html). google.com. . Retrieved 2007-04-18.
[27] "Guidelines for Successful Indexing" (http://onlinehelp.microsoft.com/en-us/bing/hh204434.aspx). bing.com. . Retrieved 2011-09-07.
[28] "Google Webmaster Tools" (http://web.archive.org/web/20071102153746/http://www.google.com/webmasters/sitemaps/login). google.com. Archived from the original (http://www.google.com/webmasters/sitemaps/login) on November 2, 2007. . Retrieved 2007-05-09.
[29] "Submitting To Search Crawlers: Google, Yahoo, Ask & Microsoft's Live Search" (http://searchenginewatch.com/showPage. html?page=2167871). Search Engine Watch. 2007-03-12. . Retrieved 2007-05-15.
[30] "Search Submit" (http://searchmarketing.yahoo.com/srchsb/index.php). searchmarketing.yahoo.com. . Retrieved 2007-05-09.
[31] "Submitting To Directories: Yahoo & The Open Directory" (http://searchenginewatch.com/showPage.html?page=2167881). Search Engine Watch. 2007-03-12. . Retrieved 2007-05-15.
[32] "What is a Sitemap file and why should I have one?" (http://www.google.com/support/webmasters/bin/answer.py?answer=40318& topic=8514). google.com. . Retrieved 2007-03-19.
[33] Cho, J., Garcia-Molina, H. (1998). "Efficient crawling through URL ordering" (http://dbpubs.stanford.edu:8090/pub/1998-51). Proceedings of the seventh conference on World Wide Web, Brisbane, Australia. . Retrieved 2007-05-09.
[34] "Newspapers Amok! New York Times Spamming Google? LA Times Hijacking Cars.com?" (http://searchengineland.com/ 070508-165231.php). Search Engine Land. May 8, 2007. . Retrieved 2007-05-09.
[35] "The Most Important SEO Strategy - ClickZ" (http://www.clickz.com/3623372). www.clickz.com. . Retrieved 2010-04-18.
[36] "Bing - Partnering to help solve duplicate content issues - Webmaster Blog - Bing Community" (http://www.bing.com/community/blogs/ webmaster/archive/2009/02/12/partnering-to-help-solve-duplicate-content-issues.aspx). www.bing.com. . Retrieved 2009-10-30.
[37] Andrew Goodman. "Search Engine Showdown: Black hats vs. White hats at SES" (http://searchenginewatch.com/showPage. html?page=3483941). SearchEngineWatch. . Retrieved 2007-05-09.
[38] Jill Whalen (November 16, 2004). "Black Hat/White Hat Search Engine Optimization" (http://www.searchengineguide.com/whalen/ 2004/1116_jw1.html). searchengineguide.com. . Retrieved 2007-05-09.
[39] "What's an SEO? Does Google recommend working with companies that offer to make my site Google-friendly?" (http://www.google. com/webmasters/seo.html). google.com. . Retrieved 2007-04-18.
[40] Andy Hagans (November 8, 2005). "High Accessibility Is Effective Search Engine Optimization" (http://alistapart.com/articles/ accessibilityseo). A List Apart. . Retrieved 2007-05-09.
[41] Matt Cutts (February 4, 2006). "Ramping up on international webspam" (http://www.mattcutts.com/blog/ ramping-up-on-international-webspam/). mattcutts.com/blog. . Retrieved 2007-05-09.
[42] Matt Cutts (February 7, 2006). "Recent reinclusions" (http://www.mattcutts.com/blog/recent-reinclusions/). mattcutts.com/blog. . Retrieved 2007-05-09.
[43] "What SEO Isn't" (http://blog.v7n.com/2006/06/24/what-seo-isnt/). blog.v7n.com. June 24, 2006. . Retrieved 2007-05-16.
[44] Melissa Burdon (March 13, 2007). "The Battle Between Search Engine Optimization and Conversion: Who Wins?" (http://www. grokdotcom.com/2007/03/13/the-battle-between-search-engine-optimization-and-conversion-who-wins/). Grok.com. . Retrieved

2007-05-09.

[45] Andy Greenberg (April 30, 2007). "Condemned To Google Hell" (http://www.forbes.com/technology/2007/04/29/ sanar-google-skyfacet-tech-cx_ag_0430googhell.html?partner=rss). Forbes. . Retrieved 2007-05-09.

[46] Matt McGee (September 21, 2011). "Schmidt's testimony reveals how Google tests alorithm changes" (http://searchengineland.com/ 13000-precision-evaluations-schmidts-testimony-reveals-how-google-tests-algorithm-changes-93740). .

[47] Jakob Nielsen (January 9, 2006). "Search Engines as Leeches on the Web" (http://www.useit.com/alertbox/search_engines.html). useit.com. . Retrieved 2007-05-14.

[48] "A survey of 25 blogs in the search space comparing external metrics to visitor tracking data" (http://www.seomoz.org/article/ search-blog-stats#4). seomoz.org. . Retrieved 2007-05-31.

[49] Graham, Jefferson (2003-08-26). "The search engine that could" (http://www.usatoday.com/tech/news/2003-08-25-google_x.htm). USA Today. . Retrieved 2007-05-15.

[50] Greg Jarboe (2007-02-22). "Stats Show Google Dominates the International Search Landscape" (http://searchenginewatch.com/showPage.html?page=3625072). Search Engine Watch. . Retrieved 2007-05-15.

[51] Mike Grehan (April 3, 2006). "Search Engine Optimizing for Europe" (http://www.clickz.com/showPage.html?page=3595926). Click. . Retrieved 2007-05-14.

[52] Jack Schofield (2008-06-10). "Google UK closes in on 90% market share" (http://www.guardian.co.uk/technology/blog/2008/jun/10/ googleukclosesinon90mark). London: Guardian. . Retrieved 2008-06-10.

[53] "Search King, Inc. v. Google Technology, Inc., CIV-02-1457-M" (http://www.docstoc.com/docs/618281/Order-(Granting-Googles-Motion-to-Dismiss-Search-Kings-Complaint)) (PDF). docstoc.com. May 27, 2003. . Retrieved 2008-05-23.

[54] Stefanie Olsen (May 30, 2003). "Judge dismisses suit against Google" (http://news.com.com/2100-1032_3-1011740.html). CNET. . Retrieved 2007-05-10.

[55] "Technology & Marketing Law Blog: KinderStart v. Google Dismissed—With Sanctions Against KinderStart's Counsel" (http://blog.ericgoldman.org/archives/2007/03/kinderstart_v_g_2.htm). blog.ericgoldman.org. . Retrieved 2008-06-23.

[56] "Technology & Marketing Law Blog: Google Sued Over Rankings—KinderStart.com v. Google" (http://blog.ericgoldman.org/archives/ 2006/03/google_sued_ove.htm). blog.ericgoldman.org. . Retrieved 2008-06-23.

External links

- Google Webmaster Guidelines (http://www.google.com/support/webmasters/bin/answer.py?hl=en& answer=35769)
- Yahoo! Webmaster Guidelines (http://help.yahoo.com/l/us/yahoo/search/basics/basics-18.html)
- " The Dirty Little Secrets of Search (http://www.nytimes.com/2011/02/13/business/13search.html)," article in The New York Times (February 12, 2011)
- Google I/O 2010 - SEO site advice from the experts (https://www.youtube.com/watch?v=7Hk5uVv8JpM) on YouTube - Technical tutorial on search engine optimization, given at Google I/O 2010.

Search engine

A **web search engine** is designed to search for information on the World Wide Web and FTP servers. The search results are generally presented in a list of results often referred to as SERPS, or "search engine results pages". The information may consist of web pages, images, information and other types of files. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain real-time information by running an algorithm on a web crawler.

History

Timeline (full list)		
Year	**Engine**	**Current status**
1993	W3Catalog	Inactive
	Aliweb	Inactive
1994	WebCrawler	Active, Aggregator
	Go.com	Active, Yahoo Search
	Lycos	Active
1995	AltaVista	Inactive(URLredirectedtoYahoo!)
	Daum	Active
	Magellan	Inactive
	Excite	Active
	SAPO	Active
	Yahoo!	Active, Launched as a directory
1996	Dogpile	Active, Aggregator
	Inktomi	Acquired by Yahoo!
	HotBot	Active (lycos.com)
	Ask Jeeves	Active(ask.com,Jeeveswentaway)
1997	Northern Light	Inactive
	Yandex	Active
1998	Google	Active
	MSN Search	Active as Bing
1999	AlltheWeb	Inactive(URLredirectedtoYahoo!)
	GenieKnows	Active, rebranded Yellowee.com
	Naver	Active
	Teoma	Active
	Vivisimo	Inactive
2000	Baidu	Active
	Exalead	Acquired by Dassault Systèmes
2002	Inktomi	Acquired by Yahoo!
2003	Info.com	Active

2004	Yahoo! Search	Active, Launched own web search (see Yahoo! Directory, 1995)
	A9.com	Inactive
	Sogou	Active
2005	AOL Search	Active
	Ask.com	Active
	GoodSearch	Active
	SearchMe	Closed
2006	wikiseek	Inactive
	Quaero	Active
	Ask.com	Active
	Live Search	Active as Bing, Launched as rebranded MSN Search
	ChaCha	Active
	Guruji.com	Active
2007	wikiseek	Inactive
	Sproose	Inactive
	Wikia Search	Inactive
	Blackle.com	Active
2008	Powerset	Inactive (redirects to Bing)
	PicoIlator	Inactive
	Viewzi	Inactive
	Boogami	Inactive
	LeapFish	Inactive
	Forestle	Inactive (redirects to Ecosia)
	VADLO	Active
	Duck Duck Go	Active, Aggregator
2009	Bing	Active, Launched as rebranded Live Search
	Yebol	Active
	Megafore	Active
	Mugurdy	Inactive due to a lack of funding
	Goby	Active
2010	Black Google Mobile	Active
	Blekko	Active
	Cuil	Inactive
	Yandex	Active, Launched global (English) search
	Yummly	Active
2011	Interred	Active
2012	Volunia	Active , only *Power User*

During the early development of the web, there was a list of webservers edited by Tim Berners-Lee and hosted on the CERN webserver. One historical snapshot from 1992 remains.[1] As more webservers went online the central list could not keep up. On the NCSA site new servers were announced under the title "What's New!"[2]

The very first tool used for searching on the Internet was Archie.[3] The name stands for "archive" without the "v". It was created in 1990 by Alan Emtage, Bill Heelan and J. Peter Deutsch, computer science students at McGill University in Montreal. The program downloaded the directory listings of all the files located on public anonymous FTP (File Transfer Protocol) sites, creating a searchable database of file names; however, Archie did not index the contents of these sites since the amount of data was so limited it could be readily searched manually.

The rise of Gopher (created in 1991 by Mark McCahill at the University of Minnesota) led to two new search programs, Veronica and Jughead. Like Archie, they searched the file names and titles stored in Gopher index systems. Veronica (*V*ery *E*asy *R*odent-*O*riented *N*et-wide *I*ndex to *C*omputerized *A*rchives) provided a keyword search of most Gopher menu titles in the entire Gopher listings. Jughead (*J*onzy's *U*niversal *G*opher *H*ierarchy *E*xcavation *A*nd *D*isplay) was a tool for obtaining menu information from specific Gopher servers. While the name of the search engine "Archie" was not a reference to the Archie comic book series, "Veronica" and "Jughead" are characters in the series, thus referencing their predecessor.

In the summer of 1993, no search engine existed yet for the web, though numerous specialized catalogues were maintained by hand. Oscar Nierstrasz at the University of Geneva wrote a series of Perl scripts that would periodically mirror these pages and rewrite them into a standard format which formed the basis for W3Catalog, the web's first primitive search engine, released on September 2, 1993.[4]

In June 1993, Matthew Gray, then at MIT, produced what was probably the first web robot, the Perl-based World Wide Web Wanderer, and used it to generate an index called 'Wandex'. The purpose of the Wanderer was to measure the size of the World Wide Web, which it did until late 1995. The web's second search engine Aliweb appeared in November 1993. Aliweb did not use a web robot, but instead depended on being notified by website administrators of the existence at each site of an index file in a particular format.

JumpStation (released in December 1993[5]) used a web robot to find web pages and to build its index, and used a web form as the interface to its query program. It was thus the first WWW resource-discovery tool to combine the three essential features of a web search engine (crawling, indexing, and searching) as described below. Because of the limited resources available on the platform on which it ran, its indexing and hence searching were limited to the titles and headings found in the web pages the crawler encountered.

One of the first "full text" crawler-based search engines was WebCrawler, which came out in 1994. Unlike its predecessors, it let users search for any word in any webpage, which has become the standard for all major search engines since. It was also the first one to be widely known by the public. Also in 1994, Lycos (which started at Carnegie Mellon University) was launched and became a major commercial endeavor.

Soon after, many search engines appeared and vied for popularity. These included Magellan, Excite, Infoseek, Inktomi, Northern Light, and AltaVista. Yahoo! was among the most popular ways for people to find web pages of interest, but its search function operated on its web directory, rather than full-text copies of web pages. Information seekers could also browse the directory instead of doing a keyword-based search.

In 1996, Netscape was looking to give a single search engine an exclusive deal to be the featured search engine on Netscape's web browser. There was so much interest that instead a deal was struck with Netscape by five of the major search engines, where for $5 million per year each search engine would be in rotation on the Netscape search engine page. The five engines were Yahoo!, Magellan, Lycos, Infoseek, and Excite.[6][7]

Search engines were also known as some of the brightest stars in the Internet investing frenzy that occurred in the late 1990s.[8] Several companies entered the market spectacularly, receiving record gains during their initial public offerings. Some have taken down their public search engine, and are marketing enterprise-only editions, such as Northern Light. Many search engine companies were caught up in the dot-com bubble, a speculation-driven market

boom that peaked in 1999 and ended in 2001.

Around 2000, Google's search engine rose to prominence. The company achieved better results for many searches with an innovation called PageRank. This iterative algorithm ranks web pages based on the number and PageRank of other web sites and pages that link there, on the premise that good or desirable pages are linked to more than others. Google also maintained a minimalist interface to its search engine. In contrast, many of its competitors embedded a search engine in a web portal.

By 2000, Yahoo! was providing search services based on Inktomi's search engine. Yahoo! acquired Inktomi in2002, and Overture (which owned AlltheWeb and AltaVista) in 2003. Yahoo! switched to Google's search engine until 2004, when it launched its own search engine based on thecombined technologies of its acquisitions.

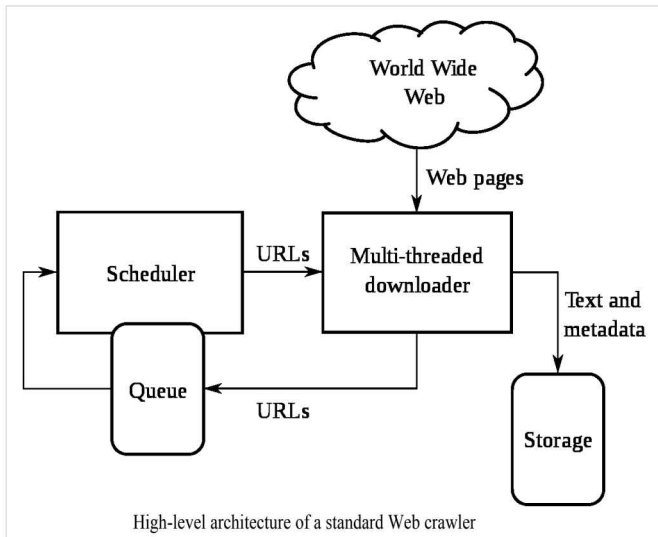
Microsoft first launched MSN Search in the fall of 1998 using search results from Inktomi. In early 1999 the site began to display listings from Looksmart blended with results from Inktomi except for a short time in 1999 when results from AltaVista were used instead. In 2004, Microsoft began a transition to its own search technology, powered by its own web crawler (called msnbot).

Microsoft's rebranded search engine, Bing, was launched on June 1, 2009. On July 29, 2009, Yahoo! and Microsoft finalized a deal in which Yahoo! Search would be powered by Microsoft Bing technology.

How web search engines work

A search engine operates in the following order:

1. Web crawling
2. Indexing
3. Searching



High-level architecture of a standard Web crawler

Web search engines work by storing information about many web pages, which they retrieve from the HTML itself. These pages are retrieved by a Web crawler (sometimes also known as aspider) — an automated Web browser which follows every link on the site. Exclusions can be made by the use of robots.txt. The contents of each page are then analyzed to determine how it should be indexed (for example, words are extracted from the titles, headings, or special fields called meta tags). Data about web pages are stored in an index database for use in later queries. A query can be a single word. The purpose of an index is to allow information to be found as quickly as possible. Some search engines, such as Google, store all or part of the source page (referred to as a cache) as well as information about the web pages, whereas others, such as AltaVista, store every word of every page they find. This cached page always holds the actual search text since it is the one that was actually indexed, so it can be very useful when the content of the current page has been updated and the search terms are no longer in it. This problem might be considered to be a mild form of linkrot, and Google's handling of it increases usability by satisfying user expectations that the search terms will be on the returned webpage. This satisfies the principle of least astonishment since the user normally expects the search terms to be on the returned pages. Increased search relevance makes these cached pages very useful, even beyond the fact that they may contain data that may no longer be available elsewhere.

When a user enters a query into a search engine (typically by using keywords), the engine examines its index and provides a listing of best-matching web pages according to its criteria, usually with a short summary containing the document's title and sometimes parts of the text. The index is built from the information stored with the data and the

method by which the information is indexed. Unfortunately, there are currently no known public search engines that allow documents to be searched by date. Most search engines support the use of the boolean operators AND, OR and NOT to further specify the search query. Boolean operators are for literal searches that allow the user to refine and extend the terms of the search. The engine looks for the words or phrases exactly as entered. Some search engines provide an advanced feature called proximity search which allows users to define the distance between keywords. There is also concept-based searching where the research involves using statistical analysis on pages containing the words or phrases you search for. As well, natural language queries allow the user to type a question in the same form one would ask it to a human. A site like this would be ask.com.

The usefulness of a search engine depends on the relevance of the **result set** it gives back. While there may be millions of web pages that include a particular word or phrase, some pages may be more relevant, popular, or authoritative than others. Most search engines employ methods to rank the results to provide the "best" results first. How a search engine decides which pages are the best matches, and what order the results should be shown in, varies widely from one engine to another. The methods also change over time as Internet usage changes and new techniques evolve. There are two main types of search engine that have evolved: one is a system of predefined and hierarchically ordered keywords that humans have programmed extensively. The other is a system that generates an "inverted index" by analyzing texts it locates. This second form relies much more heavily on the computer itself to do the bulk of the work.

Most Web search engines are commercial ventures supported by advertising revenue and, as a result, some employ the practice of allowing advertisers to pay money to have their listings ranked higher in search results. Those search engines which do not accept money for their search engine results make money by running search related ads alongside the regular search engine results. The search engines make money every time someone clicks on one of these ads.

Market share

Search engine	Market share in May 2011		Market share in December 2010[9]	
Google	82.80%		84.65%	
Yahoo!	6.42%		6.69%	
Baidu	4.89%		3.39%	
Bing	3.91%		3.29%	
Ask	0.52%		0.56%	
AOL	0.36%		0.42%	

Google's worldwide market share peaked at 86.3% in April 2010.[10] Yahoo!, Bing and other search engines are more popular in the US than in Europe.

According to Hitwise, market share in the U.S. for October 2011 was Google 65.38%, Bing-powered (Bing and Yahoo!) 28.62%, and the remaining 66 search engines 6%. However, an Experian Hit wise report released in August 2011 gave the "success rate" of searches sampled in July. Over 80 percent of Yahoo! and Bing searches resulted in the users visiting a web site, while Google's rate was just under 68 percent.[11] [12]

In the People's Republic of China, Baidu held a 61.6% market share for web search in July 2009.[13]

Search engine bias

Although search engines are programmed to rank websites based on their popularity and relevancy, empirical studies indicate various political, economic, and social biases in the information they provide.[14][15] These biases could be a direct result of economic and commercial processes (e.g., companies that advertise with a search engine can become also more popular in its organic search results), and political processes (e.g., the removal of search results in order to comply with local laws).[16] Google Bombing is one example of an attempt to manipulate search results for political, social or commercial reasons.

References

[1] World-Wide Web Servers (http://www.w3.org/History/19921103-hypertext/hypertext/DataSources/WWW/Servers.html)

[2] What's New! February 1994 (http://home.mcom.com/home/whatsnew/whats_new_0294.html)

[3] "Internet History - Search Engines" (from Search Engine Watch), Universiteit Leiden, Netherlands, September 2001, web: LeidenU-Archie (http://www.internethistory.leidenuniv.nl/index.php3?c=7).

[4] Oscar Nierstrasz (2 September 1993). "Searchable Catalog of WWW Resources (experimental)" (http://groups.google.com/group/comp.infosystems.www/browse_thread/thread/2176526a36dc8bd3/2718fd17812937ac?hl=en&lnk=gst&q=Oscar+ Nierstrasz#2718fd17812937ac). .

[5] Archive of NCSA what's new in December 1993 page(http://web.archive.org/web/20010620073530/http://archive.ncsa.uiuc.edu/SDG/Software/Mosaic/Docs/old-whats-new/whats-new-1293.html)

[6] "Yahoo! And Netscape Ink International Distribution Deal" (http://files.shareholder.com/downloads/YHOO/701084386x0x27155/ 9a3b5ed8-9e84-4cba-a1e5-77a3dc606566/YHOO_News_1997_7_8_General.pdf).

[7] *Browser Deals Push Netscape Stock Up 7.8%* (http://articles.latimes.com/1996-04-01/business/fi-53780_1_netscape-home). Los Angeles Times. 1 April1996.

[8] Gandal, Neil (2001). "The dynamics of competition in the internet search engine market". *International Journal of Industrial Organization* **19** (7): 1103–1117. doi:10.1016/S0167-7187(01)00065-0.

[9] Net Marketshare - World (http://marketshare.hitslink.com/search-engine-market-share.aspx?qprid=4)

[10] Net Market share - Google (http://marketshare.hitslink.com/report.aspx?qprid=5&qpcustom=Google - Global&qptimeframe=M& qpsp=120&qpnp=25)

[11] "Google Remains Ahead of Bing, But Relevance Drops" (http://news.yahoo.com/ google-remains-ahead-bing-relevance-drops-210457139.html). August 12, 2011..

[12] Experian Hitwise reports Bing-powered share of searches at 29 percent in October 2011 (http://www.hitwise.com/us/about-us/ press-center/press-releases/bing-powered-share-of-searches-at-29-percent), Experian Hitwise, November 16, 2011

[13] Search Engine Market Share July 2009 | Rise to the Top Blog (http://risetothetop.techwyse.com/internet-marketing/ search-engine-market-share-july-2009/)

[14] Segev, Elad (2010). Google and the Digital Divide: The Biases of Online Knowledge, Oxford: Chandos Publishing.

[15] Vaughan, L. & Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes, Information Processing & Management, 40(4), 693-707.

[16] Berkman Center for Internet & Society (2002), "Replacement of Google with Alternative Search Systems in China: Documentation and Screen Shots" (http://cyber.law.harvard.edu/filtering/china/google-replacements/), Harvard Law School.

- GBMW: Reports of 30-day punishment, re: Carmaker BMW had its German website bmw.de delisted from Google, such as: Slashdot-BMW (http://slashdot.org/article.pl?sid=06/02/05/235218) (05-Feb-2006).
- INSIZ: Maximum size of webpages indexed by MSN/Google/Yahoo! ("100-kb limit"): Max Page-size (http://www.sitepoint.com/article/indexing-limits-where-bots-stop) (28-Apr-2006).

Further reading

- For a more detailed history of early search engines, see Search Engine Birthdays (http://searchenginewatch.com/showPage.html?page=3071951)(fromSearchEngineWatch),ChrisSherman,September2003.
- Steve Lawrence; C. Lee Giles (1999). "Accessibility of information on the web". *Nature* **400** (6740): 107–9. doi:10.1038/21987. PMID 10428673.
- Bing Liu (2007), *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data* (http://www.cs.uic.edu/~liub/WebMiningBook.html). Springer, ISBN 3540378812
- Bar-Ilan, J. (2004). The use of Web search engines in information science research. ARIST, 38, 231-288.
- Levene, Mark (2005). *An Introduction to Search Engines and Web Navigation*. Pearson.
- Hock, Randolph (2007). *The Extreme Searcher's Handbook*. ISBN 978-0-910965-76-7
- Javed Mostafa (February 2005). "Seeking Better Web Searches" (http://www.sciam.com/article.cfm?articleID=0006304A-37F4-11E8-B7F483414B7F0000). *Scientific American Magazine*.
- Ross, Nancy; Wolfram, Dietmar (2000). "End user searching on the Internet: An analysis of term pair topics submitted to the Excite search engine". *Journal of the American Society for Information Science* **51** (10): 949–958. doi:10.1002/1097-4571(2000)51:10<949::AID-ASI70>3.0.CO;2-5.
- Xie, M. et al (1998). "Quality dimensions of Internet search engines". *Journal of Information Science* **24** (5): 365–372. doi:10.1177/016555159802400509.
- *Information Retrieval: Implementing and Evaluating Search Engines* (http://www.ir.uwaterloo.ca/book/). MIT Press. 2010.

External links

- Search Engines (http://www.dmoz.org/Computers/Internet/Searching/Search_Engines//) at the Open Directory Project

Search engine results page

A **search engine results page** (**SERP**), is the listing of web pages returned by a search engine in response to a keyword query. The results normally include a list of web pages with titles, a link to the page, and a short description showing where the Keywords have matched content within the page. A SERP may refer to a single page of links returned, or to the set of all links returned for a search query.

Query caching

Some search engines cache pages for frequent searches and display the cached pages instead of a live page to increase the performance of the search engine. The search engine updates the search results periodically to account for new pages, and possibly to modify the rankings of pages in the search results. Most of the results are weird and hard work is needed to make them readable.

Search result refreshing can take several days or weeks which can occasionally cause results to be inaccurate or out of date.

Different types of results

SERPs of major search engines like Google, Yahoo!, Bing, may include different types of listings: contextual, algorithmic or organic search listings, as well as sponsored listings, images, maps, definitions, videos or suggested search refinements.

The major search engines visually differentiate specific content types, such as images, news, and blogs. Many content types have specialized SERP templates and visual enhancements on the main search result page.

Advertising (sponsored listings)

SERPs may contain advertisements. This is how commercial search engines fund their operations. Common examples of these advertisements are displayed on the right hand side of the page as small classified style ads or directly above the main organic search results on the left.

Generation of SERPs

Major search engines like Google, Yahoo! and Bing primarily use content contained within the page and fallback to metadata tags of a web page to generate the content that makes up a search snippet. The html title tag will be used as the title of the snippet while the most relevant or useful contents of the web page (description tag or page copy) will be used for the description. If the web page is not available, information about the page from dmoz may be used instead.[1]

SERP tracking

Webmasters use search engine optimization (SEO) to increase their website's ranking on a specific keyword's SERP. As a result, webmasters often check SERPs to track their search engine optimization progress. To speed up the tracking process, programmers created automated software to track multiple keywords for multiple websites.

References

[1] Anatomy of a search snippet (http://www.mattcutts.com/blog/video-anatomy-of-a-search-snippet/)

Search engine marketing

Search engine marketing (**SEM**) is a form of Internet marketing that involves the promotion of websites by increasing their visibility in search engine results pages (SERPs) through the use of paid placement, contextual advertising, and paid inclusion.[1] Depending on the context, SEM can be an umbrella term for various means of marketing a website including search engine optimization (SEO), which "optimizes" website content to achieve a higher ranking in search engine results pages, or it may contrast with SEO, focusing on only paid components.[2]

Market

In 2008, North American advertisers spent US$13.5 billion on search engine marketing. The largest SEM vendors were Google AdWords, Yahoo! Search Marketing and Microsoft adCenter.[1] As of 2006, SEM was growing much faster than traditional advertising and even other channels of online marketing.[3] Because of the complex technology, a secondary "search marketing agency" market has evolved. Some marketers have difficulty understanding the intricacies of search engine marketing and choose to rely on third party agencies to manage their search marketing.

History

As the number of sites on the Web increased in the mid-to-late 90s, search engines started appearing to help people find information quickly. Search engines developed business models to finance their services, such as pay per click programs offered by Open Text[4] in 1996 and then Goto.com[5] in 1998. Goto.com later changed its name[6] to Overture in 2001, and was purchased by Yahoo! in 2003, and now offers paid search opportunities for advertisers through Yahoo! Search Marketing. Google also began to offer advertisements on search results pages in 2000 through the Google AdWords program. By 2007, pay-per-click programs proved to be primary money-makers[7] for search engines. In a market dominated by Google, in 2009 Yahoo! and Microsoft announced the intention to forge an alliance. The Yahoo! & Microsoft Search Alliance eventually received approval from regulators in the US and Europe in February 2010.[8]

Search engine optimization consultants expanded their offerings to help businesses learn about and use the advertising opportunities offered by search engines, and new agencies focusing primarily upon marketing and advertising through search engines emerged. The term "Search Engine Marketing" was proposed by Danny Sullivan in 2001[9] to cover the spectrum of activities involved in performing SEO, managing paid listings at the search engines, submitting sites to directories, and developing online marketing strategies for businesses, organizations, and individuals.

SEM methods and metrics

There are four categories of methods and metrics used to optimize websites through search engine marketing.[10][11][12][13]

1. **Keyword research and analysis** involves three "steps:" (a) Ensuring the site can be indexed in the search engines; (b) finding the most relevant and popular keywords for the site and its products; and (c) using those keywords on the site in a way that will generate and convert traffic.
2. **Website saturation and popularity**, how much presence a website has on search engines, can be analyzed through the number of pages of the site that are indexed on search engines (saturation) and how many backlinks the site has (popularity). It requires your pages containing those keywords people are looking for and ensure that they rank high enough in search engine rankings. Most search engines include some form of link popularity in their ranking algorithms. The followings are major tools measuring various aspects of saturation and link popularity: Link Popularity, Top 10 Google Analysis, and Marketleap's Link Popularity and Search Engine

Saturation.[11]

3. **Back end tools**, including Web analytic tools and HTML validators, provide data on a website and its visitors and allow the success of a website to be measured. They range from simple traffic counters to tools that work with log files[10] and to more sophisticated tools that are based on page tagging (putting JavaScript or an image on a page to track actions). These tools can deliver conversion-related information. There are three major tools used by EBSCO: (a) log file analyzing tool: WebTrends by NetiQ; (b) tag-based analytic programs WebSideStory's Hitbox; (c) transaction-based tool: TeaLeaf RealiTea. Validators check the invisible parts of websites, highlighting potential problems and many usability issues ensure your website meets W3C code standards. Try to use more than one HTML validator or spider simulator because each tests, highlights, and reports on slightly different aspects of your website.
4. **Whois tools** reveal the owners of various websites, and can provide valuable information relating to copyright and trademark issues.[12]

Paid inclusion

Paid inclusion involves a search engine company charging fees for the inclusion of a website in their results pages. Also known as sponsored listings, paid inclusion products are provided by most search engine companies, the most notable being Google.

The fee structure is both a filter against superfluous submissions and a revenue generator. Typically, the fee covers an annual subscription for one webpage, which will automatically be catalogued on a regular basis. However, some companies are experimenting with non-subscription based fee structures where purchased listings are displayed permanently.[14] A per-click fee may also apply. Each search engine is different. Some sites allow only paid inclusion, although these have had little success. More frequently, many search engines, like Yahoo!,[15] mix paid inclusion (per-page and per-click fee) with results from web crawling. Others, like Google (and as of 2006, Ask.com[16][17]), do not let webmasters pay to be in their search engine listing (advertisements are shown separately and labeled as such).

Some detractors of paid inclusion allege that it causes searches to return results based more on the economic standing of the interests of a web site, and less on the relevancy of that site to end-users.

Often the line between pay per click advertising and paid inclusion is debatable. Some have lobbied for any paid listings to be labeled as an advertisement, while defenders insist they are not actually ads since the webmasters do not control the content of the listing, its ranking, or even whether it is shown to any users. Another advantage of paid inclusion is that it allows site owners to specify particular schedules for crawling pages. In the general case, one has no control as to when their page will be crawled or added to a search engine index. Paid inclusion proves to be particularly useful for cases where pages are dynamically generated and frequently modified.

Paid inclusion is a search engine marketing method in itself, but also a tool of search engine optimization, since experts and firms can test out different approaches to improving ranking, and see the results often within a couple of days, instead of waiting weeks or months. Knowledge gained this way can be used to optimize other web pages, without paying the search engine company.

Comparison with SEO

SEM is the wider discipline that incorporates SEO. SEM includes both paid search results (Adwords) and organic search results (SEO). SEM uses AdWords,[18] pay per click (particularly beneficial for local providers as it enables potential consumers to contact a company directly with one click), article submissions, advertising and making sure SEO has been done. A keyword analysis is performed for both SEO and SEM, but not necessarily at the same time. SEM and SEO both need to be monitored and updated frequently to reflect evolving best practices.

In some contexts, the term *SEM* is used exclusively to mean *pay per click advertising*,[2] particularly in the commercial advertising and marketing communities which have a vested interest in this narrow definition. Such usage excludes the wider search marketing community that is engaged in other forms of SEM such as search engine optimization and search retargeting.

Another part of SEM is social media marketing (SMM). SMM is a type of marketing that involves exploiting social media to influence consumers that one company's products and/or services are valuable.[19] Some of the latest theoretical advances include search engine marketing management (SEMM). SEMM relates to activities including SEO but focuses on return on investment (ROI) management instead of relevant traffic building (as is the case of mainstream SEO). SEMM also integrates organic SEO, trying to achieve top ranking without using paid means of achieving top in search engines, and pay per click SEO. For example some of the attention is placed on the web page layout design and how content and information is displayed to the website visitor.

Ethical questions

Paid search advertising has not been without controversy, and the issue of how search engines present advertising on their search result pages has been the target of a series of studies and reports[20][21][22] by Consumer Reports WebWatch. The Federal Trade Commission (FTC) also issued a letter[23] in 2002 about the importance of disclosure of paid advertising on search engines, in response to a complaint from Commercial Alert, a consumer advocacy group with ties to Ralph Nader.

Another ethical controversy associated with search marketing has been the issue of trademark infringement. The debate as to whether third parties should have the right to bid on their competitors' brand names has been underway for years. In 2009 Google changed their policy, which formerly prohibited these tactics, allowing 3rd parties to bid on branded terms as long as their landing page in fact provides information on the trademarked term.[24] Though the policy has been changed this continues to be a source of heated debate.[25]

At the end of February 2011 many started to see that Google has started to penalize companies that are buying links for the purpose of passing off the rank. SEM has however nothing to do with link buying and focuses on organic SEO and PPC management.

Examples

A successful SEM project was undertaken by one of London's top SEM companies involving AdWords. AdWords is recognised as a web-based advertising utensil since it adopts keywords which can deliver adverts explicitly to web users looking for information in respect to a certain product or service. This project is highly practical for advertisers as the project hinges on cost per click (CPC) pricing, thus the payment of the service only applies if their advert has been clicked on. SEM companies have embarked on AdWords projects as a way to publicize their SEM and SEO services. This promotion has helped their business elaborate, offering added value to consumers who endeavor to employ AdWords for promoting their products and services. One of the most successful approaches to the strategy of this project was to focus on making sure that PPC advertising funds were prudently invested. Moreover, SEM companies have described AdWords as a fine practical tool for increasing a consumer's investment earnings on Internet advertising. The use of conversion tracking and Google Analytics tools was deemed to be practical for presenting to clients the performance of their canvass from click to conversion. AdWords project has enabled SEM companies to train their clients on the utensil and delivers better performance to the canvass. The assistance of AdWord canvass could contribute to the huge success in the growth of web traffic for a number of its consumer's website, by as much as 250% in only nine months.[18]

Another way Search Engine Marketing is managed is by contextual advertising. Here marketers place ads on other sites or portals that carry information relevant to their products so that the ads jump into the circle of vision of browsers who are seeking information from those sites. A successful SEM plan is the approach to capture the

relationships amongst information searchers, businesses, and search engines. Search engines were not important to some industries in the past but over the past years, the use of search engines for accessing information has become vital to increase business opportunities.[26] The use of SEM strategic tools for businesses such as tourism can attract potential consumers to view their products but it could also pose various challenges.[26] These challenges could be the competition that companies face amongst their industry and other sources of information that could draw the attention of online consumers.[26] To assist the combat of challenges, the main objective for businesses applying SEM is to improve and maintain their ranking as high as possible on SERPs so that they can gain visibility. Therefore search engines are adjusting and developing algorithms and the shifting criteria by which web pages are ranked sequentially to combat against search engine misuse and spamming, and to supply the most relevant information to searchers.[26] This could enhance the relationship amongst information searchers, businesses, and search engines by understanding the strategies of marketing to attract business.

References

[1] "The State of Search Engine Marketing 2006" (http://searchengineland.com/070208-095009.php). Search Engine Land. February 8, 2007. . Retrieved 2007-06-07.

[2] "Does SEM = SEO + CPC Still Add Up?" (http://searchengineland.com/does-sem-seo-cpc-still-add-up-37297). searchengineland.com. . Retrieved 2010-03-05.

[3] Elliott, Stuart (March 14, 2006). "More Agencies Investing in Marketing With a Click" (http://www.nytimes.com/2006/03/14/business/media/14adco.html?ex=1299992400&en=6fcd30b948dd1312&ei=5088). New York Times. . Retrieved 2007-06-07.

[4] "Engine sells results, draws fire" (http://news.com.com/2100-1023-215491.html). news.com.com. June 21, 1996. . Retrieved 2007-06-09.

[5] "GoTo Sells Positions" (http://searchenginewatch.com/showPage.html?page=2165971). searchenginewatch.com. March 3, 1998. . Retrieved 2007-06-09.

[6] "GoTo gambles with new name" (http://news.com.com/GoTo+gambles+with+new+name/2100-1023_3-272795.html). news.com.com. September 10, 2001. . Retrieved 2007-06-09.

[7] Jansen, B. J. (May 2007). "The Comparative Effectiveness of Sponsored and Nonsponsored Links for Web E-commerce Queries" (http://ist.psu.edu/faculty_pages/jjansen/academic/pubs/jansen_tweb_sponsored_links.pdf) (PDF). ACM Transactions on the Web, . . Retrieved 2007-06-09.

[8] "Microsoft-Yahoo Deal Gets Green Light" (http://www.informationweek.com/news/windows/microsoft_news/showArticle.jhtml?articleID=223000099). informationweek.com. February 18, 2010. . Retrieved 2010-07-15.

[9] "Congratulations! You're A Search Engine Marketer!" (http://searchenginewatch.com/showPage.html?page=2164351). searchenginewatch.com. November 5, 2001. . Retrieved 2007-06-09.

[10] Chadwick, Terry Brainerd (July 2005). "How search engine marketing tools can work for you: or, searching is really all about finding, first of three articles" (http://findarticles.com/p/articles/mi_m0FWE/is_7_9/ai_n14889940/). Information Outlook. . Retrieved 2011-03-21.

[11] Chadwick, Terry Brainerd (October 2005). "How search engine marketing tools can work for you: or searching is really all about finding" (http://findarticles.com/p/articles/mi_m0FWE/is_10_9/ai_n15890934/?tag=content;col1). Information Outlook. . Retrieved 2011-03-21.

[12] Chadwick, Terry Brainerd (November 2005). "How search engine marketing tools can work for you; or, searching is really all about finding, third of three articles" (http://findarticles.com/p/articles/mi_m0FWE/is_11_9/ai_n15966835/?tag=content;col). Information Outlook. . Retrieved 2011-03-21.

[13] "Search Engine Web Marketing Tips" (http://www.lewesseo.com/search-engine-web-marketing-tips/). LewesSEO.com. . Retrieved 22 February 2012.

[14] "FAQ #1: What are PermAds?" (http://www.permads.com/page.php?2#1). PermAds.com. . Retrieved 2010-09-12.

[15] Zawodny, Jeremy (2004-03-01). "Defending Paid Inclusions" (http://jeremy.zawodny.com/blog/archives/001671.html). .

[16] Ulbrich, Chris (2004-07-06). "Paid Inclusion Losing Charm?" (http://www.wired.com/news/business/0,1367,64092,00. html?tw=wn_tophead_1). Wired News. .

[17] "FAQ #18: How do I register my site/URL with Ask so that it will be indexed?" (http://about.ask.com/en/docs/about/webmasters. shtml#18). Ask.com. . Retrieved 2008-12-19.

[18] "Google Adwords Case Study" (http://www.accuracast.com/images/case-google.pdf). AccuraCast. 2007. . Retrieved 2011-03-30.

[19] Susan Ward (2011). "Social Media Marketing" (http://sbinfocanada.about.com/od/socialmedia/g/socmedmarketing.htm). About.com. . Retrieved 2011-04-22.

[20] "False Oracles: Consumer Reaction to Learning the Truth About How Search Engines Work (Abstract)" (http://www.consumerwebwatch. org/dynamic/search-report-false-oracles-abstract.cfm). consumerwebwatch.org. June 30, 2003. . Retrieved 2007-06-09.

[21] "Searching for Disclosure: How Search Engines Alert Consumers to the Presence of Advertising in Search Results" (http://www.consumerwebwatch.org/dynamic/search-report-disclosure-abstract.cfm). consumerwebwatch.org. November 8, 2004. . Retrieved 2007-06-09.

[22] "Still in Search of Disclosure: Re-evaluating How Search Engines Explain the Presence of Advertising in Search Results" (http://www.consumerwebwatch.org/dynamic/search-report-disclosure-update-abstract.cfm). consumerwebwatch.org. June 9, 2005. . Retrieved 2007-06-09.

[23] "Re: Complaint Requesting Investigation of Various Internet Search Engine Companies for Paid Placement or ([[Pay per click (http://www.ftc.gov/os/closings/staff/commercialalertletter.shtm)])"]. ftc.gov. June 22, 2002. . Retrieved 2007-06-09.

[24] "Update to U.S. ad text trademark policy" (http://adwords.blogspot.com/2009/05/update-to-us-ad-text-trademark-policy.html). adwords.blogspot.com. May 14, 2009. . Retrieved 2010-07-15.

[25] Rosso, Mark; Jansen, Bernard (Jim) (August 2010), "Brand Names as Keywords in Sponsored Search Advertising" (http://aisel.aisnet.org/ cais/vol27/iss1/6), *Communications of the Association for Information Systems* **27** (1),

[26] Zheng Xiang, Bing Pan, Rob Law, and Daniel R. Fesenmaier (June 7, 2010). "Assessing the Visibility of Destination Marketing Organizations in Google: A Case Study of Convention and Visitor Bureau Websites in the United States" (http://www.informaworld.com/ index/929886653.pdf). Journal of Travel and Tourism Marketing. . Retrieved 2011-04-22.

Image search

An **image retrieval** system is a computer system for browsing, searching and retrieving images from a large database of digital images. Most traditional and common methods of image retrieval utilize some method of adding metadata such as captioning, keywords, or descriptions to the images so that retrieval can be performed over the annotation words. Manual image annotation is time-consuming, laborious and expensive; to address this, there has been a large amount of research done on automatic image annotation. Additionally, the increase in social web applications and the semantic web have inspired the development of several web-based image annotation tools.

The first microcomputer-based image database retrieval system was developed at MIT, in the 1980s, by Banireddy Prasaad, Amar Gupta, Hoo-min Toong, and Stuart Madnick.[1]

A 2008 survey article documented progresses after 2007.[2]

Search methods

Image search is a specialized data search used to find images. To search for images, a user may provide query terms such as keyword, image file/link, or click on some image, and the system will return images "similar" to the query. The similarity used for search criteria could be meta tags, color distribution in images, region/shape attributes, etc.

- Image meta search - search of images based on associated metadata such as keywords, text, etc.
- Content-based image retrieval (CBIR) – the application of computer vision to the image retrieval. CBIR aims at avoiding the use of textual descriptions and instead retrieves images based on similarities in their contents (textures, colors, shapes etc.) to a user-supplied query image or user-specified image features.
  - List of CBIR Engines - list of engines which search for images based image visual content such as color, texture, shape/object, etc.

Data Scope

It is crucial to understand the scope and nature of image data in order to determine the complexity of image search system design. The design is also largely influenced by factors such as the diversity of user-base and expected user traffic for a search system. Along this dimension, search data can be classified into the following categories:

- *Archives* - usually contain large volumes of structured or semi-structured homogeneous data pertaining to specific topics.
- *Domain-Specific Collection* - this is a homogeneous collection providing access to controlled users with very specific objectives. Examples of such a collection are biomedical and satellite image databases.
- *Enterprise Collection* - a heterogeneous collection of images that is accessible to users within an organization's intranet. Pictures may be stored in many different locations.

- *Personal Collection* - usually consists of a largely homogeneous collection and is generally small in size, accessible primarily to its owner, and usually stored on a local storage media.
- *Web* - World Wide Web images are accessible to everyone with an Internet connection. These image collections are semi-structured, non-homogeneous and massive in volume, and are usually stored in large disk arrays.

Evaluations

There are evaluation workshops for image retrieval systems aiming to investigate and improve the performance of such systems.

- ImageCLEF [3] - a continuing track of the Cross Language Evaluation Forum [4] that evaluates systems using both textual and pure-image retrieval methods.
- Content-based Access of Image and Video Libraries [5] - a series of IEEE workshops from 1998 to 2001.

References

[1] Prasad, B E; A Gupta, H-M Toong, S.E. Madnick (February 1987). "A microcomputer-based image database management system". *IEEE Transactions on Industrial Electronics* **IE-34** (1): 83–8. doi:10.1109/TIE.1987.350929.

[2] Datta, Ritendra; Dhiraj Joshi, Jia Li, James Z. Wang (April 2008). "Image Retrieval: Ideas, Influences, and Trends of the New Age" (http://infolab.stanford.edu/~wangz/project/imsearch/review/JOUR/). *ACM Computing Surveys* **40** (2): 1–60. doi:10.1145/1348246.1348248. .

[3] http://www.imageclef.org

[4] http://www.clef-campaign.org

[5] http://ieeexplore.ieee.org/xpl/RecentCon.jsp?punumber=4980

External links

- microsoft.com (http://research.microsoft.com/en-us/um/people/larryz/zitnickcfir03.pdf), Content-Free Image Retrieval by C. Lawrence Zitnick and Takeo Kanade, May 2003
- PDF (http://www.asis.org/Bulletin/Jun-09/JunJul09_Uzwyshyn.pdf) *Bulletin of the American Society for Information Science & Technology* Special Issue on Visual Search (http://uwf.edu/ruzwyshyn/2009PDF/Bulletin_JunJul09_Finaloptimized.pdf). June/July 2009. 35:5 ISSN: 1550-8366.
- alipr.com (http://www.alipr.com) Automatic image tagging and visual image search. Developed with Stanford and Penn State technologies.
- CIRES (http://amazon.ece.utexas.edu/~qasim/research.htm) Image retrieval system developed by the University of Texas at Austin.
- FIRE (http://thomas.deselaers.de/FIRE) Image retrieval system developed by the RWTH Aachen University, Aachen, Germany.
- GIFT (http://www.gnu.org/software/gift/) GNU Image Finding Tool, originally developed at the University of Geneva, Switzerland.
- ImageCLEF (http://www.imageclef.org) A benchmark to compare the performance of image retrieval systems.
- imgSeek (http://www.imgseek.net) Open-source desktop photo collection manager and viewer with content-based search and many other features.
- isk-daemon (http://server.imgseek.net) Open-source database server capable of adding content-based (visual) image searching to any image related website or software.
- img (Anaktisi) (http://www.anaktisi.net) This Web-Solution implements a new family of CBIR descriptors. These descriptors combine in one histogram color and texture information and are suitable for accurately retrieving images.
- Caliph & Emir (http://www.semanticmetadata.net/): Creation and Retrieval of images based on MPEG-7 (GPL).
- img (Rummager) (http://www.img-rummager.com): Image retrieval Engine (Freeware Application).
- PicsLikeThat (http://www.picslikethat.com): Visual and semantic image retrieval.

Video search

A **video search engine** is a web-based search engine which crawls the web for video content. Some video search engines parse externally hosted content while others allow content to be uploaded and hosted on their own servers. Some engines also allow users to search by video format type and by length of the clip. Search results are usually accompanied by a thumbnail view of the video.

Video search engines are computer programs designed to find videos stored on digital devices, either through Internet servers or in storage units from the same computer. These searches can be made through audiovisual indexing, which can extract information from audiovisual material and record it as metadata, which will be tracked by search engines.

Utility

The main use of these search engines is the increasing creation of audiovisual content and the need to manage it properly. The digitization of audiovisual archives and the establishment of the Internet, has led to large quantities of video files stored in big databases, whose recovery can be very difficult because of the huge volumes of data and the existence of a semantic gap.

Search criterion

The search criterion used by each search engine depends on its nature and purpose of the searches.

Metadata

Metadata is information about facts. It could be information about who is the author of the video, creation date, duration, and all the information you would like to extract and include in the same files. Internet is often used in a language called XML to encode metadata, which works very well through the web and is readable by people. Thus, through this information contained in these files is the easiest way to find data of interest to us.

In the videos there are two types of metadata, that we can integrate in the video code itself and external metadata from the page where the video is. In both cases we optimize them to make them ideal when indexed.

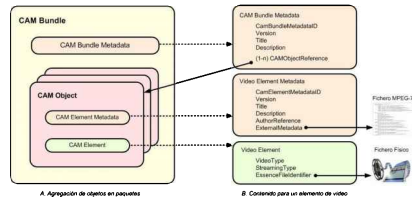
Internal metadata

All video formats incorporate their own metadata. The title, description, coding quality or transcription of the content are possible. To review these data exist programs like FLV MetaData Injector, Sorenson Squeeze or Castfire. Each one has some utilities and special specifications.

Keep in mind that converting from one format to another can lose much of this data, so check that the new format information is correct. It is therefore advisable to have the video in lots of formats, so that all search robots will be able to find and index.

External metadata

In most cases you must apply the same mechanisms as in the positioning of an image or text content.



Video metadata

Title and Description

They are the most important factors when positioning a video, because there you will find most of the necessary information. The titles have
to be clearly descriptive and should be removed every word or phrase that is not useful.

Filename

It should be descriptive, including keywords that describe the video with no need to see their title or description. Ideally, separate the words by dashes "-".

Tags

On the page where the video is, it should be a list of keywords linked to the microformat "rel-tag". These words will be used by search engines as a basis for organizing information.

Transcription and subtitles

Although not completely standard, there are two formats that store information in a temporal component that is specified, one for subtitles and another for transcripts, which can also be used for subtitles.

The formats are SRT or SUB for subtitles and TTXT for transcripts. To manage this type of formats it is interesting to use MP4Box program with which you can get this kind of files and formats.

Speech Recognition

Speech recognition consists of a transcript of the speech of the audio track of the videos, creating a text file. In this way and with the help of a phrase extractor can easily search if the video content is of our interest.

Sample search for videos by Google voice recognition

Some search engines apart from using speech recognition to search for videos, also use it to find the specific point of a multimedia file in which you cite a specific word or phrase and so go directly to this point. Gaudi (Google Audio Indexing), a project developed by Google Labs, uses voice recognition technology to locate the exact moment that one or more words have been spoken within an audio, allowing the user to go directly to exact moment that the words were spoken. If the search query matches some videos from YouTube[1], the positions are indicated by yellow markers, and must pass the mouse over to read the transcribed text.

Sample search for videos with text recognition "TalkMiner"

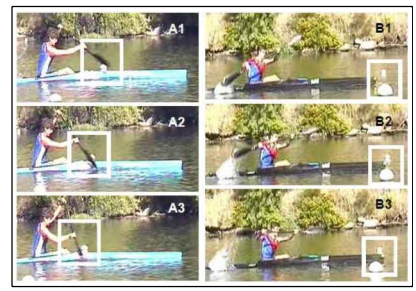
Text Recognition

The text recognition can be very useful to recognize characters in the videos through "chyrons". As with speech recognizers, there are search engines that allow, through character recognition, to play a video from a particular point where you see the word you want.

TalkMiner [2], an example of search of specific fragments from videos by text recognition, analyzes each video once per second looking for indetifier signs of a slide, such as its shape and static nature, captures the image of the slide and uses Optical Character Recognition (OCR) to detect the words on the slides. Then, these words are indexed in the search engine of TalkMiner [2], which currently offers to users more than 20,000 videos from institutions such as Stanford University, the University of California at Berkeley, and TED.

Frame Analysis

Through the visual descriptors we can analyze the frames of a video and extract information that can be scored as metadata. Descriptions are generated automatically and can describe different aspects of the frames, such as color, texture, shape, motion, and the situation.



Sample search for videos by frame analysis

Ranking criterion

The usefulness of a search engine depends on the relevance of the result set returned. While there may be millions of videos that include a particular word or phrase, some videos may be more relevant, popular or have more authority than others. This arrangement has a lot to do with search engine optimization.

Most search engines use different methods to classify the results and provide the best video in the first results. However, most programs allow you to sort the results by several criterions.

Order by relevance

This criterion is more ambiguous and less objective, but sometimes it is the closest to what we want; depends entirely on the searcher and the algorithm that the owner has chosen. That's why it has always been discussed and now that search results are so ingrained into our society it has been discussed even more. This type of management often depends on the number of times that the searched word comes out, the number of viewings of this, the number of pages that link to this content and ratings given by users who have seen it. [3]

Order by date of upload

This is a criterion based totally on the timeline where you can sort the results according to their seniority in the repository.

Order by number of views

It can give us an idea of the popularity of each video.

Order by user rating

It is common practice in repositories let the users rate the videos, so that a content of quality and relevance will have a high rank on the list of results gaining visibility. This practice is closely related to virtual communities.

Interfaces

We can distinguish two basic types of interfaces, some are web pages hosted on servers which are accessed by Internet and searched through the network, and the others are computer programs that search within a private network.

Internet

Within Internet interfaces we can find repositories that host video files which incorporate a search engine that searches only their own databases, and video searchers without repository that search in sources of external software.

Repositories with video searcher

Provides accommodation in video files stored on its servers and usually has an integrated search engine that searches through videos uploaded by its users. One of the first web repositories, or at least the most famous are the portals Vimeo, Dailymotion and YouTube.

Their searches are often based on reading the metadata tags, titles and descriptions that users assign to their videos. The disposal and order criterion of the results of these searches are usually selectable between the file upload date, the number of viewings or what they call the relevance. Still, sorting criterion are now a days the main weapon of these websites, because in terms of promotion is very important the positioning that they can give to your video.

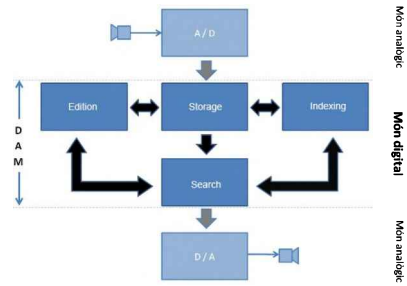
Repository with video searcher "Dailymotion"

Video searchers repositories

They are websites specialized in searching videos across the network or certain pre-selected repositories. They work by web spiders that inspect the network in an automated way to create copies of the visited websites, which will then be indexed by search engines, so they can provide faster searches.

Private Network

You can also find the case where a search engine only searches in audiovisual files stored within a computer or, as it happens in televisions, on a private server where users access through a local area network. These searchers are usually softwares or rich Internet applications with a very specific search options for maximum speed and efficiency when presenting the results. They are typically used for large databases and are therefore highly focused to satisfy the needs of television companies. An example of this type of software would be the Digition Suite, which apart from being a benchmark in this kind of interfaces is very close to us as for the storage and retrieval files system from the Corporació Catalana de Mitjans Audiovisuals. [4]



Functioning scheme

This particular suite and perhaps in its strongest point is that it integrates the entire process of creating, indexing, storing, searching, editing, and a recovery. Once we have a digitized audiovisual content is indexed with different techniques of different level depending on the importance of content and it's stored. The user, when he wants to retrieve a particular file, has to fill a search fields such as program title, issue date, characters who act or the name of the producer, and the robot starts the search. Once the results appear and they arranged according to preferences, the user can play the low quality videos to work as quickly as possible. When he finds the desired content, it is

downloaded with good definition, it's edited and reproduced. [5]

Design and algorithms

Video search has evolved slowly through several basic search formats which exist today and all use keywords. The keywords for each search can be found in the title of the media, any text attached to the media and content linked web pages, also defined by authors and users of video hosted resources.

Some video search is performed using human powered search, others create technological systems that work automatically to detect what is in the video and match the searchers needs. Many efforts to improve video search including both human powered search as well as writing algorithm that recognize what's inside the video have meant complete redevelopment of search efforts.

It is generally acknowledged that speech to text is possible, though recently Thomas Wilde, the new CEO of Everyzing, acknowledged that Everyzing works 70% of the time when there is music, ambient noise or more than one person speaking. If newscast style speaking (one person, speaking clearly, no ambient noise) is available, that can rise to 93%. (From the Web Video Summit, San Jose, CA, June 27, 2007).

Around 40 phonemes exist in every language with about 400 in all spoken languages. Rather than applying a text search algorithm after speech-to-text processing is completed, some engines use a phonetic search algorithm to find results within the spoken word. Others work by literally listening to the entire podcast and creating a text transcription using a sophisticated speech-to-text process. Once the text file is created, the website lets you search the file for any number of search words and phrases.

It is generally acknowledged that visual search into video does not work well and that no company is using it publicly. Researchers at UC San Diego and Carnegie Mellon University have been working on the visual search problem for more than 15 years, and admitted at a "Future of Search" conference at UC Berkeley in the Spring of 2007 that it was years away from being viable even in simple search.

Popular video search engines

Agnostic search

Search that is not affected by the hosting of video, where results are agnostic no matter where the video is located:

- **AltaVista Video Search** had one of the first video search engines with easy accessible use. Is found on a direct link called "Video" off the main page above the text block. [Since 2 February 2009 this feature has not been available from Altavista.com]
- **blinkx** was launched in 2004 and uses speech recognition and visual analysis to process spidered video rather than rely on metadata alone. blinkx claims to have the largest archive of video on the web and puts its collection at around 26,000,000 hours of content.
- **CastTV** is a Web-wide video search engine that was founded in 2006 and funded by Draper Fisher Jurvetson, Ron Conway, and Marc Andreessen.
- **Clipta** is a deep crawling video search engine that indexes millions of videos from across the Internet. Clipta was founded and launched in 2008.
- **Munax** released their first version all-content search engine in 2005 and powers both nation-wide and worldwide search engines with videosearch.
- **Picsearch Video Search** has been licensed to search portals since 2006. Picsearch is a search technology provider who powers image, video and audio search for over 100 major search engines around the world.
- **ScienceStage** is an integrated universal search engine for science-oriented video (lectures, conferences, documentaries, webinars, tutorials, demonstrations, grand rounds, etc.). All videos are also semantically matched to millions of research documents from open-access databases.

- **Truveo** is a Web-wide video search engine that was founded in 2004 and launched in September 2005. Truveo claims to index over 650 million videos from thousands of sources across the Web, and uses speech recognition and visual analysis in its search technology.
- **Vedeo.tv** Spanish site, but allows search in English and shows results from many video sites, including local news websites.
- **VideoSurf**[6] uses computer vision techniques to enhance its search results, and has mobile applications that query based on video captured with the phone camera.
- **yovisto** is an academic video search engine for lecture recordings and scientific conference talks based on speech processing, OCR, and user annotation.

Non-agnostic search

Search results are modified, or suspect, due to the large hosted video being given preferential treatment in search results:

- **AOL Video** offers a leading video search engine that can be used to find video located on popular video destinations across the web. In December 2005, AOL acquired Truveo Video Search.
- **Google Videos** is a popular video search engine which used to permit its visitors to upload videos. It searches YouTube and many other video hosting sites.
- **Yahoo! Video Search** Yahoo!'s search engine examines video files on the internet using its Media RSS standard. Is found on a direct link called "Video" off the main page above the text block.

References

[1] http://www.youtube.com/
[2] http://talkminer.com/
[3] **(English)** SEO by Google central webmaster (http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=35291)
[4] **(Catalan)** Digitalize or die (Alícia Conesa) (http://escac.documentacioinformativa.com/sessio1/lectures/digition.pdf)
[5] **(Catalan)** Digition Suite from Activa Multimedia (http://www.activamultimedia.com/am/v_portal/herramientas/generarpdf. php?te=26&idm=2)
[6] http://www.videosurf.com

External links

- Process of search engines *How Stuff Works* (http://www.howstuffworks.com/search-engine.htm) **(English)**
- Text query in Corporació Catalana de Mitjans Audiovisuals (Ramon Salla) (http://escac.documentacioinformativa.com/sessio1/lectures/digition.pdf) **(Catalan)**
- *Blinkx* search engine (http://websearch.about.com/od/imagesearch/a/blinkx.htm) **(English)**

Local search

Local search is the use of specialized Internet search engines that allow users to submit geographically constrained searches against a structured database of local business listings. Typical local search queries include not only information about "what" the site visitor is searching for (such as keywords, a business category, or the name of a consumer product) but also "where" information, such as a street address, city name, postal code, or geographic coordinates like latitude and longitude. Examples of local searches include "Hong Kong hotels", "Manhattan restaurants", and "Dublin Hertz". Local searches exhibit explicit or implicit local intent. A search that includes a location modifier, such as "Bellevue, WA" or "14th arrondissement", is an explicit local search. A search that references a product or service that is typically consumed locally, such as "restaurant" or "nail salon", is an implicit local search.

Local search sites are primarily supported by advertising from businesses that wish to be prominently featured when users search for specific products and services in specific locations. Local search advertising can be highly effective because it allows ads to be targeted very precisely to the search terms and location provided by the user.

Evolution

Local search is the natural evolution of traditional off-line advertising, typically distributed by newspaper publishers and TV and radio broadcasters, to the Web. Historically, consumers relied on local newspapers and local TV and radio stations to find local product and services. With the advent of the Web, consumers are increasingly using search engines to find these local products and services online. In recent years, the number of local searches online has grown rapidly while off-line information searches, such as print Yellow Page lookups, have declined. As a natural consequence of this shift in consumer behavior, local product and service providers are slowly shifting their advertising investments from traditional off-line media to local search engines.

A variety of search engines are currently providing local search, including efforts backed by the largest search engines, and new start-ups. Some of these efforts are further targeted to specific vertical segments while others are tied to mapping products.

Various geolocation techniques may be used to match visitors' queries with information of interest. The sources and types of information and points of interest returned varies with the type of local search engine.

Google Maps (formerly Google Local) looks for physical addresses mentioned in regular web pages. It provides these results to visitors, along with business listings and maps. Product-specific search engines] use techniques such as targeted web crawling and direct feeds to collect information about products for sale in a specific geographic area.

Other local search engines adjunct to major web search portals include general Windows Live Local, Yahoo! Local, and ask.com's AskCity. Yahoo!, for example, separates its local search engine features into Yahoo! Local and Yahoo! Maps, the former being focused on business data and correlating it with web data, the latter focused primarily on the map features (e.g. directions, larger map, navigation).

Search engines offer local businesses the possibility to upload their business data to their respective local search databases.

Local search, like ordinary search, can be applied in two ways. As John Battelle coined it in his book "The Search," search can be either recovery search or discovery search.

> This perfect search also has perfect recall – it knows what you've seen, and can discern between a journey of discovery – where you want to find something new – and recovery – where you want to find something you've seen before.

This applies especially to local search. Recovery search implies, for example, that a consumer knows who she is looking for (i.e., Main Street Pizza Parlor) but she does not know where they are, or needs their phone number.

Discovery search implies that the searcher knows, for example, what she wants but not who she needs it from (i.e., pizza on Main Street in Springfield).

In February 2012, Google announced that they made 40 changes to their search algorithm, including one codenamed "Venice" which Google states will improve local search results by "relying more on the ranking of (Google's) main search results as a signal",[1] meaning local search will now rely more on organic SERPs (Search Engine Result Pages).

Private label local search

Traditional local media companies, including newspaper publishers and television and radio broadcasters, are starting to add local search to their local websites in an effort to attract their share of local search traffic and advertising revenues in the markets they serve. These local media companies either develop their own technology, or license "private label" or "white label" local search solutions from third-party local search solution providers. In either case, local media companies base their solution on business listings databases developed in-house or licensed from a third-party data publisher.

Traditional print directory publishers also provide local search portals. Most regions around print directory publishers have an online presence.

Social local search

Local search that incorporates internal or external social signals could be considered social local search driven. The first site to incorporate this type of search was Explore To Yellow Pages. Explore To uses Facebook Likes as one of the signals to increase the ranking of listings where other factors may be equal or almost equal. Typical ranking signals in local search, such as keyword relevancy and distance from centroid can therefore be layered with these social signals to give a better crowdsourced experience for users.

Mobile local search

Several providers have been experimenting with providing local search for mobile devices. Some of these are location aware. In the United States, Google previously operated an experimental voice-based locative service (1-800-GOOG-411 [2]) but terminated the service in November, 2010. Many mobile web portals require the subscriber to download a small Java application, however the recently added .mobi top level domain has given impetus to the development of mobile targeted search sites are based upon a standard mobile specific XML protocol that all modern mobile browsers understand. The advantage is that no software needs to be downloaded and installed, plus these sites may be designed to simultaneously provide conventional content to traditional PC users by means of automatic browser detection.

Business owners and local search

Electronic publishers (such as businesses or individuals) who would like information such as their name, address, phone number, website, business description and business hours to appear on local search engines have several options. The most reliable way to include accurate local business information is to claim business listings through Google's, Yahoo!'s, or Bings's respective local business centers.

It is ever so more important today that small business owners claim their free local listing with Google Places since Google Places is often one of the first listings seen on Google's search result page whenever there algorithm deems a keyword query to have local intent.

Example of google places listings in organic search[3] from Google's search engine, based on the user's IP address in Toronto.

Business listing information can also be distributed via the traditional Yellow Pages, electronic Yellow Pages aggregators, and search engine optimization services. Some search engines will pick up on web pages that contain regular street addresses displayed in machine-readable text (rather than a picture of text, which is more difficult to interpret). Web pages can also use geotagging techniques.

References

[1] *Search quality highlights: 40 changes for February* (http://insidesearch.blogspot.com/2012/02/search-quality-highlights-40-changes. html), Google, February 27, 2012,
[2] http://labs.google.com/goog411/
[3] http://www.smallbusinessonlinecoach.com/wp-content/uploads/2011/08/pure-vs-blended-google-search-results.png

External links

- http://searchengineland.com/google-confirms-panda-update-link-evaluation-local-search-rankings-113078
- http://www.localmybiz.com/2012/codename-venice-general-seo-important-local-search/

Web presence

A **digital footprint** is a trail left by an entity's interactions in a digital environment; including their usage of TV, mobile phone, internet and world wide web, mobile web and other devices and sensors. Digital footprints provide data on what an entity has performed in the digital environment; and are valuable in assisting behavioural targeting, personalisation, targeted marketing, digital reputation, and other social media or social graphing services.[1]

In social media, a *digital footprint* is the size of an individual's online presence; as it relates to the number of individuals they interact with.

Description

A *digital footprint* is a collection of activities and behaviours recorded when an entity (such as a person) interacts in a digital environment. It may include the recording of activities such as system login and logouts, visits to a web-page, accessed or created files, or emails and chat messages. The digital footprint allows interested parties to access this data; possibly for data mining, or profiling purposes.

One of the first references to a digital footprint was by Nicholas Negroponte, naming it the *slug trail* in his book Being Digital in 1996. John Battelle called digital footprints the *clickstream exhaust*, while Tim O'Reilly and Esther Dyson titled it the *data exhaust*.[2] Early usage of the term focused on information left by web activity alone, but came to represent data created and consumed by all devices and sensors.[3]

Footprinting process

Inputs to digital footprint include attention, location, time of day, search results and key words, content created and consumed, digital activity and data from sensor, and from the users social crowd. Some data can come from deep IP and Internet data, such as footprinting. Value created from the collection of inputs and analysis of the data are recommendation, protection, personalisation, ability to trade or barter and contextual adaptation. Part of the analysis phase is *Reality mining*

Feedback loop

In an *open system*, data is collected from a user, which is used to build a profile (see Profiling practices); becoming usable by interested third parties to improve recommendation. Collection of data from multiple user interactions and purchases generates improved recommendations. If the same parties collect data on how that user interacts with, or influences others interactions, the service, there is an additional component of data - the output of one process becomes the input to the next.

The *closed loop digital footprint* was first explained by Tony Fish[4] in his book on digital footprints in January 2010. The closed loop takes data from the open loop and provides this as a new data input. This new data determines what the user has reacted to, or how they have been influenced. The feedback then builds a digital footprint based on social data, and the controller of the social digital footprint data can determine who and why people purchase and behave. According to a Pew Internet report published in 2007, there are two main classifications for digital footprints: passive and active. A passive digital footprint is created when data is collected about an action without any client activation, whereas active digital footprints are created when personal data is released deliberately by a user for the purpose of sharing information about oneself.[5]

Passive digital footprints can be stored in many ways depending on the situation. In an online environment a footprint may be stored in an online data base as a *hit*. This footprint may track the user IP address, when it was created, and where they came from; with the footprint later being analyzed. In an offline environment, a footprint may be stored in files, which can be accessed by administrators to view the actions performed on the machine, without being able to see who performed them.

Active digital footprints can be also be stored in many ways depending on the situation. In an online environment, a footprint can be stored by a user being logged into a site when making a post or edit, with the registered name being connected to the edit. In an off line environment a footprint may be stored in files, when the owner of the computer uses a keylogger, so logs can show the actions performed on the machine, and who performed them.

Web browsing

The digital footprint applicable specifically to the World Wide Web is the *internet footprint*;[6] also known as *cyber shadow* or *digital shadow*, information is left behind as a result of a user's web-browsing activities, including through the use of cookies. The term usually applies to an individual person, but can also refer to a business, organization, corporation or object.

Information may be intentionally or unintentionally left behind by the user; with it being either passively or actively collected by other interested parties. Depending on the amount of information left behind, it may be simple for other parties to gather large amounts of information on that individual using simple search engines. Internet footprints are used by interested parties for several reasons; including *cyber-vetting*, where interviewers could research applicants based on their online activities. Internet footprints are also used by law enforcement agencies, to provide information that would be unavailable otherwise due to a lack of probable cause.

Social networking systems may record activities of individuals, with data becoming a *life stream*. Such usage of social media and roaming services allow digital tracing data to include individual interests, social groups, behaviours, and location. Such data can be gathered from sensors within devices, and collected and analyzed without user awareness.

Privacy issues

Digital footprints are not a digital identity or passport, but the meta data collected impacts upon internet privacy, trust, security, digital reputation, and recommendation. As the digital world expands and integrates with more aspects of life, ownership and rights of data becomes important. Digital footprints are controversial in that privacy and openness are in competition.[7] Scott McNealy said in 1999 *Get Over It* when referring to privacy on the internet,[8] becoming a commonly used quote in relationship to private data and what companies do with it.

While a digital footprint can be used to infer personal information without their knowledge, it also exposes individuals private psychological sphere into the social sphere (see Bruno Latour's article (Latour 2007)). *Lifelogging* is an example of indiscriminate collection of information concerning an individuals life and behaviour (Kieron, Tuffield & Shadbolt 2009).

Notes

- Kieron, O'Hara; Tuffield, Mischa M.; Shadbolt, Nigel (2009), "Lifelogging: Privacy and empowerment with memories for life", *Identity in the Information Society* (Springer) **1**: 155, doi:10.1007/s12394-009-0008-4
- Latour, Bruno (2007), "Beware your imagination leaves digital traces"[9], *Times Higher Literary Supplement*, 6th April 2007

References

[1] Mobile Marketing (http://www.forbes.com/2009/01/12/mobile-marketing-privacy-tech-security-cx_ag_0113mobilemarket.html)
[2] Data Exhaust (http://oreilly.com/web2/archive/what-is-web-20.html?page=1)
[3] the wider Definition (http://www.mydigitalfootprint.com/footprint-cms/DIGITAL_FOOTPRINTS.html)
[4] Tony Fish (http://www.amazon.com/Tony-Fish/e/B0036U3800)
[5] Pew Internet: Digital Footprints (http://www.pewinternet.org/PPF/r/229/report_display.asp)
[6] Garfinkel, Simson; Cox, David. "Finding and Archiving the Internet Footprint" (http://simson.net/clips/academic/2009.BL. InternetFootprint.pdf). *Presented at the first Digital Lives Research Conference*. London, England. .
[7] Telegraph UK article (http://www.telegraph.co.uk/news/newstopics/politics/4339771/ Threat-to-privacy-under-data-law-campaigners-warn.html)
[8] Scott NcNealy 'get over it' (http://www.wired.com/politics/law/news/1999/01/17538)
[9] http://www.bruno-latour.fr/poparticles/poparticle/P-129-THES-GB.doc

Internet marketing

Internet marketing, also known as **web marketing**, **online marketing**, **webvertising**, or **e-marketing**, is referred to as the marketing (generally promotion) of products or services over the Internet. iMarketing is used as an abbreviated form for Internet Marketing.

Internet marketing is considered to be broad in scope because it not only refers to marketing on the Internet, but also includes marketing done via e-mail and wireless media. Digital customer data and electronic customer relationship management (ECRM) systems are also often grouped together under internet marketing.[1]

Internet marketing ties together the creative and technical aspects of the Internet, including design, development, advertising and sales.[2] Internet marketing also refers to the placement of media along many different stages of the customer engagement cycle through search engine marketing (SEM), search engine optimization (SEO), banner ads on specific websites, email marketing, mobile advertising, and Web 2.0 strategies.

In 2008, *The New York Times*, working with comScore, published an initial estimate to quantify the user data collected by large Internet-based companies. Counting four types of interactions with company websites in addition to the hits from advertisements served from advertising networks, the authors found that the potential for collecting data was up to 2,500 times per user per month.[3]

Types of Internet marketing

Internet marketing is broadly divided in to the following[4] types:

- Display Advertising: the use of web banners or banner ads placed on a third-party website to drive traffic to a company's own website and increase product awareness.[4]
- Search Engine Marketing (SEM): a form of marketing that seeks to promote websites by increasing their visibility in search engine result pages (SERPs) through the use of either paid placement, contextual advertising, and paid inclusion, or through the use of free search engine optimization techniques.[5]
- Search Engine Optimization (SEO): the process of improving the visibility of a website or a webpage in search engines via the "natural" or un-paid ("organic" or "algorithmic") search results.[6]
- Social Media Marketing: the process of gaining traffic or attention through social media websites such as Facebook, Twitter and LinkedIn.[7]
- Email Marketing: involves directly marketing a commercial message to a group of people using electronic mail.[8]
- Referral Marketing: a method of promoting products or services to new customers through referrals, usually word of mouth.[9]
- Affiliate Marketing: a marketing practice in which a business rewards one or more affiliates for each visitor or customer brought about by the affiliate's own marketing efforts.[10]
- Content Marketing: involves creating and freely sharing informative content as a means of converting prospects into customers and customers into repeat buyers.[11]

Business models

Internet marketing is associated with several business models:

- E-commerce: a model whereby goods and services are sold directly to consumers (B2C), businesses (B2B), or from consumer to consumer (C2C) using computers connected to a network.[12]
- Lead-based websites: a strategy whereby an organization generates value by acquiring sales leads from its website. Similar to walk-in customers in retail world. These prospects are often referred to as organic leads.
- Affiliate Marketing: a process wherein a product or service developed by one entity is sold by other active sellers for a share of profits. The entity that owns the product may provide some marketing material (e.g., sales letters, affiliate links, tracking facilities, etc.); however, the vast majority of affiliate marketing relationships come from

e-commerce businesses that offer affiliate programs.

- Local Internet marketing: a strategy through which a small company utilizes the Internet to find and to nurture relationships that can be used for real-world advantages. Local Internet marketing uses tools such as social media marketing, local directory listing,[13] and targeted online sales promotions.

One-to-one approaches

In a one-to-one approach, marketers target a user browsing the Internet alone and so that the marketers'messages reach the user personally.[14] This approach is used in search marketing, for which the advertisements are based on search engine keywords entered by the users. This approach usually works under the pay per click (PPC) method.

Appeal to specific interests

When appealing to specific interests, marketers place an emphasis on appealing to a specific behavior or interest, rather than reaching out to a broadly defined demographic. These marketers typically segment their markets according to age group, gender, geography, and other general factors.

Niche marketing

Niche and hyper-niche internet marketing put further emphasis on creating destinations for web users and consumers on specific topics and products. Niche marketers differ from traditional Internet marketers as they have a more specialized topic knowledge. For example, whereas in traditional Internet marketing a website would be created and promoted on a high-level topic such as kitchen appliances, niche marketing would focus on more specific topics such as 4-slice toasters.

Niche marketing provides end users of such sites very targeted information, and allows the creators to establish themselves as authorities on the topic or product.

Geo-targeting

In Internet marketing, geo targeting and geo marketing are the methods of determining the geolocation of a website visitor with geolocation software, and delivering different content to that visitor based on his or her location, such as latitude and longitude, country, region or state, city, metro code or zip code, organization, Internet Protocol (IP) address, ISP, and other criteria.

Advantages and limitations of Internet marketing

Advantages

Internet marketing is inexpensive when examining the ratio of cost to the reach of the target audience. Companies can reach a wide audience for a small fraction of traditional advertising budgets. The nature of the medium allows consumers to research and to purchase products and services conveniently. Therefore, businesses have the advantage of appealing to consumers in a medium that can bring results quickly. The strategy and overall effectiveness of marketing campaigns depend on business goals and cost-volume-profit (CVP) analysis.

Internet marketers also have the advantage of measuring statistics easily and inexpensively; almost all aspects of an Internet marketing campaign can be traced, measured, and tested, in many cases through the use of an ad server. The advertisers can use a variety of methods, such as pay per impression, pay per click, pay per play, and pay per action. Therefore, marketers can determine which messages or offerings are more appealing to the audience. The results of campaigns can be measured and tracked immediately because online marketing initiatives usually require users to click on an advertisement, to visit a website, and to perform a targeted action.

Limitations

However, from the buyer's perspective, the inability of shoppers to touch, to smell, to taste, and "to try on" tangible goods before making an online purchase can be limiting. However, there is an industry standard for e-commerce vendors to reassure customers by having liberal return policies as well as providing in-store pick-up services.

Security concerns

Information security is important both to companies and consumers that participate in online business. Many consumers are hesitant to purchase items over the Internet because they do not believe that their personal information will remain private. Some companies that purchase customer information offer the option for individuals to have their information removed from their promotional redistribution, also known as opting out. However, many customers are unaware if and when their information is being shared, and are unable to stop the transfer of their information between companies if such activity occurs. Additionally, companies holding private information are vulnerable to data attacks and leaks.

Internet browsing privacy is a related consumer concern. Web sites routinely capture browsing and search history which can be used to provide targeted advertising. Privacy policies can provide transparency to these practices. Spyware prevention software can also be used to shield the consumer.

Another consumer e-commerce concern is whether or not they will receive exactly what they purchase. Online merchants have attempted to address this concern by investing in and building strong consumer brands (e.g., Amazon.com, eBay, and Overstock.com), and by leveraging merchant and feedback rating systems and e-commerce bonding solutions. All these solutions attempt to assure consumers that their transactions will be free of problems because the merchants can be trusted to provide reliable products and services. Additionally, several major online payment mechanisms (credit cards, PayPal, Google Checkout, etc.) have provided back-end buyer protection systems to address problems if they occur.

Usage trends

Technological advancements in the telecommunications industry have dramatically affected online advertising techniques. Many firms are embracing a paradigm that is shifting the focus of advertising methodology from traditional text and image advertisements to those containing more recent technologies like JavaScript and Adobe Flash. As a result, advertisers can more effectively engage and connect their audience with their campaigns that seek to shape consumer attitudes and feelings towards specific products and services.

Effects on industries

The number of banks offering the ability to perform banking tasks over the internet has increased. Online banking appeals to customers because it is often faster and considered more convenient than visiting bank branches.[15]

Internet auctions

Internet auctions have become a multi-billion dollar business. Unique items that could only previously be found at flea markets are now being sold on Internet auction websites such as eBay. Specialized e-stores sell a vast amount of items like antiques, movie props, clothing, gadgets, and so on.[16][17]

As the premier online reselling platform, eBay is often used as a price-basis for specialized items. Buyers and sellers often look at prices on the website before going to flea markets; the price shown on eBay often becomes the item's selling price.

Advertising industry

In addition to the major effect internet marketing has had on the technology industry, the effect on the advertising industry itself has been profound. In just a few years, online advertising has grown to be worth tens of billions of dollars annually.[18][19][20] PricewaterhouseCoopers reported that US$16.9 billion was spent on Online marketing in the U.S. in 2006.[21]

This has caused a growing impact on the United States' electoral process. In 2008, candidates for President heavily utilized Internet marketing strategies to reach constituents. During the 2007 primaries candidates added, on average, over 500 social network supporters per day to help spread their message.[22] President Barack Obama raised over US$1 million in one day during his extensive Democratic candidacy campaign, largely due to online donors.[23]

Several industries have heavily invested in and benefited from internet marketing and online advertising. Some of them were originally brick and mortar businesses such as publishing, music, automotive or gambling, while others have sprung up as purely online businesses, such as digital design and media, blogging, and internet service hosting.

References

[1] Jaakko Sinisalo et al. (2007). "Mobile customer relationship management: underlying issues and challenges". *Business Process Management Journal* **13** (6): 772. doi:10.1108/14637150710834541.

[2] Charlesworth, Alan (2009). *Internet marketing: a practical approach*. Butterworth-Heinemann. p. 49.

[3] Story, Louise and comScore (March 10, 2008). "They Know More Than You Think" (http://www.nytimes.com/imagepages/2008/03/10/technology/20080310_PRIVACY_GRAPHIC.html)(JPEG). *The New York Times*. . in Story, Louise (March 10, 2008). "To Aim Ads, Web Is Keeping Closer Eye on You" (http://www.nytimes.com/2008/03/10/technology/10privacy.html). *The New York Times* (The New York Times Company). . Retrieved 2008-03-09.

[4] "Define Online Marketing" (http://reference.yourdictionary.com/word-definitions/define-online-marketing.html). Yourdictionary.com. . Retrieved 9 January 2012.

[5] "What Is SEM / Search Engine Marketing?" (http://searchengineland.com/guide/what-is-sem). Search Engine Land. February 1, 2012. .

[6] Gurevych, Vadym (2007). *OsCommerce Webmaster's Guide to Selling Online: Increase Your Sales and Profits with Experts Tips on SEO, Marketing, Design, Selling Strategies Etc*. Birmingham. p. 49. ISBN 978-1-847192-02-8.

[7] "What Is Social Media Marketing" (http://searchengineland.com/guide/what-is-social-media-marketing). Search Engine Land. . Retrieved 9 January 2012.

[8] Miller, Michael (2009). *Selling Online 2.0: Migrating from EBay to Amazon, Craigslist, and Your Own E-commerce Website*. Indianapolis, IN. p. 287. ISBN 978-0-7897-3974-2.

[9] Thurow, Shari (2008). *Search Engine Visibility*. Berkeley, CA. p. 2. ISBN 978-0-321-50324-4.

[10] Croll, Alistair, and Sean Power (2009). *Complete Web Monitoring*. Beijing: O'Reilly. p. 97. ISBN 978-0-596-15513-1.

[11] "Content Marketing 101: How to Build Your Business With Content" (http://www.copyblogger.com/content-marketing/). copyblogger. .

[12] Peacock, Michael (2008). *Selling Online with Drupal E-Commerce: Walk through the Creation of an Online Store with Drupal's E-Commerce Module*. Birmingham. p. 2. ISBN 978-1-84719-406-0.

[13] Rayner, Andrew (April 21, 2010). "Put the E-mphasis on Local Internet Marketing and reach first page on Google" (http://www.prlog.org/ 10638959-put-the-mphasis-on-local-internet-marketing-and-reach-first-page-on-google.html). . Retrieved August 15, 2010.

[14] One to One Marketing – Overview (http://www.managingchange.com/onetoone/overview.htm). Managingchange.com (1995-04-05). Retrieved on 2011-10-30.

[15] Study Finds Convenience Drives Online Banking (http://www.consumeraffairs.com/news04/online_banking_survey.html). Consumeraffairs.com (2004-12-29). Retrieved on 2011-10-30.

[16] Mohr, Ian (February 27, 2006). "Movie props on the block: Mouse to auction Miramax leftovers" (http://www.variety.com/article/VR1117938954.html?categoryid=1238&cs=1). pReed Business Information. .

[17] James, David (February 24, 2007). "Bid on Dreamgirls Costumes for Charity"" (http://offtherack.people.com/2007/02/dress_like_a_dr. html#comment-66215834). Time, Inc. . .

[18] eMarketer – Online Ad Spending to Total $19.5 Billion in 2007 (http://www.emarketer.com/Article.aspx?1004635) (2007-2-28)

[19] The Register – Internet advertising shoots past estimates (http://www.theregister.co.uk/2006/09/29/internet_advertising_booms/) (2006-09-29)

[20] Internet Advertising Bureau – Online Adspend (http://www.iabuk.net/en/1/iabknowledgebankadspend.html) (2007-06-18)

[21] PricewaterhouseCoopers reported U.S. Internet marketing spend totaled $16.9 billion (http://www.directtraffic.org/OnlineNews/Internet_marketing_20075115473.html) in 2006" (Accessed 18-June-2007)

[22] "Spartan Internet Consulting – Political Performance Index (http://spartaninternet.com/2008/news.asp?id=9) (SIPP)" (Accessed 28-June-2008)

[23] "Center For Responsive Politics (http://www.opensecrets.org/pres08/summary.php?id=N00009638) Fundraising Profile Barack Obama" (Accessed 28-June-2008)

Web crawler

A **Web crawler** is a computer program that browses the World Wide Web in a methodical, automated manner or in an orderly fashion. Other terms for Web crawlers are *ants*, *automatic indexers*, *bots*,[1] *Web spiders*,[2] *Web robots*,[2] or—especially in the FOAF community—*Web scutters*.[3]

This process is called *Web crawling* or *spidering*. Many sites, in particular search engines, use spidering as a means of providing up-to-date data. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches. Crawlers can also be used for automating maintenance tasks on a Web site, such as checking links or validating HTML code. Also, crawlers can be used to gather specific types of information from Web pages, such as harvesting e-mail addresses (usually for sending spam).

A Web crawler is one type of bot, or software agent. In general, it starts with a list of URLs to visit, called the *seeds*. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the *crawl frontier*. URLs from the frontier are recursively visited according to a set of policies.

The large volume implies that the crawler can only download a fraction of the Web pages within a given time, so it needs to prioritize its downloads. The high rate of change implies that the pages might have already been updated or even deleted.

The number of possible crawlable URLs being generated by server-side software has also made it difficult for web crawlers to avoid retrieving duplicate content. Endless combinations of HTTP GET (URL-based) parameters exist, of which only a small selection will actually return unique content. For example, a simple online photo gallery may offer three options to users, as specified through HTTP GET parameters in the URL. If there exist four ways to sort images, three choices of thumbnail size, two file formats, and an option to disable user-provided content, then the same set of content can be accessed with 48 different URLs, all of which may be linked on the site. This mathematical combination creates a problem for crawlers, as they must sort through endless combinations of relatively minor scripted changes in order to retrieve unique content.

As Edwards *et al.* noted, "Given that the bandwidth for conducting crawls is neither infinite nor free, it is becoming essential to crawl the Web in not only a scalable, but efficient way, if some reasonable measure of quality or freshness is to be maintained."[4] A crawler must carefully choose at each step which pages to visit next.

The behavior of a Web crawler is the outcome of a combination of policies:[5]

- a *selection policy* that states which pages to download,
- a *re-visit policy* that states when to check for changes to the pages,
- a *politeness policy* that states how to avoid overloading Web sites, and
- a *parallelization policy* that states how to coordinate distributed Web crawlers.

Selection policy

Given the current size of the Web, even large search engines cover only a portion of the publicly-available part. A 2005 study showed that large-scale search engines index no more than 40%-70% of the indexable Web;[6] a previous study by Dr. Steve Lawrence and Lee Giles showed that no search engine indexed more than 16% of the Web in 1999.[7] As a crawler always downloads just a fraction of the Web pages, it is highly desirable that the downloaded fraction contains the most relevant pages and not just a random sample of the Web.

This requires a metric of importance for prioritizing Web pages. The importance of a page is a function of its intrinsic quality, its popularity in terms of links or visits, and even of its URL (the latter is the case of vertical search

engines restricted to a single top-level domain, or search engines restricted to a fixed Web site). Designing a good selection policy has an added difficulty: it must work with partial information, as the complete set of Web pages is not known during crawling.

Cho *et al.* made the first study on policies for crawling scheduling. Their data set was a 180,000-pages crawl from the `stanford.edu` domain, in which a crawling simulation was done with different strategies.[8] The ordering metrics tested were breadth-first, backlink-count and partial Pagerank calculations. One of the conclusions was that if the crawler wants to download pages with high Pagerank early during the crawling process, then the partial Pagerank strategy is the better, followed by breadth-first and backlink-count. However, these results are for just a single domain. Cho also wrote his Ph.D. dissertation at Stanford on web crawling.[9]

Najork and Wiener performed an actual crawl on 328 million pages, using breadth-first ordering.[10] They found that a breadth-first crawl captures pages with high Pagerank early in the crawl (but they did not compare this strategy against other strategies). The explanation given by the authors for this result is that "the most important pages have many links to them from numerous hosts, and those links will be found early, regardless of on which host or page the crawl originates."

Abiteboul designed a crawling strategy based on an algorithm called OPIC (On-line Page Importance Computation).[11] In OPIC, each page is given an initial sum of "cash" that is distributed equally among the pages it points to. It is similar to a Pagerank computation, but it is faster and is only done in one step. An OPIC-driven crawler downloads first the pages in the crawling frontier with higher amounts of "cash". Experiments were carried in a 100,000-pages synthetic graph with a power-law distribution of in-links. However, there was no comparison with other strategies nor experiments in the real Web.

Boldi *et al.* used simulation on subsets of the Web of 40 million pages from the `.it` domain and 100 million pages from the WebBase crawl, testing breadth-first against depth-first, random ordering and an omniscient strategy. The comparison was based on how well PageRank computed on a partial crawl approximates the true PageRank value. Surprisingly, some visits that accumulate PageRank very quickly (most notably, breadth-first and the omniscent visit) provide very poor progressive approximations.[12][13]

Baeza-Yates *et al.* used simulation on two subsets of the Web of 3 million pages from the `.gr` and `.cl` domain, testing several crawling strategies.[14] They showed that both the OPIC strategy and a strategy that uses the length of the per-site queues are better than breadth-first crawling, and that it is also very effective to use a previous crawl, when it is available, to guide the current one.

Daneshpajouh *et al.* designed a community based algorithm for discovering good seeds.[15] Their method crawls web pages with high PageRank from different communities in less iteration in comparison with crawl starting from random seeds. One can extract good seed from a previously-crawled-Web graph using this new method. Using these seeds a new crawl can be very effective.

Focused crawling

The importance of a page for a crawler can also be expressed as a function of the similarity of a page to a given query. Web crawlers that attempt to download pages that are similar to each other are called **focused crawler** or **topical crawlers**. The concepts of topical and focused crawling were first introduced by Menczer[16][17] and by Chakrabarti *et al.*[18]

The main problem in focused crawling is that in the context of a Web crawler, we would like to be able to predict the similarity of the text of a given page to the query before actually downloading the page. A possible predictor is the anchor text of links; this was the approach taken by Pinkerton[19] in the first web crawler of the early days of the Web. Diligenti *et al.* [20] propose using the complete content of the pages already visited to infer the similarity between the driving query and the pages that have not been visited yet. The performance of a focused crawling depends mostly on the richness of links in the specific topic being searched, and a focused crawling usually relies on a general Web search engine for providing starting points..

Restricting followed links

A crawler may only want to seek out HTML pages and avoid all other MIME types. In order to request only HTML resources, a crawler may make an HTTP HEAD request to determine a Web resource's MIME type before requesting the entire resource with a GET request. To avoid making numerous HEAD requests, a crawler may examine the URL and only request a resource if the URL ends with certain characters such as .html, .htm, .asp, .aspx, .php, .jsp,
.jspx or a slash. This strategy may cause numerous HTML Web resources to be unintentionally skipped.

Some crawlers may also avoid requesting any resources that have a "?" in them (are dynamically produced) in order to avoid spider traps that may cause the crawler to download an infinite number of URLs from a Web site. This strategy is unreliable if the site uses URL rewriting to simplify its URLs.

URL normalization

Crawlers usually perform some type of URL normalization in order to avoid crawling the same resource more than once. The term *URL normalization*, also called *URL canonicalization*, refers to the process of modifying and standardizing a URL in a consistent manner. There are several types of normalization that may be performed including conversion of URLs to lowercase, removal of "." and ".." segments, and adding trailing slashes to the non-empty path component.[21]

Path-ascending crawling

Some crawlers intend to download as many resources as possible from a particular web site. So *path-ascending crawler* was introduced that would ascend to every path in each URL that it intends to crawl.[22] For example, when given a seed URL of http://llama.org/hamster/monkey/page.html, it will attempt to crawl /hamster/monkey/,
/hamster/, and /. Cothey found that a path-ascending crawler was very effective in finding isolated resources, or resources for which no inbound link would have been found in regular crawling.

Many path-ascending crawlers are also known as Web harvesting software, because they're used to "harvest" or collect all the content—perhaps the collection of photos in a gallery—from a specific page or host.

Re-visit policy

The Web has a very dynamic nature, and crawling a fraction of the Web can take weeks or months. By the time a Web crawler has finished its crawl, many events could have happened, including creations, updates and deletions.

From the search engine's point of view, there is a cost associated with not detecting an event, and thus having an outdated copy of a resource. The most-used cost functions are freshness and age.[23]

Freshness: This is a binary measure that indicates whether the local copy is accurate or not. The freshness of a page p in the repository at time t is defined as:

$$F_p(t) = \begin{cases} 1 & \text{if } p \text{ is equal to the local copy at time } t \\ 0 & \text{otherwise} \end{cases}$$

Age: This is a measure that indicates how outdated the local copy is. The age of a page p in the repository, at time t is defined as:

$$A_p(t) = \begin{cases} 0 & \text{if } p \text{ is not modified at time } t \\ t - \text{modification time of } p & \text{otherwise} \end{cases}$$

Coffman *et al.* worked with a definition of the objective of a Web crawler that is equivalent to freshness, but use a different wording: they propose that a crawler must minimize the fraction of time pages remain outdated. They also noted that the problem of Web crawling can be modeled as a multiple-queue, single-server polling system, on which the Web crawler is the server and the Web sites are the queues. Page modifications are the arrival of the customers, and switch-over times are the interval between page accesses to a single Web site. Under this model, mean waiting

time for a customer in the polling system is equivalent to the average age for the Web crawler.[24]

The objective of the crawler is to keep the average freshness of pages in its collection as high as possible, or to keep the average age of pages as low as possible. These objectives are not equivalent: in the first case, the crawler is just concerned with how many pages are out-dated, while in the second case, the crawler is concerned with how old the local copies of pages are.

Two simple re-visiting policies were studied by Cho and Garcia-Molina:[25]

Uniform policy: This involves re-visiting all pages in the collection with the same frequency, regardless of their rates of change.

Proportional policy: This involves re-visiting more often the pages that change more frequently. The visiting frequency is directly proportional to the (estimated) change frequency.

(In both cases, the repeated crawling order of pages can be done either in a random or a fixed order.)

Cho and Garcia-Molina proved the surprising result that, in terms of average freshness, the uniform policy outperforms the proportional policy in both a simulated Web and a real Web crawl. Intuitively, the reasoning is that, as web crawlers have a limit to how many pages they can crawl in a given time frame, (1) they will allocate too many new crawls to rapidly changing pages at the expense of less frequently updating pages, and (2) the freshness of rapidly changing pages lasts for shorter period than that of less frequently changing pages. In other words, a proportional policy allocates more resources to crawling frequently updating pages, but experiences less overall freshness time from them.

To improve freshness, the crawler should penalize the elements that change too often.[26] The optimal re-visiting policy is neither the uniform policy nor the proportional policy. The optimal method for keeping average freshness high includes ignoring the pages that change too often, and the optimal for keeping average age low is to use access frequencies that monotonically (and sub-linearly) increase with the rate of change of each page. In both cases, the optimal is closer to the uniform policy than to the proportional policy: as Coffman *et al.* note, "in order to minimize the expected obsolescence time, the accesses to any particular page should be kept as evenly spaced as possible".[24] Explicit formulas for the re-visit policy are not attainable in general, but they are obtained numerically, as they depend on the distribution of page changes. Cho and Garcia-Molina show that the exponential distribution is a good fit for describing page changes,[26] while Ipeirotis *et al.* show how to use statistical tools to discover parameters that affect this distribution.[27] Note that the re-visiting policies considered here regard all pages as homogeneous in terms of quality ("all pages on the Web are worth the same"), something that is not a realistic scenario, so further information about the Web page quality should be included to achieve a better crawling policy.

Politeness policy

Crawlers can retrieve data much quicker and in greater depth than human searchers, so they can have a crippling impact on the performance of a site. Needless to say, if a single crawler is performing multiple requests per second and/or downloading large files, a server would have a hard time keeping up with requests from multiple crawlers.

As noted by Koster, the use of Web crawlers is useful for a number of tasks, but comes with a price for the general community.[28] The costs of using Web crawlers include:

- network resources, as crawlers require considerable bandwidth and operate with a high degree of parallelism during a long period of time;
- server overload, especially if the frequency of accesses to a given server is too high;
- poorly-written crawlers, which can crash servers or routers, or which download pages they cannot handle; and
- personal crawlers that, if deployed by too many users, can disrupt networks and Web servers.

A partial solution to these problems is the robots exclusion protocol, also known as the robots.txt protocol that is a standard for administrators to indicate which parts of their Web servers should not be accessed by crawlers.[29] This standard does not include a suggestion for the interval of visits to the same server, even though this interval is the

most effective way of avoiding server overload. Recently commercial search engines like Ask Jeeves, MSN and Yahoo are able to use an extra "Crawl-delay:" parameter in the robots.txt file to indicate the number of seconds to delay between requests.

The first proposed interval between connections was 60 seconds.[30] However, if pages were downloaded at this rate from a website with more than 100,000 pages over a perfect connection with zero latency and infinite bandwidth, it would take more than 2 months to download only that entire Website; also, only a fraction of the resources from that Web server would be used. This does not seem acceptable.

Cho uses 10 seconds as an interval for accesses,[25] and the WIRE crawler uses 15 seconds as the default.[31] The MercatorWeb crawler follows an adaptive politeness policy: if it took t seconds to download a document from a given server, the crawler waits for $10t$ seconds before downloading the next page.[32] Dill *et al.* use 1 second.[33]

For those using Web crawlers for research purposes, a more detailed cost-benefit analysis is needed and ethical considerations should be taken into account when deciding where to crawl and how fast to crawl.[34]

Anecdotal evidence from access logs shows that access intervals from known crawlers vary between 20 seconds and 3–4 minutes. It is worth noticing that even when being very polite, and taking all the safeguards to avoid overloading Web servers, some complaints from Web server administrators are received. Brin and Page note that: "... running a crawler which connects to more than half a million servers (...) generates a fair amount of e-mail and phone calls. Because of the vast number of people coming on line, there are always those who do not know what a crawler is, because this is the first one they have seen."[35]

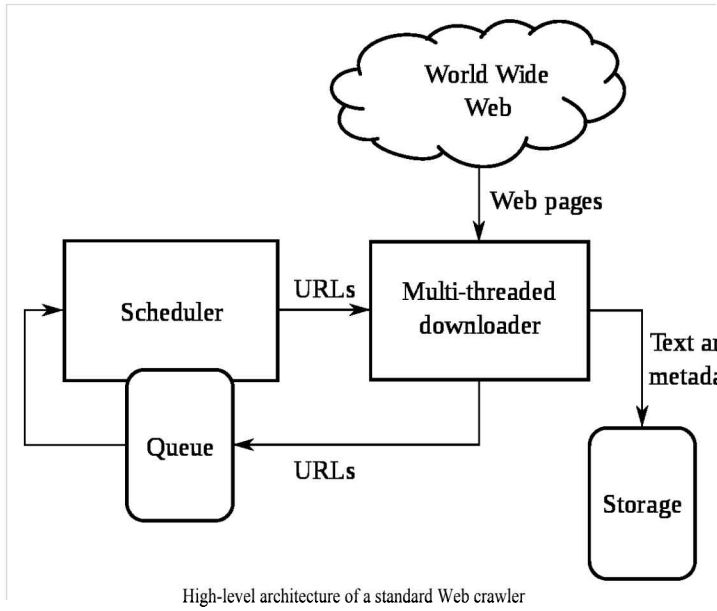
Parallelization policy

A parallel crawler is a crawler that runs multiple processes in parallel. The goal is to maximize the download rate while minimizing the overhead from parallelization and to avoid repeated downloads of the same page. To avoid downloading the same page more than once, the crawling system requires a policy for assigning the new URLs discovered during the crawling process, as the same URL can be found by two different crawling processes.

Architectures

A crawler must not only have a good crawling strategy, as noted in the previous sections, but it should also have a highly optimized architecture. Shkapenyuk and Suel noted that:[36]

> While it is fairly easy to build a slow crawler that downloads a few pages per second for a short period of time, building a high-performance system that can download hundreds of millions of pages over several weeks presents a number of challenges in system design, I/O and network efficiency, and robustness and manageability.



High-level architecture of a standard Web crawler

Web crawlers are a central part of search engines, and details on their algorithms and architecture are kept as business secrets. When crawler designs are published, there is often an important lack of detail that prevents others from reproducing the work. There are also emerging concerns about "search engine spamming", which prevent major search engines from publishing their ranking algorithms.

Crawler identification

Web crawlers typically identify themselves to a Web server by using the User-agent field of an HTTP request. Web site administrators typically examine their Web servers' log and use the user agent field to determine which crawlers have visited the web server and how often. The user agent field may include a URL where the Web site administrator may find out more information about the crawler. Spambots and other malicious Web crawlers are unlikely to place identifying information in the user agent field, or they may mask their identity as a browser or other well-known crawler.

It is important for Web crawlers to identify themselves so that Web site administrators can contact the owner if needed. In some cases, crawlers may be accidentally trapped in a crawler trap or they may be overloading a Web server with requests, and the owner needs to stop the crawler. Identification is also useful for administrators that are interested in knowing when they may expect their Web pages to be indexed by a particular search engine.

Examples

The following is a list of published crawler architectures for general-purpose crawlers (excluding focused web crawlers), with a brief description that includes the names given to the different components and outstanding features:

- **Yahoo! Slurp** is the name of the Yahoo Search crawler.
- **Bingbot** is the name of Microsoft's Bing webcrawler. It replaced **Msnbot**.
- **FAST Crawler**[37] is a distributed crawler, used by Fast Search & Transfer, and a general description of its architecture is available.
- **Googlebot**[35] is described in some detail, but the reference is only about an early version of its architecture, which was based in C++ and Python. The crawler was integrated with the indexing process, because text parsing was done for full-text indexing and also for URL extraction. There is a URL server that sends lists of URLs to be fetched by several crawling processes. During parsing, the URLs found were passed to a URL server that checked if the URL have been previously seen. If not, the URL was added to the queue of the URL server.
- **PolyBot**[36] is a distributed crawler written in C++ and Python, which is composed of a "crawl manager", one or more "downloaders" and one or more "DNS resolvers". Collected URLs are added to a queue on disk, and processed later to search for seen URLs in batch mode. The politeness policy considers both third and second level domains (e.g.: www.example.com and www2.example.com are third level domains) because third level domains are usually hosted by the same Web server.
- **RBSE**[38] was the first published web crawler. It was based on two programs: the first program, "`spider`" maintains a queue in a relational database, and the second program "`mite`", is a modified `www` ASCII browser that downloads the pages from the Web.
- **WebCrawler**[19] was used to build the first publicly-available full-text index of a subset of the Web. It was based on lib-WWW to download pages, and another program to parse and order URLs for breadth-first exploration of the Web graph. It also included a real-time crawler that followed links based on the similarity of the anchor text with the provided query.
- **World Wide Web Worm**[39] was a crawler used to build a simple index of document titles and URLs. The index could be searched by using the `grep` Unix command.
- **WebFountain**[4] is a distributed, modular crawler similar to Mercator but written in C++. It features a "controller" machine that coordinates a series of "ant" machines. After repeatedly downloading pages, a change

rate is inferred for each page and a non-linear programming method must be used to solve the equation system for maximizing freshness. The authors recommend to use this crawling order in the early stages of the crawl, and then switch to a uniform crawling order, in which all pages are being visited with the same frequency.

- **WebRACE**[40] is a crawling and caching module implemented in Java, and used as a part of a more generic system called eRACE. The system receives requests from users for downloading web pages, so the crawler acts in part as a smart proxy server. The system also handles requests for "subscriptions" to Web pages that must be monitored: when the pages change, they must be downloaded by the crawler and the subscriber must be notified. The most outstanding feature of WebRACE is that, while most crawlers start with a set of "seed" URLs, WebRACE is continuously receiving new starting URLs to crawl from.

In addition to the specific crawler architectures listed above, there are general crawler architectures published by Cho[41] and Chakrabarti.[42]

Open-source crawlers

- **Aspseek** is a crawler, indexer and a search engine written in C++ and licensed under the GPL
- **DataparkSearch** is a crawler and search engine released under the GNU General Public License.
- **GNU Wget** is a command-line-operated crawler written in C and released under the GPL. It is typically used to mirror Web and FTP sites.
- **GRUB** is an open source distributed search crawler that Wikia Search used to crawl the web.
- **Heritrix** is the Internet Archive's archival-quality crawler, designed for archiving periodic snapshots of a large portion of the Web. It was written in Java.
- **ht://Dig** includes a Web crawler in its indexing engine.
- **HTTrack** uses a Web crawler to create a mirror of a web site for off-line viewing. It is written in C and released under the GPL.
- **ICDL Crawler** is a cross-platform web crawler written in C++ and intended to crawl Web sites based on Web-site Parse Templates using computer's free CPU resources only.
- **mnoGoSearch** is a crawler, indexer and a search engine written in C and licensed under the GPL (Linux machines only)
- **Nutch** is a crawler written in Java and released under an Apache License. It can be used in conjunction with the Lucene text-indexing package.
- **Open Search Server** is a search engine and web crawler software release under the GPL.
- **Pavuk** is a command-line Web mirror tool with optional X11 GUI crawler and released under the GPL. It has bunch of advanced features compared to wget and httrack, e.g., regular expression based filtering and file creation rules.
- **PHP-Crawler** is a simple PHP and MySQL based crawler released under the BSD. Easy to install it became popular for small MySQL-driven websites on shared hosting.
- the **tkWWW Robot**, a crawler based on the tkWWW web browser (licensed under GPL).
- **YaCy**, a free distributed search engine, built on principles of peer-to-peer networks (licensed under GPL).
- **Seeks**, a free distributed search engine (licensed under Affero General Public License).

Crawling the Deep Web

A vast amount of Web pages lie in the deep or invisible Web.[43] These pages are typically only accessible by submitting queries to a database, and regular crawlers are unable to find these pages if there are no links that point to them. Google's Sitemap Protocol and mod oai[44] are intended to allow discovery of these deep-Web resources.

Deep Web crawling also multiplies the number of Web links to be crawled. Some crawlers only take some of the `<a href="URL">`-shaped URLs. In some cases, such as the Googlebot, Web crawling is done on all text contained inside the hypertext content, tags, or text.

Crawling Web 2.0 Applications

- Sheeraj Shah provides insight into Crawling Ajax-driven Web 2.0 Applications[45].
- Interested readers might wish to read AJAXSearch: Crawling, Indexing and Searching Web 2.0 Applications[46].
- Making AJAX Applications Crawlable[47], from Google Code. It defines an agreement between web servers and search engine crawlers that allows for dynamically created content to be visible to crawlers. Google currently supports this agreement.[48]

References

[1] Kobayashi, M. and Takeda, K. (2000). "Information retrieval on the web" (http://doi.acm.org/10.1145/358923.358934). *ACM Computing Surveys* (ACM Press) **32** (2): 144–173. doi:10.1145/358923.358934. .

[2] Spetka, Scott. "The TkWWW Robot: Beyond Browsing" (http://web.archive.org/web/20040903174942/archive.ncsa.uiuc.edu/SDG/IT94/Proceedings/Agents/spetka/spetka.html). NCSA. Archived from the original (http://archive.ncsa.uiuc.edu/SDG/IT94/Proceedings/Agents/spetka/spetka.html) on 3 September 2004. . Retrieved 21 November 2010.

[3] See definition of scutter on FOAF Project's wiki (http://wiki.foaf-project.org/w/Scutter)

[4] Edwards, J., McCurley, K. S., and Tomlin, J. A. (2001). "An adaptive model for optimizing performance of an incremental web crawler" (http://www10.org/cdrom/papers/210/index.html). *In Proceedings of the Tenth Conference on World Wide Web* (Hong Kong: Elsevier Science): 106–113. doi:10.1145/371920.371960. .

[5] Castillo, Carlos (2004). *Effective Web Crawling* (http://chato.cl/research/crawling_thesis) (Ph.D. thesis). University of Chile. . Retrieved 2010-08-03.

[6] Gulli, A.; Signorini, A. (2005). "The indexable web is more than 11.5 billion pages" (http://doi.acm.org/10.1145/1062745.1062789). *Special interest tracks and posters of the 14th international conference on World Wide Web*. ACM Press.. pp. 902–903. doi:10.1145/1062745.1062789. .

[7] Lawrence, Steve; C. Lee Giles (1999-07-08). "Accessibility of information on the web". *Nature* **400** (6740): 107. doi:10.1038/21987. PMID 10428673.

[8] Cho, J.; Garcia-Molina, H.; Page, L. (1998-04). "Efficient Crawling Through URL Ordering" (http://ilpubs.stanford.edu:8090/347/). *Seventh International World-Wide Web Conference*. Brisbane, Australia. . Retrieved 2009-03-23.

[9] Cho, Junghoo, "Crawling the Web: Discovery and Maintenance of a Large-Scale Web Data" (http://oak.cs.ucla.edu/~cho/papers/ cho-thesis.pdf), Ph.D. dissertation, Department of Computer Science, Stanford University, November 2001

[10] Marc Najork and Janet L. Wiener. Breadth-first crawling yields high-quality pages (http://www10.org/cdrom/papers/pdf/p208.pdf). In Proceedings of the Tenth Conference on World Wide Web, pages 114–118, Hong Kong, May 2001. Elsevier Science.

[11] Abiteboul, Serge; Mihai Preda, Gregory Cobena (2003). "Adaptive on-line page importance computation" (http://www2003.org/cdrom/ papers/refereed/p007/p7-abiteboul.html). *Proceedings of the 12th international conference on World Wide Web*. Budapest, Hungary: ACM. pp. 280–290. doi:10.1145/775152.775192. ISBN 1-58113-680-3. . Retrieved 2009-03-22.

[12] Boldi, Paolo; Bruno Codenotti, Massimo Santini, Sebastiano Vigna (2004). "UbiCrawler: a scalable fully distributed Web crawler" (http://vigna.dsi.unimi.it/ftp/papers/UbiCrawler.pdf). *Software: Practice and Experience* **34** (8): 711–726. doi:10.1002/spe.587. . Retrieved 2009-03-23.

[13] Boldi, Paolo; Massimo Santini, Sebastiano Vigna (2004). "Do Your Worst to Make the Best: Paradoxical Effects in PageRank Incremental Computations" (http://vigna.dsi.unimi.it/ftp/papers/ParadoxicalPageRank.pdf). *Algorithms and Models for the Web-Graph* (http://springerlink.com/content/g10m122f9hb6). pp. 168–180. . Retrieved 2009-03-23.

[14] Baeza-Yates, R., Castillo, C., Marin, M. and Rodriguez, A. (2005). Crawling a Country: Better Strategies than Breadth-First for Web Page Ordering (http://www.dcc.uchile.cl/~ccastill/papers/baeza05_crawling_country_better_breadth_first_web_page_ordering.pdf). In Proceedings of the Industrial and Practical Experience track of the 14th conference on World Wide Web, pages 864–872, Chiba, Japan. ACM Press.

[15] Shervin Daneshpajouh, Mojtaba Mohammadi Nasiri, Mohammad Ghodsi, A Fast Community Based Algorithm for Generating Crawler Seeds Set (http://ce.sharif.edu/~daneshpajouh/publications/A Fast Community Based Algorithm for Generating Crawler Seeds Set.pdf),

In proceeding of 4th International Conference on Web Information Systems and Technologies (WEBIST-2008 (http://www.webist.org/)), Funchal, Portugal, May 2008.

[16] Menczer, F. (1997). ARACHNID: Adaptive Retrieval Agents Choosing Heuristic Neighborhoods for Information Discovery (http://informatics.indiana.edu/fil/Papers/ICML.ps). In D. Fisher, ed., Machine Learning: Proceedings of the 14th International Conference (ICML97). Morgan Kaufmann

[17] Menczer, F. and Belew, R.K. (1998). Adaptive Information Agents in Distributed Textual Environments (http://informatics.indiana.edu/ fil/Papers/AA98.ps). In K. Sycara and M. Wooldridge (eds.) Proc. 2nd Intl. Conf. on Autonomous Agents (Agents '98). ACM Press

[18] Chakrabarti, S., van den Berg, M., and Dom, B. (1999). Focused crawling: a new approach to topic-specific web resource discovery (http://web.archive.org/web/20040317210216/http://www.fxpal.com/people/vdberg/pubs/www8/www1999f.pdf). Computer Networks, 31(11–16):1623–1640.

[19] Pinkerton, B. (1994). Finding what people want: Experiences with the WebCrawler (http://web.archive.org/web/20010904075500/http://archive.ncsa.uiuc.edu/SDG/IT94/Proceedings/Searching/pinkerton/WebCrawler.html). In Proceedings of the First World Wide Web Conference, Geneva, Switzerland.

[20] Diligenti, M., Coetzee, F., Lawrence, S., Giles, C. L., and Gori, M. (2000). Focused crawling using context graphs (http://nautilus.dii.unisi.it/pubblicazioni/files/conference/2000-Diligenti-VLDB.pdf). In Proceedings of 26th International Conference on Very Large Databases (VLDB), pages 527-534, Cairo, Egypt.

[21] Pant, Gautam; Srinivasan, Padmini; Menczer, Filippo (2004). "Crawling the Web" (http://dollar.biz.uiowa.edu/~pant/Papers/crawling. pdf). In Levene, Mark; Poulovassilis, Alexandra. *Web Dynamics: Adapting to Change in Content, Size, Topology and Use*. Springer. pp. 153–178. ISBN 9783540406761. . Retrieved 2009-03-22.

[22] Cothey, Viv (2004). "Web-crawling reliability". *Journal of the American Society for Information Science and Technology* **55** (14): 1228–1238. doi:10.1002/asi.20078.

[23] Cho, Junghoo; Hector Garcia-Molina (2000). "Synchronizing a database to improve freshness" (http://www.cs.brown.edu/courses/ cs227/2002/cache/Cho.pdf). *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. Dallas, Texas, United States: ACM. pp. 117–128. doi:10.1145/342009.335391. ISBN 1-58113-217-4. . Retrieved 2009-03-23.

[24] Jr, E. G. Coffman; Zhen Liu, Richard R. Weber (1998). "Optimal robot scheduling for Web search engines". *Journal of Scheduling* **1** (1): 15–29. doi:10.1002/(SICI)1099-1425(199806)1:1<15::AID-JOS3>3.0.CO;2-K.

[25] Cho, J. and Garcia-Molina, H. (2003). Effective page refresh policies for web crawlers (http://portal.acm.org/citation.cfm?doid=958942. 958945). ACM Transactions on Database Systems, 28(4).

[26] Cho, Junghoo; Hector Garcia-Molina (2003). "Estimating frequency of change" (http://portal.acm.org/citation.cfm?doid=857166. 857170). *ACM Trans. Interet Technol.* **3** (3): 256–290. doi:10.1145/857166.857170. . Retrieved 2009-03-22.

[27] Ipeirotis, P., Ntoulas, A., Cho, J., Gravano, L. (2005) Modeling and managing content changes in text databases (http://pages.stern.nyu.edu/~panos/publications/icde2005.pdf). In Proceedings of the 21st IEEE International Conference on Data Engineering, pages 606-617, April 2005, Tokyo.

[28] Koster, M. (1995). Robots in the web: threat or treat? ConneXions, 9(4).

[29] Koster, M. (1996). A standard for robot exclusion (http://www.robotstxt.org/wc/exclusion.html).

[30] Koster, M. (1993). Guidelines for robots writers (http://www.robotstxt.org/wc/guidelines.html).

[31] Baeza-Yates, R. and Castillo, C. (2002). Balancing volume, quality and freshness in Web crawling (http://www.chato.cl/papers/ baeza02balancing.pdf). In Soft Computing Systems – Design, Management and Applications, pages 565–572, Santiago, Chile. IOS Press Amsterdam.

[32] Heydon, Allan; Najork, Marc (1999-06-26) (PDF). *Mercator: A Scalable, Extensible Web Crawler* (http://www.cindoc.csic.es/cybermetrics/pdf/68.pdf). . Retrieved 2009-03-22.

[33] Dill, S., Kumar, R., Mccurley, K. S., Rajagopalan, S., Sivakumar, D., and Tomkins, A. (2002). Self-similarity in the web (http://www.mccurley.org/papers/fractal.pdf). ACM Trans. Inter. Tech., 2(3):205–223.

[34] "Web crawling ethics revisited: Cost, privacy and denial of service" (http://www.scit.wlv.ac.uk/~cm1993/papers/Web_Crawling_Ethics_preprint.doc). *Journal of the American Society for Information Science and Technology*. 2006. .

[35] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine (http://infolab.stanford.edu/~backrub/google. html). Computer Networks and ISDN Systems, 30(1-7):107–117.

[36] Shkapenyuk, V. and Suel, T. (2002). Design and implementation of a high performance distributed web crawler (http://cis.poly.edu/tr/tr-cis-2001-03.pdf). In Proceedings of the 18th International Conference on Data Engineering (ICDE), pages 357-368, San Jose, California. IEEE CS Press.

[37] Risvik, K. M. and Michelsen, R. (2002). Search Engines and Web Dynamics (http://citeseer.ist.psu.edu/rd/1549722,509701,1,0.25,Download/http://citeseer.ist.psu.edu/cache/papers/cs/26004/http:zSzzSzwww.idi.ntnu.nozSz~algkonzSzgenereltzSzse-dynamicweb1.pdf/risvik02search.pdf). Computer Networks, vol. 39, pp. 289–302, June 2002.

[38] Eichmann, D. (1994). The RBSE spider: balancing effective search against Web load (http://mingo.info-science.uiowa.edu/eichmann/ www94/Spider.ps). In Proceedings of the First World Wide Web Conference, Geneva, Switzerland.

[39] McBryan, O. A. (1994). GENVL and WWWW: Tools for taming the web. In Proceedings of the First World Wide Web Conference, Geneva, Switzerland.

[40] Zeinalipour-Yazti, D. and Dikaiakos, M. D. (2002). Design and implementation of a distributed crawler and filtering processor (http://www.cs.ucr.edu/~csyiazti/downloads/papers/ngits02/ngits02.pdf). In Proceedings of the Fifth Next Generation Information Technologies

and Systems (NGITS), volume 2382 of Lecture Notes in Computer Science, pages 58–74, Caesarea, Israel. Springer.

[41] Cho, Junghoo; Hector Garcia-Molina (2002). "Parallel crawlers" (http://portal.acm.org/citation.cfm?id=511464). *Proceedings of the 11th international conference on World Wide Web*. Honolulu, Hawaii, USA: ACM. pp. 124–135. doi:10.1145/511446.511464. ISBN 1-58113-449-5. . Retrieved 2009-03-23.

[42] Chakrabarti, S. (2003). Mining the Web (http://www.cs.berkeley.edu/~soumen/mining-the-web/). Morgan Kaufmann Publishers. ISBN 1-55860-754-4

[43] Shestakov, Denis (2008). Search Interfaces on the Web: Querying and Characterizing (https://oa.doria.fi/handle/10024/38506). *TUCS Doctoral Dissertations 104, University of Turku*

[44] Nelson, Michael L; Herbert Van de Sompel, Xiaoming Liu, Terry L Harrison, Nathan McFarland (2005-03-24). "mod_oai: An Apache Module for Metadata Harvesting". *Eprint arXiv:cs/0503069*: 3069. arXiv:cs/0503069. Bibcode 2005cs.....3069N.

[45] http://www.infosecwriters.com/text_resources/pdf/Crawling_AJAX_SShah.pdf

[46] http://www.dbis.ethz.ch/research/publications/AjaxSearchVLDB08.pdf

[47] http://code.google.com/web/ajaxcrawling/index.html

[48] Making AJAX Applications Crawlable: Full Specification (http://code.google.com/web/ajaxcrawling/docs/specification.html)

Further reading

- Cho, Junghoo, "Web Crawling Project" (http://oak.cs.ucla.edu/~cho/research/crawl.html), UCLA Computer Science Department.

Backlinks

Backlinks, also known as **incoming links**, **inbound links**, **inlinks**, and **inward links**, are incoming links to a website or web page. In basic link terminology, a **backlink** is any link received by a web node (web page, directory, website, or top level domain) from another web node.[1]

Inbound links were originally important (prior to the emergence of search engines) as a primary means of web navigation; today, their significance lies in search engine optimization (SEO). The number of backlinks is one indication of the popularity or importance of that website or page (for example, this is used by Google to determine the PageRank of a webpage). Outside of SEO, the backlinks of a webpage may be of significant personal, cultural or semantic interest: they indicate who is paying attention to that page.

Search engine rankings

Search engines often use the number of backlinks that a website has as one of the most important factors for determining that website's search engine ranking, popularity and importance. Google's description of their PageRank system, for instance, notes that *Google interprets a link from page A to page B as a vote, by page A, for page B.*[2] Knowledge of this form of search engine rankings has fueled a portion of the SEO industry commonly termed linkspam, where a company attempts to place as many inbound links as possible to their site regardless of the context of the originating site.

Websites often employ various search engine optimization techniques to increase the number of backlinks pointing to their website. Some methods are free for use by everyone whereas some methods like linkbaiting requires quite a bit of planning and marketing to work. Some websites stumble upon "linkbaiting" naturally; the sites that are the first with a tidbit of 'breaking news' about a celebrity are good examples of that. When "linkbait" happens, many websites will link to the 'baiting' website because there is information there that is of extreme interest to a large number of people.

There are several factors that determine the value of a backlink. Backlinks from authoritative sites on a given topic are highly valuable.[3] If both sites have content geared toward the keyword topic, the backlink is considered relevant and believed to have strong influence on the search engine rankings of the webpage granted the backlink. A backlink represents a favorable 'editorial vote' for the receiving webpage from another granting webpage. Another important

factor is the anchor text of the backlink. Anchor text is the descriptive labeling of the hyperlink as it appears on a webpage. Search engine bots (i.e., spiders, crawlers, etc.) examine the anchor text to evaluate how relevant it is to the content on a webpage. Anchor text and webpage content congruency are highly weighted in search engine results page (SERP) rankings of a webpage with respect to any given keyword query by a search engine user.

Increasingly, inbound links are being weighed against link popularity and originating context. This transition is reducing the notion of *one link, one vote* in SEO, a trend proponents hope will help curb link spam as a whole.

Technical

When HTML (HyperTextMarkupLanguage) was designed, there was no explicit mechanism in the design to keep track of backlinks in software, as this carried additional logistical and network overhead.

Most Content management systems include features to track backlinks, provided the external site linking in sends notification to the target site. Most wiki systems include the capability of determining what pages link internally to any given page, but do not track external links to any given page.

Most commercial search engines provide a mechanism to determine the number of backlinks they have recorded to a particular web page. For example, Google can be searched using Google:link:http://www.wikipedia.org|link:wikipedia.org to find the number of pages on the Web pointing to http:// wikipedia.org/.Google only shows a small fraction of the number of links pointing to a site. It credits many more backlinks than it shows for each website.

Other mechanisms have been developed to track backlinks between disparate webpages controlled by organizations that aren't associated with each other. The most notable example of this is TrackBacks between blogs.

References

[1] Lennart Björneborn and Peter Ingwersen (2004). "Toward a Basic Framework for Webometrics" (http://www3.interscience.wiley.com/ cgi-bin/abstract/109594194/ABSTRACT). *Journal of the American Society for Information Science and Technology* **55** (14): 1216–1227. doi:10.1002/asi.20077. .

[2] Google's overview of PageRank (http://www.google.com/intl/en/technology/)

[3] "Does Backlink Quality Matter?" (http://www.adgooroo.com/backlink_quality.php). *Adgooroo*. 2010-04-21. . Retrieved 2010-04-21.

[[de:Rückverweis]

Keyword stuffing

Keyword stuffing is considered to be an unethical search engine optimization (SEO) technique. Keyword stuffing occurs when a web page is loaded with keywords in the meta tags or in content. The repetition of words in meta tags may explain why many search engines no longer use these tags.

Keyword stuffing had been used in the past to obtain maximum search engine ranking and visibility for particular phrases. This method is completely outdated and adds no value to rankings today. In particular, Google no longer gives good rankings to pages employing this technique.

Hiding text from the visitor is done in many different ways. Text colored to blend with the background, CSS "Z" positioning to place text "behind" an image — and therefore out of view of the visitor — and CSS absolute positioning to have the text positioned far from the page center are all common techniques. By 2005, many invisible text techniques were easily detected by major search engines.

"Noscript" tags are another way to place hidden content within a page. While they are a valid optimization method for displaying an alternative representation of scripted content, they may be abused, since search engines may index content that is invisible to most visitors.

Sometimes inserted text includes words that are frequently searched (such as "sex"), even if those terms bear little connection to the content of a page, in order to attract traffic to advert-driven pages.

In the past, keyword stuffing was considered to be either a white hat or a black hat tactic, depending on the context of the technique, and the opinion of the person judging it. While a great deal of keyword stuffing was employed to aid in spamdexing, which is of little benefit to the user, keyword stuffing in certain circumstances was not intended to skew results in a deceptive manner. Whether the term carries a pejorative or neutral connotation is dependent on whether the practice is used to pollute the results with pages of little relevance, or to direct traffic to a page of relevance that would have otherwise been de-emphasized due to the search engine's inability to interpret and understand related ideas. This is no longer the case. Search engines now employ themed, related keyword techniques to interpret the intent of the content on a page.

With relevance to keyword stuffing, it is quoted by the largest of search engines that they recommend Keyword Research [1] and use (with respect to the quality content you have to offer the web), to aid their visitors in the search of your valuable material. To prevent Keyword Stuffing you should wisely use keywords in respect with SEO, Search Engine Optimization. It could be best described as keywords should be reasonable and necessary, yet it is acceptable to assist with proper placement and your targeted effort to achieve search results. Placement of such words in the provided areas of HTML are perfectly allowed and reasonable. Google discusses keyword stuffing as Randomly Repeated Keywords [2].

In online journalism

Headlines in online news sites are increasingly packed with just the search-friendly keywords that identify the story. Puns and plays on words have gone by the wayside. Overusing this strategy is also called keyword stuffing. Old-school reporters and editors frown on the practice, but it is effective in optimizing news stories for search.[3]

References

[1] http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=66358

[2] http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=35291

[3] *On Language, The Web Is At War With Itself* (http://www.npr.org/templates/story/story.php?storyId=128356609), Linton Weeks, for National Public Radio, July 15, 2010.

External links

- Google Guidelines (http://www.google.com/support/webmasters/bin/answer.py?answer=35769)
- Yahoo! Guidelines (http://help.yahoo.com/l/us/yahoo/search/basics/basics-18.html)
- Live Search (MSN Search) Guidelines (http://help.live.com/Help.aspx?market=en-US&project=WL_Webmasters&querytype=topic&query=WL_WEBMASTERS_REF_GuidelinesforSuccessfulIndexing.htm)

Article spinning

Article spinning is a search engine optimization technique by which blog or website owners post a unique version of relevant content on their sites. It works by rewriting existing articles, or parts of articles, and replacing elements to provide a slightly different perspective on the topic. Many article marketers believe that article spinning helps avoid the feared penalties in the Search Engine Results pages (SERP) for using duplicate content. If the original articles are plagiarized from other websites or if the original article was used without the copyright owner's permission, such copyright infringements may result in the writer facing a legal challenge, while writers producing multiple versions of their own original writing need not worry about such things.

Website owners may pay writers to perform spinning manually, rewriting all or parts of articles. Writers also spin their own articles, manually or automatically, allowing them to sell the same articles with slight variations to a number of clients or to use the article for multiple purposes, for example as content and also for article marketing. There are a number of software applications which will automatically replace words or phrases in articles. Automatic rewriting can change the meaning of a sentence through the use of words with similar but subtly different meaning to the original. For example, the word "picture" could be replaced by the word "image" or "photo." Thousands of word-for-word combinations are stored in either a text file or database thesaurus to draw from. This ensures that a large percentage of words are different from the original article.

Article spinning requires "spintax." Spintax (or spin syntax) is the list of text, sentences, or synonyms that are embedded into an article. The spinning software then substitutes your synonym choices into the article to create new, unique variations of the base article.

Contrary to popular opinion, Google until 2010 did not penalize web sites that have duplicated content on them, but the advances in filtering techniques mean that duplicate content will rarely feature well in SERPS.[1] [2]. In 2010 and 2011, changes to Google's search algorithm targeting content farms aim to penalize sites containing significant duplicate content.[3]

The duplication of web content may also break copyright law in many countries. See, for example, the United States Online Copyright Infringement Liability Limitation Act (OCILLA) and the Digital Millennium Copyright Act.

References

[1] Lasnik, Adam (2006-12-18). "Deftly dealing with duplicate content" (http://googlewebmastercentral.blogspot.com/2006/12/ deftly-dealing-with-duplicate-content.html). *Google Webmaster Central Blog*. Google Inc.. . Retrieved 2007-09-18.

[2] "Webmaster Help Centre: Little or no original content" (http://www.google.com/support/webmasters/bin/answer.py?answer=66361). Google Inc.. . Retrieved 2007-09-18.

[3] http://googleblog.blogspot.com/2011/02/finding-more-high-quality-sites-in.html

Link farm

For the discussion about Wikipedia, see Wikipedia is not a link farm.

On the World Wide Web, a **link farm** is any group of web sites that all hyperlink to every other site in the group. Although some link farms can be created by hand, most are created through automated programs and services. A link farm is a form of spamming the index of a search engine (sometimes called spamdexing or spamexing). Other link exchange systems are designed to allow individual websites to selectively exchange links with other relevant websites and are not considered a form of spamdexing.

Search engines require ways to confirm page relevancy. A known method is to examine for one-way links coming directly from relevant websites. The process of building links should not be confused with being listed on link farms, as the latter requires reciprocal return links, which often renders the overall backlink advantage useless. This is due to oscillation, causing confusion over which is the vendor site and which is the promoting site.

History

Link farms were developed by search engine optimizers in 1999 to take advantage of the Inktomi search engine's dependence upon link popularity. Although link popularity is used by some search engines to help establish a ranking order for search results, the Inktomi engine at the time maintained two indexes. Search results were produced from the primary index which was limited to approximately 100 million listings. Pages with few inbound links fell out of the Inktomi index on a monthly basis.

Inktomi was targeted for manipulation through link farms because it was then used by several independent but popular search engines. Yahoo!, then the most popular search service, also used Inktomi results to supplement its directory search feature. The link farms helped stabilize listings primarily for online business Web sites that had few natural links from larger, more stable sites in the Inktomi index.

Link farm exchanges were at first handled on an informal basis, but several service companies were founded to provide automated registration, categorization, and link page updates to member Web sites.

When the Google search engine became popular, search engine optimizers learned that Google's ranking algorithm depended in part on a link weighting scheme called PageRank. Rather than simply count all inbound links equally, the PageRank algorithm determines that some links may be more valuable than others, and therefore assigns them more weight than others. Link farming was adapted to help increase the PageRank of member pages.

However, the link farms became susceptible to manipulation by unscrupulous webmasters who joined the services, received inbound linkage, and then found ways to hide their outbound links or to avoid posting any links on their sites at all. Link farm managers had to implement quality controls and monitor member compliance with their rules to ensure fairness.

Alternative link farm products emerged, particularly link-finding software that identified potential reciprocal link partners, sent them template-based emails offering to exchange links, and created directory-like link pages for Web sites, in the hope of building their link popularity and PageRank.

Search engines countered the link farm movement by identifying specific attributes associated with link farm pages and filtering those pages from indexing and search results. In some cases, entire domains were removed from the search engine indexes in order to prevent them from influencing search results.

External links

- Google Information for Webmasters[1]
- Yahoo!'s Search Content Quality Guidelines [2]
- MSN Guidelines for Successful Indexing [3]
- The Dirty Little Secrets of Search [4] at The New York Times

References

[1] http://www.google.com/support/webmasters/bin/answer.py?answer=35769&hl=en
[2] http://help.yahoo.com/l/us/yahoo/search/basics/basics-18.html
[3] http://help.live.com/help.aspx?mkt=en-us&project=wl_webmasters
[4] http://www.nytimes.com/2011/02/13/business/13search.html

Spamdexing

In computing, **spamdexing** (also known as **search spam**, **search engine spam**, **web spam** or **search engine poisoning**)[1] is the deliberate manipulation of search engine indexes. It involves a number of methods, such as repeating unrelated phrases, to manipulate the relevance or prominence of resources indexed in a manner inconsistent with the purpose of the indexing system.[2][3] Some consider it to be a part of search engine optimization, though there are many search engine optimization methods that improve the quality and appearance of the content of web sites and serve content useful to many users.[4] Search engines use a variety of algorithms to determine relevancy ranking. Some of these include determining whether the search term appears in the META keywords tag, others whether the search term appears in the body text or URL of a web page. Many search engines check for instances of spamdexing and will remove suspect pages from their indexes. Also, people working for a search-engine organization can quickly block the results-listing from entire websites that use spamdexing, perhaps alerted by user complaints of false matches. The rise of spamdexing in the mid-1990s made the leading search engines of the time less useful.

Common spamdexing techniques can be classified into two broad classes: *content spam*[4] (or *term spam*) and *link spam*.[3]

History

The earliest known reference[2] to the term *spamdexing* is by Eric Convey in his article "Porn sneaks way back on Web," The Boston Herald, May 22, 1996, where he said:

> The problem arises when site operators load their Web pages with hundreds of extraneous terms so search engines will list them among legitimate addresses.
>
> The process is called "spamdexing," a combination of spamming — the Internet term for sending users unsolicited information — and "indexing." [2]

Content spam

These techniques involve altering the logical view that a search engine has over the page's contents. They all aim at variants of the vector space model for information retrieval on text collections.

Keyword stuffing

Keyword stuffing involves the calculated placement of keywords within a page to raise the keyword count, variety, and density of the page. This is useful to make a page appear to be relevant for a web crawler in a way that makes it more likely to be found. Example: A promoter of a Ponzi scheme wants to attract web surfers to a site where he advertises his scam. He places hidden text appropriate for a fan page of a popular music group on his page, hoping that the page will be listed as a fan site and receive many visits from music lovers. Older versions of indexing programs simply counted how often a keyword appeared, and used that to determine relevance levels. Most modern search engines have the ability to analyze a page for keyword stuffing and determine whether the frequency is consistent with other sites created specifically to attract search engine traffic. Also, large webpages are truncated, so that massive dictionary lists cannot be indexed on a single webpage.

Hidden or invisible text

Unrelated hidden text is disguised by making it the same color as the background, using a tiny font size, or hiding it within HTML code such as "no frame" sections, alt attributes, zero-sized DIVs, and "no script" sections. People screening websites for a search-engine company might temporarily or permanently block an entire website for having invisible text on some of its pages. However, hidden text is not always spamdexing: it can also be used to enhance accessibility.

Meta-tag stuffing

This involves repeating keywords in the Meta tags, and using meta keywords that are unrelated to the site's content. This tactic has been ineffective since 2005.

Doorway pages

"Gateway" or doorway pages are low-quality web pages created with very little content but are instead stuffed with very similar keywords and phrases. They are designed to rank highly within the search results, but serve no purpose to visitors looking for information. A doorway page will generally have "click here to enter" on the page.

Scraper sites

Scraper sites are created using various programs designed to "scrape" search-engine results pages or other sources of content and create "content" for a website.[5] The specific presentation of content on these sites is unique, but is merely an amalgamation of content taken from other sources, often without permission. Such websites are generally full of advertising (such as pay-per-click ads[2]), or they redirect the user to other sites. It is even feasible for scraper sites to outrank original websites for their own information and organization names.

Article spinning

Article spinning involves rewriting existing articles, as opposed to merely scraping content from other sites, to avoid penalties imposed by search engines for duplicate content. This process is undertaken by hired writers or automated using a thesaurus database or a neural network.

Link spam

Link spam is defined as links between pages that are present for reasons other than merit.[6] Link spam takes advantage of link-based ranking algorithms, which gives websites higher rankings the more other highly ranked websites link to it. These techniques also aim at influencing other link-based ranking techniques such as the HITS algorithm.

Link-building software

A common form of link spam is the use of link-building software to automate the search engine optimization process.

Link farms

Link farms are tightly-knit communities of pages referencing each other, also known facetiously as *mutual admiration societies*[7].

Hidden links

Putting hyperlinks where visitors will not see them to increase link popularity. Highlighted link text can help rank a webpage higher for matching that phrase.

Sybil attack

A Sybil attack is the forging of multiple identities for malicious intent, named after the famous multiple personality disorder patient "Sybil" (Shirley Ardell Mason). A spammer may create multiple web sites at different domain names that all link to each other, such as fake blogs (known as spam blogs).

Spam blogs

Spam blogs are blogs created solely for commercial promotion and the passage of link authority to target sites. Often these "splogs" are designed in a misleading manner that will give the effect of a legitimate website but upon close inspection will often be written using spinning software or very poorly written and barely readable content. They are similar in nature to link farms.

Page hijacking

Page hijacking is achieved by creating a rogue copy of a popular website which shows contents similar to the original to a web crawler but redirects web surfers to unrelated or malicious websites.

Buying expired domains

Some link spammers monitor DNS records for domains that will expire soon, then buy them when they expire and replace the pages with links to their pages. *See* Domaining. However Google resets the link data on expired domains. Some of these techniques may be applied for creating a Google bomb — that is, to cooperate with other users to boost the ranking of a particular page for a particular query.

Cookie stuffing

Cookie stuffing involves placing an affiliate tracking cookie on a website visitor's computer without their knowledge, which will then generate revenue for the person doing the cookie stuffing. This not only generates fraudulent affiliate sales, but also has the potential to overwrite other affiliates' cookies, essentially stealing their legitimately earned commissions.

Using world-writable pages

Web sites that can be edited by users can be used by spamdexers to insert links to spam sites if the appropriate anti-spam measures are not taken.

Automated spambots can rapidly make the user-editable portion of a site unusable. Programmers have developed a variety of automated spam prevention techniques to block or at least slow down spambots.

Spam in blogs

Spam in blogs is the placing or solicitation of links randomly on other sites, placing a desired keyword into the hyperlinked text of the inbound link. Guest books, forums, blogs, and any site that accepts visitors' comments are particular targets and are often victims of drive-by spamming where automated software creates nonsense posts with links that are usually irrelevant and unwanted.

Comment spam

Comment spam is a form of link spam that has arisen in web pages that allow dynamic user editing such as wikis, blogs, and guestbooks. It can be problematic because agents can be written that automatically randomly select a user edited web page, such as a Wikipedia article, and add spamming links.[8]

Wiki spam

Wiki spam is a form of link spam on wiki pages. The spammer uses the open editability of wiki systems to place links from the wiki site to the spam site. The subject of the spam site is often unrelated to the wiki page where the link is added. In early 2005, Wikipedia implemented a default "nofollow" value for the "rel" HTML attribute. Links with this attribute are ignored by Google's PageRank algorithm. Forum and Wiki admins can use these to discourage Wiki spam.

Referrer log spamming

Referrer spam takes place when a spam perpetrator or facilitator accesses a web page (the *referee*), by following a link from another web page (the *referrer*), so that the referee is given the address of the referrer by the person's Internet browser. Some websites have a referrer log which shows which pages link to that site. By having a robot randomly access many sites enough times, with a message or specific address given as the referrer, that message or Internet address then appears in the referrer log of those sites that have referrer logs. Since some Web search engines base the importance of sites on the number of different sites linking to them, referrer-log spam may increase the search engine rankings of the spammer's sites. Also, site administrators who notice the referrer log entries in their logs may follow the link back to the spammer's referrer page.

Other types of spamdexing

Mirror websites

A mirror site is the hosting of multiple websites with conceptually similar content but using different URLs. Some search engines give a higher rank to results where the keyword searched for appears in the URL.

URL redirection

URL redirection is the taking of the user to another page without his or her intervention, *e.g.*, using META refresh tags, Flash, JavaScript, Java or Server side redirects.

Cloaking

Cloaking refers to any of several means to serve a page to the search-engine spider that is different from that seen by human users. It can be an attempt to mislead search engines regarding the content on a particular web site. Cloaking, however, can also be used to ethically increase accessibility of a site to users with disabilities or provide human users with content that search engines aren't able to process or parse. It is also used to deliver content based on a user's location; Google itself uses IP delivery, a form of cloaking, to deliver results. Another form of cloaking is *code swapping*, *i.e.*, optimizing a page for top ranking and then swapping another page in its place once a top ranking is achieved.

References

[1] *SearchEngineLand*, Danny Sullivan's video explanation of Search Engine Spam, October 2008 (http://searchengineland.com/ what-is-search-engine-spam-the-video-edition-15202.php) . Retrieved 2008-11-13.

[2] "WordSpy - spamdexing" (definition), March 2003, webpage: WordSpy-spamdexing (http://www.wordspy.com/words/spamdexing.asp).

[3] Gyöngyi, Zoltán; Garcia-Molina, Hector (2005), "Web spam taxonomy" (http://airweb.cse.lehigh.edu/2005/gyongyi.pdf), *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), 2005 in The 14th International World Wide Web Conference (WWW 2005) May 10, (Tue)-14 (Sat), 2005, Nippon Convention Center (Makuhari Messe), Chiba, Japan.*, New York, NY: ACM Press, ISBN 1-59593-046-9,

[4] Ntoulas, Alexandros; Manasse, Mark; Najork, Marc; Fetterly, Dennis (2006), "Detecting Spam Web Pages through Content Analysis", *The 15th International World Wide Web Conference (WWW 2006) May 23–26, 2006, Edinburgh, Scotland.*, New York, NY: ACM Press, ISBN 1-59593-323-9

[5] "Scraper sites, spam and Google" (tactics/motives), Googlerankings.com diagnostics, 2007, webpage: GR-SS (http://diagnostics. googlerankings.com/scraper-sites.html).

[6] Davison, Brian (2000), "Recognizing Nepotistic Links on the Web" (http://www.cse.lehigh.edu/~brian/pubs/2000/aaaiws/aaai2000ws. pdf), *AAAI-2000 workshop on Artificial Intelligence for Web Search*, Boston: AAAI Press, pp. 23–28,

[7] Search Engines: Technology, Society, and Business - Marti Hearst, Aug 29, 2005 (http://www2.sims.berkeley.edu/courses/is141/f05/ lectures/se-course-intro.pdf)

[8] Mishne, Gilad; David Carmel and Ronny Lempel (2005). "Blocking Blog Spam with Language Model Disagreement" (http://airweb.cse. lehigh.edu/2005/mishne.pdf) (PDF). *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web.*. Retrieved 2007-10-24.

External links

To report spamdexed pages

- Found on Google search engine (http://www.google.com/contact/spamreport.html) results
- Found on Yahoo! search engine (http://help.yahoo.com/l/us/yahoo/search/spam_abuse.html) results

Search engine help pages for webmasters

- Google's Webmaster Guidelines page (http://www.google.com/support/webmasters/bin/answer. py?answer=35769)
- Yahoo!'s Search Engine Indexing page (http://help.yahoo.com/help/us/ysearch/indexing/index.html)

Other tools and information for webmasters

- AIRWeb series of workshops on Adversarial Information Retrieval on the Web (http://airweb.cse.lehigh.edu/)
- Popular Black Hat SEO Techniques (http://sitemonetized.com/Top-Black-Hat-SEO-Techniques/)
- CMMS (http://winmain.vn)

Index

Search engine indexing collects, parses, and stores data to facilitate fast and accurate information retrieval. Index design incorporates interdisciplinary concepts from linguistics, cognitive psychology, mathematics, informatics, physics, and computer science. An alternate name for the process in the context of search engines designed to find web pages on the Internet is *web indexing*.

Popular engines focus on the full-text indexing of online, natural language documents.[1] Media types such as video and audio[2] and graphics[3][4] are also searchable.

Meta search engines reuse the indices of other services and do not store a local index, whereas cache-based search engines permanently store the index along with the corpus. Unlike full-text indices, partial-text services restrict the depth indexed to reduce index size. Larger services typically perform indexing at a predetermined time interval due to the required time and processing costs, while agent-based search engines index in real time.

Indexing

The purpose of storing an index is to optimize speed and performance in finding relevant documents for a search query. Without an index, the search engine would scan every document in the corpus, which would require considerable time and computing power. For example, while an index of 10,000 documents can be queried within milliseconds, a sequential scan of every word in 10,000 large documents could take hours. The additional computer storage required to store the index, as well as the considerable increase in the time required for an update to take place, are traded off for the time saved during information retrieval.

Index design factors

Major factors in designing a search engine's architecture include: Merge factors

How data enters the index, or how words or subject features are added to the index during text corpus traversal, and whether multiple indexers can work asynchronously. The indexer must first check whether it is updating old content or adding new content. Traversal typically correlates to the data collection policy. Search engine index merging is similar in concept to the SQL Merge command and other merge algorithms.[5]

Storage techniques

How to store the index data, that is, whether information should be data compressed or filtered.

Index size

How much computer storage is required to support the index.

Lookup speed

How quickly a word can be found in the inverted index. The speed of finding an entry in a data structure, compared with how quickly it can be updated or removed, is a central focus of computer science.

Maintenance

How the index is maintained over time.[6] Fault tolerance

How important it is for the service to be reliable. Issues include dealing with index corruption, determining whether bad data can be treated in isolation, dealing with bad hardware, partitioning, and schemes such as hash-based or composite partitioning,[7] as well as replication.

Index data structures

Search engine architectures vary in the way indexing is performed and in methods of index storage to meet the various design factors. Types of indices include:

Suffix tree

Figuratively structured like a tree, supports linear time lookup. Built by storing the suffixes of words. The suffix tree is a type of trie. Tries support extendable hashing, which is important for search engine indexing.[8] Used for searching for patterns in DNA sequences and clustering. A major drawback is that storing a word in the tree may require space beyond that required to store the word itself.[9] An alternate representation is a suffix array, which is considered to require less virtual memory and supports data compression such as the BWT algorithm.

Inverted index

Stores a list of occurrences of each atomic search criterion,[10] typically in the form of a hash table or binary tree.[11][12]

Citation index

Stores citations or hyperlinks between documents to support citation analysis, a subject of Bibliometrics.

Ngram index

Stores sequences of length of data to support other types of retrieval or text mining.[13] Document-term matrix

Used in latent semantic analysis, stores the occurrences of words in documents in a two-dimensional sparse matrix.

Challenges in parallelism

A major challenge in the design of search engines is the management of serial computing processes. There are many opportunities for race conditions and coherent faults. For example, a new document is added to the corpus and the index must be updated, but the index simultaneously needs to continue responding to search queries. This is a collision between two competing tasks. Consider that authors are producers of information, and a web crawler is the consumer of this information, grabbing the text and storing it in a cache (or corpus). The forward index is the consumer of the information produced by the corpus, and the inverted index is the consumer of information produced by the forward index. This is commonly referred to as a **producer-consumer model**. The indexer is the producer of searchable information and users are the consumers that need to search. The challenge is magnified when working with distributed storage and distributed processing. In an effort to scale with larger amounts of indexed information, the search engine's architecture may involve distributed computing, where the search engine consists of several machines operating in unison. This increases the possibilities for incoherency and makes it more difficult to maintain a fully synchronized, distributed, parallel architecture.[14]

Inverted indices

Many search engines incorporate an inverted index when evaluating a search query to quickly locate documents containing the words in a query and then rank these documents by relevance. Because the inverted index stores a list of the documents containing each word, the search engine can use direct access to find the documents associated with each word in the query in order to retrieve the matching documents quickly. The following is a simplified illustration of an invertedindex:

Inverted Index

Word	Documents
the	Document1, Document3, Document4, Document5
cow	Document 2, Document 3, Document 4
says	Document 5
moo	Document 7

This index can only determine whether a word exists within a particular document, since it stores no information regarding the frequency and position of the word; it is therefore considered to be a boolean index. Such an index determines which documents match a query but does not rank matched documents. In some designs the index includes additional information such as the frequency of each word in each document or the positions of a word in each document.[15] Position information enables the search algorithm to identify word proximity to support searching for phrases; frequency can be used to help in ranking the relevance of documents to the query. Such topics are the central research focus of information retrieval.

The inverted index is a sparse matrix, since not all words are present in each document. To reduce computer storage memory requirements, it is stored differently from a two dimensional array. The index is similar to the term document matrices employed by latent semantic analysis. The inverted index can be considered a form of a hash table. In some cases the index is a form of a binary tree, which requires additional storage but may reduce the lookup time. In larger indices the architecture is typically a distributed hash table.[16]

Index merging

The inverted index is filled via a merge or rebuild. A rebuild is similar to a merge but first deletes the contents of the inverted index. The architecture may be designed to support incremental indexing,[17][18] where a merge identifies the document or documents to be added or updated and then parses each document into words. For technical accuracy, a merge conflates newly indexed documents, typically residing in virtual memory, with the index cache residing on one or more computer hard drives.

After parsing, the indexer adds the referenced document to the document list for the appropriate words. In a larger search engine, the process of finding each word in the inverted index (in order to report that it occurred within a document) may be too time consuming, and so this process is commonly split up into two parts, the development of a forward index and a process which sorts the contents of the forward index into the inverted index. The inverted index is so named because it is an inversion of the forward index.

The forward index

The forward index stores a list of words for each document. The following is a simplified form of the forward index:

Forward Index

Document	Words
Document 1	the,cow,says,moo
Document 2	the,cat,and,the,hat
Document 3	the,dish,ran,away,with,the,spoon

The rationale behind developing a forward index is that as documents are parsing, it is better to immediately store the words per document. The delineation enables Asynchronous system processing, which partially circumvents the inverted index update bottleneck.[19] The forward index is sorted to transform it to an inverted index. The forward index is essentially a list of pairs consisting of a document and a word, collated by the document. Converting the forward index to an inverted index is only a matter of sorting the pairs by the words. In this regard, the inverted index is a word-sorted forward index.

Compression

Generating or maintaining a large-scale search engine index represents a significant storage and processing challenge. Many search engines utilize a form of compression to reduce the size of the indices on disk.[20] Consider the following scenario for a full text, Internet search engine.

- An estimated 2,000,000,000 different web pages exist as of the year 2000[21]
- Suppose there are 250 words on each webpage (based on the assumption they are similar to the pages of a novel.[22]
- It takes 8 bits (or 1 byte) to store a single character. Some encodings use 2 bytes per character[23][24]
- The average number of characters in any given word on a page may be estimated at 5 (Wikipedia:Size comparisons)
- The average personal computer comes with 100 to 250 gigabytes of usable space[25]

Given this scenario, an uncompressed index (assuming a non-conflated, simple, index) for 2 billion web pages would need to store 500 billion word entries. At 1 byte per character, or 5 bytes per word, this would require 2500 gigabytes of storage space alone, more than the average free disk space of 25 personal computers. This space requirement may be even larger for a fault-tolerant distributed storage architecture. Depending on the compression technique chosen, the index can be reduced to a fraction of this size. The tradeoff is the time and processing power required to perform compression and decompression.

Notably, large scale search engine designs incorporate the cost of storage as well as the costs of electricity to power the storage. Thus compression is a measure of cost.

Document parsing

Document parsing breaks apart the components (words) of a document or other form of media for insertion into the forward and inverted indices. The words found are called *tokens*, and so, in the context of search engine indexing and natural language processing, parsing is more commonly referred to as tokenization. It is also sometimes called word boundary disambiguation, tagging, text segmentation, content analysis, text analysis, text mining, concordance generation, speech segmentation, lexing, or lexical analysis. The terms 'indexing', 'parsing', and 'tokenization' are used interchangeably in corporate slang.

Natural language processing, as of 2006, is the subject of continuous research and technological improvement. Tokenization presents many challenges in extracting the necessary information from documents for indexing to support quality searching. Tokenization for indexing involves multiple technologies, the implementation of which are commonly kept as corporate secrets.

Challenges in natural language processing

Word Boundary Ambiguity

> Native English speakers may at first consider tokenization to be a straightforward task, but this is not the case with designing a multilingual indexer. In digital form, the texts of other languages such as Chinese, Japanese or Arabic represent a greater challenge, as words are not clearly delineated by whitespace. The goal during tokenization is to identify words for which users will search. Language-specific logic is employed to properly identify the boundaries of words, which is often the rationale for designing a parser for each language supported (or for groups of languages with similar boundary markers and syntax).

Language Ambiguity

> To assist with properly ranking matching documents, many search engines collect additional information about each word, such as its language or lexical category (part of speech). These techniques are language-dependent, as the syntax varies among languages. Documents do not always clearly identify the language of the document or represent it accurately. In tokenizing the document, some search engines attempt to automatically identify the language of the document.

Diverse File Formats

> In order to correctly identify which bytes of a document represent characters, the file format must be correctly handled. Search engines which support multiple file formats must be able to correctly open and access the document and be able to tokenize the characters of the document.

Faulty Storage

> The quality of the natural language data may not always be perfect. An unspecified number of documents, particular on the Internet, do not closely obey proper file protocol. binary characters may be mistakenly encoded into various parts of a document. Without recognition of these characters and appropriate handling, the index quality or indexer performance could degrade.

Tokenization

Unlike literate humans, computers do not understand the structure of a natural language document and cannot automatically recognize words and sentences. To a computer, a document is only a sequence of bytes. Computers do not 'know' that a space character separates words in a document. Instead, humans must program the computer to identify what constitutes an individual or distinct word, referred to as a token. Such a program is commonly called a tokenizer or parser or lexer. Many search engines, as well as other natural language processing software, incorporate specialized programs for parsing, such as YACC or Lex.

During tokenization, the parser identifies sequences of characters which represent words and other elements, such as punctuation, which are represented by numeric codes, some of which are non-printing control characters. The parser can also identify entities such as email addresses, phone numbers, and URLs. When identifying each token, several characteristics may be stored, such as the token's case (upper, lower, mixed, proper), language or encoding, lexical category (part of speech, like 'noun' or 'verb'), position, sentence number, sentence position, length, and line number.

Language recognition

If the search engine supports multiple languages, a common initial step during tokenization is to identify each document's language; many of the subsequent steps are language dependent (such as stemming and part of speech tagging). Language recognition is the process by which a computer program attempts to automatically identify, or categorize, the language of a document. Other names for language recognition include language classification, language analysis, language identification, and language tagging. Automated language recognition is the subject of ongoing research in natural language processing. Finding which language the words belongs to may involve the use of a language recognition chart.

Format analysis

If the search engine supports multiple document formats, documents must be prepared for tokenization. The challenge is that many document formats contain formatting information in addition to textual content. For example, HTML documents contain HTML tags, which specify formatting information such as new line starts, **bold** emphasis, and font size or style. If the search engine were to ignore the difference between content and 'markup', extraneous information would be included in the index, leading to poor search results. Format analysis is the identification and handling of the formatting content embedded within documents which controls the way the document is rendered on a computer screen or interpreted by a software program. Format analysis is also referred to as structure analysis, format parsing, tag stripping, format stripping, text normalization, text cleaning, and text preparation. The challenge of format analysis is further complicated by the intricacies of various file formats. Certain file formats are proprietary with very little information disclosed, while others are well documented. Common, well-documented file formats that many search engines support include:

- HTML
- ASCII text files (a text document without specific computer readable formatting)
- Adobe's Portable Document Format (PDF)
- PostScript (PS)
- LaTeX
- UseNet netnews server formats
- XML and derivatives like RSS
- SGML
- Multimedia meta data formats like ID3
- Microsoft Word
- Microsoft Excel
- Microsoft Powerpoint

- IBM Lotus Notes

Options for dealing with various formats include using a publicly available commercial parsing tool that is offered by the organization which developed, maintains, or owns the format, and writing a custom parser.

Some search engines support inspection of files that are stored in a compressed or encrypted file format. When working with a compressed format, the indexer first decompresses the document; this step may result in one or more files, each of which must be indexed separately. Commonly supported compressed file formats include:

- ZIP - Zip archive file
- RAR - Roshal ARchive File
- CAB - Microsoft Windows Cabinet File
- Gzip - File compressed with gzip
- BZIP - File compressed using bzip2
- Tape ARchive (TAR), Unix archive file, not (itself) compressed
- TAR.Z, TAR.GZ or TAR.BZ2 - Unix archive files compressed with Compress, GZIP or BZIP2

Format analysis can involve quality improvement methods to avoid including 'bad information' in the index. Content can manipulate the formatting information to include additional content. Examples of abusing document formatting for spamdexing:

- Including hundreds or thousands of words in a section which is hidden from view on the computer screen, but visible to the indexer, by use of formatting (e.g. hidden "div" tag in HTML, which may incorporate the use of CSS or Javascript to do so).
- Setting the foreground font color of words to the same as the background color, making words hidden on the computer screen to a person viewing the document, but not hidden to the indexer.

Section recognition

Some search engines incorporate section recognition, the identification of major parts of a document, prior to tokenization. Not all the documents in a corpus read like a well-written book, divided into organized chapters and pages. Many documents on the web, such as newsletters and corporate reports, contain erroneous content and side-sections which do not contain primary material (that which the document is about). For example, this article displays a side menu with links to other web pages. Some file formats, like HTML or PDF, allow for content to be displayed in columns. Even though the content is displayed, or rendered, in different areas of the view, the raw markup content may store this information sequentially. Words that appear sequentially in the raw source content are indexed sequentially, even though these sentences and paragraphs are rendered in different parts of the computer screen. If search engines index this content as if it were normal content, the quality of the index and search quality may be degraded due to the mixed content and improper word proximity. Two primary problems are noted:

- Content in different sections is treated as related in the index, when in reality it is not
- Organizational 'side bar' content is included in the index, but the side bar content does not contribute to the meaning of the document, and the index is filled with a poor representation of its documents.

Section analysis may require the search engine to implement the rendering logic of each document, essentially an abstract representation of the actual document, and then index the representation instead. For example, some content on the Internet is rendered via Javascript. If the search engine does not render the page and evaluate the Javascript within the page, it would not 'see' this content in the same way and would index the document incorrectly. Given that some search engines do not bother with rendering issues, many web page designers avoid displaying content via Javascript or use the Noscript tag to ensure that the web page is indexed properly. At the same time, this fact can also be exploited to cause the search engine indexer to 'see' different content than the viewer.

Meta tag indexing

Specific documents often contain embedded meta information such as author, keywords, description, and language. For HTML pages, the meta tag contains keywords which are also included in the index. Earlier Internet search engine technology would only index the keywords in the meta tags for the forward index; the full document would not be parsed. At that time full-text indexing was not as well established, nor was the hardware able to support such technology. The design of the HTML markup language initially included support for meta tags for the very purpose of being properly and easily indexed, without requiring tokenization.[26]

As the Internet grew through the 1990s, many brick-and-mortar corporations went 'online' and established corporate websites. The keywords used to describe webpages (many of which were corporate-oriented webpages similar to product brochures) changed from descriptive to marketing-oriented keywords designed to drive sales by placing the webpage high in the search results for specific search queries. The fact that these keywords were subjectively specified was leading to spamdexing, which drove many search engines to adopt full-text indexing technologies in the 1990s. Search engine designers and companies could only place so many 'marketing keywords' into the content of a webpage before draining it of all interesting and useful information. Given that conflict of interest with the business goal of designing user-oriented websites which were 'sticky', the customer lifetime value equation was changed to incorporate more useful content into the website in hopes of retaining the visitor. In this sense, full-text indexing was more objective and increased the quality of search engine results, as it was one more step away from subjective control of search engine result placement, which in turn furthered research of full-text indexing technologies.

In Desktop search, many solutions incorporate meta tags to provide a way for authors to further customize how the search engine will index content from various files that is not evident from the file content. Desktop search is more under the control of the user, while Internet search engines must focus more on the full text index.

References

[1] Clarke, C., Cormack, G.: Dynamic Inverted Indexes for a Distributed Full-Text Retrieval System. Tech Rep MT-95-01, University of Waterloo, February 1995.

[2] Stephen V. Rice, Stephen M. Bailey. Searching for Sounds (http://www.comparisonics.com/SearchingForSounds.html). Comparisonics Corporation. May 2004. Verified Dec 2006

[3] Charles E. Jacobs, Adam Finkelstein, David H. Salesin. Fast Multiresolution Image Querying (http://grail.cs.washington.edu/projects/ query/mrquery.pdf). Department of Computer Science and Engineering, University of Washington. 1995. Verified Dec 2006

[4] Lee, James. Software Learns to Tag Photos (http://www.technologyreview.com/read_article.aspx?id=17772&ch=infotech). MIT Technology Review. November 09, 2006. Pg 1-2. Verified Dec 2006. Commercial external link

[5] Brown, E.W.: Execution Performance Issues in Full-Text Information Retrieval. Computer Science Department, University of Massachusetts at Amherst, Technical Report 95-81, October 1995.

[6] Cutting, D., Pedersen, J.: Optimizations for dynamic inverted index maintenance. Proceedings of SIGIR, 405-411, 1990.

[7] Linear Hash Partitioning (http://dev.mysql.com/doc/refman/5.1/en/partitioning-linear-hash.html). MySQL 5.1 Reference Manual. Verified Dec 2006

[8] trie (http://www.nist.gov/dads/HTML/trie.html), Dictionary of Algorithms and Data Structures (http://www.nist.gov/dads), U.S. National Institute of Standards and Technology (http://www.nist.gov).

[9] Gusfield, Dan (1999) [1997]. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. USA: Cambridge University Press. ISBN 0-521-58519-8..

[10] Black, Paul E., inverted index (http://www.nist.gov/dads/HTML/invertedIndex.html), Dictionary of Algorithms and Data Structures (http://www.nist.gov/dads), U.S. National Institute of Standards and Technology (http://www.nist.gov) Oct 2006. Verified Dec 2006.

[11] C. C. Foster, Information retrieval: information storage and retrieval using AVL trees, Proceedings of the 1965 20th national conference, p.192-205, August 24–26, 1965, Cleveland, Ohio, United States

[12] Landauer, W. I.: The balanced tree and its utilization in information retrieval. IEEE Trans. on Electronic Computers, Vol. EC-12, No. 6, December 1963.

[13] Google Ngram Datasets (http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13) for sale at LDC (http://www. ldc.upenn.edu/) Catalog

[14] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. Google, Inc. OSDI. 2004.

[15] Grossman, Frieder, Goharian. IR Basics of Inverted Index (http://www.cs.clemson.edu/~juan/CPSC862/Concept-50/ IR-Basics-of-Inverted-Index.pdf). 2002. Verified Aug 2011.

[16] Tang, Hunqiang. Dwarkadas, Sandhya. "Hybrid Global Local Indexing for Efficient Peer to Peer Information Retrieval". University of Rochester. Pg 1. http://www.cs.rochester.edu/u/sandhya/papers/nsdi04.ps

[17] Tomasic, A., et al.: Incremental Updates of Inverted Lists for Text Document Retrieval. Short Version of Stanford University Computer Science Technical Note STAN-CS-TN-93-1, December, 1993.

[18] Luk, R. W. P. and W. Lam (2007) Efficient in-memory extensible inverted file. Information Systems 32(5):733-754. doi:10.1016/j.is.2006.06.001

[19] Sergey Brin and Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine (http://infolab.stanford.edu/~backrub/ google.html). Stanford University. 1998. Verified Dec 2006.

[20] H.S. Heaps. Storage analysis of a compression coding for a document database. 1NFOR, I0(i):47-61, February 1972.

[21] Murray, Brian H. Sizing the Internet (http://www.cyveillance.com/web/downloads/Sizing_the_Internet.pdf). Cyveillance, Inc. Pg 2. July 2000. Verified Dec 2006.

[22] Blair Bancroft. Word Count:A Highly Personal-and Probably Controversial-Essay on Counting Words (http://www.blairbancroft.com/ word_count.htm). Personal Website. Verified Dec 2006.

[23] The Unicode Standard - Frequently Asked Questions (http://www.unicode.org/faq/basic_q.html#15). Verified Dec 2006.

[24] Storage estimates (http://www.uplink.freeuk.com/data.html). Verified Dec 2006.

[25] Average Total Hard Drive Size by Global Region (http://www.pcpitstop.com/research/regionaltech.asp), February 2008. Verified May 2008.

[26] Berners-Lee, T., "Hypertext Markup Language - 2.0", RFC 1866, Network Working Group, November 1995.

Further reading

- R. Bayer and E. McCreight. Organization and maintenance of large ordered indices. Acta Informatica, 173-189, 1972.
- Donald E. Knuth. The art of computer programming, volume 1 (3rd ed.): fundamental algorithms, Addison Wesley Longman Publishing Co. Redwood City, CA, 1997.
- Donald E. Knuth. The art of computer programming, volume 3: (2nd ed.) sorting and searching, Addison Wesley Longman Publishing Co. Redwood City, CA, 1998.
- Gerald Salton. Automatic text processing, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 1988.
- Gerard Salton. Michael J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, Inc., New York, NY, 1986.
- Gerard Salton. Lesk, M.E.: Computer evaluation of indexing and text processing. Journal of the ACM. January 1968.
- Gerard Salton. The SMART Retrieval System - Experiments in Automatic Document Processing. Prentice Hall Inc., Englewood Cliffs, 1971.
- Gerard Salton. The Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley, Reading, Mass., 1989.
- Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Chapter 8. ACM Press 1999.
- G. K. Zipf. Human Behavior and the Principle of Least Effort. Addison-Wesley, 1949.
- Adelson-Velskii, G.M., Landis, E. M.: An information organization algorithm. DANSSSR, 146, 263-266 (1962).
- Edward H. Sussenguth, Jr., Use of tree structures for processing files, Communications of the ACM, v.6 n.5, p. 272-279, May 1963
- Harman, D.K., et al.: Inverted files. In Information Retrieval: Data Structures and Algorithms, Prentice-Hall, pp 28–43, 1992.
- Lim, L., et al.: Characterizing Web Document Change, LNCS 2118, 133–146, 2001.
- Lim, L., et al.: Dynamic Maintenance of Web Indexes Using Landmarks. Proc. of the 12th W3 Conference, 2003.
- Moffat, A., Zobel, J.: Self-Indexing Inverted Files for Fast Text Retrieval. ACM TIS, 349–379, October 1996, Volume 14, Number 4.
- Mehlhorn, K.: Data Structures and Efficient Algorithms, Springer Verlag, EATCS Monographs, 1984.
- Mehlhorn, K., Overmars, M.H.: Optimal Dynamization of Decomposable Searching Problems. IPL 12, 93–98, 1981.

- Mehlhorn, K.: Lower Bounds on the Efficiency of Transforming Static Data Structures into Dynamic Data Structures. Math. Systems Theory 15, 1–16, 1981.
- Koster, M.: ALIWEB: Archie-Like indexing in the Web. Computer Networks and ISDN Systems, Vol. 27, No. 2 (1994) 175-182 (also see Proc. First Int'l World Wide Web Conf., Elsevier Science, Amsterdam, 1994, pp. 175–182)
- Serge Abiteboul and Victor Vianu. Queries and Computation on the Web (http://dbpubs.stanford.edu:8090/pub/showDoc.Fulltext?lang=en&doc=1996-20&format=text&compression=&name=1996-20.text). Proceedings of the International Conference on Database Theory. Delphi, Greece 1997.
- Ian H Witten, Alistair Moffat, and Timothy C. Bell. Managing Gigabytes: Compressing and Indexing Documents and Images. New York: Van Nostrand Reinhold, 1994.
- A. Emtage and P. Deutsch, "Archie--An Electronic Directory Service for the Internet." Proc. Usenix Winter 1992 Tech. Conf., Usenix Assoc., Berkeley, Calif., 1992, pp. 93–110.
- M. Gray, World Wide Web Wanderer (http://www.mit.edu/people/mkgray/net/).
- D. Cutting and J. Pedersen. "Optimizations for Dynamic Inverted Index Maintenance." Proceedings of the 13th International Conference on Research and Development in Information Retrieval, pp. 405–411, September 1990.
- Stefan Büttcher, Charles L. A. Clarke, and Gordon V. Cormack. Information Retrieval: Implementing and Evaluating Search Engines (http://www.ir.uwaterloo.ca/book/). MIT Press, Cambridge, Mass., 2010.

Black hat

A **black hat** is the villain or *bad guy*, especially in a western movie in which such a character would stereotypically wear a black hat in contrast to the hero's white hat, especially in black and white movies.[1] In Jewish circles, it refers to the more orthodox Jews who wear large-brimmed black hats. The phrase is often used figuratively, especially in computing slang, where it refers to a computer security hacker who breaks into networks or computers, or creates computer viruses.[2]

Notable actors playing black hat villains in movies

- Jack Palance
- Lee Marvin
- Lee Van Cleef
- Leo Gordon
- Wallace Beery
- Henry Fonda
- Ben Kingsley

References

[1] George N. Fenin, William K. Everson (1973), *The Western, from silents to the seventies*

[2] *Oxford English Dictionary* (http://dictionary.oed.com). Oxford University Press. 2008. . "black hat n. colloq. (orig. U.S.) (a) a villain or criminal, esp. one in a film or other work of fiction; a 'bad guy'; (b) Computing slang a person who engages in illegal or malicious hacking, creates or distributes computer viruses, etc."

Danny Sullivan

Danny Sullivan is the editor-in-chief of *Search Engine Land*, a blog that covers news and information about search engines, and search marketing.

Search Engine Land is owned by Third Door Media, of which Danny Sullivan is partner and chief content officer. Third Door Media also owns and operates other search related companies, including Search Marketing Now, which provides webcasts and webinars, both live and on demand, about web marketing; and Search Marketing Expo, a search engine marketing conference.[1][2]


Danny Sullivan

Biography

Sullivan was born in 1965 and raised in Newport Beach, California. His name is Danny, not Daniel.[3] He graduated from the University of California, Irvine and spent a year in England working for the BBC.
He married his wife while there and moved back to California where he worked in the graphic design department and was a reporter for the *Los Angeles Times* and *The Orange County Register*.[4][5] He got his start in technology when he helped found Maximized Online with programmer Ken Spreitzer.

He then moved to Chitterne, a small village in England with his wife and two sons.[4] His family recently returned to Newport Beach.

Affiliated websites

Search Engine Watch

Sullivan started *Search Engine Watch* in June 1997 after he posted research about search engines, called *A Webmaster's Guide To Search Engines*, in April 1996. *Search Engine Watch* was a website with tips on how to get good search engine results. Shortly after beginning in November that year, he sold it for an undisclosed amount to MecklerMedia (now Jupitermedia). He stayed on to maintain the site, and be the editor-in-chief.[6] In 2006, it was sold to Incisive Media for $43 million. *Search Engine Watch* was considered by Matt Cutts of Google as "must reading", and Tim Mayer of Yahoo! as the "most authoritative source on search."[4]

He has also staged the Search Engine Strategies conference six times each year, attracting 1,500 to 6,000 attendees each time.[4] On August 29, 2006, Sullivan announced he would be leaving *Search Engine Watch* on November 30, 2006. He later came to an agreement with Jupitermedia to continue participating in SES through 2007.[7][8][9]

Search Engine Land

Search Engine Land is a news web site that covers search engine marketing and search engine optimization. It was founded in 2006 by Sullivan after he left *Search Engine Watch*. *Search Engine Land* stories have been cited numerous times by other media outlets.[10][11][12]

Search Engine Land reports that it attracted 95,000 unique users in February, 2007, representing 15% of the world market for search advertising. This user group spends approximately US$1.5 billion per annum.[13]

References

[1] SMX: The Search Marketing Expo Conference Series (http://searchmarketingexpo.com/), SMX

[2] Our Brands (http://thirddoormedia.com/brands.shtml), Third Door Media

[3] 25 Things I Hate About Google (http://blog.searchenginewatch.com/blog/060313-161500). blog.searchenginewatch.com. Retrieved July 3, 2007.

[4] Jefferson Graham: Got a search engine question? Ask Mr. Sullivan (http://www.usatoday.com/tech/news/2006-08-01-sullivan-search_x. htm), 8/1/2006, USA Today

[5] Elizabeth Stone: Helping Webmasters Land in Search Engines' Nets (http://www.nytimes.com/library/tech/00/03/circuits/articles/ 23sull.html), March 23, 2000, *The New York Times*

[6] History of Search Engine Watch (http://searchenginewatch.com/showPage.html?page=2155701), Search Engine Watch, November 20, 1997

[7] Stepping Down From Search Engine Watch (http://blog.searchenginewatch.com/blog/060829-063058), SEW Blog

[8] Leaving Search Engine Watch (http://daggle.com/060829-112950.html), Daggle

[9] Daggle (http://daggle.com/061023-150510.html), Oct 23, 2006, Danny Sullivan

[10] Google offers free voice-activated local search (http://news.com.com/2060-10800_3-0.html), CNET News.com, April 6, 2007

[11] After Long Delays, Yahoo Launches Panama Ranking System (http://www.marketingvox.com/archives/2007/02/06/ after-long-delays-yahoo-launches-panama-ranking-system/), MarketingVOX, Feb 6, 2007

[12] AFP and Google News settle lawsuit (http://today.reuters.co.uk/news/articlenews.aspx?type=internetNews& storyID=2007-04-07T201241Z_01_N07281154_RTRIDST_0_OUKIN-UK-GOOGLE-AFP.XML), Reuters UK, Apr 7, 2007

[13] Search Engine Land.com Readers Spend More than $1.5 Billion on Interactive Marketing for Their Companies and Clients (http://www. financevisor.com/market/news_detail.aspx?rid=55613), FinanceVisor, April 16, 2007

External links

- Search Engine Land(http://www.searchengineland.com)
- Daggle (http://www.daggle.com), Sullivan's personal blog

Meta element

Meta elements are the HTML or XHTML `<meta ... >` element used to provide structured metadata about a Web page. Multiple elements are often used on the same page: the element is the same, but its attributes are different. Meta elements can be used to specify page description, keywords and any other metadata not provided through the other `head` elements and attributes.

The meta element has two uses: either to emulate the use of the HTTP response header, or to embed additional metadata within the HTML document.

With HTML up to and including HTML 4.01 and XHTML, there were four valid attributes: `content`, `http-equiv`, `name` and `scheme`. Under HTML 5 there are now five valid attributes: `charset` having been added. `http-equiv` is used to emulate the HTTP header. `name` to embed metadata. The value of the statement, in either case, is contained in the `content` attribute, which is the only required attribute unless `charset` is given. `charset` is used to indicate the character set of the document, and is available in HTML5.

Such elements must be placed as tags in the `head` section of an HTML or XHTML document.

An example of the use of the `meta` element

In one form, `meta` elements can specify HTTP headers which should be sent before the actual content when the HTML page is served from Web server to client. For example:

```
<meta http-equiv="Content-Type" content="text/html" >
```

This specifies that the page should be served with an HTTP header called 'Content-Type' that has a value 'text/html'. In the general form, a `meta` element specifies `name` and associated `content` attributes describing aspects of the HTML page. For example:

```
<meta name="keywords" content="wikipedia,encyclopedia" >
```

In this example, the `meta` element identifies itself as containing the 'keywords' relevant to the document, Wikipedia and encyclopedia.

Meta tags can be used to indicate the location a business serves:

```
<meta name="zipcode" content="45212,45208,45218" >
```

In this example, geographical information is given according to ZIP codes. Default charset

for plain text is simply set with meta:

```
<meta http-equiv="Content-Type" content="text/html; charset=UTF-8" >
```

Meta element used in search engine optimization

Meta elements provide information about a given Web page, most often to help search engines categorize them correctly. They are inserted into the HTML document, but are often not directly visible to a user visiting the site.

They have been the focus of a field of marketing research known as search engine optimization (SEO), where different methods are explored to provide a user's site with a higher ranking on search engines. In the mid to late 1990s, search engines were reliant on meta data to correctly classify a Web page and webmasters quickly learned the commercial significance of having the right meta element, as it frequently led to a high ranking in the search engines

— and thus, high traffic to the website.

As search engine traffic achieved greater significance in online marketing plans, consultants were brought in who were well versed in how search engines perceive a website. These consultants used a variety of techniques

(legitimate and otherwise) to improve ranking for their clients.

Meta elements have significantly less effect on search engine results pages today than they did in the 1990s and their utility has decreased dramatically as search engine robots have become more sophisticated. This is due in part to the nearly infinite re-occurrence (keyword stuffing) of meta elements and/or to attempts by unscrupulous website placement consultants to manipulate (spamdexing) or otherwise circumvent search engine ranking algorithms.

While search engine optimization can improve search engine ranking, consumers of such services should be careful to employ only reputable providers. Given the extraordinary competition and technological craftsmanship required for top search engine placement, the implication of the term "search engine optimization" has deteriorated over the last decade. Where it once implied bringing a website to the top of a search engine's results page, for some consumers it now implies a relationship with keyword spamming or optimizing a site's internal search engine for improved performance.

Major search engine robots are more likely to quantify such extant factors as the volume of incoming links from related websites, quantity and quality of content, technical precision of source code, spelling, functional v. broken hyperlinks, volume and consistency of searches and/or viewer traffic, time within website, page views, revisits, click-throughs, technical user-features, uniqueness, redundancy, relevance, advertising revenue yield, freshness, geography, language and other intrinsic characteristics.

The `keywords` attribute

The `keywords` attribute was popularized by search engines such as Infoseek and AltaVista in 1995, and its popularity quickly grew until it became one of the most commonly used `meta` elements.[1] By late 1997, however, search engine providers realized that information stored in `meta` elements, especially the `keywords` attribute, was often unreliable and misleading, and at worst, used to draw users into spam sites. (Unscrupulous webmasters could easily place false `keywords` into their `meta` elements in order to draw people to their site.)

Search engines began dropping support for metadata provided by the `meta` element in 1998, and by the early 2000s, most search engines had veered completely away from reliance on `meta` elements. In July 2002, AltaVista, one of the last major search engines to still offer support, finally stopped considering them.[2]

No consensus exists whether or not the `keywords` attribute has any effect on ranking at any of the major search engines today. It is speculated that it does, if the keywords used in the `meta` can also be found in the page copy itself. With respect to Google, thirty-seven leaders in search engine optimization concluded in April 2007 that the relevance of having your keywords in the `meta`-attribute `keywords` is little to none[3] and in September 2009 Matt Cutts of Google announced that they are no longer taking keywords into account whatsoever.[4] However, both these articles suggest that Yahoo! still makes use of the keywords meta tag in some of its rankings. Yahoo! itself claims support for the keywords meta tag in conjunction with other factors for improving search rankings.[5] In Oct 2009 Search Engine Round Table announced that "Yahoo Drops The Meta Keywords Tag Also"[6] but informed us that the announcement made by Yahoo!'s Senior Director of Search was incorrect.[7] In the corrected statement Yahoo! Senior Director of Search states that "...What changed with Yahoo's ranking algorithms is that while we still index the meta keyword tag, the ranking importance given to meta keyword tags receives the lowest ranking signal in our system.... it will actually have less effect than introducing those same words in the body of the document, or any other section."[7]

The `description` attribute

Unlike the `keywords` attribute, the `description` attribute is supported by most major search engines, like Yahoo! and Bing, while Google will fall back on this tag when information about the page itself is requested (e.g. using the `related:` query) keywords are very important in description to increase the ranking of site in search engine. The `description` attribute provides a concise explanation of a Web page's content. This allows the Web page authors to give a more meaningful description for listings than might be displayed if the search engine was

unable to automatically create its own description based on the page content. The description is often, but not always, displayed on search engine results pages, so it can affect click-through rates. Industry commentators have *suggested* that major search engines also consider keywords located in the `description` attribute when ranking pages.[8] W3C doesn't specify the size of this description meta tag, but almost all search engines recommend it to be shorter than 155 characters of plain text.

The `language` attribute

The `language` attribute tells search engines what natural language the website is written in (e.g. English, Spanish or French), as opposed to the coding language (e.g. HTML). It is normally an IETF language tag for the language name. It is of most use when a website is written in multiple languages and can be included on each page to tell search engines in which language a particular page is written.[9]

The `robots` attribute

The `robots` attribute, supported by several major search engines,[10] controls whether search engine spiders are allowed to index a page, or not, and whether they should follow links from a page, or not. The attribute can contain one or more comma-separate values. The `noindex` value prevents a page from being indexed, and `nofollow` prevents links from being crawled. Other values recognized by one or more search engines can influence how the engine indexes pages, and how those pages appear on the search results. These include `noarchive`, which instructs a search engine not to store an archived copy of the page, and `nosnippet`, which asks that the search engine not include a snippet from the page along with the page's listing in search results.[11]

Meta tags are not the best option to prevent search engines from indexing content of a website. A more reliable and efficient method is the use of the robots.txt file (robots exclusion standard).

Additional attributes for search engines

NOODP

The search engines Google, Yahoo! and MSN use in some cases the title and abstract of the Open Directory Project (ODP) listing of a website for the title and/or description (also called snippet or abstract) in the search engine results pages (SERP). To give webmasters the option to specify that the ODP content should not be used for listings of their website, Microsoft introduced in May 2006 the new "`NOODP`" value for the "`robots`" element of the meta tags.[12] Google followed in July 2006[13] and Yahoo! in October 2006.[14]

The syntax is the same for all search engines who support the tag.

```
<meta name="robots" content="noodp" >
```

Webmasters can decide if they want to disallow the use of their ODP listing on a per search engine basis Google:

```
<meta name="googlebot" content="noodp" >
```

Yahoo!

```
<meta name="Slurp" content="noodp" >
```

MSN and Live Search:

```
<meta name="msnbot" content="noodp" >
```

NOYDIR

Yahoo! puts content from their own Yahoo! directory next to the ODP listing. In 2007 they introduced a meta tag that lets web designers opt-out of this.[15]

If you add the `NOYDIR` tag to a page, Yahoo! won't display the Yahoo! Directory titles and abstracts.

```
<meta name="robots" content="noydir" >
<meta name="Slurp" content="noydir" >
```

Robots-NoContent

Yahoo! also introduced in May 2007 the attribute value: `class="robots-nocontent"`.[16] This is not a meta tag, but an attribute and value, which can be used throughout Web page tags where needed. Content of the page where this attribute is being used will be ignored by the Yahoo! crawler and not included in the search engine's index.

Examples for the use of the `robots-nocontent` tag:

```
<div class="robots-nocontent">excluded content</div>
<span class="robots-nocontent">excluded content</span>
<p class="robots-nocontent">excluded content</p>
```

Academic studies

Google does not use HTML keyword or meta tag elements for indexing. The Director of Research at Google, Monika Henzinger, was quoted (in 2002) as saying, "Currently we don't trust metadata because we are afraid of being manipulated." [17] Other search engines developed techniques to penalize Web sites considered to be "cheating the system". For example, a Web site repeating the same meta keyword several times may have its ranking *decreased* by a search engine trying to eliminate this practice, though that is unlikely. It is more likely that a search engine will ignore the meta keyword element completely, and most do regardless of how many words used in the element.

Google does, however, use meta tag elements for displaying site links. The title tags are used to create the link in search results:

The meta description often appears in Google search results to describe the link:

```
<meta name="description" content="A blurb to describe the content of the page appears here" >
```

Redirects

Meta refresh elements can be used to instruct a Web browser to automatically refresh a Web page after a given time interval. It is also possible to specify an alternative URL and use this technique in order to redirect the user to a different location. Auto refreshing via a META element has been deprecated for more than ten years,[18] and recognized as problematic before that.[19]

The W3C suggests that user agents should allow users to disable it, otherwise META refresh should not be used by web pages. For Internet Explorer's security settings, under the miscellaneous category, meta refresh can be turned off by the user, thereby disabling its redirect ability. In Mozilla Firefox it can be disabled in the configuration file under the key name "accessibility.blockautorefresh".[20]

Many web design tutorials also point out that client-side redirecting tends to interfere with the normal functioning of a Web browser's "back" button. After being redirected, clicking the back button will cause the user to go back to the redirect page, which redirects them again. Some modern browsers seem to overcome this problem however,

including Safari, Mozilla Firefox and Opera.

Auto-redirects via markup (versus server-side redirects) are not in compliance with the W3C's - Web Content Accessibility Guidelines (WCAG) 1.0 (guideline 7.5).[21]

HTTP message headers

Meta elements of the form `<meta http-equiv="foo" content="bar">` can be used as alternatives to http headers. For example, `<meta http-equiv="expires" content="Wed, 21 June 2006 14:25:27 GMT">` would tell the browser that the page "expires" on June 21, 2006 at 14:25:27 GMT and that it may safely cache the page until then.

Alternative to `meta` elements

An alternative to `meta` elements for enhanced subject access within a website is the use of a back-of-book-style index for the website. See the American Society of Indexers [22] website for an example.

In 1994, ALIWEB, also used an index file to provide the type of information commonly found in meta keywords attributes.

References

[1] Statistic (June 4, 1997), META attributes by count (http://vancouver-webpages.com/META/bycount.shtml), *Vancouver Webpages*, retrieved June 3, 2007

[2] Danny Sullivan (October 1, 2002), Death Of A Meta Tag (http://searchenginewatch.com/showPage.html?page=2165061), *SearchEngineWatch.com*, retrieved June 03, 2007

[3] [http://sangers.nu/blog/tech/20080909-the-meta-tag-attribute--keywords "In 2007, 37 leaders in search engine optimisation concluded that having keywords in the keywords attribute is little to none." Sanger.nu blog, September 9 2008, retrieved August 22 011

[4] "Google does not use the keywords meta tag in web ranking" (http://googlewebmastercentral.blogspot.com/2009/09/google-does-not-use-keywords-meta-tag.html) Google Webmaster Central Blog, September 21 2009, retrieved September 21 2009

[5] Yahoo! FAQs, How do I improve the ranking of my web site in the search results? (http://web.archive.org/web/20071015182848/http://help.yahoo.com/l/us/yahoo/search/ranking/ranking-02.html), *Yahoo.com*, retrieved November 12, 2008

[6] "Yahoo Drops The Meta Keywords Tag Also" (http://www.seroundtable.com/archives/020918.html) SEO Roundtable, October 8 2009, retrieved April 22 2011

[7] "Yahoo's Senior Director of Search Got It Wrong, Yahoo Uses Meta Keywords Still" (http://www.seroundtable.com/archives/020964. html) SEO Roundtable, October 16 2009, retrieved April 22 2011

[8] Danny Sullivan, How To Use HTML Meta Tags (http://searchenginewatch.com/showPage.html?page=2167931), Search Engine Watch, December 5, 2002

[9] 1 Website Designer Using language metatags in websites (http://www.1websitedesigner.com/language-metatags) February 19, 2008

[10] Vanessa Fox, Using the robots meta tag (http://googlewebmastercentral.blogspot.com/2007/03/using-robots-meta-tag.html), Official Google Webmaster Central Blog, 3/05/2007

[11] Danny Sullivan (March 5, 2007), Meta Robots Tag 101: Blocking Spiders, Cached Pages & More (http://searchengineland.com/ 070305-204850.php), *SearchEngineLand.com*, retrieved June 3, 2007

[12] Betsy Aoki (May 22, 2006), Opting Out of Open Directory Listings for Webmasters (http://blogs.msdn.com/livesearch/archive/2006/ 05/22/603917.aspx), *Live Search Blog*, retrieved June 3, 2007

[13] Vanessa Fox (July 13, 2006), More control over page snippets (http://sitemaps.blogspot.com/2006/07/more-control-over-page-snippets. html), *Inside Google Sitemaps*, retrieved June 3, 2007

[14] Yahoo! Search (October 24, 2006), Yahoo! Search Weather Update and Support for 'NOODP' (http://www.ysearchblog.com/archives/ 000368.html), *Yahoo! Search Blog*, retrieved June 3, 2007

[15] Yahoo! Search (February 28, 2007), Yahoo! Search Support for 'NOYDIR' Meta Tags and Weather Update (http://www.ysearchblog. com/archives/000418.html), *Yahoo! Search Blog*, retrieved June 3, 2007

[16] Yahoo! Search (May 02, 2007), Introducing Robots-Nocontent for Page Sections (http://www.ysearchblog.com/archives/000444.html), *Yahoo! Search Blog*, retrieved June 3, 2007

[17] Greta de Groat (2002). "Perspectives on the Web and Google: Monika Henzinger, Director of Research, Google", *Journal of Internet Cataloging*, Vol. 5(1), pp. 17-28, 2002.

[18] W3C Techniques for Web Content Accessibility Guidelines (http://www.w3.org/TR/1999/WD-WAI-PAGEAUTH-19990226/ wai-pageauth-tech) W3C Working Draft 26-Feb-1999

[19] Techniques for Web Content Accessibility Guidelines (http://www.w3.org/TR/1999/WD-WAI-PAGEAUTH-19990217/ wai-pageauth-tech) W3C Working Draft 17-Feb-1999
[20] Accessibility.blockautorefresh (http://web.archive.org/web/20090602225157/http://kb.mozillazine.org/Accessibility. blockautorefresh) mozillaZine, archived June 2 2009 from the original(http://kb.mozillazine.org/Accessibility.blockautorefresh)
[21] W3CRecommendation(May5,1999),WebContentAccessibilityGuidelines1.0-Guideline7(http://www.w3.org/TR/1999/ WAI-WEBCONTENT-19990505/#gl-movement). *W3.org*, retrieved September 28, 2007
[22] http://www.asindexing.org

External links

- W3C HTML 4.01 Specification: section 7.4.4, Meta data (http://www.w3.org/TR/html4/struct/global. html#h-7.4.4)

Meta tags

Meta elements are the HTML or XHTML `<meta ... >` element used to provide structured metadata about a Web page. Multiple elements are often used on the same page: the element is the same, but its attributes are different. Meta elements can be used to specify page description, keywords and any other metadata not provided through the other `head` elements and attributes.

The meta element has two uses: either to emulate the use of the HTTP response header, or to embed additional metadata within the HTML document.

With HTML up to and including HTML 4.01 and XHTML, there were four valid attributes: `content`, `http-equiv`, `name` and `scheme`. Under HTML 5 there are now five valid attributes: `charset` having been added. `http-equiv` is used to emulate the HTTP header. `name` to embed metadata. The value of the statement, in either case, is contained in the `content` attribute, which is the only required attribute unless `charset` is given. `charset` is used to indicate the character set of the document, and is available in HTML5.

Such elements must be placed as tags in the `head` section of an HTML or XHTML document.

An example of the use of the `meta` element

In one form, `meta` elements can specify HTTP headers which should be sent before the actual content when the HTML page is served from Web server to client. For example:

```
<meta http-equiv="Content-Type" content="text/html" >
```

This specifies that the page should be served with an HTTP header called 'Content-Type' that has a value 'text/html'. In the general form, a `meta` element specifies `name` and associated `content` attributes describing aspects of the HTML page. For example:

```
<meta name="keywords" content="wikipedia,encyclopedia" >
```

In this example, the `meta` element identifies itself as containing the 'keywords' relevant to the document, Wikipedia and encyclopedia.

Meta tags can be used to indicate the location a business serves:

```
<meta name="zipcode" content="45212,45208,45218" >
```

In this example, geographical information is given according to ZIP codes. Default charset

for plain text is simply set with meta:

```
<meta http-equiv="Content-Type" content="text/html; charset=UTF-8" >
```

Meta element used in search engine optimization

Meta elements provide information about a given Web page, most often to help search engines categorize them correctly. They are inserted into the HTML document, but are often not directly visible to a user visiting the site.

They have been the focus of a field of marketing research known as search engine optimization (SEO), where different methods are explored to provide a user's site with a higher ranking on search engines. In the mid to late 1990s, search engines were reliant on meta data to correctly classify a Web page and webmasters quickly learned the commercial significance of having the right meta element, as it frequently led to a high ranking in the search engines
— and thus, high traffic to the website.

As search engine traffic achieved greater significance in online marketing plans, consultants were brought in who were well versed in how search engines perceive a website. These consultants used a variety of techniques (legitimate and otherwise) to improve ranking for their clients.

Meta elements have significantly less effect on search engine results pages today than they did in the 1990s and their utility has decreased dramatically as search engine robots have become more sophisticated. This is due in part to the nearly infinite re-occurrence (keyword stuffing) of meta elements and/or to attempts by unscrupulous website placement consultants to manipulate (spamdexing) or otherwise circumvent search engine ranking algorithms.

While search engine optimization can improve search engine ranking, consumers of such services should be careful to employ only reputable providers. Given the extraordinary competition and technological craftsmanship required for top search engine placement, the implication of the term "search engine optimization" has deteriorated over the last decade. Where it once implied bringing a website to the top of a search engine's results page, for some consumers it now implies a relationship with keyword spamming or optimizing a site's internal search engine for improved performance.

Major search engine robots are more likely to quantify such extant factors as the volume of incoming links from related websites, quantity and quality of content, technical precision of source code, spelling, functional v. broken hyperlinks, volume and consistency of searches and/or viewer traffic, time within website, page views, revisits, click-throughs, technical user-features, uniqueness, redundancy, relevance, advertising revenue yield, freshness, geography, language and other intrinsic characteristics.

The `keywords` attribute

The `keywords` attribute was popularized by search engines such as Infoseek and AltaVista in 1995, and its popularity quickly grew until it became one of the most commonly used `meta` elements.[1] By late 1997, however, search engine providers realized that information stored in `meta` elements, especially the `keywords` attribute, was often unreliable and misleading, and at worst, used to draw users into spam sites. (Unscrupulous webmasters could easily place false `keywords` into their `meta` elements in order to draw people to their site.)

Search engines began dropping support for metadata provided by the `meta` element in 1998, and by the early 2000s, most search engines had veered completely away from reliance on `meta` elements. In July 2002, AltaVista, one of the last major search engines to still offer support, finally stopped considering them.[2]

No consensus exists whether or not the `keywords` attribute has any effect on ranking at any of the major search engines today. It is speculated that it does, if the keywords used in the `meta` can also be found in the page copy itself. With respect to Google, thirty-seven leaders in search engine optimization concluded in April 2007 that the relevance of having your keywords in the `meta`-attribute `keywords` is little to none[3] and in September 2009 Matt Cutts of Google announced that they are no longer taking keywords into account whatsoever.[4] However, both these articles suggest that Yahoo! still makes use of the keywords meta tag in some of its rankings. Yahoo! itself claims support for the keywords meta tag in conjunction with other factors for improving search rankings.[5] In Oct 2009 Search Engine Round Table announced that "Yahoo Drops The Meta Keywords Tag Also"[6] but informed us that the announcement made by Yahoo!'s Senior Director of Search was incorrect.[7] In the corrected statement

Yahoo! Senior Director of Search states that "...What changed with Yahoo's ranking algorithms is that while we still index the meta keyword tag, the ranking importance given to meta keyword tags receives the lowest ranking signal in our system.... it will actually have less effect than introducing those same words in the body of the document, or any other section."[7]

The `description` attribute

Unlike the `keywords` attribute, the `description` attribute is supported by most major search engines, like Yahoo! and Bing, while Google will fall back on this tag when information about the page itself is requested (e.g. using the `related:` query) keywords are very important in description to increase the ranking of site in search engine. The `description` attribute provides a concise explanation of a Web page's content. This allows the Web page authors to give a more meaningful description for listings than might be displayed if the search engine was unable to automatically create its own description based on the page content. The description is often, but not always, displayed on search engine results pages, so it can affect click-through rates. Industry commentators have *suggested* that major search engines also consider keywords located in the `description` attribute when ranking pages.[8] W3C doesn't specify the size of this description meta tag, but almost all search engines recommend it to be shorter than 155 characters of plain text.

The `language` attribute

The `language` attribute tells search engines what natural language the website is written in (e.g. English, Spanish or French), as opposed to the coding language (e.g. HTML). It is normally an IETF language tag for the language name. It is of most use when a website is written in multiple languages and can be included on each page to tell search engines in which language a particular page is written.[9]

The `robots` attribute

The `robots` attribute, supported by several major search engines,[10] controls whether search engine spiders are allowed to index a page, or not, and whether they should follow links from a page, or not. The attribute can contain one or more comma-separate values. The `noindex` value prevents a page from being indexed, and `nofollow` prevents links from being crawled. Other values recognized by one or more search engines can influence how the engine indexes pages, and how those pages appear on the search results. These include `noarchive`, which instructs a search engine not to store an archived copy of the page, and `nosnippet`, which asks that the search engine not include a snippet from the page along with the page's listing in search results.[11]

Meta tags are not the best option to prevent search engines from indexing content of a website. A more reliable and efficient method is the use of the robots.txt file (robots exclusion standard).

Additional attributes for search engines

NOODP

The search engines Google, Yahoo! and MSN use in some cases the title and abstract of the Open Directory Project (ODP) listing of a website for the title and/or description (also called snippet or abstract) in the search engine results pages (SERP). To give webmasters the option to specify that the ODP content should not be used for listings of their website, Microsoft introduced in May 2006 the new "`NOODP`" value for the "`robots`" element of the meta tags.[12] Google followed in July 2006[13] and Yahoo! in October 2006.[14]

The syntax is the same for all search engines who support the tag.

```
<meta name="robots" content="noodp" >
```

Webmasters can decide if they want to disallow the use of their ODP listing on a per search engine basis

Google:

```
<meta name="googlebot" content="noodp" >
```

Yahoo!

```
<meta name="Slurp" content="noodp" >
```

MSN and Live Search:

```
<meta name="msnbot" content="noodp" >
```

NOYDIR

Yahoo! puts content from their own Yahoo! directory next to the ODP listing. In 2007 they introduced a meta tag that lets web designers opt-out of this.[15]

If you add the `NOYDIR` tag to a page, Yahoo! won't display the Yahoo! Directory titles and abstracts.

```
<meta name="robots" content="noydir" >
<meta name="Slurp" content="noydir" >
```

`Robots-NoContent`

Yahoo! also introduced in May 2007 the attribute value: `class="robots-nocontent"`.[16] This is not a meta tag, but an attribute and value, which can be used throughout Web page tags where needed. Content of the page where this attribute is being used will be ignored by the Yahoo! crawler and not included in the search engine's index.

Examples for the use of the `robots-nocontent` tag:

```
<div class="robots-nocontent">excluded content</div>
<span class="robots-nocontent">excluded content</span>
<p class="robots-nocontent">excluded content</p>
```

Academic studies

Google does not use HTML keyword or meta tag elements for indexing. The Director of Research at Google, Monika Henzinger, was quoted (in 2002) as saying, "Currently we don't trust metadata because we are afraid of being manipulated." [17] Other search engines developed techniques to penalize Web sites considered to be "cheating the system". For example, a Web site repeating the same meta keyword several times may have its ranking *decreased* by a search engine trying to eliminate this practice, though that is unlikely. It is more likely that a search engine will ignore the meta keyword element completely, and most do regardless of how many words used in the element.

Google does, however, use meta tag elements for displaying site links. The title tags are used to create the link in search results:

The meta description often appears in Google search results to describe the link:

```
<meta name="description" content="A blurb to describe the content of the page appears here" >
```

Redirects

Meta refresh elements can be used to instruct a Web browser to automatically refresh a Web page after a given time interval. It is also possible to specify an alternative URL and use this technique in order to redirect the user to a different location. Auto refreshing via a META element has been deprecated for more than ten years,[18] and recognized as problematic before that.[19]

The W3C suggests that user agents should allow users to disable it, otherwise META refresh should not be used by web pages. For Internet Explorer's security settings, under the miscellaneous category, meta refresh can be turned off by the user, thereby disabling its redirect ability. In Mozilla Firefox it can be disabled in the configuration file under the key name "accessibility.blockautorefresh".[20]

Many web design tutorials also point out that client-side redirecting tends to interfere with the normal functioning of a Web browser's "back" button. After being redirected, clicking the back button will cause the user to go back to the redirect page, which redirects them again. Some modern browsers seem to overcome this problem however, including Safari, Mozilla Firefox and Opera.

Auto-redirects via markup (versus server-side redirects) are not in compliance with the W3C's - Web Content Accessibility Guidelines (WCAG) 1.0 (guideline 7.5).[21]

HTTP message headers

Meta elements of the form `<meta http-equiv="foo" content="bar">` can be used as alternatives to http headers. For example, `<meta http-equiv="expires" content="Wed, 21 June 2006 14:25:27 GMT">` would tell the browser that the page "expires" on June 21, 2006 at 14:25:27 GMT and that it may safely cache the page until then.

Alternative to `meta` elements

An alternative to `meta` elements for enhanced subject access within a website is the use of a back-of-book-style index for the website. See the American Society of Indexers [22] website for an example.

In 1994, ALIWEB, also used an index file to provide the type of information commonly found in meta keywords attributes.

References

[1] Statistic (June 4, 1997), META attributes by count (http://vancouver-webpages.com/META/bycount.shtml), *Vancouver Webpages*, retrieved June 3, 2007

[2] Danny Sullivan (October 1, 2002), Death Of A Meta Tag (http://searchenginewatch.com/showPage.html?page=2165061), *SearchEngineWatch.com*, retrieved June 03, 2007

[3] [http://sangers.nu/blog/tech/20080909-the-meta-tag-attribute--keywords "In 2007, 37 leaders in search engine optimisation concluded that having keywords in the keywords attribute is little to none." Sanger.nu blog, September 9 2008, retrieved August 22 2011

[4] "Google does not use the keywords meta tag in web ranking" (http://googlewebmastercentral.blogspot.com/2009/09/google-does-not-use-keywords-meta-tag.html) Google Webmaster Central Blog, September 21 2009, retrieved September 21 2009

[5] Yahoo! FAQs, How do I improve the ranking of my web site in the search results? (http://web.archive.org/web/20071015182848/http://help.yahoo.com/l/us/yahoo/search/ranking/ranking-02.html), *Yahoo.com*, retrieved November 12, 2008

[6] "Yahoo Drops The Meta Keywords Tag Also" (http://www.seroundtable.com/archives/020918.html) SEO Roundtable, October 8 2009, retrieved April 22 2011

[7] "Yahoo's Senior Director of Search Got It Wrong, Yahoo Uses Meta Keywords Still" (http://www.seroundtable.com/archives/020964. html) SEO Roundtable, October 16 2009, retrieved April 22 2011

[8] Danny Sullivan, How To Use HTML Meta Tags (http://searchenginewatch.com/showPage.html?page=2167931), Search Engine Watch, December 5, 2002

[9] 1 Website Designer Using language metatags in websites (http://www.1websitedesigner.com/language-metatags) February 19, 2008

[10] Vanessa Fox, Using the robots meta tag (http://googlewebmastercentral.blogspot.com/2007/03/using-robots-meta-tag.html), Official Google Webmaster Central Blog, 3/05/2007

[11] Danny Sullivan (March 5, 2007), Meta Robots Tag 101: Blocking Spiders, Cached Pages & More (http://searchengineland.com/ 070305-204850.php), *SearchEngineLand.com*, retrieved June 3, 2007
[12] Betsy Aoki (May 22, 2006), Opting Out of Open Directory Listings for Webmasters (http://blogs.msdn.com/livesearch/archive/2006/ 05/22/603917.aspx), *Live Search Blog*, retrieved June 3, 2007
[13] Vanessa Fox (July 13, 2006), More control over page snippets (http://sitemaps.blogspot.com/2006/07/more-control-over-page-snippets. html), *Inside Google Sitemaps*, retrieved June 3,2007
[14] Yahoo! Search (October 24, 2006), Yahoo! Search Weather Update and Support for 'NOODP' (http://www.ysearchblog.com/archives/ 000368.html), *Yahoo! Search Blog*, retrieved June 3,2007
[15] Yahoo! Search (February 28, 2007), Yahoo! Search Support for 'NOYDIR' Meta Tags and Weather Update (http://www.ysearchblog. com/archives/000418.html), *Yahoo! Search Blog*, retrieved June 3, 2007
[16] Yahoo! Search (May 02, 2007), Introducing Robots-Nocontent for Page Sections (http://www.ysearchblog.com/archives/000444.html), *Yahoo! Search Blog*, retrieved June 3, 2007
[17] Greta de Groat (2002). "Perspectives on the Web and Google: Monika Henzinger, Director of Research, Google", *Journal of Internet Cataloging*, Vol. 5(1), pp. 17-28, 2002.
[18] W3C Techniques for Web Content Accessibility Guidelines (http://www.w3.org/TR/1999/WD-WAI-PAGEAUTH-19990226/ wai-pageauth-tech) W3C Working Draft 26-Feb-1999
[19] Techniques for Web Content Accessibility Guidelines (http://www.w3.org/TR/1999/WD-WAI-PAGEAUTH-19990217/ wai-pageauth-tech) W3C Working Draft 17-Feb-1999
[20] Accessibility.blockautorefresh (http://web.archive.org/web/20090602225157/http://kb.mozillazine.org/Accessibility. blockautorefresh) mozillaZine, archived June 2 2009 from the original (http://kb.mozillazine.org/Accessibility.blockautorefresh)
[21] W3C Recommendation (May 5, 1999), Web Content Accessibility Guidelines 1.0 - Guideline 7 (http://www.w3.org/TR/1999/ WAI-WEBCONTENT-19990505/#gl-movement). *W3.org*, retrieved September 28, 2007

External links

- W3C HTML 4.01 Specification: section 7.4.4, Meta data (http://www.w3.org/TR/html4/struct/global. html#h-7.4.4)

Inktomi

Inktomi Corporation was a California company that provided software for Internet service providers. It was founded in 1996 by UC Berkeley professor Eric Brewer and graduate student Paul Gauthier. The company was initially founded based on the real-world success of the web search engine they developed at the university. After the bursting of the dot-com bubble, Inktomi was acquired by Yahoo!

History

Inktomi's software was incorporated in the widely-used HotBot search engine, which displaced AltaVista as the leading web-crawler-based search engine, itself to be displaced later by Google. In a talk given to a UC Berkeley seminar on Search Engines[1] in October 2005, Eric Brewer credited much of the AltaVista displacement to technical differences of scale.

The company went on to develop Traffic Server, a proxy cache for web traffic and on-demand streaming media. Traffic Server found a limited marketplace due to several factors, but was deployed by several large service providers including AOL. One of the things that Traffic Server did was to transcode images down to a smaller size for AOL dialup users, leading many websites to provide special noncacheable pages with the phrase, "AOL Users Click Here" to navigate to these pages.

In November 1999 Inktomi acquired Webspective; in August 2000 Inktomi acquired Ultraseek Server from Disney's Go.com; in September, 2000, Inktomi acquired FastForward Networks;[2] in December 2000, Inktomi acquired the Content Bridge Business Unit from Adero, a content delivery network, which had formed the Content Bridge Alliance with Inktomi, AOL and a number of other ISPs, hosting providers and IP transport providers; and in June 2001 Inktomi acquired eScene Networks. Webspective developed technology for synchronizing and managing

content across a host of distributed servers to be used in clustered or distributed load-balancing. Fast Forward developed software for the distribution of live streaming media over the Internet using "app-level" multicast technology. eScene Networks developed software that provided an integrated workflow for the management and publishing of video content (now owned by Media Publisher, Inc.). With this combination of technologies, Inktomi became an "arms merchant" to a growing number of content delivery network (CDN) service providers. Inktomi stock peaked with a split-adjusted price of $241 a share in March 2000.

In earlier acquisitions Inktomi acquired C2B and Impulse Buy Networks, both companies which had pioneered the comparison shopping space and that had pioneered the performance-based marketing market, with over 4 million products registered in the service in 1998, and serving millions of merchandise product offers daily across 20,000 websites including Yahoo!, MSN, and AOL shopping. Merchants paid a percentage of sales and or a cost per click for traffic sent to their websites—ultimately this model became known as pay per click and was perfected by Google and Overture Services, Inc.

With the financial collapse of the service provider industry and overall burst of the dot-com bubble, Inktomi lost most of its customer base. In 2002, Inktomi board brought in turnaround expert and long term media investor Keyur Patel and restructured the organization to focus back on search and divest from non core assets. This move turned out to be brilliant and led ultimately to be acquired by Yahoo! in 2002 for $1.63 a share (or $235 million). In a separate transaction, the Ultraseek Server product (renamed Inktomi Enterprise Search) was sold to competitor Verity, Inc. in late 2002.

In 2006, the technology behind the Inktomi Proxy Server was acquired by Websense, which has modified it and included it their Websense Security Gateway solution.

In 2009, Yahoo! asked to enter Traffic Server into incubation with the Apache Incubator, which was accepted in July. The original Inktomi Traffic Server source, with additional Yahoo! modifications, was donated to the open source community that same year. In April 2010, the Apache Traffic Server[3] top-level project was officially created, marking the official acceptance of the new project.

Acquisitions

In September 1998 Inktomi acquired C2B Technologies,[4] adding a shopping engine technology to its portfolio; In April 1999 Inktomi acquired Impulse Buy Network, adding 400 merchants to its shopping engine and performance based business shopping model. [5]; in November 1999 Inktomi acquired Webspective; in August 2000 Inktomi acquired Ultraseek Server from Disney's Go.com; in September, 2000, Inktomi acquired FastForward Networks;[2] in December 2000, Inktomi acquired the Content Bridge Business Unit from Adero, a content delivery network, which had formed the Content Bridge Alliance with Inktomi, AOL and a number of other ISPs, hosting providers and IP transport providers; and in June 2001 Inktomi acquired eScene Networks. Webspective developed technology for synchronizing and managing content across a host of distributed servers to be used in clustered or distributed load-balancing. Fast Forward developed software for the distribution of live streaming media over the Internet using "app-level" multicast technology. eScene Networks developed software that provided an integrated workflow for the management and publishing of video content (now owned by Qumu, Inc.). With this combination of technologies, Inktomi became an "arms merchant" to a growing number of Content Delivery Network (CDN) service providers.

Inktomi name and logo

According to the Inktomi website, "The company's name, pronounced 'INK-tuh-me', is derived from a Lakota Indian legend about a trickster spider character. Inktomi is known for his ability to defeat larger adversaries through wit and cunning."[6] The tri-color, nested cube logo was created by Tom Lamar in 1996.

Executives

Prior to the acquisition of Inktomi by Yahoo! in 2002:

Corporate officers

- David C. Peterschmidt – Chairman, President and Chief Executive Officer
- Dr. Eric A. Brewer – Chief Scientist
- Keyur A. Patel - Chief Strategy Officer (turnaround)
- Timothy J. Burch – Vice President of Human Resources
- Ted Hally – Senior Vice President and General Manager of Network Products
- Jerry M. Kennelly – Executive Vice President, Chief Financial Officer and Secretary
- Al Shipp – Senior Vice President of Worldwide Field Operations
- Timothy Stevens – Senior Vice President of Business Affairs, General Counsel and Assistant Secretary
- Steve Hill – Vice President of Europe

Board of directors

- David C. Peterschmidt – Chairman, President and Chief Executive Officer, Inktomi Corporation
- Dr. Eric A. Brewer – Chief Scientist, Inktomi Corporation
- Frank Gill – Executive Vice President, Intel Corporation
- Fredric W. Harman – General Partner, Oak Investment Partners
- Alan Shugart – Chief Executive Officer, Al Shugart International

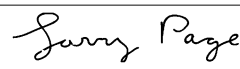
Mission statement

"The Inktomi mission is to build scalable software applications that are core to the Internet infrastructure."[6]

References

[1] SIMS 141: Search Engines: Technology, Society, and Business (http://www.sims.berkeley.edu/courses/is141/f05/). Course Syllabus, Fall 2005.
[2] Inktomi to buy FastForward Networks for $1.3 billion (http://www.news.com/2100-1023-245643.html). CNET news.com, September 13, 2000.
[3] "Apache Traffic Server website" (http://trafficserver.apache.org). 2010-04-22. . Retrieved 2010-06-18.
[4] Inktomi to buy C2B (http://news.cnet.com/2100-1001-215079.html). CNET"
[5] Inktomi buys Impulse Buy (http://news.cnet.com/Inktomi-buys-Impulse-Buy/2100-1017_3-224817.html) CNET.com
[6] Inktomi website, April 28,1999.

Larry Page

Larry Page	
	
Born	Lawrence Page March 26, 1973 Lansing, Michigan, U.S.
Residence	Palo Alto, California, U.S.
Nationality	American
Alma mater	University of Michigan (B.S.) Stanford University (M.S.)
Occupation	Computer scientist, internet entrepreneur
Known for	Co-founder and CEO of Google Inc.
Salary	1 [1]
Net worth	US$ 18.7 billion (2012)[1]
Title	CEO of Google
Spouse	Lucinda Southworth (m. 2007)
Children	1
Signature	
Website	
Google.com - Larry Page [2]	

Lawrence "Larry" Page[3] (born March 26, 1973) is an American computer scientist and internet entrepreneur who, with Sergey Brin, is best known as the co-founder of Google. On April 4, 2011, he took on the role of chief executive officer of Google, replacing Eric Schmidt.[4][5] As of 2012, his personal wealth is estimated to be $18.7 billion.[1] He is the inventor of PageRank, which became the foundation of Google's search ranking algorithm.

Early life and education

Larry Page was born in Lansing, Michigan.[6][7] His father, Carl Page, earned a Ph.D. in computer science in 1965 when the field was in its infancy, and is considered a "pioneer in computer science and artificial intelligence." Both he and Page's mother were computer science professors at Michigan State University.[8][9] Gloria Page, his mother, is Jewish but he was raised without religion.[10]

Page attended the Okemos Montessori School (now called Montessori Radmoor) in Okemos, Michigan from 1975 to 1979, and graduated from East Lansing High School in 1991.[11] He holds a Bachelor of Science in computer engineering from the University of Michigan with honors and a Master of Science in computer science from Stanford University. While at the University of Michigan, "Page created an inkjet printer made of Lego bricks" (actually a line plotter),[12] served as the president of the Eta Kappa Nu in Fall 1994,[13] and was a member of the 1993 "Maize & Blue" University of Michigan Solar Car team.

During an interview, Page recalled his childhood, noting that his house "was usually a mess, with computers and *Popular Science* magazines all over the place". His attraction to computers started when he was six years old when he got to "play with the stuff lying around". He became the "first kid in his elementary school to turn in an assignment from a word processor."[14] His older brother also taught him to take things apart, and before long he was taking "everything in his house apart to see how it worked". He said that "from a very early age, I also realized I wanted to invent things. So I became really interested in technology...and business . . . probably from when I was 12, I knew I was going to start a company eventually".[14]

After enrolling for a Ph.D. program in computer science at Stanford University, Larry Page was in search of a dissertation theme and considered exploring the mathematical properties of the World Wide Web, understanding its link structure as a huge graph.[15][16] His supervisor Terry Winograd encouraged him to pursue this idea, which Page later recalled as "the best advice I ever got".[17] Page then focused on the problem of finding out which web pages link to a given page, considering the number and nature of such backlinks to be valuable information about that page (with the role of citations in academic publishing in mind).[16] In his research project, nicknamed "BackRub", he was soon joined by Sergey Brin, a fellow Stanford Ph.D. student.[16]

John Battelle, co-founder of *Wired* magazine, wrote of Page that he had reasoned that the "entire Web was loosely based on the premise of citation – after all, what is a link but a citation? If he could devise a method to count and qualify each backlink on the Web, as Page puts it 'the Web would become a more valuable place'."[16] Battelle further described how Page and Brin began working together on the project:

> "At the time Page conceived of BackRub, the Web comprised an estimated 10 million documents, with an untold number of links between them. The computing resources required to crawl such a beast were well beyond the usual bounds of a student project. Unaware of exactly what he was getting into, Page began building out his crawler.
>
> "The idea's complexity and scale lured Brin to the job. A polymath who had jumped from project to project without settling on a thesis topic, he found the premise behind BackRub fascinating. "I talked to lots of research groups" around the school, Brin recalls, "and this was the most exciting project, both because it tackled the Web, which represents human knowledge, and because I liked Larry".[16]

Brin and Page originally met in March 1995, during a spring orientation of new computer Ph.D. candidates. Brin, who had already been in the program for two years, was assigned to show some students, including Page, around campus, and they later became good friends.[18]

To convert the backlink data gathered by BackRub's web crawler into a measure of importance for a given web page, Brin and Page developed the PageRank algorithm, and realized that it could be used to build a search engine far superior to existing ones.[16] It relied on a new kind of technology that analyzed the relevance of the back links that connected one Web page to another.[18] In August 1996, the initial version of Google was made available, still on the Stanford University Web site.[16]

Business

In 1998, Brin and Page founded Google, Inc.[19] Page ran Google as co-president along with Brin until 2001 when they hired Eric Schmidt as Chairman and CEO of Google. In January 2011 Google announced that Page would replace Schmidt as CEO in April the same year.[20] Both Page and Brin earn an annual compensation of one dollar. On April 4, 2011, Page officially became the chief executive of Google, while Schmidt stepped down to become executive chairman.

Page at MIT's Tesla Day in 2009

Personal life

Page married Lucinda Southworth at Richard Branson's Caribbean island, Necker Island in 2007.[21] Southworth is a research scientist and sister of actress and model Carrie Southworth.[22][23][24] They have one child.

Other interests

Page is an active investor in alternative energy companies, such as Tesla Motors, which developed the Tesla Roadster, a 244-mile (**unknown operator: u'strong'** km) range battery electric vehicle.[25] He continues to be committed to renewable energy technology, and with the help of Google.org, Google's philanthropic arm, promotes the adoption of plug-in hybrid electric cars and other alternative energy investments.[14]

Brin and Page are the executive producers of the 2007 film *Broken Arrows*.[26]

Awards and recognition

PC Magazine has praised Google as among the Top 100 Web Sites and Search Engines (1998) and awarded Google the Technical Excellence Award, for Innovation in Web Application Development in 1999. In 2000, Google earned a Webby Award, a People's Voice Award for technical achievement, and in 2001, was awarded Outstanding Search Service, Best Image Search Engine, Best Design, Most Webmaster Friendly Search Engine, and Best Search Feature at the Search Engine Watch Awards."[27]

In 2002, Page, along with Sergey Brin, was named to the MIT Technology Review TR100, as one of the top 100 innovators in the world under the age of 35.[28]

In 2003, both Brin and Page received an honorary MBA from IE Business School "for embodying the entrepreneurial spirit and lending momentum to the creation of new businesses...."[29] And in 2004, they received the Marconi Foundation Prize, the "Highest Award in Engineering," and were elected Fellows of the Marconi Foundation at Columbia University. "In announcing their selection, John Jay Iselin, the Foundation's president, congratulated the two men for their invention that has fundamentally changed the way information is retrieved today." They joined a "select cadre of 32 of the world's most influential communications technology pioneers...."[30] He was elected to the National Academy of Engineering in 2004. In 2005, Brin and Page were elected Fellows of the American Academy of Arts and Sciences.[31] In 2002 the World Economic Forum named Page a Global Leader for Tomorrow and in 2004 the X PRIZE chose Page as a trustee for their board.[12]

In 2004, Page and Brin were named "Persons of the Week" by *ABC World News Tonight*. Page received an honorary doctorate from the University of Michigan in 2009 during graduation commencement ceremonies.[32]

In 2011, he was ranked 24th on the Forbes list of billionaires and as the 11th richest person in the United States.[1]

References

[1] Forbes (2011). "Larry Page" (http://www.forbes.com/profile/larry-page/). Forbes. . Retrieved November 2011.

[2] https://www.google.com/about/corporate/company/execs.html#larry

[3] Larry Page (1999). "Lawrence or Larry Page's Page" (http://infolab.stanford.edu/~page/). Stanford Web Site. . Retrieved May 18, 2010.

[4] Google, Inc. (2011-01-20). "Google Announces Fourth Quarter and Fiscal Year 2010 Results and Management Changes" (http://investor.google.com/earnings/2010/Q4_google_earnings.html). Google Investor Relations. . Retrieved 2011-05-28.

[5] "An update from the Chairman" (http://googleblog.blogspot.com/2011/01/update-from-chairman.html). 2011-01-20. . Retrieved 2011-05-28.

[6] Stross, Randall. *Planet Google: One Company's Audacious Plan to Organize Everything We Know*, Simon & Schuster (2008) p. 75.

[7] Brandt, Richard L. *Inside Larry and Sergey's Brain*, Penguin (2009).

[8] Smale, Will. "Profile: The Google founders" (http://news.bbc.co.uk/2/hi/business/3666241.stm) *BBC*, April 30, 2004

[9] http://www.cse.msu.edu/endowment/carl_page.php

[10] Malseed, Mark (February 2007). The Story of Sergey Brin (http://www.momentmag.com/Exclusive/2007/2007-02/ 200702-BrinFeature.html) *Momentmag.Com* Retrieved (2011-06-06).

[11] "Google Chooses Michigan for Expansion," Office of the Governor, State of Michigan, July 11, 2006 (http://www.michigan.gov/gov/ 0,1607,7-168-23442_21974-147032--,00.html) [retrieved March 6, 2010]

[12] Google Corporate Information: Management: Larry Page (http://www.google.com/corporate/execs.html#larry)

[13] "HKN College Chapter Directory" (http://www.hkn.org/admin/chapters/beta_epsilon.html). Eta Kappa Nu. January 15, 2007...

[14] Scott, Virginia. *Google: Corporations That Changed the World*, Greenwood Publishing Group (2008)

[15] Brin, S.; Page, L. (1998). "The anatomy of a large-scale hypertextual Web search engine" (http://infolab.stanford.edu/pub/papers/ google.pdf). *Computer Networks and ISDN Systems* **30**: 107–117. doi:10.1016/S0169-7552(98)00110-X. ISSN 0169-7552. .

[16] Battelle, John. " The Birth of Google (http://www.wired.com/wired/archive/13.08/battelle.html?tw=wn_tophead_4)." *Wired Magazine.* August 2005.

[17] The best advice I ever got (http://money.cnn.com/galleries/2008/fortune/0804/gallery.bestadvice.fortune/2.html) (*Fortune*, April 2008)

[18] Moschovitis Group. *The Internet: A Historical Encyclopedia*, ABC-CLIO (2005)

[19] "Larry Page Profile" (http://www.google.com/corporate/execs.html#larry). Google. .

[20] Efrati, Amir (January 21, 2011). "Google's Page to Replace Schmidt as CEO" (http://online.wsj.com/article/SB10001424052748704881304576094340081291776.html?mod=googlenews_wsj). *The Wall Street Journal.* .

[21] Google founder Larry Page to marry (http://www.reuters.com/article/mediaNews/idUSN1360879220071114), Reuters.

[22] McCarthy, Megan. "President Bush, Clintons to meet at Googler wedding?" ValleyWag.com (http://valleywag.com/tech/larry-and-lucy/ president-bush-clintons-to-meet-at-googler-wedding-331398.php) December 7, 2007.

[23] Coleridge, Daniel R. "*Night Shift*'s Model MD." SOAPnet.com. (http://soapnet.go.com/soapnet/article/path-articleNum_8389/ category_shows) July 16, 2008. Retrieved September 10, 2008.

[24] Google Co-Founder Page to Wed (http://ap.google.com/article/ALeqM5hsrCJztj9j0xezqYCWBi8Dg2ifbQD8ST2SL82), The Associated Press.

[25] SiliconBeat: Tesla Motors New Electric Sports Car (http://www.siliconbeat.com/entries/2006/06/01/tesla_motors_new_electric_sportscar_company_raises_40m_from_google_guys_others.html)

[26] Google founders help fund classmate's independent film (http://www.usatoday.com/life/movies/news/ 2005-12-30-google-founders-film_x.htm)

[27] National Science Foundation (http://www.nsfgrfp.org/why_apply/fellow_profiles/sergey_brin), Fellow Profiles.

[28] "2002 Young Innovators Under 35: Larry Page, 29" (http://www.technologyreview.com/tr35/profile.aspx?trid=380). Technology Review. 2002. . Retrieved August 14, 2011.

[29] Brin and Page Awarded MBAs (http://www.ie.edu/IE/php/en/noticia.php?id=225), Press Release, Sept. 9, 2003

[30] Brin and Page Receive Marconi Foundation's Highest Honor (http://findarticles.com/p/articles/mi_m0EIN/is_2004_Sept_23/ ai_n6208748), Press Release, Sept. 23, 2004

[31] Academy Elects 225th Class of Fellows and Foreign Honorary Members (http://www.amacad.org/news\new2005.aspx)

[32] "Larry Page's University of Michigan 2009 Spring Commencement Address=2009-10-6" (http://www.google.com/intl/en/press/annc/ 20090502-page-commencement.html). .

External links

- Larry Page (https://plus.google.com/106189723444098348646/about) on Google+
- Google Corporate Information: Management (http://www.google.com/corporate/execs.html)
- America's Best Leaders (http://www.usnews.com/usnews/news/articles/051031/31google.htm)
- Profile (http://news.bbc.co.uk/2/hi/business/3666241.stm) at *BBC News*
- Larry Page's Alumni Profile Composite (http://www.engin.umich.edu/alumni/engineer/04SS/achievements/ advances.html)
- Channel 4 News: " Google's Vision (http://www.channel4.com/news/special-reports/ special-reports-storypage.jsp?id=2419)", May 23, 2006
- "The Searchmeisters" (http://bnaibrith.org/pubs/bnaibrith/spring2006bbm/searchmeisters.cfm) – profile on Page and Brin from the *B'nai B'rith Magazine* (Spring 2006)
- Video: Google Founders – Charlie Rose interview from 2001 (14min) (http://video.google.com/videoplay?docid=6958201596441974119#0h37m34s)
- On the Origins of Google (http://www.nsf.gov/discoveries/disc_summ.jsp?cntn_id=100660&org=NSF)
- Why you should ignore VCs (http://edcorner.stanford.edu/authorMaterialInfo.html?mid=1076&author=149), Larry Page speaks at Stanford
- Why you should ignore experience (http://marchingcubes.org/index.php/The_future_ain't_what_it_used_to_be), Experience is (sometimes) overrated.
- San Francisco Chronicle, Google Co-Founder's Search Ends (http://www.sfgate.com/cgi-bin/article.cgi?f=/c/a/2007/11/13/BUV6TB3HH.DTL)
- The Page Girl (http://valleywag.com/tech/larry-page/ny-post-outs-lucy-southworth-larry-pages-girlfriend-157256.php)

Sergey Brin

Sergey Brin	
	
Born	Sergey Mikhaylovich Brin Russian: Сергей Михайлович Брин August 21, 1973 Moscow, RSFSR, Soviet Union
Residence	Los Altos, California, U.S.
Nationality	American
Ethnicity	Jewish
Alma mater	UniversityofMaryland(B.S.1993) Stanford University (M.S. 1995)
Occupation	Computer scientist, Internet entrepreneur
Known for	Co-founder of Google
Salary	$1[1]
Net worth	US$ 18.7 billion (2012)[1]
Spouse	Anne Wojcicki[2]
Children	1
Signature	
Website	
Google.com - Sergey Brin [3]	

Sergey Mikhaylovich Brin (Russian: Сергей Михайлович Брин; born August 21, 1973) is a Russian-born American computer scientist and Internet entrepreneur who, with Larry Page, co-founded Google, one of the most profitable Internet companies.[4][5] As of 2012, his personal wealth is estimated to be $18.7 billion.[1]

Brin immigrated to the United States from the Soviet Union at the age of six. He earned his undergraduate degree at the University of Maryland, following in his father's and grandfather's footsteps by studying mathematics, as well as computer science. After graduation, he moved to Stanford University to acquire a Ph.D in computer science. There he met Larry Page, with whom he later became friends. They crammed their dormitory room with inexpensive computers and applied Brin's data mining system to build a superior search engine. The program became popular at Stanford and they suspended their PhD studies to start up Google in a rented garage.

The Economist magazine referred to Brin as an "Enlightenment Man", and someone who believes that "knowledge is always good, and certainly always better than ignorance", a philosophy that is summed up by Google's motto of

making all the world's information "universally accessible and useful"[6] and "Don't be evil".

Early life and education

Sergey Brin was born in Moscow to Russian Jewish parents, Michael Brin and Eugenia Brin, both graduates of Moscow State University.[7] His father is a mathematics professor at the University of Maryland, and his mother is a research scientist at NASA's Goddard Space Flight Center.[8][9]

Childhood in the Soviet Union

In 1979, when Brin was six, his family felt compelled to emigrate to the United States. In an interview with Mark Malseed, author of *The Google Story*,[10] Sergey's father explains how he was "forced to abandon his dream of becoming an astronomer even before he reached college". Although an official policy of anti-Semitism did not exist in the Soviet Union, Michael Brin claims Communist Party heads barred Jews from upper professional ranks by denying them entry to universities: "Jews were excluded from the physics departments, in particular..." Michael Brin therefore changed his major to mathematics where he received nearly straight A's. He said, "Nobody would even consider me for graduate school because I was Jewish."[11] At Moscow State University, Jews were required to take their entrance exams in different rooms from non-Jewish applicants, which were nicknamed "gas chambers", and they were marked on a harsher scale.[12]

The Brin family lived in a three-room apartment in central Moscow, which they also shared with Sergey's paternal grandmother.[11] Sergey told Malseed, "I've known for a long time that my father wasn't able to pursue the career he wanted", but Sergey only picked up the details years later after they had settled in the United States. He learned how, in 1977, after his father returned from a mathematics conference in Warsaw, Poland, he announced that it was time for the family to emigrate. "We cannot stay here any more", he told his wife and mother. At the conference, he was able to "mingle freely with colleagues from the United States, France, England and Germany, and discovered that his intellectual brethren in the West were 'not monsters.'" He added, "I was the only one in the family who decided it was really important to leave..."[11]

Sergey's mother was less willing to leave their home in Moscow, where they had spent their entire lives. Malseed writes, "For Genia, the decision ultimately came down to Sergey. While her husband admits he was thinking as much about his own future as his son's, for her, 'it was 80/20' about Sergey." They formally applied for their exit visa in September 1978, and as a result his father "was promptly fired". For related reasons, his mother also had to leave her job. For the next eight months, without any steady income, they were forced to take on temporary jobs as they waited, afraid their request would be denied as it was for many refuseniks. During this time his parents shared responsibility for looking after him and his father taught himself computer programming. In May 1979, they were granted their official exit visas and were allowed to leave the country.[11]

At an interview in October, 2000, Brin said, "I know the hard times that my parents went through there, and am very thankful that I was brought to the States."[13] A decade earlier, in the summer of 1990, a few weeks before his 17th birthday, his father led a group of gifted high school math students, including Sergey, on a two-week exchange program to the Soviet Union. "As Sergey recalls, the trip awakened his childhood fear of authority" and he remembers that his first "impulse on confronting Soviet oppression had been to throw pebbles at a police car." Malseed adds, "On the second day of the trip, while the group toured a sanitarium in the countryside near Moscow, Sergey took his father aside, looked him in the eye and said, 'Thank you for taking us all out of Russia.'"[11]

Education in the United States

Brin attended grade school at Paint Branch Montessori School in Adelphi, Maryland, but he received further education at home; his father, a professor in the department of mathematics at the University of Maryland, nurtured his interest in mathematics and his family helped him retain his Russian-language skills. In September 1990, after having attended Eleanor Roosevelt High School in Greenbelt, Maryland, Brin enrolled in the University of Maryland to study computer science and mathematics, where he received his Bachelor of Science in May 1993 with honors.[14]

Brin began his graduate study in computer science at Stanford University on a graduate fellowship from the National Science Foundation. In 1993, he interned at Wolfram Research, makers of Mathematica.[14] He is on leave from his Ph.D. studies at Stanford.[15]

Search engine development

During an orientation for new students at Stanford, he met Larry Page. In a recent interview for *The Economist*, Brin jokingly said: "We're both kind of obnoxious." They seemed to disagree on most subjects. But after spending time together, they "became intellectual soul-mates and close friends". Brin's focus was on developing data mining systems while Page's was in extending "the concept of inferring the importance of a research paper from its citations in other papers."[6] Together, the pair authored what is widely considered their seminal contribution, a paper entitled "The Anatomy of a Large-Scale Hypertextual Web Search Engine."[16]

Combining their ideas, they "crammed their dormitory room with cheap computers" and tested their new search engine designs on the web. Their project grew quickly enough "to cause problems for Stanford's computing infrastructure." But they realized they had succeeded in creating a superior engine for searching the web and suspended their PhD studies to work more on their system.[6]

As Mark Malseed wrote, "Soliciting funds from faculty members, family and friends, Sergey and Larry scraped together enough to buy some servers and rent that famous garage in Menlo Park. ... [soon after], Sun Microsystems co-founder Andy Bechtolsheim wrote a $100,000 check to "Google, Inc." The only problem was, "Google, Inc." did not yet exist—the company hadn't yet been incorporated. For two weeks, as they handled the paperwork, the young men had nowhere to deposit the money."[11]

The Economist magazine describes Brin's approach to life, like Page's, as based on a vision summed up by Google's motto, "of making all the world's information 'universally accessible and useful.'" Others have compared their vision to the impact of Johannes Gutenberg, the inventor of modern printing:

> "In 1440, Johannes Gutenberg introduced Europe to the mechanical printing press, printing Bibles for mass consumption. The technology allowed for books and manuscripts – originally replicated by hand – to be printed at a much faster rate, thus spreading knowledge and helping to usher in the European Renaissance. . . Google has done a similar job."[17]

The comparison was likewise noted by the authors of *The Google Story*: "Not since Gutenberg . . . has any new invention empowered individuals, and transformed access to information, as profoundly as Google."[10]:1

Not long after the two "cooked up their new engine for web searches, they began thinking about information that is today beyond the web", such as digitizing books, and expanding health information.[6]

Personal life

In May 2007, Brin married Anne Wojcicki in The Bahamas. Wojcicki is a biotech analyst and a 1996 graduate of Yale University with a B.S. in biology.[2][18] She has an active interest in health information, and together she and Brin are developing new ways to improve access to it. As part of their efforts, they have brainstormed with leading researchers about the human genome project. "Brin instinctively regards genetics as a database and computing problem. So does his wife, who co-founded the firm, 23andMe", which lets people analyze and compare their own

genetic makeup (consisting of 23 pairs of chromosomes).[6] In a recent announcement at Google's Zeitgeist conference, he said he hoped that some day everyone would learn their genetic code in order to help doctors, patients, and researchers analyze the data and try to repair bugs.[6]

Brin's mother, Eugenia, has been diagnosed with Parkinson's disease. In 2008, he decided to make a donation to the University of Maryland School of Medicine, where his mother is being treated.[19] Brin used the services of 23andMe and discovered that although Parkinson's is generally not hereditary, both he and his mother possess a mutation of the LRRK2 gene (G2019S) that puts the likelihood of his developing Parkinson's in later years between 20 and 80%.[6] When asked whether ignorance was not bliss in such matters, he stated that his knowledge means that he can now take measures to ward off the disease. An editorial in *The Economist* magazine states that "Mr Brin regards his mutation of LRRK2 as a bug in his personal code, and thus as no different from the bugs in computer code that Google's engineers fix every day. By helping himself, he can therefore help others as well. He considers himself lucky.... But Mr. Brin was making a much bigger point. Isn't knowledge always good, and certainty always better than ignorance?"[6]

Brin and his wife run The Brin Wojcicki Foundation.[20]

In November, 2011 Brin and his wife's foundation, The Brin Wojcicki Foundation, awarded 500,000 dollars to the Wikimedia Foundation as it started its eighth annual fundraising campaign.[21]

Censorship of Google in China

Remembering his youth and his family's reasons for leaving the Soviet Union, he "agonized over Google's decision to appease the communist government of China by allowing it to censor search engine results", but decided that the Chinese would still be better off than without having Google available.[6] He explained his reasoning to *Fortune* magazine:

> "We felt that by participating there, and making our services more available, even if not to the 100 percent that we ideally would like, that it will be better for Chinese web users, because ultimately they would get more information, though not quite all of it."[22]

On January 12, 2010, Google reported a large cyber attack on its computers and corporate infrastructure that began a month earlier, which included accessing numerous Gmail accounts and the theft of Google's intellectual property. After the attack was determined to have originated in China, the company stated that it would no longer agree to censor its search engine in China and may exit the country altogether. The *New York Times* reported that "a primary goal of the attackers was accessing the Gmail accounts of Chinese human rights activists, but that the attack also targeted 20 other large companies in the finance, technology, media and chemical sectors."[23][24] It was later reported that the attack included "one of Google's crown jewels, a password system that controls access by millions of users worldwide."[25]

In late March, 2010, it officially discontinued its China-based search engine while keeping its uncensored Hong Kong site in operation. Speaking for Google, Brin stated during an interview, "One of the reasons I am glad we are making this move in China is that the China situation was really emboldening other countries to try and implement their own firewalls."[26] During another interview with *Spiegel*, he added, "For us it has always been a discussion about how we can best fight for openness on the Internet. We believe that this is the best thing that we can do for preserving the principles of the openness and freedom of information on the Internet."[27]

While only a few large companies so far pledged their support for the move, many Internet "freedom proponents are cheering the move," and it is "winning it praise in the U.S." from lawmakers.[26][28] Senator Byron Dorgan stated that "Google's decision is a strong step in favor of freedom of expression and information."[29] And Congressman Bob Goodlatte said, "I applaud Google for its courageous step to stop censoring search results on Google.com. Google has drawn a line in the sand and is shining a light on the very dark area of individual liberty restrictions in China."[30] From the business perspective, many recognize that the move is likely to affect Google's profits: "Google

is going to pay a heavy price for its move, which is why it deserves praise for refusing to censor its service in China."[31] *The New Republic* adds that "Google seems to have arrived at the same link that was obvious to Andrei Sakharov: the one between science and freedom," referring to the move as "heroism."[32]

Awards and recognition

In 2002, Brin, along with Larry Page, was named to the MIT Technology Review TR100, as one of the top 100 innovators in the world under the age of 35.[33]

In 2003, both Brin and Page received an honorary MBA from IE Business School "for embodying the entrepreneurial spirit and lending momentum to the creation of new businesses...".[34] And in 2004, they received the Marconi Foundation Prize, the "Highest Award in Engineering", and were elected Fellows of the Marconi Foundation at Columbia University. "In announcing their selection, John Jay Iselin, the Foundation's president, congratulated the two men for their invention that has fundamentally changed the way information is retrieved today."They joined a "select cadre of 32 of the world's most influential communications technology pioneers..."[35]

In 2004, Brin received the Academy of Achievement's Golden Plate Award with Larry Page at a ceremony in Chicago, Illinois.

In November 2009, *Forbes magazine* decided Brin and Larry Page were the fifth most powerful people in the world.[36] Earlier that same year, in February, Brin was inducted into the National Academy of Engineering, which is "among the highest professional distinctions accorded to an engineer ... [and] honors those who have made outstanding contributions to engineering research, practice...". He was selected specifically, "for leadership in development of rapid indexing and retrieval of relevant information from the World Wide Web."[37]

In their "Profiles" of Fellows, the National Science Foundation included a number of earlier awards:

> "he has been a featured speaker at the World Economic Forum and the Technology, Entertainment and Design Conference. ... *PC Magazine* has praised Google [of] the Top 100 Web Sites and Search Engines (1998) and awarded Google the Technical Excellence Award, for Innovation in Web Application Development in 1999. In 2000, Google earned a Webby Award, a People's Voice Award for technical achievement, and in 2001, was awarded Outstanding Search Service, Best Image Search Engine, Best Design, Most Webmaster Friendly Search Engine, and Best Search Feature at the Search Engine Watch Awards."[38]

According to *Forbes* he is the 15th richest person in the world with a personal wealth of US$16.7 billion in 2012.[39]

Other interests

Brin is working on other, more personal projects that reach beyond Google. For example, he and Page are trying to help solve the world's energy and climate problems at Google's philanthropic arm Google.org, which invests in the alternative energy industry to find wider sources of renewable energy. The company acknowledges that its founders want "to solve really big problems using technology."[40]

In October 2010, for example, they invested in a major offshore wind power development to assist the East coast power grid,[41] which may eventually become the first "offshore wind farm" in the United States.[42] A week earlier they introduced a car that, with "artificial intelligence," can drive itself using video cameras and radar sensors.[40] In the future, drivers of cars with similar sensors would have fewer accidents. These safer vehicles could therefore be built lighter and require less fuel consumption.[43]

They are trying to get companies to create innovative solutions to increasing the world's energy supply.[44] He is an investor in Tesla Motors, which has developed the Tesla Roadster, a 244-mile (**unknown operator: u'strong'** km) range battery electric vehicle.

Brin has appeared on television shows and many documentaries, including *Charlie Rose*, CNBC, and CNN. In 2004, he and Larry Page were named "Persons of the Week" by *ABC World News Tonight*. In January 2005 he was

nominated to be one of the World Economic Forum's "Young Global Leaders". He and Page are also the executive producers of the 2007 film *Broken Arrows*.

In June 2008, Brin invested $4.5 million in Space Adventures, the Virginia-based space tourism company. His investment will serve as a deposit for a reservation on one of Space Adventures' proposed flights in 2011. So far, Space Adventures has sent seven tourists into space.[45]

He and Page co-own a customized Boeing 767–200 and a Dornier Alpha Jet, and pay $1.4 million a year to house them and two Gulfstream V jets owned by Google executives at Moffett Federal Airfield. The aircraft have had scientific equipment installed by NASA to allow experimental data to be collected in flight.[46][47]

Brin is a member of AmBAR, a networking organization for Russian-speaking business professionals (both expatriates and immigrants) in the United States. He has made many speaking appearances.[48]

References

[1] "Sergey Brin" (http://www.forbes.com/profile/sergey-brin/). Forbes. . Retrieved February, 2011.

[2] Argetsinger, Amy; Roberts, Roxanne (May 13, 2007). "Amy Argetsinger and Roxanne Roberts - Oprah Winfrey's Degrees of Communication at Howard" (http://www.washingtonpost.com/wp-dyn/content/article/2007/05/12/AR2007051201168.html). *The Washington Post*. . Retrieved October 20, 2007.

[3] https://www.google.com/about/corporate/company/execs.html#sergey

[4] "Play it loud: Google employee 59 on the Bob Dylan attitude of Google's early days" (http://www.silicon.com/management/ceo-essentials/2011/08/05/play-it-loud-google-employee-59-on-the-bob-dylan-attitude-of-googles-early-days-39747760/). . Retrieved August 18, 2011.

[5] "Daily Market Beat - Trinity Investment Research" (http://www.trinityinvestmentresearch.com/daily-market-beat/1106). . Retrieved August 18, 2011.

[6] "Enlightenment Man" (http://www.economist.com/science/tq/displaystory.cfm?story_id=12673407). *The Economist*. Dec. 6, 2008. .

[7] "Dominic Lawson: More migrants please, especially the clever ones" (http://www.independent.co.uk/opinion/commentators/dominic-lawson/dominic-lawson-more-migrants-please-especially-the-clever-ones-2368622.html), *The Independent*, U.K., October 11, 2011

[8] Smale, Will (April 30, 2004). " Profile: The Google founders (http://news.bbc.co.uk/2/hi/business/3666241.stm)". BBC News. Retrieved 2010-01-07.

[9] " Sergey Brin (http://www.nndb.com/people/826/000044694/)". NNDB. Retrieved 2010-01-07.

[10] Vise, David, and Malseed, Mark. *The Google Story*, Delta Publ. (2006)

[11] Malseed, Mark (February 2007). " The Story of Sergey Brin (http://www.momentmag.com/Exclusive/2007/2007-02/ 200702-BrinFeature.html)". *Moment Magazine*. Retrieved2010-01-07.

[12] *The Independent* (London). October 11, 2011. http://www.independent.co.uk/opinion/commentators/dominic-lawson/ dominic-lawson-more-migrants-please-especially-the-clever-ones-2368622.html.

[13] Scott, Virginia. *Google: Corporations That Changed the World*, Greenwood Publishing Group (2008)

[14] Brin, Sergey (January 7, 1997). "Resume" (http://infolab.stanford.edu/~sergey/resume.html). . Retrieved March 9, 2008.

[15] "Sergey Brin: Executive Profile & Biography – BusinessWeek" (http://investing.businessweek.com/businessweek/research/stocks/people/person.asp?personId=534604&symbol=GOOG). *Business Week*. . Retrieved March 9, 2008. "He is currently on leave from the PhD program in computer science at Stanford university..."

[16] Brin, S.; Page, L. (1998). "The anatomy of a large-scale hypertextual Web search engine" (http://infolab.stanford.edu/pub/papers/ google.pdf). *Computer Networks and ISDN Systems* **30**: 107–117. doi:10.1016/S0169-7552(98)00110-X. ISSN 0169-7552. .

[17] *Information Technology* (http://www.librarystuff.net/2009/10/01/google-the-gutenberg/), Oct. 1, 2009

[18] "Anne Wojcicki Marries the Richest Bachelor" (http://cosmetic-makeovers.com/2007/05/18/ anne-wojcicki-marries-the-richest-bachelor). Cosmetic Makovers. . Retrieved October 20, 2007.

[19] Helft, Miguel (September 19, 2008). "Google Co-Founder Has Genetic Code Linked to Parkinson's" (http://www.nytimes.com/2008/09/19/technology/19google.html?_r=1&partner=rssnyt&emc=rss&oref=slogin). *The New York Times*. . Retrieved September 18, 2008.

[20] http://dynamodata.fdncenter.org/990s/990search/ffindershow.cgi?id=RINF001

[21] http://www.mercurynews.com/business-headlines/ci_19369678

[22] Martin, Dick. *Rebuilding Brand America: hat We Must Do to Restore Our Reputation and Safeguard the Future of American Business Abroad*, AMACOM Div. American Mgmt. Assn. (2007)

[23] "Google, Citing Cyber Attack, Threatens to Exit China" (http://www.nytimes.com/2010/01/13/world/asia/13beijing.html), *New York Times*, January 12, 2010

[24] A new approach to China (http://googleblog.blogspot.com/2010/01/new-approach-to-china.html)

[25] "Cyberattack on Google Said to Hit Password System" (http://www.nytimes.com/2010/04/20/technology/20google.html?nl=technology&emc=techupdateema1) *New York Times*, April 19, 2010

[26] "Brin Drove Google to Pull Back in China" (http://online.wsj.com/article/SB10001424052748704266504575141064259998090.html) *Wall Street Journal*, March 24, 2010
[27] "Google Co-Founder on Pulling out of China" (http://www.spiegel.de/international/business/0,1518,686269,00.html) *Spiegel Online*, March 30, 2010
[28] "Congress slams China and Microsoft, praises Google" (http://money.cnn.com/2010/03/24/technology/china_google_hearing/index. htm) *CNN Money*, March 24, 2010
[29] "Google's deals in doubt amid spat with Beijing" (http://news.yahoo.com/s/ap/20100325/ap_on_hi_te/as_china_google) Yahoo News, March 25, 2010
[30] "GOODLATTE STATEMENT IN SUPPORT OF GOOGLE'S DECISION TO STOP CENSORING IN CHINA" (http://goodlatte.house. gov/2010/03/goodlatte-statement-in-support-of-googles-decision-to-stop-censoring-in-china.shtml) March 23, 2010
[31] "Google's strategy in China deserves praise" (http://www.kansascity.com/2010/03/28/1842611/googles-strategy-in-china-deserves. html) *Kansas City Star*, March 28, 2010
[32] "Don't Be Evil" (http://www.tnr.com/article/dont-be-evil), "The Heroism of Google," *The New Republic*, April 21, 2010
[33] "2002 Young Innovators Under 35: Sergey Brin, 28" (http://www.technologyreview.com/tr35/profile.aspx?TRID=238). Technology Review. 2002. . Retrieved August 14, 2011.
[34] Brin and Page Awarded MBAs (http://www.ie.edu/IE/php/en/noticia.php?id=225), Press Release, Sept. 9, 2003
[35] " Brin and Page Receive Marconi Foundation's Highest Honor (http://findarticles.com/p/articles/mi_m0EIN/is_2004_Sept_23/ ai_n6208748)". Press Release, September 23, 2004.
[36] "The World's Most Powerful People: #5 Sergey Brin and Larry Page" (http://www.forbes.com/lists/2009/20/ power-09_Sergey-Brin-and-Larry-Page_D664.html) *Forbes* magazine, Nov. 11, 2009
[37] National Academy of Engineering (http://www8.nationalacademies.org/onpinews/newsitem.aspx?RecordID=02062009), Press Release, Feb. 6, 2009
[38] National Science Foundation (http://www.nsfgrfp.org/why_apply/fellow_profiles/sergey_brin), Fellow Profiles
[39] "Topic page on Sergey Brin" (http://billionaires.forbes.com/topic/Sergey_Brin). Forbes. . Retrieved April 5, 2010.
[40] "Cars and Wind: What's next for Google as it pushes beyond the Web?" (http://voices.washingtonpost.com/posttech/2010/10/google_a_vanguard_of_the.html) *Washington Post*, Oct. 12, 2010
[41] "The wind cries transmission" (http://googleblog.blogspot.com/2010/10/wind-cries-transmission.html) Official Google Blog, Oct. 11, 2010
[42] "Google joins $5 billion U.S. offshore wind grid project" (http://www.reuters.com/article/idUSTRE69B4NU20101012?pageNumber=2) *Reuters* Oct. 12, 2010
[43] Markoff, John. "Google Cars Drive Themselves, in Traffic" (http://www.nytimes.com/2010/10/10/science/10google.html?_r=2& hp=&pagewanted=all) *New York Times,* Oct. 9, 2010
[44] Guynn, Jessica (September 17, 2008). " Google's Schmidt, Page and Brin hold court at Zeitgeist (http://latimesblogs.latimes.com/ technology/2008/09/googles-schmidt.html)". *Los Angeles Times*. Retrieved 2010-01-07.
[45] Schwartz, John (June 11, 2008). "Google Co-Founder Books a Space Flight" (http://www.nytimes.com/2008/06/11/technology/ 11soyuz.html?hp). The New York Times Online. . Retrieved June 11, 2008.
[46] Helft, Miguel (September 13, 2007). "Google Founders' Ultimate Perk: A NASA Runway" (http://www.nytimes.com/2007/09/13/ technology/13google.html). *The New York Times*. . Retrieved September 13, 2007.
[47] Kopytoff, Verne (September 13, 2007). "Google founders pay NASA $1.3 million to land at Moffett Airfield" (http://www.sfgate.com/ cgi-bin/article.cgi?f=/c/a/2007/09/13/BUPRS4MHA.DTL). *San Francisco Chronicle*. . Retrieved September 13, 2007.
[48] *American Business Association of Russian Professionals* (http://www.svod.org/2006/sponsors/ambar)

External links

- Sergey Brin (https://plus.google.com/109813896768294978296/about) on Google+
- List of scientific publications by Sergey Brin (http://en.scientificcommons.org/sergey_brin)
- Sergey Brin and Larry Page (http://www.ted.com/speakers/sergey_brin_and_larry_page.html/) at TED Conferences
- Appearances (http://www.c-spanvideo.org/sergeybrin) on C-SPAN
- Sergey Brin (http://www.charlierose.com/guest/view/2593) on *Charlie Rose*
- Sergey Brin (http://www.imdb.com/name/nm1962236/) at the Internet Movie Database
- Works by or about Sergey Brin (http://worldcat.org/identities/lccn-no2005-73928) in libraries (WorldCat catalog)
- Sergey Brin (http://topics.bloomberg.com/sergey-brin/) collected news and commentary at *Bloomberg News*
- Sergey Brin (http://www.guardian.co.uk/media/sergeybrin) collected news and commentary at *The Guardian*

- Sergey Brin (http://topics.nytimes.com/top/reference/timestopics/people/b/sergey_brin/) collected news and commentary at *The New York Times*
- Sergey Brin (http://topics.wsj.com/person/B/sergey-brin/584) collected news and commentary at *The Wall Street Journal*
- Profile: Sergey Brin (http://news.bbc.co.uk/2/hi/business/3666241.stm) at *BBC News*
- Sergey Brin (http://www.forbes.com/profile/sergey-brin) at *Forbes*

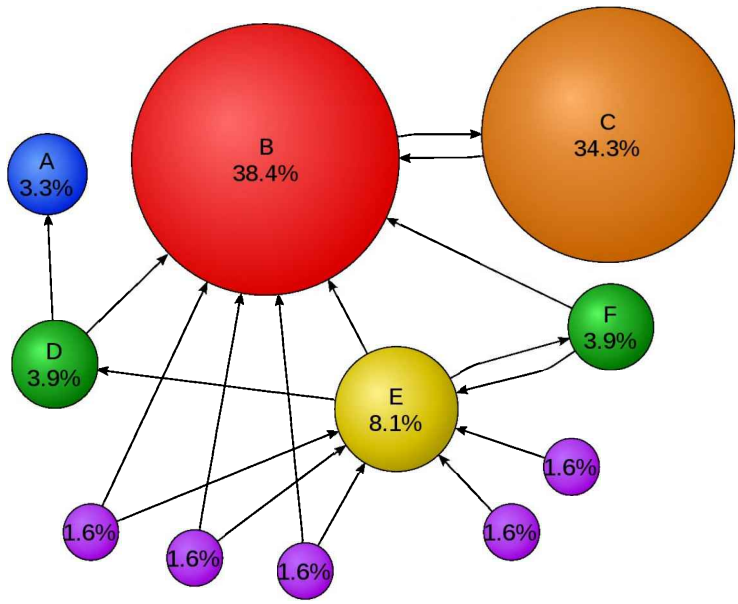
Interviews and articles

- Linux Journal interview (http://www.linuxjournal.com/article/4196) – August 31, 2000
- Net Café Television Interview (http://www.archive.org/details/Newwebsi01) – October 6, 2000. Interview starts around 18 minutes and 15 seconds in.
- 14, 2003 Fresh Air radio interview (http://www.npr.org/templates/rundowns/rundown.php?prgId=13& prgDate=October) - October 14, 2003
- Search Engine Watch interview (http://searchenginewatch.com/searchday/article.php/3081081) – October 16, 2003
- Video of Brin giving a lecture at UC Berkeley (http://video.google.com/ videoplay?docid=7582902000166025817) – Mentions Wikipedia and discusses development of search engines, Google and its evolution, Q&A (Fall 2005)
- Time Magazine Podcast about Google and its founders (http://www.time.com/time/podcast/business/ In_Search_of_the_Real_Google.mp3)
- "The Searchmeisters" (http://bnaibrith.org/pubs/bnaibrith/spring2006bbm/searchmeisters.cfm) – profile on Brin and Page from the *B'nai B'rith Magazine* (Spring 2006)
- Forbes.com: Fortunes That Roared In 2004 (http://forbes.com/2004/12/23/cz_pn_fortuneslide_2.html)
- Momentmag.com: The Story of Sergey Brin (http://www.momentmag.com/Exclusive/2007/2007-02/ 200702-BrinFeature.html)
- On the Origins of Google (http://www.nsf.gov/discoveries/disc_summ.jsp?cntn_id=100660&org=NSF), National Science Foundation

PageRank

PageRank is a link analysis algorithm, named after Larry Page[1] and used by the Google Internet search engine, that assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set. The algorithm may be applied to any collection of entities with reciprocal quotations and references. The numerical weight that it assigns to any given element E is referred to as the *PageRank of E* and denoted by $PR(E)$.

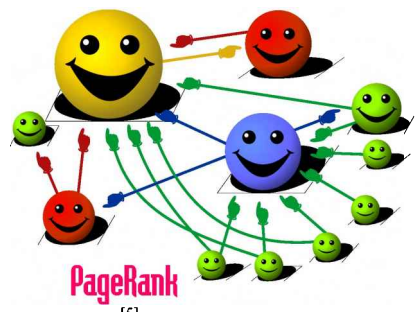
The name "PageRank" is a trademark of Google, and the PageRank process has been patented (U.S. Patent 6285999 [2]). However, the patent is assigned to Stanford University and not to Google. Google has exclusive license rights on the patent from Stanford University. The university received 1.8 million shares of Google in exchange for use of the patent; the shares were sold in 2005 for $336 million.[3][4]



Mathematical **PageRanks** for a simple network, expressed as percentages. (Google uses a logarithmic scale.) Page C has a higher PageRank than Page E, even though there are fewer links to C; the one link to C comes from an important page and hence is of high value. If web surfers who start on a random page have an 85% likelihood of choosing a random link from the page they are currently visiting, and a 15% likelihood of jumping to a page chosen at random from the entire web, they will reach Page E 8.1% of the time. (The 15% likelihood of jumping to an arbitrary page corresponds to a damping factor of 85%.) Without damping, all web surfers would eventually end up on Pages A, B, or C, and all other pages would have PageRank zero. In the presence of damping, Page A effectively links to all pages in the web, even though it has no outgoing links of its own.

Description

A PageRank results from a mathematical algorithm based on the graph, the webgraph, created by all World Wide Web pages as nodes and hyperlinks as edges, taking into consideration authority hubs such as cnn.com or usa.gov. The rank value indicates an importance of a particular page. A hyperlink to a page counts as a vote of support. The PageRank of a page is defined recursively and depends on the number and PageRank metric of all pages that link to it ("incoming links"). A page that is linked to by many pages with high PageRank receives a high rank itself. If there are no links to a web page there is no support for that page.



[5] Principles of PageRank

Numerous academic papers concerning PageRank have been published since Page and Brin's original paper. In practice, the PageRank concept has proven to be vulnerable to manipulation, and extensive research has been devoted to identifying falsely inflated PageRank and ways to ignore links from documents with falsely inflated PageRank.

Other link-based ranking algorithms for Web pages include the HITS algorithm invented by Jon Kleinberg (used by Teoma and now Ask.com), the IBM CLEVER project, and the TrustRank algorithm.

History

PageRank was developed at Stanford University by Larry Page (hence the name *Page*-Rank[6]) and Sergey Brin as part of a research project about a new kind of search engine.[7] Sergey Brin had the idea that information on the web could be ordered in a hierarchy by "link popularity": a page is ranked higher as there are more links to it.[8] It was co-authored by Rajeev Motwani and Terry Winograd. The first paper about the project, describing PageRank and the initial prototype of the Google search engine, was published in 1998:[5] shortly after, Page and Brin founded Google Inc., the company behind the Google search engine. While just one of many factors that determine the ranking of Google search results, PageRank continues to provide the basis for all of Google's web search tools.[9]

PageRank has been influenced by citation analysis, early developed by Eugene Garfield in the 1950s at the University of Pennsylvania, and by Hyper Search, developed by Massimo Marchiori at the University of Padua. In the same year PageRank was introduced (1998), Jon Kleinberg published his important work on HITS. Google's founders cite Garfield, Marchiori, and Kleinberg in their original paper.[5]

A small search engine called "RankDex" from IDD Information Services designed by Robin Li was, since 1996, already exploring a similar strategy for site-scoring and page ranking.[10] The technology in RankDex would be patented by 1999[11] and used later when Li founded Baidu in China.[12][13] Li's work would be referenced by some of Larry Page's U.S. patents for his Google search methods.[14]

Algorithm

PageRank is a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. PageRank can be calculated for collections of documents of any size. It is assumed in several research papers that the distribution is evenly divided among all documents in the collection at the beginning of the computational process. The PageRank computations require several passes, called "iterations", through the collection to adjust approximate PageRank values to more closely reflect the theoretical true value.

A probability is expressed as a numeric value between 0 and 1. A 0.5 probability is commonly expressed as a "50% chance" of something happening. Hence, a PageRank of 0.5 means there is a 50% chance that a person clicking on a random link will be directed to the document with the 0.5 PageRank.

Simplified algorithm

Assume a small universe of four web pages: **A**, **B**, **C** and **D**. Links from a page to itself, or multiple outbound links from one single page to another single page, are ignored. PageRank is initialized to the same value for all pages. In the original form of PageRank, the sum of PageRank over all pages was the total number of pages on the web at that time, so each page in this example would have an initial PageRank of 1. However, later versions of PageRank, and the remainder of this section, assume a probability distribution between 0 and 1. Hence the initial value for each page is 0.25.

The PageRank transferred from a given page to the targets of its outbound links upon the next iteration is divided equally among all outbound links.

If the only links in the system were from pages **B**, **C**, and **D** to **A**, each link would transfer 0.25 PageRank to **A** upon the next iteration, for a total of 0.75.

$$PR(A) = PR(B) + PR(C) + PR(D).$$

Suppose instead that page **B** had a link to pages **C** and **A**, while page **D** had links to all three pages. Thus, upon the next iteration, page **B** would transfer half of its existing value, or 0.125, to page **A** and the other half, or 0.125, to page **C**. Since **D** had three outbound links, it would transfer one third of its existing value, or approximately 0.083, to

A.

$$PR(A) = \frac{PR(B)}{2} + \frac{PR(C)}{1} + \frac{PR(D)}{3}.$$

In other words, the PageRank conferred by an outbound link is equal to the document's own PageRank score divided by the number of outbound links **L( )**.

$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)}.$$

In the general case, the PageRank value for any page **u** can be expressed as:

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)},$$

i.e. the PageRank value for a page **u** is dependent on the PageRank values for each page **v** contained in the set $\mathbf{B_u}$ (the set containing all pages linking to page **u**), divided by the number $L(v)$ of links from page **v**.

Damping factor

The PageRank theory holds that even an imaginary surfer who is randomly clicking on links will eventually stop clicking. The probability, at any step, that the person will continue is a damping factor *d*. Various studies have tested different damping factors, but it is generally assumed that the damping factor will be set around 0.85.[5]

The damping factor is subtracted from 1 (and in some variations of the algorithm, the result is divided by the number of documents (*N*) in the collection) and this term is then added to the product of the damping factor and the sum of the incoming PageRank scores. That is,

So any page's PageRank is derived in large part from the PageRanks of other pages. The damping factor adjusts the derived value downward. The

$$PR(A) = \frac{1-d}{N} + d\left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \cdots\right).$$

original paper, however, gave the following formula, which has led to some confusion:

$$PR(A) = 1 - d + d\left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \cdots\right).$$

The difference between them is that the PageRank values in the first formula sum to one, while in the second formula each PageRank is multiplied by *N* and the sum becomes *N*. A statement in Page and Brin's paper that "the sum of all PageRanks is one"[5] and claims by other Google employees[15] support the first variant of the formula above.

Page and Brin confused the two formulas in their most popular paper "The Anatomy of a Large-Scale Hypertextual Web Search Engine", where they mistakenly claimed that the latter formula formed a probability distribution over web pages.[5]

Google recalculates PageRank scores each time it crawls the Web and rebuilds its index. As Google increases the number of documents in its collection, the initial approximation of PageRank decreases for all documents.

The formula uses a model of a *random surfer* who gets bored after several clicks and switches to a random page. The PageRank value of a page reflects the chance that the random surfer will land on that page by clicking on a link. It can be understood as a Markov chain in which the states are pages, and the transitions, which are all equally probable, are the links between pages.

If a page has no links to other pages, it becomes a sink and therefore terminates the random surfing process. If the random surfer arrives at a sink page, it picks another URL at random and continues surfing again.

When calculating PageRank, pages with no outbound links are assumed to link out to all other pages in the collection. Their PageRank scores are therefore divided evenly among all other pages. In other words, to be fair with

pages that are not sinks, these random transitions are added to all nodes in the Web, with a residual probability usually set to $d = 0.85$, estimated from the frequency that an average surfer uses his or her browser's bookmark feature.

So, the equation is as follows:

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

where $p_1, p_2, \cdots, p_N$ are the pages under consideration, $M(p_i)$ is the set of pages that link to p_i, $L(p_j)$ is the number of outbound links on page p_j, and N is the total number of pages.

The PageRank values are the entries of the dominant eigenvector of the modified adjacency matrix. This makes PageRank a particularly elegant metric: the eigenvector is

$$\mathbf{R} = \begin{bmatrix} PR(p_1) \\ PR(p_2) \\ \vdots \\ PR(p_N) \end{bmatrix}$$

where **R** is the solution of the equation

$$\mathbf{R} = \begin{bmatrix} (1-d)/N \\ (1-d)/N \\ \vdots \\ (1-d)/N \end{bmatrix} + d \begin{bmatrix} \ell(p_1,p_1) & \ell(p_1,p_2) & \cdots & \ell(p_1,p_N) \\ \ell(p_2,p_1) & \ddots & & \vdots \\ \vdots & & \ell(p_i,p_j) & \\ \ell(p_N,p_1) & \cdots & & \ell(p_N,p_N) \end{bmatrix} \mathbf{R}$$

where the adjacency function $\ell(p_i, p_j)$ is 0 if page p_j does not link to p_i, and normalized such that, for each j

$$\sum_{i=1}^{N} \ell(p_i, p_j) = 1,$$

i.e. the elements of each column sum up to 1, so the matrix is a stochastic matrix (for more details see the computation section below). Thus this is a variant of the eigenvector centrality measure used commonly in network analysis.

Because of the large eigengap of the modified adjacency matrix above,[16] the values of the PageRank eigenvector can be approximated to within a high degree of accuracy within only a few iterations.

As a result of Markov theory, it can be shown that the PageRank of a page is the probability of arriving at that page after a large number of clicks. This happens to equal t^{-1} where t is the expectation of the number of clicks (or random jumps) required to get from the page back to itself.

One main disadvantage of PageRank is that it favors older pages. A new page, even a very good one, will not have many links unless it is part of an existing site (a site being a densely connected set of pages, such as Wikipedia).

The Google Directory (itself a derivative of the Open Directory Project) allows users to see results sorted by PageRank within categories. The Google Directory is the only service offered by Google where PageRank fully determines display order. In Google's other search services (such as its primary Web search), PageRank is only used to weight the relevance scores of pages shown in search results.

Several strategies have been proposed to accelerate the computation of PageRank.[17]

Various strategies to manipulate PageRank have been employed in concerted efforts to improve search results rankings and monetize advertising links. These strategies have severely impacted the reliability of the PageRank concept, which purports to determine which documents are actually highly valued by the Web community.

Since December 2007, when it started *actively* penalizing sites selling paid text links, Google has combatted link farms and other schemes designed to artificially inflate PageRank. How Google identifies link farms and other PageRank manipulation tools is among Google's trade secrets.

Computation

PageRank can be computed either iteratively or algebraically. The iterative method can be viewed as the power iteration method[18][19] or the power method. The basic mathematical operations performed are identical.

Iterative

At $t = 0$, an initial probability distribution is assumed, usually

$$PR(p_i; 0) = \frac{1}{N}.$$

At each time step, the computation, as detailed above, yields

$$PR(p_i; t+1) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j; t)}{L(p_j)},$$

or in matrix notation

$$\mathbf{R}(t+1) = d\mathcal{M}\mathbf{R}(t) + \frac{1-d}{N}\mathbf{1}, \qquad (*)$$

where $\mathbf{R}_i(t) = PR(p_i; t)$ and $\mathbf{1}$ is the column vector of length N containing only ones. The matrix $\mathcal{M}$ is defined as

$$\mathcal{M}_{ij} = \begin{cases} 1/L(p_j), & \text{if } j \text{ links to } i \\ 0, & \text{otherwise} \end{cases}$$

i.e.,

$$\mathcal{M} := (K^{-1}A)^T,$$

where A denotes the adjacency matrix of the graph and K is the diagonal matrix with the outdegrees in the diagonal.

The computation ends when for some small ϵ

$$|\mathbf{R}(t+1) - \mathbf{R}(t)| < \epsilon,$$

i.e., when convergence is assumed.

Algebraic

For $t \to \infty$ (i.e., in the steady state), the above equation (*) reads

$$\mathbf{R} = d\mathcal{M}\mathbf{R} + \frac{1-d}{N}\mathbf{1}. \qquad (**)$$

The solution is given by

$$\mathbf{R} = (\mathbf{I} - d\mathcal{M})^{-1}\frac{1-d}{N}\mathbf{1},$$

with the identity matrix $\mathbf{I}$.

The solution exists and is unique for $0 < d < 1$. This can be seen by noting that $\mathcal{M}$ is by construction a stochastic matrix and hence has an eigenvalue equal to one as a consequence of the Perron–Frobenius theorem.

Power Method

If the matrix $\mathcal{M}$ is a transition probability, i.e., column-stochastic with no columns consisting of just zeros and $\mathbf{R}$ is a probability distribution (i.e., $|\mathbf{R}| = 1$, $\mathbf{E}\mathbf{R} = 1$ where $\mathbf{E}$ is matrix of all ones), Eq. (**) is equivalent to

$$\mathbf{R} = \left(d\mathcal{M} + \frac{1-d}{N}\mathbf{E} \right) \mathbf{R} =: \widehat{\mathcal{M}}\mathbf{R}. \qquad (***)$$

Hence PageRank $\mathbf{R}$ is the principal eigenvector of $\widehat{\mathcal{M}}$. A fast and easy way to compute this is using the power method: starting with an arbitrary vector $x(0)$, the operator $\widehat{\mathcal{M}}$ is applied in succession, i.e.,

$$x(t+1) = \widehat{\mathcal{M}}x(t),$$

until

$$|x(t+1) - x(t)| < \epsilon.$$

Note that in Eq. (***) the matrix on the right-hand side in the parenthesis can be interpreted as

$$\frac{1-d}{N}\mathbf{I} = (1-d)\mathbf{P}\mathbf{1}^t,$$

where $\mathbf{P}$ is an initial probability distribution. In the current case

$$\mathbf{P} := \frac{1}{N}\mathbf{1}.$$

Finally, if $\mathcal{M}$ has columns with only zero values, they should be replaced with the initial probability vector $\mathbf{P}$. In other words

$$\mathcal{M}' := \mathcal{M} + \mathcal{D},$$

where the matrix $\mathcal{D}$ is defined as

$$\mathcal{D} := \mathbf{P}\mathbf{D}^t,$$

with

$$\mathbf{D}_i = \begin{cases} 1, & \text{if } L(p_i) = 0 \\ 0, & \text{otherwise} \end{cases}$$

In this case, the above two computations using $\mathcal{M}$ only give the same PageRank if their results are normalized:

$$\mathbf{R}_{\text{power}} = \frac{\mathbf{R}_{\text{iterative}}}{|\mathbf{R}_{\text{iterative}}|} = \frac{\mathbf{R}_{\text{algebraic}}}{|\mathbf{R}_{\text{algebraic}}|}.$$

PageRank MATLAB/Octave implementation

```
% Parameter M adjacency matrix where M_i,j represents the link from 'j'
 to 'i', such that for all 'j' sum(i, M_i,j) = 1
% Parameter d damping factor
% Parameter v_quadratic_error quadratic error for v
% Return v, a vector of ranks such that v_i is the i-th rank from [0,
1]

function [v] = rank(M, d, v_quadratic_error)
N = size(M, 2);  % N is equal to half the size of M
v = rand(N, 1);
v = v ./ norm(v, 2);
last_v = ones(N, 1) * inf;
M_hat = (d .* M) + (((1 - d) / N) .* ones(N, N));
while(norm(v - last_v, 2) > v_quadratic_error)
      last_v = v;
```

```
        v = M_hat * v;
        v = v ./ norm(v, 2);
end
```

Example of code calling the rank function defined above:

```
M = [0 0 0 0 1 ; 0.5 0 0 0 0 ; 0.5 0 0 0 0 ; 0 1 0.5 0 0 ; 0 0 0.5 1 0];
rank(M, 0.80, 0.001)
```

Efficiency

Depending on the framework used to perform the computation, the exact implementation of the methods, and the required accuracy of the result, the computation time of the these methods can vary greatly.

Variations

Google Toolbar

The Google Toolbar's PageRank feature displays a visited page's PageRank as a whole number between 0 and 10. The most popular websites have a PageRank of 10. The least have a PageRank of 0. Google has not disclosed the specific method for determining a Toolbar PageRank value, which is to be considered only a rough indication of the value of a website.

PageRank measures the number of sites that link to a particular page.[20] The PageRank of a particular page is roughly based upon the quantity of inbound links as well as the PageRank of the pages providing the links. The algorithm also includes other factors, such as the size of a page, the number of changes, the time since the page was updated, the text in headlines and the text in hyperlinked anchor texts.[8]

The Google Toolbar's PageRank is updated infrequently, so the values it shows are often out of date.

SERP Rank

The search engine results page (SERP) is the actual result returned by a search engine in response to a keyword query. The SERP consists of a list of links to web pages with associated text snippets. The SERP rank of a web page refers to the placement of the corresponding link on the SERP, where higher placement means higher SERP rank. The SERP rank of a web page is a function not only of its PageRank, but of a relatively large and continuously adjusted set of factors (over 200),[21][22] commonly referred to by internet marketers as "Google Love".[23] Search engine optimization (SEO) is aimed at achieving the highest possible SERP rank for a website or a set of web pages.

After the introduction of Google Places into the mainstream organic SERP, PageRank played little to no role in ranking a business in the Local Business Results.[24] While the theory of citations still plays a role in the algorithm, PageRank is not a factor since business listings, rather than web pages, are ranked.

Google directory PageRank

The Google Directory PageRank is an 8-unit measurement. Unlike the Google Toolbar, which shows a numeric PageRank value upon mouseover of the green bar, the Google Directory only displays the bar, never the numeric values.

False or spoofed PageRank

In the past, the PageRank shown in the Toolbar was easily manipulated. Redirection from one page to another, either via a HTTP 302 response or a "Refresh" meta tag, caused the source page to acquire the PageRank of the destination page. Hence, a new page with PR 0 and no incoming links could have acquired PR 10 by redirecting to the Google home page. This spoofing technique, also known as 302 Google Jacking, was a known vulnerability. Spoofing can generally be detected by performing a Google search for a source URL; if the URL of an entirely different site is displayed in the results, the latter URL may represent the destination of a redirection.

Manipulating PageRank

For search engine optimization purposes, some companies offer to sell high PageRank links to webmasters.[25] As links from higher-PR pages are believed to be more valuable, they tend to be more expensive. It can be an effective and viable marketing strategy to buy link advertisements on content pages of quality and relevant sites to drive traffic and increase a webmaster's link popularity. However, Google has publicly warned webmasters that if they are or were discovered to be selling links for the purpose of conferring PageRank and reputation, their links will be devalued (ignored in the calculation of other pages' PageRanks). The practice of buying and selling links is intensely debated across the Webmaster community. Google advises webmasters to use the nofollow HTML attribute value on sponsored links. According to Matt Cutts, Google is concerned about webmasters who try to game the system, and thereby reduce the quality and relevancy of Google search results.[25]

The intentional surfer model

The original PageRank algorithm reflects the so-called random surfer model, meaning that the PageRank of a particular page is derived from the theoretical probability of visiting that page when clicking on links at random. However, real users do not randomly surf the web, but follow links according to their interest and intention. A page ranking model that reflects the importance of a particular page as a function of how many actual visits it receives by real users is called the *intentional surfer model*.[26] The Google toolbar sends information to Google for every page visited, and thereby provides a basis for computing PageRank based on the intentional surfer model. The introduction of the nofollow attribute by Google to combat Spamdexing has the side effect that webmasters commonly use it on outgoing links to increase their own PageRank. This causes a loss of actual links for the Web crawlers to follow, thereby making the original PageRank algorithm based on the random surfer model potentially unreliable. Using information about users' browsing habits provided by the Google toolbar partly compensates for the loss of information caused by the nofollow attribute. The SERP rank of a page, which determines a page's actual placement in the search results, is based on a combination of the random surfer model (PageRank) and the intentional surfer model (browsing habits) in addition to other factors.[27]

Other uses

A version of PageRank has recently been proposed as a replacement for the traditional Institute for Scientific Information (ISI) impact factor,[28] and implemented at eigenfactor.org [29]. Instead of merely counting total citation to a journal, the "importance" of each citation is determined in a PageRank fashion.

A similar new use of PageRank is to rank academic doctoral programs based on their records of placing their graduates in faculty positions. In PageRank terms, academic departments link to each other by hiring their faculty from each other (and from themselves).[30]

PageRank has been used to rank spaces or streets to predict how many people (pedestrians or vehicles) come to the individual spaces or streets.[31][32] In lexical semantics it has been used to perform Word Sense Disambiguation[33] and to automatically rank WordNet synsets according to how strongly they possess a given semantic property, such as positivity or negativity.[34]

A dynamic weighting method similar to PageRank has been used to generate customized reading lists based on the link structure of Wikipedia.[35]

A Web crawler may use PageRank as one of a number of importance metrics it uses to determine which URL to visit during a crawl of the web. One of the early working papers [36] that were used in the creation of Google is *Efficient crawling through URL ordering*,[37] which discusses the use of a number of different importance metrics to determine how deeply, and how much of a site Google will crawl. PageRank is presented as one of a number of these importance metrics, though there are others listed such as the number of inbound and outbound links for a URL, and the distance from the root directory on a site to the URL.

The PageRank may also be used as a methodology [38] to measure the apparent impact of a community like the Blogosphere on the overall Web itself. This approach uses therefore the PageRank to measure the distribution of attention in reflection of the Scale-free network paradigm.

In any ecosystem, a modified version of PageRank may be used to determine species that are essential to the continuing health of the environment.[39]

An application of PageRank to the analysis of protein networks in biology is reported recently.[40]

nofollow

In early 2005, Google implemented a new value, "nofollow",[41] for the rel attribute of HTML link and anchor elements, so that website developers and bloggers can make links that Google will not consider for the purposes of PageRank—they are links that no longer constitute a "vote" in the PageRank system. The nofollow relationship was added in an attempt to help combat spamdexing.

As an example, people could previously create many message-board posts with links to their website to artificially inflate their PageRank. With the nofollow value, message-board administrators can modify their code to automatically insert "rel='nofollow'" to all hyperlinks in posts, thus preventing PageRank from being affected by those particular posts. This method of avoidance, however, also has various drawbacks, such as reducing the link value of legitimate comments. (See: Spam in blogs#nofollow)

In an effort to manually control the flow of PageRank among pages within a website, many webmasters practice what is known as PageRank Sculpting[42]—which is the act of strategically placing the nofollow attribute on certain internal links of a website in order to funnel PageRank towards those pages the webmaster deemed most important. This tactic has been used since the inception of the nofollow attribute, but may no longer be effective since Google announced that blocking PageRank transfer with nofollow does not redirect that PageRank to other links.[43]

Deprecation

PageRank was once available for the verified site maintainers through the Google Webmaster Tools interface. However on October 15, 2009, a Google employee confirmed[44] that the company had removed PageRank from its *Webmaster Tools* section, explaining that "We've been telling people for a long time that they shouldn't focus on PageRank so much; many site owners seem to think it's the most important metric for them to track, which is simply not true."[44] The PageRank indicator is not available in Google's own Chrome browser.

The visible page rank is updated very infrequently.

On 6 October 2011, many users mistakenly thought Google PageRank was gone. As it turns out, it was simply an update to the URL used to query the PageRank from Google.[45]

Google now also relies on other strategies as well as PageRank, such as Google Panda[46].

Notes

[1] "Google Press Center: Fun Facts" (http://web.archive.org/web/20090424093934/http://www.google.com/press/funfacts.html). www.google.com. Archived from the original (http://www.google.com/press/funfacts.html) on 2009-04-24..

[2] http://www.google.com/patents?vid=6285999

[3] Lisa M. Krieger (1 December 2005). "Stanford Earns $336 Million Off Google Stock" (http://www.redorbit.com/news/education/318480/stanford_earns_336_million_off_google_stock/). San Jose Mercury News, cited by redOrbit.. Retrieved 2009-02-25.

[4] Richard Brandt. "Starting Up. How Google got its groove" (http://www.stanfordalumni.org/news/magazine/2004/novdec/features/ startingup.html). Stanford magazine.. Retrieved 2009-02-25.

[5] Brin, S.; Page, L. (1998). "The anatomy of a large-scale hypertextual Web search engine" (http://infolab.stanford.edu/pub/papers/google. pdf). *Computer Networks and ISDN Systems* **30**: 107–117. doi:10.1016/S0169-7552(98)00110-X. ISSN 0169-7552..

[6] David Vise and Mark Malseed (2005). *The Google Story* (http://www.thegooglestory.com/). p. 37. ISBN ISBN 0-553-80457-X..

[7] Page, Larry, "PageRank: Bringing Order to the Web" (http://web.archive.org/web/20020506051802/www-diglib.stanford.edu/cgi-bin/ WP/get/SIDL-WP-1997-0072?1), Stanford Digital Library Project, talk. August 18, 1997 (archived 2002)

[8] 187-page study from Graz University, Austria (http://www.google-watch.org/gpower.pdf), includes the note that also human brains are used when determining the page rank in Google

[9] "Google Technology" (http://www.google.com/technology/). Google.com.. Retrieved 2011-05-27.

[10] Li, Yanhong (August 6, 2002). "Toward a qualitative search engine". *Internet Computing, IEEE* (IEEE Computer Society) **2** (4): 24–29. doi:10.1109/4236.707687.

[11] USPTO, "Hypertext Document Retrieval System and Method" (http://www.google.com/patents?hl=en&lr=&vid=USPAT5920859&id=x04ZAAAAEBAJ&oi=fnd&dq=yanhong+li&printsec=abstract#v=onepage&q=yanhong li&f=false), U.S. Patent number:5920859, Inventor: Yanhong Li, Filing date: Feb 5, 1997, Issue date: Jul 6, 1999

[12] Greenberg, Andy, "The Man Who's Beating Google" (http://www.forbes.com/forbes/2009/1005/ technology-baidu-robin-li-man-whos-beating-google_2.html), *Forbes* magazine, October 05, 2009

[13] "About: RankDex" (http://www.rankdex.com/about.html), *rankdex.com*

[14] Cf. especially Lawrence Page, U.S. patents 6,799,176 (2004) "Method for scoring documents in a linked database", 7,058,628 (2006) "Method for node ranking in a linked database", and 7,269,587 (2007) "Scoring documents in a linked database" 2011

[15] Matt Cutts's blog: Straight from Google: What You Need to Know (http://www.mattcutts.com/blog/seo-for-bloggers/), see page 15 of his slides.

[16] Taher Haveliwala and Sepandar Kamvar. (March 2003). "The Second Eigenvalue of the Google Matrix" (http://www-cs-students.stanford.edu/~taherh/papers/secondeigenvalue.pdf) (PDF). *Stanford University Technical Report*: 7056. arXiv:math/0307056. Bibcode 2003math.....7056N..

[17] Gianna M. Del Corso, Antonio Gullí, Francesco Romani (2005). "Fast PageRank Computation via a Sparse Linear System". *Internet Mathematics* **2** (3). doi:10.1.1.118.5422.

[18] Arasu, A. and Novak, J. and Tomkins, A. and Tomlin, J. (2002). "PageRank computation and the structure of the web: Experiments and algorithms". *Proceedings of the Eleventh International World Wide Web Conference, Poster Track*. Brisbane, Australia. pp. 107–117. doi:10.1.1.18.5264.

[19] Massimo Franceschet (2010). "PageRank: Standing on the shoulders of giants". arXiv:1002.2858 [cs.IR].

[20] Google Webmaster central (http://www.google.com/support/forum/p/Webmasters/thread?tid=4aeb4d5fce33350b&hl=en) discussion on PR

[21] Aubuchon, Vaughn. "Google Ranking Factors - SEO Checklist" (http://www.vaughns-1-pagers.com/internet/google-ranking-factors. htm)..

[22] Fishkin, Rand; Jeff Pollard (April 2, 2007). "Search Engine Ranking Factors - Version 2" (http://www.seomoz.org/article/ search-ranking-factors). seomoz.org.. Retrieved May 11, 2009.

[23] http://www.infoworld.com/t/search-engines/google-corrupt-search-me-428

[24] "Ranking of listings : Ranking - Google Places Help" (http://google.com/support/places/bin/answer.py?hl=en&answer=7091). Google.com. . Retrieved 2011-05-27.

[25] "How to report paid links" (http://www.mattcutts.com/blog/how-to-report-paid-links/). mattcutts.com/blog. April 14, 2007. . Retrieved 2007-05-28.

[26] Jøsang, A. (2007). "Trust and Reputation Systems" (http://www.unik.no/people/josang/papers/Jos2007-FOSAD.pdf). In Aldini, A. (PDF). *Foundations of Security Analysis and Design IV, FOSAD 2006/2007 Tutorial Lectures*. . **4677**. Springer LNCS 4677. pp. 209–245. doi:10.1007/978-3-540-74810-6. .

[27] SEOnotepad. "Myth of the Google Toolbar Ranking" (http://www.seonotepad.com/search-engines/google-seo/ myth-of-the-google-toolbar-ranking/). .

[28] Johan Bollen, Marko A. Rodriguez, and Herbert Van de Sompel. (December 2006). "Journal Status". *Scientometrics* **69** (3): 1030. arXiv:cs.GL/0601030. Bibcode 2006cs.....1030B.

[29] http://www.eigenfactor.org

[30] Benjamin M. Schmidt and Matthew M. Chingos (2007). "Ranking Doctoral Programs by Placement: A New Method" (http://www.people.fas.harvard.edu/~gillum/rankings_paper.pdf) (PDF). *PS: Political Science and Politics* **40** (July): 523–529. .

[31] B. Jiang (2006). "Ranking spaces for predicting human movement in an urban environment". *International Journal of Geographical Information Science* **23** (7): 823–837. arXiv:physics/0612011. doi:10.1080/13658810802022822.

[32] Jiang B., Zhao S., and Yin J. (2008). "Self-organized natural roads for predicting traffic flow: a sensitivity study". *Journal of Statistical Mechanics: Theory and Experiment* **P07008** (07): 008. arXiv:0804.1630. Bibcode 2008JSMTE..07..008J. doi:10.1088/1742-5468/2008/07/P07008.

[33] Roberto Navigli, Mirella Lapata. "An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation" (http://www.dsi.uniroma1.it/~navigli/pubs/PAMI_2010_Navigli_Lapata.pdf). IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 32(4), IEEE Press, 2010, pp. 678–692.

[34] Andrea Esuli and Fabrizio Sebastiani. "PageRanking WordNet synsets: An Application to Opinion-Related Properties" (http://nmis.isti.cnr.it/sebastiani/Publications/ACL07.pdf) (PDF). *In Proceedings of the 35th Meeting of the Association for Computational Linguistics, Prague, CZ, 2007, pp. 424–431*. . Retrieved June 30, 2007.

[35] Wissner-Gross, A. D. (2006). "Preparation of topical readings lists from the link structure of Wikipedia" (http://www.alexwg.org/ publications/ProcIEEEICALT_6-825.pdf). *Proceedings of the IEEE International Conference on Advanced Learning Technology* (Rolduc, Netherlands): 825. doi:10.1109/ICALT.2006.1652568. .

[36] "Working Papers Concerning the Creation of Google" (http://dbpubs.stanford.edu:8091/diglib/pub/projectdir/google.html). *Google*. . Retrieved November 29, 2006.

[37] Cho, J., Garcia-Molina, H., and Page, L. (1998). "Efficient crawling through URL ordering" (http://dbpubs.stanford.edu:8090/pub/ 1998-51). *Proceedings of the seventh conference on World Wide Web* (Brisbane, Australia). .

[38] http://de.scientificcommons.org/23846375

[39] Burns, Judith (2009-09-04). "Google trick tracks extinctions" (http://news.bbc.co.uk/2/hi/science/nature/8238462.stm). BBC News. . Retrieved 2011-05-27.

[40] G. Ivan and V. Grolmusz (2011). "When the Web meets the cell: using personalized PageRank for analyzing protein interaction networks" (http://bioinformatics.oxfordjournals.org/content/27/3/405). *Bioinformatics* (Vol. 27, No. 3. pp. 405-407) **27** (3): 405–7. doi:10.1093/bioinformatics/btq680. PMID 21149343. .

[41] "Preventing Comment Spam" (http://googleblog.blogspot.com/2005/01/preventing-comment-spam.html). *Google*. . Retrieved January 1, 2005.

[42] "PageRank Sculpting: Parsing the Value and Potential Benefits of Sculpting PR with Nofollow" (http://www.seomoz.org/blog/ pagerank-sculpting-parsing-the-value-and-potential-benefits-of-sculpting-pr-with-nofollow). SEOmoz. . Retrieved 2011-05-27.

[43] "PageRank sculpting" (http://www.mattcutts.com/blog/pagerank-sculpting/). Mattcutts.com. 2009-06-15. . Retrieved 2011-05-27.

[44] Susan Moskwa. "PageRank Distribution Removed From WMT" (http://www.google.com/support/forum/p/Webmasters/thread?tid=6a1d6250e26e9e48&hl=en). . Retrieved October 16, 2009

[45] *WhatCulture!*. 6 October 2011. http://whatculture.com/technology/google-pagerank-is-not-dead.php. Retrieved 7 October 2011.

[46] *Google Panda Update: Say Goodbye to Low-Quality Link Building* (http://searchenginewatch.com/article/2067687/ Google-Panda-Update-Say-Goodbye-to-Low-Quality-Link-Building), Search Engine Watch, 08.02.11,

References

- Altman, Alon; Moshe Tennenholtz (2005). "Ranking Systems: The PageRank Axioms" (http://stanford.edu/~epsalon/pagerank.pdf) (PDF). *Proceedings of the 6th ACM conference on Electronic commerce (EC-05)*. Vancouver, BC. Retrieved 2008-02-05.
- Cheng, Alice; Eric J. Friedman (2006-06-11). "Manipulability of PageRank under Sybil Strategies" (http://www.cs.duke.edu/nicl/netecon06/papers/ne06-sybil.pdf) (PDF). *Proceedings of the First Workshop on the Economics of Networked Systems (NetEcon06)*. Ann Arbor, Michigan. Retrieved 2008-01-22.
- Farahat, Ayman; LoFaro, Thomas; Miller, Joel C.; Rae, Gregory and Ward, Lesley A. (2006). "Authority Rankings from HITS, PageRank, and SALSA: Existence, Uniqueness, and Effect of Initialization". *SIAM Journal on Scientific Computing* **27** (4): 1181–1201. doi:10.1137/S1064827502412875.
- Haveliwala, Taher; Jeh, Glen and Kamvar, Sepandar (2003). "An Analytical Comparison of Approaches to Personalizing PageRank" (http://www-cs-students.stanford.edu/~taherh/papers/comparison.pdf) (PDF). *Stanford University Technical Report*.
- Langville, Amy N.; Meyer, Carl D. (2003). "Survey: Deeper Inside PageRank". *Internet Mathematics* **1** (3).
- Langville, Amy N.; Meyer, Carl D. (2006). *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press. ISBN 0-691-12202-4.
- Page, Lawrence; Brin, Sergey; Motwani, Rajeev and Winograd, Terry (1999). *The PageRank citation ranking: Bringing order to the Web* (http://dbpubs.stanford.edu:8090/pub/showDoc.Fulltext?lang=en&doc=1999-66& format=pdf&compression=).
- Richardson, Matthew; Domingos, Pedro (2002). "The intelligent surfer: Probabilistic combination of link and content information in PageRank" (http://www.cs.washington.edu/homes/pedrod/papers/nips01b.pdf) (PDF). *Proceedings of Advances in Neural Information Processing Systems*. **14**.

Relevant patents

- Original PageRank U.S. Patent—Method for node ranking in a linked database (http://patft.uspto.gov/netacgi/ nph-Parser?patentnumber=6,285,999)—Patent number 6,285,999—September 4, 2001
- PageRank U.S. Patent—Method for scoring documents in a linked database (http://patft1.uspto.gov/netacgi/ nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=/netahtml/PTO/srchnum.htm&r=1&f=G&l=50&s1=6,799,176.PN.&OS=PN/6,799,176&RS=PN/6,799,176)—Patent number 6,799,176—September 28, 2004
- PageRank U.S. Patent—Method for node ranking in a linked database (http://patft.uspto.gov/netacgi/ nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=/netahtml/PTO/search-adv.htm&r=1&p=1&f=G&l=50&d=PTXT&S1=7,058,628.PN.&OS=pn/7,058,628&RS=PN/7,058,628)—Patent number 7,058,628—June 6, 2006
- PageRank U.S. Patent—Scoring documents in a linked database (http://patft.uspto.gov/netacgi/ nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=/netahtml/PTO/search-adv.htm&r=1&p=1&f=G&l=50&d=PTXT&S1=7,269,587.PN.&OS=pn/7,269,587&RS=PN/7,269,587)—Patent number 7,269,587—September 11, 2007

External links

- Our Search: Google Technology (http://www.google.com/technology/) by Google
- How Google Finds Your Needle in the Web's Haystack (http://www.ams.org/featurecolumn/archive/ pagerank.html) by the American Mathematical Society
- Web PageRank prediction with Markov models (http://www.needocs.com/document/ web-pagerank-prediction-with-markov-models,10342) Michalis Vazirgiannis, Dimitris Drosos, Pierre Senellart, Akrivi Vlachou - Research paper
- How does Google rank webpages? (http://scenic.princeton.edu/network20q/lectures/Q3_notes.pdf) 20Q: About Networked Life, A class on networks
- Scientist discovers PageRank-type algorithm from the 1940s (http://www.technologyreview.com/blog/arxiv/ 24821)—February 17, 2010

Inbound link

Backlinks, also known as **incoming links**, **inbound links**, **inlinks**, and **inward links**, are incoming links to a website or web page. In basic link terminology, a **backlink** is any link received by a web node (web page, directory, website, or top level domain) from another web node.[1]

Inbound links were originally important (prior to the emergence of search engines) as a primary means of web navigation; today, their significance lies in search engine optimization (SEO). The number of backlinks is one indication of the popularity or importance of that website or page (for example, this is used by Google to determine the PageRank of a webpage). Outside of SEO, the backlinks of a webpage may be of significant personal, cultural or semantic interest: they indicate who is paying attention to that page.

Search engine rankings

Search engines often use the number of backlinks that a website has as one of the most important factors for determining that website's search engine ranking, popularity and importance. Google's description of their PageRank system, for instance, notes that *Google interprets a link from page A to page B as a vote, by page A, for page B.*[2] Knowledge of this form of search engine rankings has fueled a portion of the SEO industry commonly termed linkspam, where a company attempts to place as many inbound links as possible to their site regardless of the context of the originating site.

Websites often employ various search engine optimization techniques to increase the number of backlinks pointing to their website. Some methods are free for use by everyone whereas some methods like linkbaiting requires quite a bit of planning and marketing to work. Some websites stumble upon "linkbaiting" naturally; the sites that are the first with a tidbit of 'breaking news' about a celebrity are good examples of that. When "linkbait" happens, many websites will link to the 'baiting' website because there is information there that is of extreme interest to a large number of people.

There are several factors that determine the value of a backlink. Backlinks from authoritative sites on a given topic are highly valuable.[3] If both sites have content geared toward the keyword topic, the backlink is considered relevant and believed to have strong influence on the search engine rankings of the webpage granted the backlink. A backlink represents a favorable 'editorial vote' for the receiving webpage from another granting webpage. Another important factor is the anchor text of the backlink. Anchor text is the descriptive labeling of the hyperlink as it appears on a webpage. Search engine bots (i.e., spiders, crawlers, etc.) examine the anchor text to evaluate how relevant it is to the content on a webpage. Anchor text and webpage content congruency are highly weighted in search engine results page (SERP) rankings of a webpage with respect to any given keyword query by a search engine user.

Increasingly, inbound links are being weighed against link popularity and originating context. This transition is reducing the notion of *one link, one vote* in SEO, a trend proponents hope will help curb link spam as a whole.

Technical

When HTML (HyperTextMarkupLanguage) was designed, there was no explicit mechanism in the design to keep track of backlinks in software, as this carried additional logistical and network overhead.

Most Content management systems include features to track backlinks, provided the external site linking in sends notification to the target site. Most wiki systems include the capability of determining what pages link internally to any given page, but do not track external links to any given page.

Most commercial search engines provide a mechanism to determine the number of backlinks they have recorded to a particular web page. For example, Google can be searched using Google:link:http://www.wikipedia.org|link:wikipedia.org to find the number of pages on the Web pointing to http:// wikipedia.org/.Google only shows a small fraction of the number of links pointing to a site. It credits many more backlinks than it shows for each website.

Other mechanisms have been developed to track backlinks between disparate webpages controlled by organizations that aren't associated with each other. The most notable example of this is TrackBacks between blogs.

References

[1] Lennart Björneborn and Peter Ingwersen (2004). "Toward a Basic Framework for Webometrics" (http://www3.interscience.wiley.com/ cgi-bin/abstract/109594194/ABSTRACT). *Journal of the American Society for Information Science and Technology* **55** (14): 1216–1227. doi:10.1002/asi.20077. .

[2] Google's overview of PageRank (http://www.google.com/intl/en/technology/)

[3] "Does Backlink Quality Matter?" (http://www.adgooroo.com/backlink_quality.php). *Adgooroo*. 2010-04-21. . Retrieved 2010-04-21.

[[de:Rückverweis]

Matt Cutts

Matt Cutts works for the Search Quality group in Google, specializing in search engine optimization issues.[1]


Matt Cutts

Career

Cutts started his career in search when working on his Ph.D. at the University of North Carolina at Chapel Hill. According to quotation in a personal interview with an operator of another website, Matt got his Bachelor's degree at the University of Kentucky and Master's degree from the University of North Carolina, Chapel Hill. In the interview he was quoted his field of study was computer graphics and movement tracking, then moved into the field of information retrieval, and search engines[2] after taking two required outside classes from the university's Information and Library Science department.[2]

Before working at the Search Quality group at Google, Cutts worked at the ads engineering group and SafeSearch, Google's family filter. There he earned the nickname "porn cookie guy" by giving his wife's homemade cookies to any Googler who provided an example of unwanted pornography in the search results.[3]

Cutts is one of the co-inventors listed upon a Google patent related to search engines and web spam,[4] which was the first to publicly propose using historical data to identify link spam.

In November 2010, Cutts started a contest challenging developers to make Microsoft Kinect more compatible with the Linux operating system. At the time, Microsoft had stated that the use of Kinect with devices other than the Xbox 360 was not supported by them.[5]

Cutts has given advice and made statements on help related to the use of the Google search engine and related issues.[6]

In January 2012, on the news that Google had violated its quality guidelines, Cutts defended the downgrading of Chrome homepage results noting that it was not given special dispensation.[7]

References

[1] Ward, Mark (2004-06-08). "Inside the Google search machine" (http://news.bbc.co.uk/2/hi/technology/3783507.stm). *BBC News Online*. . Retrieved 2008-05-04.

[2] Wall, Aaron (2005). Interview of Matt Cutts (http://www.search-marketing.info/newsletter/articles/matt-cutts.htm). Retrieved December 15, 2006

[3] 'Google': An interesting read on a powerhouse company (http://www.usatoday.com/money/books/reviews/2005-11-13-google-book_x. htm), USA Today, November 13, 2005

[4] Acharya, A., et al., (2005) Information retrieval based on historical data (http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PG01&p=1&u=/netahtml/PTO/srchnum.html&r=1&f=G&l=50&s1="20050071741".PGNR.&OS=DN/20050071741&RS=DN/20050071741)

[5] "Kinect hacked days after release" (http://www.bbc.co.uk/news/technology-11742236). *BBC News* (BBC). 12 November 2010. . Retrieved 15 November 2010.

[6] http://www.brafton.com/news/cutts-urges-content-writers-to-use-keywords-naturally

- http://www.readwriteweb.com/enterprise/2011/12/googles-matt-cutts-good-conten.php
- http://www.webpronews.com/matt-cutts-talks-keyword-density-2011-12
- http://blogs.ft.com/fttechhub/2012/01/roundup-soundcloud-apple-google/

[7] Arthur, Charles. "Google shoves Chrome down search rankings after sponsored blog mixup" (http://www.guardian.co.uk/technology/ 2012/jan/04/google-chrome-browser-search-rankings?newsfeed=true). The Guardian. .

- http://news.cnet.com/8301-30685_3-57351920-264/two-days-after-google-flub-unruly-raises-$25-million/

Further reading

- David Vise and Mark Malseed (2005-11-15). *The Google Story*. Delacorte Press. ISBN 0-553-80457-X.

External links

- Matt Cutts: Gadgets, Google, and SEO (http://www.mattcutts.com/blog/) – his personal blog
- Matt Cutts (https://twitter.com/mattcutts) on Twitter
- Matt Cutts (http://www.ted.com/speakers/matt_cutts.html/) at TED Conferences
- 2009 BusinessWeek profile (http://www.businessweek.com/the_thread/techbeat/archives/2009/10/ matt_cutts_goog.html)
- Philipp Lenssen (2005). "Matt Cutts, Google's Gadgets Guy" (http://blog.outer-court.com/archive/ 2005-11-17-n52.html). blog.outer-court.com, Personal Blog. Retrieved December 15, 2006.

nofollow

nofollow is a value that can be assigned to the `rel` attribute of an HTML `a` element to instruct some search engines that a hyperlink should not influence the link target's ranking in the search engine's index. It is intended to reduce the effectiveness of certain types of search engine spam, thereby improving the quality of search engine results and preventing spamdexing from occurring.

Concept and specification

The `nofollow` value was originally suggested to stop comment spam in blogs. Believing that comment spam affected the entire blogging community, in early 2005 Google's Matt Cutts and Blogger's Jason Shellen proposed the value to address the problem.[1][2]

The specification for `nofollow` is copyrighted 2005-2007 by the authors and subject to a royalty free patent policy, e.g. per the W3C Patent Policy 20040205,[3] and IETF RFC 3667 & RFC 3668. The authors intend to submit this specification to a standards body with a liberal copyright/licensing policy such as the GMPG, IETF, and/or W3C.[2]

Example

```
<a href="http://www.example.com/" rel="nofollow">Link text</a>
```

Introduction and support

Google announced in early 2005 that hyperlinks with `rel="nofollow"`[4] would not influence the link target's PageRank.[5] In addition, the Yahoo and Bing search engines also respect this attribute value.[6]

On June 15, 2009, Matt Cutts, a well-known software engineer of Google, announced on his blog that GoogleBot will no longer treat nofollowed links in the same way, in order to prevent webmasters from using nofollow for PageRank sculpting. As a result of this change the usage of nofollow leads to evaporation of pagerank. In order to avoid the above, SEOs developed alternative techniques that replace nofollowed tags with obfuscated JavaScript code and thus permit PageRank sculpting. Additionally several solutions have been suggested that include the usage of iframes, Flash and JavaScript.

Interpretation by the individual search engines

While all engines that use the `nofollow` value exclude links that use it from their ranking calculation, the details about the exact interpretation of it vary from search engine to search engine.[7][8]

- Google states that their engine takes "nofollow" literally and does not "follow" the link at all. However, experiments conducted by SEOs show conflicting results. These studies reveal that Google does follow the link, but it does not index the linked-to page, unless it was in Google's index already for other reasons (such as other, non-nofollow links that point to the page).[8][9]
- Yahoo! "follows it", but excludes it from their ranking calculation.
- Bing respects "nofollow" as regards not counting the link in their ranking, but it is not proven whether or not Bing follows the link.
- Ask.com also respects the attribute.[10]

rel="nofollow" Action	Google	Yahoo!	Bing	Ask.com
Uses the link for ranking	No	No	No	?
Follows the link	No	Yes	?	No
Indexes the "linked to" page	No	Yes	No	No
Shows the existence of the link	Only for a previously indexed page	Yes	Yes	Yes
In results pages for anchor text	Only for a previously indexed page	Yes	Only for a previously indexed page	Yes

Use by weblog software

Many weblog software packages mark reader-submitted links this way[11] by default (often with no option to disable it, except for modification of the software's code).

More sophisticated server software could spare the nofollow for links submitted by trusted users like those registered for a long time, on a whitelist, or with an acceptable karma level. Some server software adds `rel="nofollow"` to pages that have been recently edited but omits it from stable pages, under the theory that stable pages will have had offending links removed by human editors.

The widely used blogging platform WordPress versions 1.5 and above automatically assign the `nofollow` attribute to all user-submitted links (comment data, commenter URI, etc.).[12] However, there are several free plugins available that automatically remove the `nofollow` attribute value.[13]

Use on other websites

MediaWiki software, which powers Wikipedia, was equipped with nofollow support soon after initial announcement in 2005. The option was enabled on most Wikipedias. One of the prominent exceptions was the English Wikipedia. Initially, after a discussion, it was decided not to use `rel="nofollow"` in articles and to use a URL blacklist instead. In this way, English Wikipedia contributed to the scores of the pages it linked to, and expected editors to link to relevant pages.

In May 2006, a patch to MediaWiki software allowed to enable nofollow selectively in namespaces. This functionality was used on pages that are not considered to be part of the actual encyclopedia, such as discussion pages and resources for editors.[14] Following increasing spam problems and a within-Foundation request from founder Jimmy Wales, `rel="nofollow"` was added to article-space links in January 2007.[15][16] However, the various interwiki templates and shortcuts that link to other Wikimedia Foundation projects and many external wikis such as Wikia are not affected by this policy.

Other websites like Slashdot, with high user participation, add `rel="nofollow"` only for potentially misbehaving users. Potential spammers posing as users can be determined through various heuristics like age of

registered account and other factors. Slashdot also uses the poster's karma as a determinant in attaching a nofollow tag to user submitted links.

Social bookmarking and photo sharing websites that use the `rel="nofollow"` tag for their outgoing links include YouTube and Digg.com[17] (for most links); websites that don't use the `rel="nofollow"` tag include Propeller.com (no longer an active website) (formerly Netscape.com), Yahoo! My Web 2.0, and TechnoratiFavs.[18]

Repurpose

Paid links

Search engines have attempted to repurpose the nofollow attribute for something different. Google began suggesting the use of `nofollow` also as a machine-readable disclosure for paid links, so that these links do not get credit in search engines' results.

The growth of the link buying economy, where companies' entire business models are based on paid links that affect search engine rankings,[19] caused the debate about the use of `nofollow` in combination with paid links to move into the center of attention of the search engines, who started to take active steps against link buyers and sellers. This triggered a very strong response from web masters.[20]

Control internal PageRank flow

Search engine optimization professionals started using the `nofollow` attribute to control the flow of PageRank within a website, but Google since corrected this error, and any link with a nofollow attribute decreases the PageRank that the page can pass on. This practice is known as "PageRank sculpting". This is an entirely different use than originally intended. `nofollow` was designed to control the flow of PageRank from one website to another. However, some SEOs have suggested that a `nofollow` used for an internal link should work just like `nofollow` used for external links.

Several SEOs have suggested that pages such as "About Us", "Terms of Service", "Contact Us", and "Privacy Policy" pages are not important enough to earn PageRank, and so should have `nofollow` on internal links pointing to them. Google employee Matt Cutts has provided indirect responses on the subject, but has never publicly endorsed this point ofview.[21]

The practice is controversial and has been challenged by some SEO professionals, including Shari Thurow[22] and Adam Audette.[23] Site search proponents have pointed out that visitors do search for these types of pages, so using `nofollow` on internal links pointing to them may make it difficult or impossible for visitors to find these pages in site searches powered by major search engines.

Although proponents of use of `nofollow` on internal links have cited an inappropriate attribution to Matt Cutts[24] (see Matt's clarifying comment, rebutting the attributed statement)[25] as support for using the technique, Cutts himself never actually endorsed the idea. Several Google employees (including Matt Cutts) have urged Webmasters not to focus on manipulating internal PageRank. Google employee Adam Lasnik[26] has advised webmasters that there are better ways (e.g. click hierarchy) than `nofollow` to "sculpt a bit of PageRank", but that it is available and "we're not going to frown upon it".

No reliable data has been published on the effectiveness or potential harm that use of `nofollow` on internal links may provide. Unsubstantiated claims have been challenged throughout the debate and some early proponents of the idea have subsequently cautioned people not to view the use of `nofollow` on internal links as a silver bullet or quick-success solution.

More general consensus seems to favor the use of `nofollow` on internal links pointing to user-controlled pages which may be subjected to spam link practices, including user profile pages, user comments, forum signatures and posts, calendar entries, etc.

YouTube, a Google company, uses nofollow on a number of internal 'help' and 'share' links.[27]

Criticism

Employment of the nofollow attribute by Wikipedia on all external links has been criticized for not passing the deserved rank to referenced pages which are the original source of each Wikipedia article's content. This was done to combat spamdexing on Wikipedia pages, which are an otherwise tempting target for spammers as Wikipedia is a very high ranking site on most search engines. It's argued that the referencing Wikipedia article may show in search results before its own sources; this may be seen as unfair and discourage contributions.[28][29]

Use of nofollow where comments or other user content is posted (e.g. Wikipedia) not only depreciates the links of spammers but also of users that might be constructively contributing to a discussion, and preventing such legitimate links from influencing the page ranking of the websites they target.[30]

References

[1] "The nofollow Attribute and SEO" (http://www.published-articles.com/Art/18844/136/The-nofollow-Attribute-and-SEO.html). *Published-Articles.com*. May 22, 2009. . Retrieved September 8, 2009.

[2] rel="nofollow" Specification (http://microformats.org/wiki/rel-nofollow), *Microformats.org*, retrieved June 17, 2007

[3] [[W3C (http://www.w3.org/Consortium/Patent-Policy-20040205/)] Patent Policy 20040205], *W3.ORG*

[4] W3C (December 24, 1999), HTML 4.01 Specification (http://www.w3.org/TR/REC-html40/struct/links.html#adef-rel), *W3C.org*, retrieved May 29, 2007

[5] Google (January 18, 2006), Preventing comment spam (http://googleblog.blogspot.com/2005/01/preventing-comment-spam.html), *Official Google Blog*, retrieved on May 29, 2007

[6] Microsoft (June 3, 2008), Bing.com (http://www.bing.com/community/blogs/webmaster/archive/2008/06/03/ robots-exclusion-protocol-joining-together-to-provide-better-documentation.aspx), "Bing Community", retrieved on June 11, 2009

[7] Loren Baker (April 29, 2007), How Google, Yahoo & Ask.com Treat the NoFollow Link Attribute (http://www.searchenginejournal.com/ how-google-yahoo-askcom-treat-the-no-follow-link-attribute/4801/), *Search Engine Journal*, retrieved May 29, 2007

[8] Michael Duz (December 2, 2006), rel="nofollow" Google, Yahoo and MSN (http://www.seo-blog.com/rel-nofollow.php), *SEO Blog*, retrieved May 29, 2007

[9] Rel Nofollow Test (http://www.bubub.org/rel_nofollow_test.html) from August 2007

[10] "Webmasters" (http://about.ask.com/en/docs/about/webmasters.shtml#10). About Ask.com. . Retrieved 2012-01-09.

[11] Google Blog (January 18, 2005), (http://googleblog.blogspot.com/2005/01/preventing-comment-spam.html), *The Official Google Blog*, retrieved September 28, 2010

[12] Codex Documentation, Nofollow (http://codex.wordpress.org/Nofollow), *Wordpress.org Documentation*, retrieved May 29, 2007

[13] WordPress Plugins, Plugins tagged as Nofollow (http://wordpress.org/extend/plugins/tags/nofollow), *WordPress Extensions*, retrieved March 10, 2008

[14] Wikipedia (May 29, 2006), Wikipedia Signpost/2006-05-29/Technology report, *Wikipedia.org*, retrieved May 29, 2007

[15] Brion Vibber (January 20, 2007), Nofollow back on URL links on en.wikipedia.org articles for now (http://lists.wikimedia.org/pipermail/ wikien-l/2007-January/061137.html), *Wikimedia List WikiEN-l*, retrieved May 29, 2007

[16] Wikipedia:Wikipedia Signpost/2007-01-22/Nofollow

[17] John Quinn (September 2, 2009), Recent Changes to NOFOLLOW on External Links (http://blog.digg.com/?p=864), *Digg the Blog*, retrieved on September 3, 2009

[18] Loren Baker (November 15, 2007), Social Bookmarking Sites Which Don't Use NoFollow Bookmarks and Search Engines (http://www. searchenginejournal.com/social-bookmarking-sites-which-dont-use-nofollow-bookmarks-and-search-engines/5985/), *Search Engine Journal*, retrieved on December 16, 2007

[19] Philipp Lenssen (April 19, 2007), The Paid Links Economy (http://blog.outer-court.com/archive/2007-04-19-n50.html), *Google Blogoscoped*, retrieved June 17, 2007

[20] Carsten Cumbrowski (May 14th, 2007), Matt Cutts on Paid Links Discussion - Q&A (http://www.searchenginejournal.com/ matt-cutts-update-on-paid-links-discussion-qa/4907/), *SearchEngineJournal.com*, retrieved June 17, 2007

[21] October 8, 2007, Eric Enge Interviews Google's Matt Cutts (http://www.stonetemple.com/articles/interview-matt-cutts.shtml), *Stone Temple Consulting*, retrieved on January 20, 2008.

[22] March 6, 2008, You'd be wise to "nofollow" this dubious advice (http://searchengineland.com/080306-083414.php), *Search Engine Land*.

[23] June 3, 2008 8 Arguments Against Sculpting PageRank With Nofollow (http://www.audettemedia.com/blog/ arguments-against-nofollow), *Audette Media*.

[24] August 29, 2007 Matt Cutts on Nofollow, Links-Per-Page and the Value of Directories (http://www.seomoz.org/blog/ questions-answers-with-googles-spam-guru), *SEomoz*.

[25] August 29, 2007 Seomoz.org (http://www.seomoz.org/blog/questions-answers-with-googles-spam-guru#jtc33536), *SEOmoz comment by Matt Cutts*.

[26] February 20, 2008 Interview with Adam Lasnik of Google (http://www.toprankblog.com/2008/02/adam-lasnik-video/)

[27] "NofollowReciprocity" (http://www.inverudio.com/programs/WordPressBlog/NofollowReciprocity.php). Inverudio.com. 2010-01-28. . Retrieved 2012-01-09.

[28] ""nofollow" criticism at" (http://blogoscoped.com/archive/2007-01-22-n21.html). Blogoscoped.com. 2007-01-25. . Retrieved 2012-01-09.

[29] ""nofollow" criticism at www.marketingpilgrim.com" (http://www.marketingpilgrim.com/2007/01/campaign-to-reduce-wikipedias-pagerank-to-zero.html). Marketingpilgrim.com. 2007-01-23. . Retrieved 2012-01-09.

[30] Official Google Blog: Preventing comment spam http://googleblog.blogspot.com/2005/01/preventing-comment-spam.html

Open Directory Project

Open Directory Project

URL	dmoz.org [1]
Commercial?	No
Type of site	Web directory
Registration	Optional
Content license	Creative Commons Attribution 3.0 Unported
Owner	Netscape (AOL)
Created by	Netscape
Launched	June 5, 1998
Alexa rank	▼730 (March 2012)[2]

The **Open Directory Project** (**ODP**), also known as **Dmoz** (from *directory.mozilla.org*, its original domain name), is a multilingual open content directory of World Wide Web links. It is owned by Netscape but it is constructed and maintained by a community of volunteer editors.

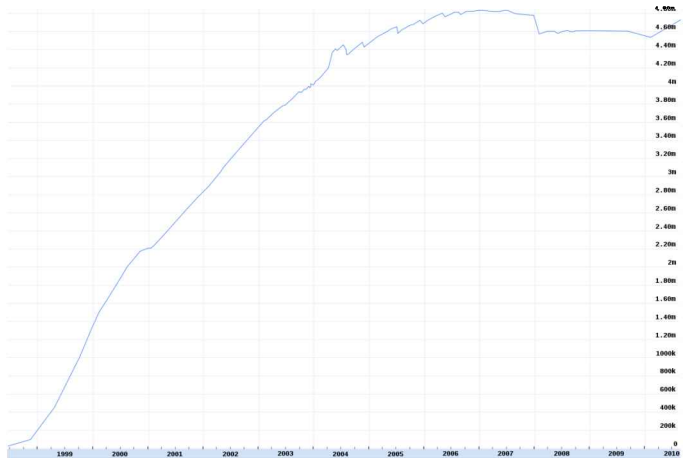
ODP uses a hierarchical ontology scheme for organizing site listings. Listings on a similar topic are grouped into categories which can then include smaller categories.

Project information

ODP was founded in the United States as **Gnuhoo** by Rich Skrenta and Bob Truel in 1998 while they were both working as engineers for Sun Microsystems. Chris Tolles, who worked at Sun Microsystems as the head of marketing for network security products, also signed on in 1998 as a co-founder of Gnuhoo along with co-founders Bryn Dole and Jeremy Wenokur. Skrenta had developed TASS, an ancestor of tin, the popular threaded Usenet newsreader for Unix systems. Coincidentally, the original category structure of the Gnuhoo directory was based loosely on the structure of Usenet newsgroups then in existence.

The Gnuhoo directory went live on June 5, 1998. After a *Slashdot* article suggested that Gnuhoo had nothing in common with the spirit of free software,[3] for which the GNU project was known, Richard Stallman and the Free Software Foundation objected to the use of Gnu. So Gnuhoo was changed to **NewHoo**. Yahoo! then objected to the use of "Hoo" in the name, prompting them to switch the name again. **ZURL** was the likely choice.[4] However, before the switch to ZURL, NewHoo was acquired by Netscape Communications Corporation in October 1998 and became the **Open Directory Project**. Netscape released the ODP data under the Open Directory License. Netscape was acquired by AOL shortly thereafter and ODP was one of the assets included in the acquisition. AOL later merged with Time-Warner.

By the time Netscape assumed stewardship, the Open Directory Project had about 100,000 URLs indexed with contributions from about 4500 editors. On October 5, 1999, the number of URLs indexed by ODP reached one million. According to an unofficial estimate, the URLs in the Open Directory numbered 1.6 million in April 2000, surpassing those in the Yahoo! Directory.[5] ODP achieved the milestones of indexing two million URLs on August 14, 2000, three million listings on November 18, 2001 and four million on December 3, 2003.



ODP's growth over time

From January 2006 the Open Directory published online reports to inform the public about the development of the project. The first report covered the year 2005. Monthly reports were issued subsequently until September 2006.[6] These reports gave greater insight into the functioning of the directory than the simplified statistics given on the front page of the directory. The number of listings and categories cited on the front page include "Test" and "Bookmarks" categories but these are not included in the RDF dump offered to users. The total number of editors who have contributed to the directory as of March 31, 2007 was 75,151.[7] There were about 7330 active editors during August 2006.[6]

System failure and editing outage, October to December 2006

On October 20, 2006, the ODP's main server suffered a catastrophic failure of the system[8] that prevented editors from working on the directory until December 18, 2006. During that period, an older build of the directory was visible to the public. On January 13, 2007, the Site Suggestion and Update Listings forms were again made available.[9] On January 26, 2007, weekly publication of RDF dumps resumed. To avoid future outages, the system now resides on a redundant configuration of two Intel-based servers.[10]

Competing and spinoff projects

As the ODP became more widely known, two other major web directories edited by volunteers and sponsored by Go.com and Zeal emerged, both now defunct. These directories did not license their content for open content distribution.[11][12]

The concept of using a large-scale community of editors to compile online content has been successfully applied to other types of projects. ODP's editing model directly inspired three other open content volunteer projects: an open content restaurant directory known as ChefMoz,[13] an open content music directory known as MusicMoz,[14] and an encyclopedia known as OpenSite.[15]

Content

Gnuhoo borrowed the basic outline for its initial ontology from Usenet. In 1998, Rich Skrenta said, "I took a long list of groups and hand-edited them into a hierarchy."[16] For example, the topic covered by the comp.ai.alife newsgroup was represented by the category Computers/AI/Artificial_Life. The original divisions were for *Adult*, *Arts*, *Business*, *Computers*, *Games*, *Health*, *Home*, *News*, *Recreation*, *Reference*, *Regional*, *Science*, *Shopping*, *Society* and *Sports*. While these fifteen *top-level* categories have remained intact, the ontology of second- and lower-level categories has undergone a gradual evolution; significant changes are initiated by discussion among editors and then implemented

when consensus has been reached.

In July 1998, the directory became multilingual with the addition of the *World* top-level category. The remainder of the directory lists only English language sites. By May 2005, seventy-five languages were represented. The growth rate of the non-English components of the directory has been greater than the English component since 2002. While the English component of the directory held almost 75% of the sites in 2003, the *World* level grew to over 1.5 million sites as of May 2005, forming roughly one-third of the directory. The ontology in non-English categories generally mirrors that of the English directory, although exceptions which reflect language differences are quite common.

Several of the top-level categories have unique characteristics. The *Adult* category is not present on the directory homepage but it is fully available in the RDF dump that ODP provides. While the bulk of the directory is categorized primarily by topic, the Regional category is categorized primarily by region. This has led many to view ODP as two parallel directories: *Regional* and *Topical*.

On November 14, 2000, a special directory within the Open Directory was created for people under 18 years of age.[17] Key factors distinguishing this "Kids and Teens" area from the main directory are:

- stricter guidelines which limit the listing of sites to those which are targeted or "appropriate" for people under 18 years of age;[18]
- category names as well as site descriptions use vocabulary which is "age appropriate";
- *age tags* on each listing distinguish content appropriate for kids (age 12 and under), teens (13 to 15 years old) and mature teens (16 to 18 years old);
- Kids and Teens content is available as a separate RDF dump;
- editing permissions are such that the community is parallel to that of the Open Directory. By May 2005, this

portion of the Open Directory included over 32,000 site listings.

Since early 2004, the whole site has been in UTF-8 encoding. Prior to this, the encoding used to be ISO 8859-1 for English language categories and a language-dependent character set for other languages. The RDF dumps have been encoded in UTF-8 since early 2000.

Maintenance

Directory listings are maintained by editors. While some editors focus on the addition of new listings, others focus on maintaining the existing listings. This includes tasks such as the editing of individual listings to correct spelling and/or grammatical errors, as well as monitoring the status of linked sites. Still others go through site submissions to remove spam and duplicate submissions.

Robozilla is a Web crawler written to check the status of all sites listed in ODP. Periodically, Robozilla will flag sites which appear to have moved or disappeared and editors follow up to check the sites and take action. This process is critical for the directory in striving to achieve one of its founding goals: to reduce the link rot in web directories. Shortly after each run, the sites marked with errors are automatically moved to the unreviewed queue where editors may investigate them when time permits.

Due to the popularity of the Open Directory and its resulting impact on search engine rankings (See PageRank), domains with lapsed registration that are listed on ODP have attracted domain hijacking, an issue that has been addressed by regularly removing expired domains from the directory.

While corporate funding and staff for the ODP have diminished in recent years, volunteers have created editing tools such as linkcheckers to supplement Robozilla, category crawlers, spellcheckers, search tools that directly sift a recent RDF dump, bookmarklets to help automate some editing functions, mozilla based add-ons,[19] and tools to help work through unreviewed queues.

License and requirements

ODP data is made available for open content distribution under the terms of the Open Directory License, which requires a specific ODP attribution table on every Web page that uses the data.

The Open Directory License also includes a requirement that users of the data continually check the ODP site for updates and discontinue use and distribution of the data or works derived from the data once an update occurs. This restriction prompted the Free Software Foundation to refer to the Open Directory License as a non-free documentation license, citing the right to redistribute a given version not being permanent and the requirement to check for changes to the license.[20]

RDF dumps

ODP data is made available through an RDF-like dump that is published on a dedicated download server, where an archive of previous versions is also available.[21] New versions are usually generated weekly. An ODP editor has catalogued a number of bugs that are/were encountered when implementing the ODP RDF dump, including UTF-8 encoding errors (fixed since August 2004) and an RDF format that does not comply with the final RDF specification because ODP RDF generation was implemented before the RDF specification was finalized.[22] So while today the so-called RDF dump is valid XML, it is not strictly RDF but an ODP-specific format and as such, software to process the ODP RDF dump needs to take account of this.

Content users

ODP data powers the core directory services for many of the Web's largest search engines and portals, including Netscape Search, AOL Search, and Alexa. Google Directory used ODP information, until being shuttered in July 2011.[23]

Other uses are also made of ODP data. For example, in the spring of 2004 Overture announced a search service for third parties combining Yahoo! Directory search results with ODP titles, descriptions and category metadata. The search engine Gigablast announced on May 12, 2005 its searchable copy of the Open Directory. The technology permits search of websites listed in specific categories, "in effect, instantly creating over 500,000 vertical search engines".[24]

As of 8 September 2006, the ODP listed 313 English-language Web sites that use ODP data as well as 238 sites in other languages.[25] However, these figures do not reflect the full picture of use, as those sites that use ODP data without following the terms of the ODP license are not listed.

Policies and procedures

There are restrictions imposed on who can become an ODP editor. The primary gatekeeping mechanism is an editor application process wherein editor candidates demonstrate their editing abilities, disclose affiliations that might pose a conflict of interest and otherwise give a sense of how the applicant would likely mesh with the ODP culture and mission.[26] A majority of applications are rejected but reapplying is allowed and sometimes encouraged. The same standards apply to editors of all categories and subcategories.

ODP's editing model is a hierarchical one. Upon becoming editors, individuals will generally have editing permissions in only a small category. Once they have demonstrated basic editing skills in compliance with the Editing Guidelines, they are welcome to apply for additional editing privileges in either a broader category or in a category elsewhere in the directory. Mentorship relationships between editors are encouraged and internal forums provide a vehicle for new editors to ask questions.

ODP has its own internal forums, the contents of which are intended only for editors to communicate with each other primarily about editing topics. Access to the forums requires an editor account and editors are expected to keep the contents of these forumsprivate.[27]

Over time, senior editors may be granted additional privileges which reflect their editing experience and leadership within the editing community. The most straightforward are *editall* privileges which allow an editor to access all categories in the directory. *Meta* privileges additionally allow editors to perform tasks such as reviewing editor applications, setting category features and handling external and internal abuse reports. *Cateditall* privileges are similar to *editall* but only for a single directory category. Similarly, *catmod* privileges are similar to *meta* but only for a single directory category. *Catmv* privileges allow editors to make changes to directory ontology by moving or renaming categories. All of these privileges are granted by admins and staff, usually after discussion with *meta* editors.

In August 2004, a new level of privileges called *admin* was introduced. Administrator status was granted to a number of long serving metas by staff. Administrators have the ability to grant editall+ privileges to other editors and to approve new directory-wide policies, powers which had previously only been available to root (staff) editors.[28] A full list of senior editors is available to the public,[29] as is a listing of all current editors.[30]

All ODP editors are expected to abide by ODP's Editing Guidelines. These guidelines describe editing basics: which types of sites may be listed and which may not; how site listings should be titled and described in a loosely consistent manner; conventions for the naming and building of categories; conflict of interest limitations on the editing of sites which the editor may own or otherwise be affiliated with; and a code of conduct within the community.[31] Editors who are found to have violated these guidelines may be contacted by staff or senior editors, have their editing permissions cut back or lose their editing privileges entirely. ODP Guidelines are periodically revised after discussion in editor forums.

Site submissions

One of the original motivations for forming Gnuhoo/Newhoo/ODP was the frustration that many people experienced in getting their sites listed on Yahoo! Directory. However Yahoo! has since implemented a paid service for timely consideration of site submissions. That lead has been followed by many other directories. Some accept no free submissions at all. By contrast the ODP has maintained its policy of free site submissions for all types of site—the only one of the major general directories to do so.

One result has been a gradual divergence between the ODP and other directories in the balance of content. The pay-for-inclusion model favours those able and willing to pay, so commercial sites tend to predominate in directories using it.[32] Conversely, a directory manned by volunteers will reflect the aims and interests of those volunteers. The ODP lists a high proportion of informational and non-profit sites.

Another consequence of the free submission policy is that the ODP has enormous numbers of submissions still waiting for review. In large parts those consist of spam and incorrectly submitted sites.[33] So the average processing time for a site submission has grown longer with each passing year. However the time taken cannot be predicted, since the variation is so great: a submission might be processed within hours or take several years.[34] However, site suggestions are just one of many sources of new listings. Editors are under no obligation to check them for new listings and are actually encouraged to use other sources.[34][35]

Controversy and criticism

There have long been allegations that volunteer ODP editors give favorable treatment to their own websites while concomitantly thwarting the good faith efforts of their competition.[36] Such allegations are fielded by ODP's staff and meta editors, who have the authority to take disciplinary action against volunteer editors who are suspected of engaging in abusive editing practices.[37] In 2003, ODP introduced a new *Public Abuse Report System* that allows members of the general public to report and track allegations of abusive editor conduct using an online form.[38] Uninhibited discussion of ODP's purported shortcomings has become more common on mainstream Webmaster discussion forums. Although site policies suggest that an individual site should be submitted to only one category,[39] as of October 2007, Topix.com, a news aggregation site operated by ODP founder Rich Skrenta, has more than

17,000 listings.[40]

Early in the history of the ODP, its staff gave representatives of selected companies, such as *Rolling Stone* or CNN, editing access in order to list individual pages from their websites.[41] Links to individual CNN articles have been added until 2004 and have been entirely removed from the directory in January 2008[42] due to being outdated and not considered worth the effort to maintain. Such experiments have not been repeated later.

Ownership and management

Underlying some controversy surrounding ODP is its ownership and management. Some of the original GnuHoo volunteers felt that they had been deceived into joining a commercial enterprise.[3] To varying degrees, those complaints have continued up until the present.

At ODP's inception, there was little thought given to the idea of how ODP should be managed and there were no official forums, guidelines or FAQs. In essence, ODP began as a free for all.[43]

As time went on, the ODP Editor Forums became the *de facto* ODP parliament and when one of ODP's staff members would post an opinion in the forums, it would be considered an official ruling.[27] Even so, ODP staff began to give trusted senior editors additional editing privileges, including the ability to approve new editor applications, which eventually led to a stratified hierarchy of duties and privileges among ODP editors, with ODP's paid staff having the final say regarding ODP's policies and procedures.[28][44]

Robert Keating, a principal of Touchstone Consulting Group in Washington, D.C. since 2006, has worked as AOL's Program Manager for ODP since 2004. He started working for AOL in 1999 as Senior Editor for AOL Search, then as Managing Editor, AOL Search, ODP, and then as Media Ecosystem Manager, AOL Product Marketing.[45][46]

Editor removal procedures

ODP's editor removal procedures are overseen by ODP's staff and meta editors. According to ODP's official editorial guidelines, editors are removed for abusive editing practices or uncivil behaviour. Discussions that may result in disciplinary action against volunteer editors take place in a private forum which can only be accessed by ODP's staff and meta editors. Volunteer editors who are being discussed are not given notice that such proceedings are taking place.[44] Some people find this arrangement distasteful, wanting instead a discussion modelled more like a trial held in the U.S. judicial system.[47]

In the article *Editor Removal Explained*, ODP meta editor Arlarson states that "a great deal of confusion about the removal of editors from ODP results from false or misleading statements by former editors".[48]

The ODP's confidentiality guidelines prohibit any current ODP editors in a position to know anything from discussing the reasons for specific editor removals.[44] However, a generic list of reasons is for example given in the guidelines.[49] In the past, this has led to removed ODP editors wondering why they cannot login at ODP to perform their editing work.[50][51]

Allegations that editors are removed for criticizing policies

David F. Prenatt, Jr. (former ODP editor *netesq*) and the former editor known by the alias *The Cunctator* claim to have been removed for disagreeing with staff about changes to the policies, with special regard to the ODP's copyright policies. According to their claims, staff used the excuse that their behaviour was uncivil to remove bothersome editors.[47][52][53]

Blacklisting allegations

Senior ODP editors have the ability to attach "warning" or "do not list" notes to individual domains but no editor has the unilateral ability to block certain sites from being listed. Sites with these notes might still be listed and at times notes are removed after some discussion.[54]

Hierarchical structure

Recently criticism of ODP's hierarchical structure emerged. Many believe hierarchical directories are too complicated. As the recent emergence of Web 2.0, folksonomies began to appear. These people thought folksonomies, networks and directed graph are more "natural" and easier to manage than hierarchies.[55][56][57]

Software

Search

The ODP Search software is a derivative version of Isearch which is open source, licensed under the Mozilla Public License.[58]

Editor forums

The ODP Editor Forums were originally run on software that was based on the proprietary Ultimate Bulletin Board system. In June 2003, they switched to the open source phpBB system. As of 2007, these forums are powered by a modified version of phpBB.

Bug tracking

The bug tracking software used by the ODP is Bugzilla and the web server Apache. Squid web proxy server was also used but it was removed in August 2007 when the storage servers were reorganized. All these applications are open source.

Interface

The ODP database/editing software is closed source (although Richard Skrenta of ODP did say in June 1998 that he was considering licensing it under the GNU General Public License). This has led to criticism from the aforementioned GNU project, many of whom also criticise the ODP content license.[59]

As such, there have been some efforts to provide alternatives to ODP. These alternatives would allow communities of like-minded editors to set up and maintain their own open source/open content Web directories. However, no significant open source/open content alternative to ODP has emerged.

References

[1] http://www.dmoz.org/

[2] "Dmoz.org Site Info" (http://www.alexa.com/siteinfo/dmoz.org). Alexa Internet. . Retrieved 2012-03-02.

[3] "The GnuHoo BooBoo" (http://slashdot.org/article.pl?sid=98/06/23/0849239). *Slashdot*. . Retrieved April 27, 2007.

[4] Zurl Directory (http://web.archive.org/web/20071223080027/blog.topix.com/archives/2004_05.html), archived version of Topix.com blog entry dated May 29, 2004 by Skrenta, founder of the ODP.

[5] ODP and Yahoo Size Charts (http://www.geniac.net/odp/) by ODP editor geniac

[6] ODP reports (http://freenet-homepage.de/miscellanea/odp_reports/) by ODP volunteer administrator chris2001

[7] ODP Front Page (http://www.dmoz.org/), retrieved August 15, 2006

[8] "Dmoz'sCatastrophicServer/HardwareFailure"(http://dmozgrunt.blogspot.com/2006/10/dmozs-catastrophic-serverhardware.html) October 27, 2006, retrieved November 15, 2006.

[9] dmoz.org technical problems (http://www.resource-zone.com/forum/showthread.php?t=45325) at resource-zone.com (retrieved January 13, 2007).

[10] *The Hamsters' New Home* (http://www.miscellanea.de/newsletter/2006Winter/new_servers.html), in: Open Directory newsletter issue Winter 2006, retrieved December 26, 2006.

[11] Zeal Terms of use (http://web.archive.org/web/20020202210117/http://www.zeal.com/about/terms_of_use.jhtml) taken from Archive.org

[12] GO Network Terms of Service and Conditions of use (http://web.archive.org/web/20000510161046/info.go.com/doc/policy/terms. html) taken from Archive.org

[13] *ChefMoz Fine Dining Menu* (http://www.dmoz.org/newsletter/2003Autumn/chefmoz-column.html), in: Open Directory newsletter issue Autumn 2003

[14] About MusicMoz (http://musicmoz.org/about.html), on musicmoz.org

[15] help (http://open-site.org/help/) on open-site.org

[16] Danny Sullivan, NewHoo: Yahoo Built By The Masses (http://searchenginewatch.com/showPage.html?page=2166381), Search Engine Watch, July 1, 1998

[17] Kids and Teens Launches! (http://www.dmoz.org/newsletter/2000Nov/press.html) Open Directory Project Newsletter, November/December 2000

[18] Kids&Teens Guidelines (http://www.dmoz.org/guidelines/kguidelines/)

[19] "ODPExtension" (https://addons.mozilla.org/firefox/addon/176740) Mozilla based add-on, ODP Magic.. formerly known as ODP Extension

[20] GnuProject: Non-FreeDocumentationLicenses (http://www.gnu.org/licenses/license-list.html#NonFreeDocumentationLicenses)

[21] Open Directory RDF Dump(http://rdf.dmoz.org/)

[22] ODP/dmoz Data Dump ToDo List (http://rainwaterreptileranch.org/steve/sw/odp/rdflist.html)

[23] "Google Streamlining: Say Goodbye to the Google Directory and Labs!" (http://www.pandia.com/sew/3963-google-streamlining-say-goodbye-to-the-google-directory-and-labs.html). *Pandia Search Engine News*. 21 July 2011. . Retrieved 25 July 2011.

[24] 500,000 Vertical Search Engines (http://www.gigablast.com/prdir.html), a press release from May 12, 2005

[25] Category: Sites Using ODP Data (http://www.dmoz.org/Computers/Internet/Searching/Directories/Open_Directory_Project/ Sites_Using_ODP_Data/) on www.dmoz.org. Retrieved on September 8, 2006.

[26] Become an Editor at the Open Directory Project (http://www.dmoz.org/cgi-bin/apply.cgi)

[27] ODP Communication Guidelines (http://www.dmoz.org/guidelines/communication.html)

[28] Open Directory Project Administrator Guidelines (http://www.dmoz.org/guidelines/admin/)

[29] Open Directory Meta-editor report (http://www.dmoz.org/edoc/editall.html)

[30] Open Directory Editor list (http://www.dmoz.org/edoc/editorlist.txt)

[31] ODP Directory Editorial Guidelines (http://www.dmoz.org/guidelines/)

[32] When Giant Directories Roamed the Earth (http://www.searchlounge.org/index.php?p=40), an example showing the initial impact on Looksmart.

[33] RZ-Posting by Meta-Editor hutcheson (http://www.resource-zone.com/forum/showthread.php?p=211667#post211670)

[34] FAQ: How long until my site will be reviewed? (http://www.resource-zone.com/forum/faq.php?faq=faq_site_questions#faq_how_long) on Resource-Zone.com

[35] ODP Help Resources (http://blog.dmoz.org/2009/01/21/odp-help-resources/) in the official DMOZ Blog on 21.01.2009

[36] How To: ODP Editor Is Competitor (http://www.webmasterworld.com/forum17/94.htm) posted on Webmasterworld.com in 2000

[37] ODP Meta Guidelines: Editor Abuse and Removal (http://www.dmoz.org/guidelines/meta/abuse.html), accessed October 9, 2008.

[38] Open Directory Project: Public Abuse Report System. (http://report-abuse.dmoz.org/)

[39] How to suggest a site to the Open Directory (http://www.dmoz.org/add.html)

[40] Open Directory Project Search: "topix" (http://search.dmoz.org/cgi-bin/search?search=topix) (accessed October 18, 2007)

[41] Multiple URL's in DMOZ (http://www.webmasterworld.com/forum17/1330.htm) posted on Webmasterworld.com in 2003

[42] http://www.dmoz.org/News/ (http://web.archive.org/web/*/http://www.dmoz.org/News/), taken from Archive.org

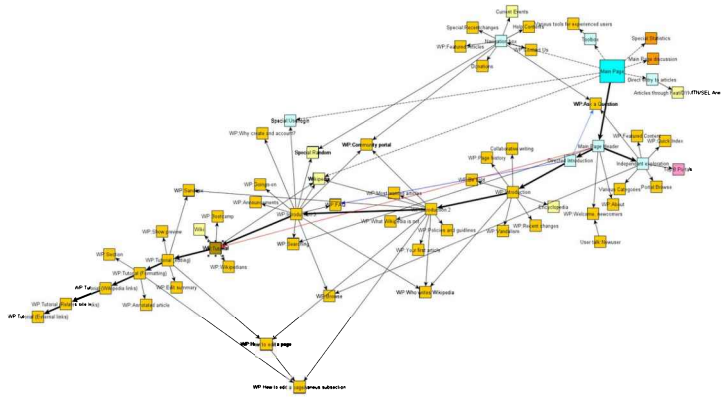
[43] Laisha, The History of the Open Directory Project (http://www.laisha.com/zine/odphistory.html) in *Search Engine Laisha's List*, June, 1999.
[44] Open Directory Project Meta Guidelines (http://www.dmoz.org/guidelines/meta/)
[45] Meet AOL's DMOZ Staff Team (http://blog.dmoz.org/2009/01/08/meet-aols-dmoz-staff-team/), DMOZ Blog, January 8, 2009
[46] Robert Keating (http://www.linkedin.com/pub/robert-keating/2/457/424),LinkedIn]
[47] XODP Yahoo! Group Message Archive (http://tech.groups.yahoo.com/group/xodp/messages/1)
[48] Arlarson, Editor Removal Explained (http://www.dmoz.org/newsletter/2000Sep/removal.html), Open Directory Project Newsletter (September 2000).
[49] Guidelines: Account Removal(http://www.dmoz.org/guidelines/accounts.html#removal)
[50] Thread: Editor account expired (http://www.resource-zone.com/forum/showthread.php?t=19936) on Resource-Zone
[51] Thread: Can't Login (http://www.resource-zone.com/forum/showthread.php?t=3655) on Resource-Zone
[52] David F. Prenatt, Jr., Life After the Open Directory Project (http://www.traffick.com/story/06-2000-xodp.asp), Traffick.com (June 1, 2000).
[53] CmdrTaco (October 24, 2000). "Dmoz (aka AOL) Changing Guidelines In Sketchy Way" (http://slashdot.org/articles/00/10/24/ 1252232.shtml). *Slashdot*. .
[54] Add Note to URL Feature (http://www.dmoz.org/urlnote.html), in ODP Documentation
[55] Hrițcu, C., (April 8, 2005). "Folksonomies vs. Ontologies" (http://hritcu.wordpress.com/2005/04/08/folksonomies-vs-ontologies/). .
[56] Shirky, C. (March 15, 2005). "Ontology is Overrated: Links, Tags and Post-hoc Metadata" (http://www.itconversations.com/shows/ detail470.html). *ITConversations*. .
[57] Hammond, T.; Hannay, T.; Lund, B. & Scott, J. (April 2005). "Social Bookmarking Tools (I)" (http://www.dlib.org/dlib/april05/ hammond/04hammond.html). *D-Lib Magazine*. .
[58] Open Directory Search Guide (http://www.dmoz.org/searchguide.html)
[59] FSF: Non-Free Documentation Licenses (http://www.fsf.org/licensing/licenses/index_html#NonFreeDocumentationLicenses): "The primary problems are that your right to redistribute any given version is not permanent and that it requires the user to keep checking back at that site, which is too restrictive of the user's freedom."

External links

- The Open Directory Project (http://www.dmoz.org/)
- Open Directory Project Blog (http://blog.dmoz.org/)
- Open Directory Project (http://www.dmoz.org/Computers/Internet/Searching/Directories/Open_Directory_Project/) at the Open Directory Project

Sitemap

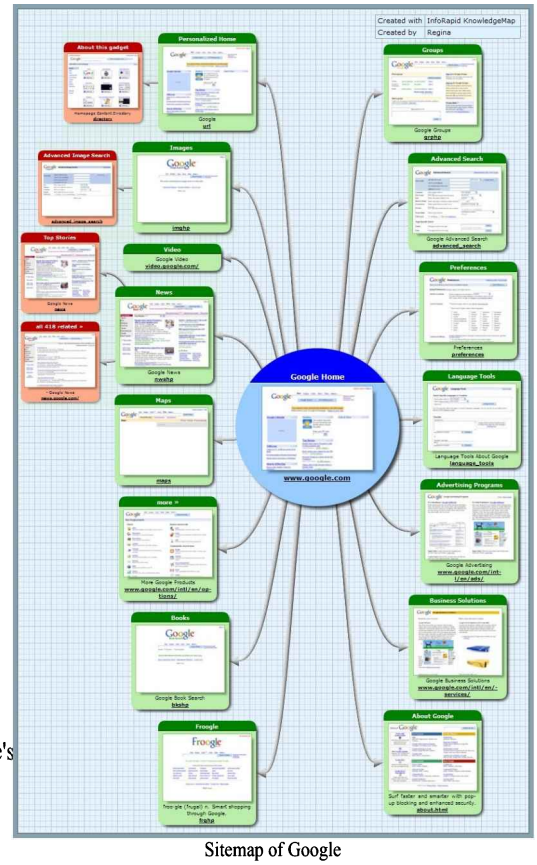
A **site map** (or **sitemap**) is a list of pages of a web site accessible to crawlers or users. It can be either a document in any form used as a planning tool for web design, or a web page that lists the pages on a web site, typically organized in hierarchical fashion. This helps visitors and search engine bots find pages on the site.



A site map of what links from the English Wikipedia's Main Page.

While some developers argue that **site index** is a more appropriately used term to relay page function, web visitors are used to seeing each term and generally associate both as one and the same. However, a site index is often used to mean an A-Z index that provides access to particular content, while a site map provides a general top-down view of the overall site contents.

XML is a document structure and encoding standard used, amongst many other things, as the standard for webcrawlers to find and parse sitemaps. There is an example of an XML sitemap below (missing link to site). The instructions to the sitemap are given to the crawler bot by a Robots Text file, an example of this is also given below. Site maps can improve search engine optimization of a site by making sure that all the pages can be found. This is especially important if a site uses a dynamic access to content such as Adobe Flash or JavaScript menus that do not include HTML links.



Sitemap of Google

They also act as a navigation aid [1] by providing an overview of a site's

Benefits of XML sitemaps to search-optimize Flash sites

Below is an example of a validated XML sitemap for a simple three page web site. Sitemaps are a useful tool for making sites built in Flash and other non-html languages searchable. Note that because the website's navigation is built with Flash (Adobe), the initial homepage of a site developed in this way would probably be found by an automated search program (ref: bot). However, the subsequent pages are unlikely to be found without an XML sitemap.

XML sitemap example:

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>http://www.example.com/?id=who</loc>
    <lastmod>2009-09-22</lastmod>
    <changefreq>monthly</changefreq>
    <priority>0.8</priority>
  </url>
  <url>
    <loc>http://www.example.com/?id=what</loc>
    <lastmod>2009-09-22</lastmod>
    <changefreq>monthly</changefreq>
    <priority>0.5</priority>
  </url>
  <url>
    <loc>http://www.example.com/?id=how</loc>
    <lastmod>2009-09-22</lastmod>
    <changefreq>monthly</changefreq>
    <priority>0.5</priority>
  </url>
</urlset>
```

XML Sitemaps

Google introduced Google Sitemaps so web developers can publish lists of links from across their sites. The basic premise is that some sites have a large number of dynamic pages that are only available through the use of forms and user entries. The Sitemap files contains URLs to these pages so that web crawlers can find them[2]. Bing, Google, Yahoo and Ask now jointly support the Sitemaps protocol.

Since Bing, Yahoo, Ask, and Google use the same protocol, having a Sitemap lets the four biggest search engines have the updated page information. Sitemaps do not guarantee all links will be crawled, and being crawled does not guarantee indexing. However, a Sitemap is still the best insurance for getting a search engine to learn about your entire site.[3]

XML Sitemaps have replaced the older method of "submitting to search engines" by filling out a form on the search engine's submission page. Now web developers submit a Sitemap directly, or wait for search engines to find it.

XML (Extensible Markup Language) is much more precise than HTML coding. Errors are not tolerated, and so syntax must be exact. It is advised to use an XML syntax validator such as the free one found at: http://validator. w3.org

There are automated XML site map generators available (both as software and web applications) for more complex sites.

More information defining the field operations and other Sitemap options are defined at http://www.sitemaps.org (Sitemaps.org: Google, Inc., Yahoo, Inc., and Microsoft Corporation)

See also Robots.txt, which can be used to identify sitemaps on the server.

References

[1] Site Map Usability (http://www.useit.com/alertbox/sitemaps.html) Jakob Nielsen's Alertbox, August 12, 2008
[2] "WordPress Plugin: Google XML Sitemaps" (http://linksku.com/10/wordpress-plugins). Linksku. .
[3] Joint announcement (http://www.google.com/press/pressrel/sitemapsorg.html) from Google, Yahoo, Bing supporting Sitemaps

External links

- CommonOfficialWebsite(http://www.sitemaps.org/)-JointlymaintainedwebsitebyGoogle,Yahoo,MSN for an XML sitemap format.
- /Sitemapgenerators(http://www.dmoz.org/Computers/Internet/Searching/Search_Engines/Sitemaps)atthe Open Directory Project
- Tools and tutorial (http://www.scriptol.com/seo/simple-map.html) Helping to build a cross-systemssitemap generator.

Robots Exclusion Standard

The **Robot Exclusion Standard**, also known as the **Robots Exclusion Protocol** or **robots.txt protocol**, is a convention to prevent cooperating web crawlers and other web robots from accessing all or part of a website which is otherwise publicly viewable. Robots are often used by search engines to categorize and archive web sites, or by webmasters to proofread source code. The standard is different from, but can be used in conjunction with, Sitemaps, a robot *inclusion* standard forwebsites.

History

The invention of "robots.txt" is attributed to Martijn Koster, when working for WebCrawler in 1994[1][2]. "robots.txt" was then popularized with the advent of AltaVista, and other popular search engines, in the following years.

About the standard

If a site owner wishes to give instructions to web robots they must place a text file called `robots.txt` in the root of the web site hierarchy (e.g. `www.example.com/robots.txt`). This text file should contain the instructions in a specific format (see examples below). Robots that *choose* to follow the instructions try to fetch this file and read the instructions before fetching any other file from the web site. If this file doesn't exist web robots assume that the web owner wishes to provide no specific instructions.

A robots.txt file on a website will function as a request that specified robots ignore specified files or directories when crawling a site. This might be, for example, out of a preference for privacy from search engine results, or the belief that the content of the selected directories might be misleading or irrelevant to the categorization of the site as a whole, or out of a desire that an application only operate on certain data. Links to pages listed in robots.txt can still appear in search results if they are linked to from a page that is crawled.[3]

For websites with multiple subdomains, each subdomain must have its own robots.txt file. If `example.com` had a robots.txt file but `a.example.com` did not, the rules that would apply for `example.com` would not apply to `a.example.com`.

Disadvantages

Despite the use of the terms "allow" and "disallow", the protocol is purely advisory. It relies on the cooperation of the web robot, so that marking an area of a site out of bounds with robots.txt does not guarantee exclusion of all web robots. In particular, malicious web robots are unlikely to honor robots.txt

There is no official standards body or RFC for the robots.txt protocol. It was created by consensus [4] in June 1994 by members of the robots mailing list (robots-request@nexor.co.uk). The information specifying the parts that should not be accessed is specified in a file called **robots.txt** in the top-level directory of the website. The robots.txt patterns are matched by simple substring comparisons, so care should be taken to make sure that patterns matching directories have the final '/' character appended, otherwise all files with names starting with that substring will match, rather than just those in the directory intended.

Examples

This example tells **all robots** to visit **all files** because the wildcard `*` specifies all robots:

```
User-agent: *
Disallow:
```

This example tells **all robots** to stay out of a website:

```
User-agent: *
Disallow: /
```

The next is an example that tells **all robots** not to enter four directories of a website:

```
User-agent: *
Disallow: /cgi-bin/
Disallow: /images/
Disallow: /tmp/
Disallow: /private/
```

Example that tells **a specific robot** not to enter one specific directory:

```
User-agent: BadBot # replace the 'BadBot' with the actual user-agent of the bot
Disallow: /private/
```

Example that tells **all robots** not to enter one specific file:

```
User-agent: *
Disallow: /directory/file.html
```

Note that all other files in the specified directory will be processed. Example

demonstrating how comments can be used:

```
# Comments appear after the "#" symbol at the start of a line, or after a directive
User-agent: * # match all bots
Disallow: / # keep them out
```

Example demonstrating how to add the parameter to tell bots where the Sitemap is located

```
User-agent: *
Sitemap: http://www.example.com/sitemap.xml  # tell the bots where your sitemap is located
```

Nonstandard extensions

Crawl-delay directive

Several major crawlers support a `Crawl-delay` parameter, set to the number of seconds to wait between successive requests to the same server:[5][6][7]

```
User-agent: *
Crawl-delay: 10
```

Allow directive

Some major crawlers support an `Allow` directive which can counteract a following `Disallow` directive.[8] [9] This is useful when one tells robots to avoid an entire directory but still wants some HTML documents in that directory crawled and indexed. While by standard implementation the first matching robots.txt pattern always wins, Google's implementation differs in that Allow patterns with equal or more characters in the directive path win over a matching Disallow pattern.[10] Bing uses the `Allow` or `Disallow` directive which is the most specific.[11]

In order to be compatible to all robots, if one wants to allow single files inside an otherwise disallowed directory, it is necessary to place the Allow directive(s) first, followed by the Disallow, for example:

```
Allow: /folder1/myfile.html
Disallow: /folder1/
```

This example will Disallow anything in /folder1/ except /folder1/myfile.html, since the latter will match first. In case of Google, though, the order is not important.

Sitemap

Some crawlers support a `Sitemap` directive, allowing multiple Sitemaps in the same robots.txt in the form:[12]

```
Sitemap: http://www.gstatic.com/s2/sitemaps/profiles-sitemap.xml
Sitemap: http://www.google.com/hostednews/sitemap_index.xml
```

Universal "*" match

the *Robot Exclusion Standard* does not mention anything about the "*" character in the `Disallow:` statement. Some crawlers like Googlebot and Slurp recognize strings containing "*", while MSNbot and Teoma interpret it in different ways.[13]

References

[1] http://www.robotstxt.org/orig.html#status

[2] http://www.robotstxt.org/norobots-rfc.txt

[3] http://www.youtube.com/watch?v=KBdEwpRQRD0#t=196s

[4] http://www.robotstxt.org/wc/norobots.html

[5] "How can I reduce the number of requests you make on my web site?" (http://help.yahoo.com/l/us/yahoo/search/webcrawler/slurp-03. html). *Yahoo! Slurp*. . Retrieved 2007-03-31.

[6] "MSNBot is crawling a site too frequently" (http://search.msn.com/docs/siteowner. aspx?t=SEARCH_WEBMASTER_FAQ_MSNBotIndexing.htm&FORM=WFDD#D). *Troubleshoot issues with MSNBot and site crawling*. . Retrieved 2007-02-08.

[7] "About Ask.com: Webmasters" (http://about.ask.com/en/docs/about/webmasters.shtml#15). .

[8] "Webmaster Help Center - How do I block Googlebot?" (http://www.google.com/support/webmasters/bin/answer.py?hl=en& answer=156449&from=40364). . Retrieved 2007-11-20.

[9] "How do I prevent my site or certain subdirectories from being crawled? - Yahoo Search Help" (http://help.yahoo.com/l/us/yahoo/ search/webcrawler/slurp-02.html). . Retrieved 2007-11-20.

[10] "Google's Hidden Interpretation of Robots.txt" (http://blog.semetrical.com/googles-secret-approach-to-robots-txt/). . Retrieved 2010-11-15.
[11] "Robots Exclusion Protocol - joining together to provide better documentation" (http://www.bing.com/community/site_blogs/b/webmaster/archive/2008/06/03/robots-exclusion-protocol-joining-together-to-provide-better-documentation.aspx). . Retrieved 2009-12-03.
[12] "Yahoo! Search Blog - Webmasters can now auto-discover with Sitemaps" (http://ysearchblog.com/2007/04/11/ webmasters-can-now-auto-discover-with-sitemaps/). . Retrieved 2009-03-23.
[13] "Search engines and dynamic content issues" (http://ghita.org/search-engines-dynamic-content-issues.html). *MSNbot issues with robots.txt*. . Retrieved 2007-04-01.

External links

- www.robotstxt.org - The Web Robots Pages (http://www.robotstxt.org/)
- History of robots.txt (http://www.antipope.org/charlie/blog-static/2009/06/ how_i_got_here_in_the_end_part_3.html) - (how Charles Stross prompted its invention; original comment (http://yro.slashdot.org/comments.pl?sid=377285&cid=21554125) on Slashdot)
- Block or remove pages using a robots.txt file - Google Webmaster Tools Help=Using the robots.txt analysis tool (http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=156449)
- About Robots.txt at the Mediawiki website (http://www.mediawiki.org/wiki/Robots.txt)
- List of Bad Bots (http://www.kloth.net/internet/badbots.php) - rogue robots and spiders which ignore these guidelines
- Wikipedia's Robots.txt - an example (http://en.wikipedia.org/robots.txt)
- Robots.txt Generator + Tutorial (http://www.mcanerin.com/EN/search-engine/robots-txt.asp)
- Robots.txt Generator Tool (http://www.howrank.com/Robots.txt-Tool.php)
- Robots.txt is not a security measure (http://www.diovo.com/2008/09/robotstxt-is-not-a-security-measure/)

Robots.txt

The **Robot Exclusion Standard**, also known as the **Robots Exclusion Protocol** or **robots.txt protocol**, is a convention to prevent cooperating web crawlers and other web robots from accessing all or part of a website which is otherwise publicly viewable. Robots are often used by search engines to categorize and archive web sites, or by webmasters to proofread source code. The standard is different from, but can be used in conjunction with, Sitemaps, a robot *inclusion* standard for websites.

History

The invention of "robots.txt" is attributed to Martijn Koster, when working for WebCrawler in 1994[1][2]. "robots.txt" was then popularized with the advent of AltaVista, and other popular search engines, in the following years.

About the standard

If a site owner wishes to give instructions to web robots they must place a text file called `robots.txt` in the root of the web site hierarchy (e.g. `www.example.com/robots.txt`). This text file should contain the instructions in a specific format (see examples below). Robots that *choose* to follow the instructions try to fetch this file and read the instructions before fetching any other file from the web site. If this file doesn't exist web robots assume that the web owner wishes to provide no specific instructions.

A robots.txt file on a website will function as a request that specified robots ignore specified files or directories when crawling a site. This might be, for example, out of a preference for privacy from search engine results, or the belief that the content of the selected directories might be misleading or irrelevant to the categorization of the site as a

whole, or out of a desire that an application only operate on certain data. Links to pages listed in robots.txt can still appear in search results if they are linked to from a page that is crawled.[3]

For websites with multiple subdomains, each subdomain must have its own robots.txt file. If `example.com` had a robots.txt file but `a.example.com` did not, the rules that would apply for `example.com` would not apply to `a.example.com`.

Disadvantages

Despite the use of the terms "allow" and "disallow", the protocol is purely advisory. It relies on the cooperation of the web robot, so that marking an area of a site out of bounds with robots.txt does not guarantee exclusion of all web robots. In particular, malicious web robots are unlikely to honor robots.txt

There is no official standards body or RFC for the robots.txt protocol. It was created by consensus [4] in June 1994 by members of the robots mailing list (robots-request@nexor.co.uk). The information specifying the parts that should not be accessed is specified in a file called **robots.txt** in the top-level directory of the website. The robots.txt patterns are matched by simple substring comparisons, so care should be taken to make sure that patterns matching directories have the final '/' character appended, otherwise all files with names starting with that substring will match, rather than just those in the directory intended.

Examples

This example tells **all robots** to visit **all files** because the wildcard `*` specifies all robots:

```
User-agent: *
Disallow:
```

This example tells **all robots** to stay out of a website:

```
User-agent: *
Disallow: /
```

The next is an example that tells **all robots** not to enter four directories of a website:

```
User-agent: *
Disallow: /cgi-bin/
Disallow: /images/
Disallow: /tmp/
Disallow: /private/
```

Example that tells **a specific robot** not to enter one specific directory:

```
User-agent: BadBot # replace the 'BadBot' with the actual user-agent of the bot
Disallow: /private/
```

Example that tells **all robots** not to enter one specific file:

```
User-agent: *
Disallow: /directory/file.html
```

Note that all other files in the specified directory will be processed. Example

demonstrating how comments can be used:

```
# Comments appear after the "#" symbol at the start of a line, or after a directive
User-agent: * # match all bots
```

```
Disallow: / # keep them out
```

Example demonstrating how to add the parameter to tell bots where the Sitemap is located

```
User-agent: *
Sitemap: http://www.example.com/sitemap.xml  # tell the bots where your sitemap is located
```

Nonstandard extensions

Crawl-delay directive

Several major crawlers support a `Crawl-delay` parameter, set to the number of seconds to wait between successive requests to the same server:[4][5][6]

```
User-agent: *
Crawl-delay: 10
```

Allow directive

Some major crawlers support an `Allow` directive which can counteract a following `Disallow` directive.[7] [8] This is useful when one tells robots to avoid an entire directory but still wants some HTML documents in that directory crawled and indexed. While by standard implementation the first matching robots.txt pattern always wins, Google's implementation differs in that Allow patterns with equal or more characters in the directive path win over a matching Disallow pattern.[9] Bing uses the `Allow` or `Disallow` directive which is the most specific.[10]

In order to be compatible to all robots, if one wants to allow single files inside an otherwise disallowed directory, it is necessary to place the Allow directive(s) first, followed by the Disallow, for example:

```
Allow: /folder1/myfile.html
Disallow: /folder1/
```

This example will Disallow anything in /folder1/ except /folder1/myfile.html, since the latter will match first. In case of Google, though, the order is not important.

Sitemap

Some crawlers support a `Sitemap` directive, allowing multiple Sitemaps in the same robots.txt in the form:[11]

```
Sitemap: http://www.gstatic.com/s2/sitemaps/profiles-sitemap.xml
Sitemap: http://www.google.com/hostednews/sitemap_index.xml
```

Universal "*" match

the *Robot Exclusion Standard* does not mention anything about the "*" character in the `Disallow:` statement. Some crawlers like Googlebot and Slurp recognize strings containing "*", while MSNbot and Teoma interpret it in different ways.[12]

References

[1] http://www.robotstxt.org/orig.html#status
[2] http://www.robotstxt.org/norobots-rfc.txt
[3] http://www.youtube.com/watch?v=KBdEwpRQRD0#t=196s
[4] "How can I reduce the number of requests you make on my web site?" (http://help.yahoo.com/l/us/yahoo/search/webcrawler/slurp-03. html). *Yahoo! Slurp*. . Retrieved2007-03-31.
[5] "MSNBot is crawling a site too frequently" (http://search.msn.com/docs/siteowner. aspx?t=SEARCH_WEBMASTER_FAQ_MSNBotIndexing.htm&FORM=WFDD#D). *Troubleshoot issues with MSNBot and site crawling*. . Retrieved 2007-02-08.
[6] "About Ask.com: Webmasters" (http://about.ask.com/en/docs/about/webmasters.shtml#15). .
[7] "Webmaster Help Center - How do I block Googlebot?" (http://www.google.com/support/webmasters/bin/answer.py?hl=en& answer=156449&from=40364). . Retrieved2007-11-20.
[8] "How do I prevent my site or certain subdirectories from being crawled? - Yahoo Search Help"(http://help.yahoo.com/l/us/yahoo/ search/webcrawler/slurp-02.html). . Retrieved2007-11-20.
[9] "Google's Hidden Interpretation of Robots.txt" (http://blog.semetrical.com/googles-secret-approach-to-robots-txt/). . Retrieved 2010-11-15.
[10] "Robots Exclusion Protocol - joining together to provide better documentation" (http://www.bing.com/community/site_blogs/b/ webmaster/archive/2008/06/03/robots-exclusion-protocol-joining-together-to-provide-better-documentation.aspx). . Retrieved 2009-12-03.
[11] "Yahoo! Search Blog - Webmasters can now auto-discover with Sitemaps" (http://ysearchblog.com/2007/04/11/ webmasters-can-now-auto-discover-with-sitemaps/). . Retrieved 2009-03-23.
[12] "Search engines and dynamic content issues" (http://ghita.org/search-engines-dynamic-content-issues.html). *MSNbot issues with robots.txt*. . Retrieved2007-04-01.

External links

- www.robotstxt.org - The Web Robots Pages (http://www.robotstxt.org/)
- History of robots.txt (http://www.antipope.org/charlie/blog-static/2009/06/ how_i_got_here_in_the_end_part_3.html) - (how Charles Stross prompted its invention; original comment (http:/ /yro.slashdot.org/comments.pl?sid=377285&cid=21554125) on Slashdot)
- Blockorremovepagesusingarobots.txtfile-GoogleWebmasterToolsHelp=Usingtherobots.txtanalysistool (http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=156449)
- About Robots.txt at the Mediawiki website (http://www.mediawiki.org/wiki/Robots.txt)
- List of Bad Bots (http://www.kloth.net/internet/badbots.php) - rogue robots and spiders which ignore these guidelines
- Wikipedia's Robots.txt - an example (http://en.wikipedia.org/robots.txt)
- Robots.txt Generator + Tutorial (http://www.mcanerin.com/EN/search-engine/robots-txt.asp)
- Robots.txt Generator Tool (http://www.howrank.com/Robots.txt-Tool.php)
- Robots.txtisnotasecuritymeasure(http://www.diovo.com/2008/09/robotstxt-is-not-a-security-measure/)

301 redirect

URL redirection, also called **URL forwarding**, is a World Wide Web technique for making a web page available under more than one URL address. When a web browser attempts to open a URL that has been redirected, a page with a different URL is opened. For example, www.example.com [1] is redirected to www.iana.org/domains/example/ [2]. Similarly, **Domain redirection** or **domain forwarding** is when all pages in a URL domain are redirected to a different domain, as when wikipedia.com [3] and wikipedia.net [4] are automatically redirected to wikipedia.org [5]. URL redirection can be used for URL shortening, to prevent broken links when web pages are moved, to allow multiple domain names belonging to the same owner to refer to a single web site, to guide navigation into and out of a website, for privacy protection, and for less innocuous purposes such as phishing attacks using URLs that are similar to a targeted web site.

Purposes

There are several reasons to use URL redirection:

Similar domain names

A user might mis-type a URL—for example, "example.com" and "exmaple.com". Organizations often register these "mis-spelled" domains and re-direct them to the "correct" location: example.com. The addresses example.com and example.net could both redirect to a single domain, or web page, such as example.org. This technique is often used to "reserve" other top-level domains (TLD) with the same name, or make it easier for a true ".edu" or ".net" to redirect to a more recognizable ".com" domain.

Moving a site to a new domain

A web page may be redirected for several reasons:

- a web site might need to change its domain name;
- an author might move his or her pages to a new domain;
- two web sites might merge.

With URL redirects, incoming links to an outdated URL can be sent to the correct location. These links might be from other sites that have not realized that there is a change or from bookmarks/favorites that users have saved in their browsers.

The same applies to search engines. They often have the older/outdated domain names and links in their database and will send search users to these old URLs. By using a "moved permanently" redirect to the new URL, visitors will still end up at the correct page. Also, in the next search engine pass, the search engine should detect and use the newer URL.

Logging outgoing links

The access logs of most web servers keep detailed information about where visitors came from and how they browsed the hosted site. They do not, however, log which links visitors left by. This is because the visitor's browser has no need to communicate with the original server when the visitor clicks on an outgoing link.

This information can be captured in several ways. One way involves URL redirection. Instead of sending the visitor straight to the other site, links on the site can direct to a URL on the original website's domain that automatically redirects to the real target. This technique bears the downside of the delay caused by the additional request to the original website's server. As this added request will leave a trace in the server log, revealing exactly which link was followed, it can also be a privacy issue.[6]

The same technique is also used by some corporate websites to implement a statement that the subsequent content is at another site, and therefore not necessarily affiliated with the corporation. In such scenarios, displaying the warning causes an additional delay.

Short aliases for long URLs

Web applications often include lengthy descriptive attributes in their URLs which represent data hierarchies, command structures, transaction paths and session information. This practice results in a URL that is aesthetically unpleasant and difficult to remember, and which may not fit within the size limitations of microblogging sites. URL shortening services provide a solution to this problem by redirecting a user to a longer URL from a shorter one..

Meaningful, persistent aliases for long or changing URLs

Sometimes the URL of a page changes even though the content stays the same. Therefore URL redirection can help users who have bookmarks. This is routinely done on Wikipedia whenever a page is renamed.

Manipulating search engines

Some years ago, redirect techniques were used to fool search engines. For example, one page could show popular search terms to search engines but redirect the visitors to a different target page. There are also cases where redirects have been used to "steal" the page rank of one popular page and use it for a different page, usually involving the 302 HTTP status code of "moved temporarily."[7][8]

Search engine providers noticed the problem and took appropriate actions . Usually, sites that employ such techniques to manipulate search engines are punished automatically by reducing their ranking or by excluding them from the search index.

As a result, today, such manipulations usually result in less rather than more site exposure.

Satire and criticism

In the same way that a Google bomb can be used for satire and political criticism, a domain name that conveys one meaning can be redirected to any other web page, sometimes with malicious intent. The website `shadyurl.com` offers a satirical service that will create an apparently "suspicious and frightening" redirection URL for even benign webpages. For example, an input of `en.wikipedia.org` generates `5z8.info/hookers_e4u5_inject_worm`.

Manipulating visitors

URL redirection is sometimes used as a part of phishing attacks that confuse visitors about which web site they are visiting . Because modern browsers always show the real URL in the address bar, the threat is lessened. However, redirects can also take you to sites that will otherwise attempt to attack in other ways. For example, a redirect might take a user to a site that would attempt to trick them into downloading antivirus software and ironically installing a trojan of some sort instead.

Removing `referer` information

When a link is clicked, the browser sends along in the HTTP request a field called referer which indicates the source of the link. This field is populated with the URL of the current web page, and will end up in the logs of the server serving the external link. Since sensitive pages may have sensitive URLs (for example, `http://company.com/plans-for-the-next-release-of-our-product`), it is not desirable for the `referer` URL to leave the organization. A redirection page that performs referrer hiding could be embedded in all external URLs, transforming for example `http://externalsite.com/page` into

`http://redirect.company.com/http://externalsite.com/page`. This technique also eliminates other potentially sensitive information from the referer URL, such as the session ID, and can reduce the chance of phishing by indicating to the end user that they passed a clear gateway to another site.

Techniques

There are several techniques to implement a redirect. In many cases, Refresh meta tag is the simplest one. However, there exist several strong opinions discouraging this method.[9]

Manual redirect

The simplest technique is to ask the visitor to follow a link to the new page, usually using an HTML anchor as such:

```
Please follow <a href="http://www.example.com/">this link</a>.
```

This method is often used as a fall-back for automatic methods — if the visitor's browser does not support the automatic redirect method, the visitor can still reach the target document by following the link.

HTTP status codes 3xx

In the HTTP protocol used by the World Wide Web, a **redirect** is a response with a status code beginning with *3* that induces a browser to go to another location, with annotation describing the reason, which allows for the correct subsequent action (such as changing links in the case of code 301, a permanent change of address)

The HTTP standard [10] defines several status codes [11] for redirection:

- 300 multiple choices (e.g. offer different languages)
- 301 moved permanently
- 302 found (originally temporary redirect, but now commonly used to specify redirection for unspecified reason)
- 303 see other (e.g. for results of cgi-scripts)
- 307 temporary redirect

All of these status codes require that the URL of the redirect target be given in the Location: header of the HTTP response. The 300 multiple choices will usually list all choices in the body of the message and show the default choice in the Location: header.

Within the 3xx range, there are also some status codes that are quite different from the above redirects (they are not discussed here with their details):

- 304 not modified
- 305 use proxy

This is a sample of an HTTP response that uses the 301 "moved permanently" redirect:

```
HTTP/1.1 301 Moved Permanently

Location: http://www.example.org/

Content-Type: text/html

Content-Length: 174

<html>

<head>
```

```
<title>Moved</title>

</head>

<body>

<h1>Moved</h1>

<p>This page has moved to <a href="http://www.example.org/">http://www.example.org/</a>.</p>

</body>

</html>
```

Using server-side scripting for redirection

Often, web authors don't have sufficient permissions to produce these status codes: The HTTP header is generated by the web server program and not read from the file for that URL. Even for CGI scripts, the web server usually generates the status code automatically and allows custom headers to be added by the script. To produce HTTP status codes with cgi-scripts, one needs to enable non-parsed-headers.

Sometimes, it is sufficient to print the "Location: 'url'" header line from a normal CGI script. Many web servers choose one of the 3xx status codes for such replies.

Frameworks for server-side content generation typically require that HTTP headers be generated before response data. As a result, the web programmer who is using such a scripting language to redirect the user's browser to another page must ensure that the redirect is the first or only part of the response. In the ASP scripting language, this can also be accomplished using the methods `response.buffer=true` and `response.redirect "http://www.example.com/"`. Using PHP, one can use the header function as follows:

```
header('HTTP/1.1 301 Moved Permanently');

header('Location: http://www.example.com/');

exit();
```

According to the HTTP protocol, the Location header must contain an absolute URI.[12] When redirecting from one page to another within the same site, it is a common mistake to use a relative URI. As a result most browsers tolerate relative URIs in the Location header, but some browsers display a warning to the end user.

There are other methods that can be used for performing redirects, but they do not offer the flexibility that mod_rewrite offers. These alternative rules use functions within mod_alias:

```
Redirect permanent /oldpage.html http://www.example.com/newpage.html

Redirect 301 /oldpage.html http://www.example.com/newpage.html
```

To redirect a requests for any non-canonical domain name using .htaccess or within a <Directory> section in an Apache config file:

```
RewriteEngine on

RewriteCond %{HTTP_HOST}
^([^.:]+\.)*oldsite\.example\.com\.?(:[0-9]*)?$ [NC]

RewriteRule ^(.*)$ http://newsite.example.net/$1 [R=301,L]
```

Use of .htaccess for this purpose usually does not require administrative permissions. However, .htaccess can be disabled by your host, and so may not work (or continue to work) if they do so.

In addition, some server configurations may require the addition of the line:

```
Options +FollowSymLinks
```

ahead of the "RewriteEngine on" directive, in order to enable the mod_rewrite module.

When you have access to the main Apache config files (such as httpd.conf), it is best to avoid the use of .htaccess files.

If the code is placed into an Apache config file and not within any <Directory> container, then the RewriteRule pattern must be changed to include a leading slash:

```
RewriteEngine on

RewriteCond %{HTTP_HOST} ^([^.:]+\.)*oldwebsite\.com\.?(:[0-9]*)?$ [NC]

RewriteRule ^/(.*)$ http://www.preferredwebsite.net/$1 [R=301,L]
```

Refresh Meta tag and HTTP refresh header

Netscape introduced a feature to refresh the displayed page after a certain amount of time. This method is often called meta refresh. It is possible to specify the URL of the new page, thus replacing one page after some time by another page:

- HTML <meta> tag[13]
- An exploration of dynamic documents [14]
- Meta refresh

A timeout of 0 seconds means an immediate redirect. Meta Refresh with a timeout of 0 seconds is accepted as a 301 permanent redirect by Google, allowing to transfer PageRank from static html files.[15]

This is an example of a simple HTML document that uses this technique:

```
<html>

<head>

<meta http-equiv="Refresh" content="0; url=http://www.example.com/" />

</head>

<body>

<p>Please follow <a href="http://www.example.com/">this link</a>.</p>
```

```
</body>

</html>
```

- This technique is usable by all web authors because the meta tag is contained inside the document itself.
- The meta tag must be placed in the "head" section of the HTML file.
- The number "0" in this example may be replaced by another number to achieve a delay of that many seconds.
- This is a proprietary extension to HTML introduced by Netscape but supported by most web browsers. The manual link in the "body" section is for users whose browsers do not support this feature.

This is an example of achieving the same effect by issuing an HTTP refresh header:

```
HTTP/1.1 200 ok

Refresh: 0; url=http://www.example.com/

Content-type: text/html

Content-length: 78

Please follow <a href="http://www.example.com/">this link</a>!
```

This response is easier to generate by CGI programs because one does not need to change the default status code. Here is a simple CGI program that effects this redirect:

```
#!/usr/bin/perl

print "Refresh: 0; url=http://www.example.com/\r\n";

print "Content-type: text/html\r\n";

print "\r\n";

print "Please follow <a href=\"http://www.example.com/\">this link</a>!"
```

Note: Usually, the HTTP server adds the status line and the Content-length header automatically.

This method is considered by the W3C to be a poor method of redirection, since it does not communicate any information about either the original or new resource, to the browser (or search engine). The W3C's Web Content Accessibility Guidelines (7.4) [16] discourage the creation of auto-refreshing pages, since most web browsers do not allow the user to disable or control the refresh rate. Some articles that they have written on the issue include W3C Web Content Accessibility Guidelines (1.0): Ensure user control of time-sensitive content changes [17] and Use standard redirects: don't break the back button! [18]

This example works best for a refresh, or in simple terms - a redirect for webpages, as follows, however, for a refresh under 4 seconds, your webpage will not be given priority listing on search engines. For some users, this is preferred not to be listed. Inline, you will find the time as in seconds:

CONTENT="2

this number can be adjusted to suit your needs. Place in

your head:

```
<HTML>

<HEAD>

   <META HTTP-EQUIV="refresh" CONTENT="2;URL=http://www.example.com/example.html">

</HEAD>
```

JavaScript redirects

JavaScript offers several ways to display a different page in the current browser window. Quite frequently, they are used for a redirect. However, there are several reasons to prefer HTTP header or the refresh meta tag (whenever it is possible) over JavaScript redirects:

- Security considerations
- Some browsers don't support JavaScript
- many web crawlers don't execute JavaScript.

Frame redirects

A slightly different effect can be achieved by creating a single HTML frame that contains the target page:

```
<frameset rows="100%">

  <frame src="http://www.example.com/">

</frameset>

<noframes>

  <body>Please follow <a href="http://www.example.com/">link</a>!</body>

</noframes>
```

One main difference to the above redirect methods is that for a frame redirect, the browser displays the URL of the frame document and not the URL of the target page in the URL bar.

This technique is commonly called *cloaking*. This may be used so that the reader sees a more memorable URL or, with fraudulent intentions, to conceal a phishing site as part of website spoofing.[19]

Redirect loops

It is quite possible that one redirect leads to another redirect. For example, the URL http://www.wikipedia.com/ wiki/URL_redirection (note the differences in the domain name) is first redirected to http://www.wikipedia.org/ wiki/URL_redirection and again redirected to the correct URL: http://en.wikipedia.org/wiki/URL_redirection. This is appropriate: the first redirection corrects the wrong domain name, the second redirection selects the correct language section, and finally, the browser displays the correct page.

Sometimes, however, a mistake can cause the redirection to point back to the first page, leading to an infinite loop of redirects. Browsers usually break that loop after a few steps and display an error message instead.

The HTTP standard [11] states:

> A client SHOULD detect infinite redirection loops, since such loops generate network traffic for each redirection.
>
> Previous versions of this specification recommended a maximum of five redirections; some clients may exist that implement such a fixed limitation.

Services

There exist services that can perform URL redirection on demand, with no need for technical work or access to the webserver your site is hosted on.

URL redirection services

A **redirect service** is an information management system, which provides an internet link that redirects users to the desired content. The typical benefit to the user is the use of a memorable domain name, and a reduction in the length of the URL or web address. A redirecting link can also be used as a permanent address for content that frequently changes hosts, similarly to the Domain Name System.

Hyperlinks involving URL redirection services are frequently used in spam messages directed at blogs and wikis. Thus, one way to reduce spam is to reject all edits and comments containing hyperlinks to known URL redirection services; however, this will also remove legitimate edits and comments and may not be an effective method to reduce spam.

Recently, URL redirection services have taken to using AJAX as an efficient, user friendly method for creating shortened URLs.

A major drawback of some URL redirection services is the use of delay pages, or frame based advertising, to generate revenue.

History

The first redirect services took advantage of top-level domains (TLD) such as ".to" (Tonga), ".at" (Austria) and ".is" (Iceland). Their goal was to make memorable URLs. The first mainstream redirect service was V3.com that boasted 4 million users at its peak in 2000. V3.com success was attributed to having a wide variety of short memorable domains including "r.im", "go.to", "i.am", "come.to" and "start.at". V3.com was acquired by FortuneCity.com, a large free web hosting company, in early 1999. In 2001 emerged .tk (Tokelau) as a TLD used for memorable names.[20] As the sales price of top level domains started falling from $70.00 per year to less than $10.00, the demand for memorable redirection services eroded.

With the launch of TinyURL in 2002 a new kind of redirecting service was born, namely URL shortening. Their goal was to make long URLs short, to be able to post them on internet forums. Since 2006, with the 140 character limit on the extremely popular Twitter service, these short URL services have seen a resurgence.

Referrer Masking

Redirection services can hide the referrer by placing an intermediate page between the page the link is on and its destination. Although these are conceptually similar to other URL redirection services, they serve a different purpose, and they rarely attempt to shorten or obfuscate the destination URL (as their only intended side-effect is to hide referrer information and provide a clear gateway between other websites.)

This type of redirection is often used to prevent potentially-malicious links from gaining information using the referrer, for example a session ID in the query string. Many large community websites use link redirection on external links to lessen the chance of an exploit that could be used to steal account information, as well as make it clear when a user is leaving a service, to lessen the chance of effective phishing .

Here is a simplistic example of such a service, written in PHP.

```
<?php

$url = htmlspecialchars($_GET['url']);

header( 'Refresh: 0; url=http://'.$url );

?>

<!-- Fallback using meta refresh. -->

<html>

 <head>

  <title>Redirecting...</title>

  <meta http-equiv="refresh" content="0;url=http://<?php echo $url; ?>">

 </head>

 <body>

 Attempting to redirect to <a href="http://<?php echo $url; ?>">http://<?php echo $url; ?></a>.

 </body>

</html>
```

References

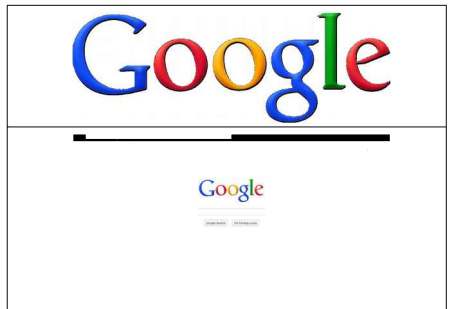
[1] http://www.example.com
[2] http://www.iana.org/domains/example/
[3] http://www.wikipedia.com
[4] http://www.wikipedia.net
[5] http://www.wikipedia.org
[6] "Google revives redirect snoopery" (http://blog.anta.net/2009/01/29/509/). *blog.anta.net*. 2009-01-29. ISSN 1797-1993. . Retrieved 2009-01-30.
[7] Google's serious hijack problem (http://www.pandia.com/sw-2004/40-hijack.html)
[8] Stop 302 Redirects and Scrapers from Hijacking Web Page PR - Page Rank (http://www.loriswebs.com/hijacking_web_pages.html)
[9] Core Techniques for Web Content Accessibility Guidelines 1.0 section 7 (http://www.w3.org/TR/WCAG10-CORE-TECHS/ #auto-page-refresh), w3.org, published 2000-11-6, fetched 2009-12-15.
[10] http://www.w3.org/Protocols/rfc2616/rfc2616.html
[11] http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html#sec10.3
[12] R. Fielding, et al., Request for Comments: 2616, Hypertext Transfer Protocol — HTTP/1.1, published 1999-07, §14.30 "Location" (http://www.w3.org/Protocols/rfc2616/rfc2616-sec14.html#sec14.30), fetched 2008-10-07
[13] http://www.w3schools.com/tags/tag_meta.asp
[14] http://web.archive.org/web/20020802170847/http://wp.netscape.com/assist/net_sites/pushpull.html
[15] Google and Yahoo accept undelayed meta refreshs as 301 redirects, 3 September 2007, http://sebastians-pamphlets.com/ google-and-yahoo-treat-undelayed-meta-refresh-as-301-redirect/
[16] http://www.w3.org/TR/WAI-WEBCONTENT/#tech-no-periodic-refresh
[17] http://www.w3.org/TR/WAI-WEBCONTENT/#gl-movement
[18] http://www.w3.org/QA/Tips/reback
[19] Anti-Phishing Technology" (http://www.sfbay-infragard.org/Documents/phishing-sfectf-report.pdf), Aaron Emigh, Radix Labs, 19 January 2005
[20] "Net gains for tiny Pacific nation" (http://news.bbc.co.uk/2/hi/technology/6991719.stm). *BBC News*. 2007-09-14. . Retrieved 2010-05-27.

External links

- Mapping URLs to Filesystem Locations (http://httpd.apache.org/docs/1.3/urlmapping.html)
- Paper on redirection spam (UC Davis) (http://www.cs.ucdavis.edu/~hchen/paper/www07.pdf)
- Servlet redirect example (http://www.jsptube.com/examples/response-sendredirect-servlet.html)
- Servlet forward example (http://www.jsptube.com/examples/servlet-forward.html)
- Security vulnerabilities in URL Redirectors (http://projects.webappsec.org/URL-Redirector-Abuse) The Web Application Security Consortium Threat Classification
- 301 Redirects for moved pages using .htaccess (http://www.dancatts.com/articles/ htaccess-301-redirects-for-moved-pages.php)
- 301-redirect.info, site summarizing redirection methods in Apache, PHP, ASP, JPs or ColdFusion programming (http://www.301-redirect.info/)
- Redirecting your visitors to your preferred domain (http://911-need-code-help.blogspot.com/2011/03/ redirecting-visitors-to-preferred.html) using 301 permanent redirects — rationale and mod_rewrite/PHP/ASP.NET implementations

Google Instant

Google Instant



Google Search homepage

URL	Google.com [1]
Commercial?	Yes
Type of site	Web search engine
Registration	Optional
Available language(s)	Multilingual (124)
Owner	Google
Created by	SergeyBrinandLarryPage
Launched	September 15, 1997[2]
Alexa rank	1 (March 2012)[3]
Revenue	From AdWords
Current status	Active

Google Search (or **Google Web Search**) is a web search engine owned by Google Inc. Google Search is the most-used search engine on the World Wide Web,[4] receiving several hundred million queries each day through its various services.[5]

The order of search results on Google's search-results pages is based, in part, on a priority rank called a "PageRank". Google Search provides many options for customized search, using Boolean operators such as: implied "AND" (if several concatenated search terms separated by spaces are given, only pages containing all of them should be returned), exclusion ("-xx"), alternatives ("xx OR yy"), and wildcard ("x * x").[6]

The main purpose of Google Search is to hunt for text in Web pages, as opposed to other data, such as with Google Image Search. Google Search was originally developed by Larry Page and Sergey Brin in 1997.[7] Google Search provides at least 22 special features beyond the original word-search capability.[8] These include synonyms, weather forecasts, time zones, stock quotes, maps, earthquake data, movie showtimes, airports, home listings, and sports scores. There are special features for numbers, including ranges (70..73),[9] prices, temperatures, money/unit conversions ("10.5 cm in inches"), calculations ("3*4+sqrt(6)-pi/2"), package tracking, patents, area codes,[8] and language translation of displayed pages. In June 2011, Google introduced "Google Voice Search" and "Search by Image" features for allowing the users to search words by speaking and by giving images.[10]

The frequency of use of many search terms have reached a volume that they may indicate broader economic, social and health trends.[11] Data about the frequency of use of search terms on Google (available through Google Adwords, Google Trends, and Google Insights for Search) have been shown to correlate with flu outbreaks and

unemployment levels and provide the information faster than traditional reporting methods and government surveys.

Search engine

PageRank

Google's rise to success was in large part due to a patented algorithm called PageRank that helps rank web pages that match a given search string.[12] When Google was a Stanford research project, it was nicknamed BackRub because the technology checks backlinks to determine a site's importance. Previous keyword-based methods of ranking search results, used by many search engines that were once more popular than Google, would rank pages by how often the search terms occurred in the page, or how strongly associated the search terms were within each resulting page. The PageRank algorithm instead analyzes human-generated links assuming that web pages linked from many important pages are themselves likely to be important. The algorithm computes a recursive score for pages, based on the weighted sum of the PageRanks of the pages linking to them. PageRank is thought to correlate well with human concepts of importance. In addition to PageRank, Google, over the years, has added many other secret criteria for determining the ranking of pages on result lists, reported to be over 200 different indicators.[13] The specifics of which are kept secret to keep spammers at bay and help Google maintain an edge over its competitors globally.

Search results

The exact percentage of the total of web pages that Google indexes are not known, as it is very difficult to accurately calculate. Google not only indexes and caches web pages, but also takes "snapshots" of other file types, which include PDF, Word documents, Excel spreadsheets, Flash SWF, plain text files, and so on.[14] Except in the case of text and SWF files, the cached version is a conversion to (X)HTML, allowing those without the corresponding viewer application to read the file. Users can customize the search engine, by setting a default language, using the "SafeSearch" filtering technology and set the number of results shown on each page. Google has been criticized for placing long-term cookies on users' machines to store these preferences, a tactic which also enables them to track a user's search terms and retain the data for more than a year. For any query, up to the first 1000 results can be shown with a maximum of 100 displayed per page. The ability to specify the number of results is available only if "Instant Search" is not enabled. If "Instant Search" is enabled, only 10 results are displayed, regardless of this setting.

Non-indexable data

Despite its immense index, there is also a considerable amount of data available in online databases which are accessible by means of queries but not by links. This so-called invisible or deep Web is minimally covered by Google and other search engines.[15] The deep Web contains library catalogs, official legislative documents of governments, phone books, and other content which is dynamically prepared to respond to a query.

Google optimization

Since Google is the most popular search engine, many webmasters have become eager to influence their website's Google rankings. An industry of consultants has arisen to help websites increase their rankings on Google and on other search engines. This field, called search engine optimization, attempts to discern patterns in search engine listings, and then develop a methodology for improving rankings to draw more searchers to their client's sites. Search engine optimization encompasses both "on page" factors (like body copy, title elements, H1 heading elements and image alt attribute values) and Off Page Optimization factors (like anchor text and PageRank). The general idea is to affect Google's relevance algorithm by incorporating the keywords being targeted in various places "on page", in particular the title element and the body copy (note: the higher up in the page, presumably the better its keyword prominence and thus the ranking). Too many occurrences of the keyword, however, cause the page to look suspect to Google's spam checking algorithms. Google has published guidelines for website owners who would like to raise their rankings when using legitimate optimization consultants.[16] It has been hypothesized, and, allegedly, is the opinion of the owner of one business about which there has been numerous complaints, that negative publicity, for example, numerous consumer complaints, may serve as well to elevate page rank on Google Search as favorable comments.[17] The particular problem addressed in *The New York Times* article, which involved DecorMyEyes, was addressed shortly thereafter by an undisclosed fix in the Google algorithm. According to Google, it was not the frequently published consumer complaints about DecorMyEyes which resulted in the high ranking but mentions on news websites of events which affected the firm such as legal actions against it.[18]

Functionality

Google search consists of a series of localized websites. The largest of those, the google.com site, is the top most-visited website in the world.[19] Some of its features include a definition link for most searches including dictionary words, the number of results you got on your search, links to other searches (e.g. for words that Google believes to be misspelled, it provides a link to the search results using its proposed spelling), and many more.

Search syntax

Google's search engine normally accepts queries as a simple text, and breaks up the user's text into a sequence of search terms, which will usually be words that are to occur in the results, but one can also use Boolean operators, such as: quotations marks (") for a phrase, a prefix such as "+", "-" for qualified terms (no longer valid, the '+' was removed from google on 10/19/11[20]), or one of several advanced operators, such as "site:". The webpages of "Google Search Basics"[21] describe each of these additional queries and options (*see below:* Search options). Google's Advanced Search web form gives several additional fields which may be used to qualify searches by such criteria as date of first retrieval. All advanced queries transform to regular queries, usually with additional qualified term.

Query expansion

Google applies query expansion to the submitted search query, transforming it into the query that will actually be used to retrieve results. As with page ranking, the exact details of the algorithm Google uses are deliberately obscure, but certainly the following transformations are among those that occur:

- Term reordering: in information retrieval this is a standard technique to reduce the work involved in retrieving results. This transformation is invisible to the user, since the results ordering uses the original query order to determine relevance.
- Stemming is used to increase search quality by keeping small syntactic variants of search terms.[22]
- There is a limited facility to fix possible misspellings in queries.

"I'm Feeling Lucky"

Google's homepage includes a button labeled "I'm Feeling Lucky". When a user types in a search and clicks on the button the user will be taken directly to the first search result, bypassing the search engine results page. The thought is that if a user is "feeling lucky", the search engine will return the perfect match the first time without having to page through the search results. However, with the introduction of Google Instant, it is not possible to use the button properly unless the Google Instant function is switched off. According to a study by Tom Chavez of "Rapt", this feature costs Google $110 million a year as 1% of all searches use this feature and bypass all advertising.[23]

On October 30, 2009, for some users, the "I'm Feeling Lucky" button was removed from Google's main page, along with the regular search button. Both buttons were replaced with a field that reads, "This space intentionally left blank." This text faded out when the mouse was moved on the page, and normal search functionality is achieved by filling in the search field with the desired terms and pressing enter. A Google spokesperson explains, "This is just a test, and a way for us to gauge whether our users will like an even simpler search interface."[24] Personalized Google homepages retained both buttons and their normal functions.

On May 21, 2010, the 30th anniversary of Pac-Man, the "I'm Feeling Lucky" button was replaced with a button reading the words "Insert Coin". After pressing the button, the user would begin a Google-themed game of Pac-Man in the area where the Google logo would normally be. Pressing the button a second time would begin a two-player version of the same game that includes Ms. Pacman for player 2. This version can be accessed at `www.google.com/pacman/`[25] as a permanent link to the page.

Rich Snippets

On 12 May 2009, Google announced that they would be parsing the hCard, hReview, and hProduct microformats and using them to populate search result pages with what they called "Rich Snippets".[26]

Special features

Besides the main search-engine feature of searching for text, Google Search has more than 22 "special features" (activated by entering any of dozens of *trigger words*) when searching:[8][9][27]

- weather – The weather conditions, temperature, wind, humidity, and forecast,[8] for many cities, can be viewed by typing "weather" along with a city for larger cities or city and state, U.S. zip code, or city and country for smaller cities (such as: weather Lawrence, Kansas; weather Paris; weather Bremen, Germany).
- stock quotes – The market data[8] for a specific company or fund can be viewed, by typing the ticker symbol (or include "stock"), such as: CSCO; MSFT; IBM stock; F stock (lists Ford Motor Co.); or AIVSX (fund). Results show inter-day changes, or 5-year graph, etc. This does not work for stock names which are one letter long, such as Citigroup (C) or Macy's (M) (Ford being an exception), or are common words, such as Diamond Offshore (DO) or Majesco (COOL).

- time – The current time in many cities (worldwide),[8] can be viewed by typing "time" and the name of the city (such as: time Cairo; time Pratt, KS).
- sports scores – The scores and schedules, for sports teams,[8] can be displayed by typing the team name or league name into the search box.
- unit conversion – Measurements can be converted,[8] by entering each phrase, such as: 10.5 cm in inches; or 90 km in miles
- currency conversion – A money or currency converter can be selected,[8] by typing the names or currency codes (listed by ISO 4217): 6789 Euro in USD; 150 GBP in USD; 5000 Yen in USD; 5000 Yuan in lira (the U.S. dollar can be *USD* or "US$" or "$", while Canadian is *CAD*, etc.).
- calculator – Calculation results can be determined,[8] as calculated live, by entering a formula in numbers or words, such as: 6*77+pi+sqrt(e^3)/888 plus 0.45. The user is given the option to search for the formula, after calculation. The calculator also uses the unit and currency conversion functions to allow unit-aware calculations. For example, "(3 EUR/liter)/(40 miles/gallon) in USD/mile" calculates the dollar cost per mile for a 40 mpg car with gas costing 3 euros a liter. The caret "^" raises a number to an exponent power, and percentages are allowed ("40% of 300").[9] There is also some debate as to Google's calculation of 0^0. Many mathematicians believe that 0^0 is undefined but Google's calculator shows the result as 1.[28]
- numeric ranges – A set of numbers can be matched by using a double-dot between range numbers (70..73 or 90..100) to match any positive number in the range, inclusive.[9] Negative numbers are treated as using exclusion-dash to not match the number.
- dictionary lookup – A definition for a word or phrase can be found,[8] by entering "define" followed by a colon and the word(s) to lookup (such as, "define:philosophy")
- maps – Some related maps can be displayed,[8] by typing in the name or U.S. ZIP code of a location and the word "map" (such as: New York map; Kansas map; or Paris map).
- movie showtimes – Reviews or film showtimes can be listed for any movies playing nearby,[8] by typing "movies" or the name of any current film into the search box. If a specific location was saved on a previous search, the top search result will display showtimes for nearby theaters for that movie.
- public data – Trends for population (or unemployment rates)[8] can be found for U.S. states & counties, by typing "population" or "unemployment rate" followed by a state or county name.
- real estate and housing – Home listings in a given area can be displayed,[8] using the trigger words "housing", "home", or "real estate" followed by the name of a city or U.S. zip code.
- travel data/airports – The flight status for arriving or departing U.S. flights can be displayed,[8] by typing in the name of the airline and the flight number into the search box (such as: American airlines 18). Delays at a specific airport can also be viewed (by typing the name of the city or three-letter airport code plus word "airport").
- package tracking – Package mail can be tracked[8] by typing the tracking number of a Royal Mail, UPS, FedEx or USPS package directly into the search box. Results will include quick links to track the status of each shipment.
- patent numbers – U.S. patents can be searched[8][27] by entering the word "patent" followed by the patent number into the search box (such as: Patent 5123123).
- area code – The geographical location (for any U.S. telephone area code)[8] can be displayed by typing a 3-digit area code (such as: 650).
- synonym search – A search can match words similar to those specified,[8] by placing the tilde sign (~) immediately in front of a search term, such as: ~fast food.

Search options

The webpages maintained by the Google Help Center have text describing more than 15 various search options.[29] The Google operators:

- OR – Search for either one, such as "price high OR low" searches for *"price"* with *"high"* or *"low"*.
- "-" – Search while excluding a word, such as "apple -tree" searches where word *"tree"* is not used.
- "+" – (Removed on 10/19/11[20]) Force inclusion of a word, such as "Name+of+the Game" to require the words *"of"* & *"the"* to appear on a matching page.
- "*" – Wildcard operator to match any words between other specific words. Some of the

query options are as follows:

- define: – The query prefix "define:" will provide a definition[29] of the words listed after it.
- stocks: – After "stocks:" the query terms are treated as stock ticker symbols[29] for lookup.
- site: – Restrict the results to those websites in the given domain,[29] such as, site:www.acmeacme.com. The option "site:com" will search all domain URLs named with ".com" (no space after "site:").
- allintitle: – Only the page titles are searched[29] (not the remaining text on each webpage).
- intitle: – Prefix to search in a webpage title,[29] such as "intitle:googlesearch" will list pages with word "google" in title, and word "search" anywhere (no space after "intitle:").
- allinurl: – Only the page URL address lines are searched[29] (not the text inside each webpage).
- inurl: – Prefix for each word to be found in the URL;[29] others words are matched anywhere, such as "inurl:acme search" matches "acme" in a URL, but matches "search" anywhere (no space after "inurl:").

The page-display options (or query types) are:

- cache: – Highlights the search-words within the cached document, such as "cache:www.google.com xxx" shows cached content with word "xxx" highlighted.
- link: – The prefix "link:" will list webpages that have links to the specified webpage, such as "link:www.google.com" lists webpages linking to the Google homepage.
- related: – The prefix "related:" will list webpages that are "similar" to a specified web page.
- info: – The prefix "info:" will display some background information about one specified webpage, such as, info:www.google.com. Typically, the info is the first text (160 bytes, about 23 words) contained in the page, displayed in the style of a results entry (for just the 1 page as matching the search).
- filetype: – results will only show files of the desired type (ex filetype:pdf will return pdf files)

Error messages

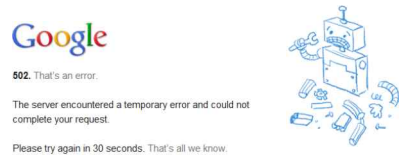
Some searches will give a 403 Forbidden error with the text "We're sorry...

```
... but your query looks similar to automated requests from a computer
virus or spyware application. To protect our users, we can't process
your request right now.

 the meantime, if you suspect that your computer or network has been
infected, you might want to run a virus checker or spyware remover to
make sure that your systems are free of viruses and other spurious
software.
```

We apologize for the inconvenience, and hope we'll see you again on Google." sometimes followed by a CAPTCHA prompt.[30]

The screen was first reported in 2005, and was a response to the heavy use of Google by search engine optimization companies to check on ranks of sites they were optimizing. The message is triggered by high volumes of requests from a single IP address. Google apparently uses the Google cookie as part of its determination of refusing service.[30]



Google's Server Error page

In June 2009, after the death of pop superstar Michael Jackson, this message appeared to many internet users who were searching Google for news stories related to the singer, and was assumed by Google to be a DDoS attack, although many queries were submitted by legitimate searchers.

January 2009 malware bug

Google flags search results with the message "This site may harm your computer" if the site is known to install malicious software in the background or otherwise surreptitiously. Google does this to protect users against visiting sites that could harm their computers. For approximately 40 minutes on January 31, 2009, *all* search results were mistakenly classified as malware and could therefore not be clicked; instead a warning message was displayed and the user was required to enter the requested URL manually. The bug was caused by human error.[31][32][33][34] The URL of "/" (which expands to all URLs) was mistakenly added to the malware patterns file.[32][33]

Google Doodles

On certain occasions, the logo on Google's webpage will change to a special version, known as a "Google Doodle". Clicking on the Doodle links to a string of Google search results about the topic. The first was a reference to the Burning Man Festival in 1998,[35][36] and others have been produced for the birthdays of notable people like Albert Einstein, historical events like the interlocking Lego block's 50th anniversary and holidays like Valentine's Day.[37] Some Google Doodles have interactivity beyond a simple search, such as the famous "Google Pacman" version that appeared on May 21,2010.

Google Caffeine

In August 2009, Google announced the rollout of a new search architecture, codenamed "Caffeine".[38] The new architecture was designed to return results faster and to better deal with rapidly updated information[39] from services including Facebook and Twitter.[38] Google developers noted that most users would notice little immediate change, but invited developers to test the new search in its sandbox.[40] Differences noted for their impact upon search engine optimization included heavier keyword weighting and the importance of the domain's age.[41][42] The move was interpreted in some quarters as a response to Microsoft's recent release of an upgraded version of its own search service, renamed Bing.[43] Google announced completion of Caffeine on 8 June 2010, claiming 50% fresher results due to continuous updating of its index.[44] With Caffeine, Google moved its back-end indexing system away from MapReduce and onto BigTable, the company's distributed database platform.[45] Caffeine is also based on Colossus, or GFS2,[46] an overhaul of the GFS distributed file system.[47]

Privacy

Searches made by search engines, including Google, leave traces, raising concerns about privacy but sometimes facilitating the administration of justice; murderers have been detected and convicted as a result of incriminating searches they made such as "tips with killing with a baseball bat".[48].

A search can be traced in several ways. When using a search engine through a browser program on a computer, search terms and other information will usually be stored on the computer by default, unless steps are taken to erase them. An Internet Service Provider may store records which relate search terms to an IP address and a time. The search engine provider (e.g., Google) may keep logs with the same information[49]. Whether such logs are kept, and access to them by law enforcement agencies, is subject to legislation and working practices; the law may mandate, prohibit, or say nothing about logging of various types of information.

The technically knowledgeable and forewarned user can avoid leaving traces.

Encrypted Search

In May 2010 Google rolled out SSL-encrypted web search.[50] The encrypted search can be accessed at `encrypted.google.com`[51]

Instant Search

Google Instant, a feature that displays suggested results while the user types, was introduced in the United States on September 8, 2010. In concert with the Google Instant launch, Google disabled the ability of users to choose to see more than 10 search results per page. At the time of the announcement Google expected Instant to save users 2 to 5 seconds in every search, collectively about 11 million seconds per hour.[52] Search engine marketing pundits speculate that Google Instant will have a great impact on local and paid search.[53]

Instant Search can be disabled via Google's "preferences" menu, but autocomplete-style search suggestions now cannot be disabled; Google confirm that this is intentional.[54]

The publication 2600: The Hacker Quarterly has compiled a list of words that are restricted by Google Instant.[55] These are terms the web giant's new instant search feature will not search.[56][57] Most terms are often vulgar and derogatory in nature, but some apparently irrelevant searches including "Myleak" are removed.[57]

Redesign

In late June 2011, Google introduced a new look to the Google home page in order to boost the use of the Google+ social tools.[58]

One of the major changes was replacing the classic navigation bar with a black one. Google's digital creative director Chris Wiggins explains: "We're working on a project to bring you a new and improved Google experience, and over the next few months, you'll continue to see more updates to our look and feel."[59] The new navigation bar has been negatively received by a vocal minority.[60]

International

Google is available in many languages and has been localized completely or partly for many countries.[61] The interface has also been made available in some languages for humorous purpose:

- Bork, bork, bork!
- Elmer Fudd
- Leetspeak
- Klingon
- Pig Latin
- Pirate

In addition to the main URL Google.com, Google Inc. owns 160 domain names for each of the countries/regions in which it has been localized.[61]

Search products

In addition to its tool for searching webpages, Google also provides services for searching images, Usenet newsgroups, news websites, videos, searching by locality, maps, and items for sale online. In 2006, Google has indexed over 25 billion web pages,[62] 400 million queries per day,[62] 1.3 billion images, and over one billion Usenet messages. It also caches much of the content that it indexes. Google operates other tools and services including Google News, Google Suggest, Google Product Search, Google Maps, Google Co-op, Google Earth, Google Docs, Picasa, Panoramio, YouTube, Google Translate, Google Blog Search and Google Desktop Search.

There are also products available from Google that are not directly search-related. Gmail, for example, is a webmail application, but still includes search features; Google Browser Sync does not offer any search facilities, although it aims to organize your browsing time.

Also Google starts many new beta products, like Google Social Search or Google Image Swirl.

Energy consumption

Google claims that a search query requires altogether about 1 kJ or 0.0003 kW·h.[63]

References

[1] https://www.google.com/

[2] "WHOIS - google.com" (http://reports.internic.net/cgi/whois?whois_nic=google.com&type=domain). . Retrieved 2009-01-27.

[3] "Google.com Site Info" (http://www.alexa.com/siteinfo/google.com). Alexa Internet. . Retrieved 2012-03-02.

[4] "Alexa Search Engine ranking" (http://www.alexa.com/siteinfo/google.com+yahoo.com+altavista.com). . Retrieved 2009-11-15.

[5] "Almost 12 Billion U.S. Searches Conducted in July" (http://searchenginewatch.com/showPage.html?page=3630718). SearchEngineWatch. 2008-09-02. .

[6] *...The *, or wildcard, is a little-known feature that can be very powerful...* (http://www.google.co.nz/support/websearch/bin/answer.py?answer=136861)

[7] "WHOIS - google.com" (http://reports.internic.net/cgi/whois?whois_nic=google.com&type=domain). . Retrieved 2009-01-27.

[8] "Search Features" (http://www.google.com/intl/en/help/features.html). Google.com. May 2009. .

[9] "Google Help : Cheat Sheet" (http://www.google.com/help/cheatsheet.html). Google. 2010. .

[10] *Voice Search for Google.com - Just click the mic and say your search. And, Search Google by giving Image* (http://qualitypoint.blogspot.com/2011/06/voice-search-for-googlecom.html)

[11] Hubbard, Douglas (2011). *Pulse: The New Science of Harnessing Internet Buzz to Track Threats and Opportunities*. John Wiley & Sons.

[12] Brin, S.; Page, L. (1998). "The anatomy of a large-scale hypertextual Web search engine" (http://infolab.stanford.edu/pub/papers/ google.pdf). *Computer Networks and ISDN Systems* **30**: 107–117. doi:10.1016/S0169-7552(98)00110-X. ISSN 0169-7552. .

[13] "Corporate Information: Technology Overview" (http://www.google.com/corporate/tech.html). Google. . Retrieved 2009-11-15. Wired.com (http://www.wired.com/magazine/2010/02/ff_google_algorithm/)

[14] "Google Frequently Asked Questions - File Types" (http://www.google.com/help/faq_filetypes.html#what). Google. . Retrieved 2011-09-12.

[15] Sherman, Chris and Price, Gary. "The Invisible Web: Uncovering Sources Search Engines Can't See, In: *Library Trends* 52 (2) 2003: Organizing the Internet:" (http://hdl.handle.net/2142/8528). pp. 282–298. .
[16] "Google Webmaster Guidelines" (http://www.google.com/webmasters/guidelines.html). Google. . Retrieved 2009-11-15.
[17] Segal, David (November 26, 2010). "A Bully Finds a Pulpit on the Web" (https://www.nytimes.com/2010/11/28/business/28borker. html). *The New York Times*. . Retrieved November 27, 2010.
[18] Blogspot.com (http://googleblog.blogspot.com/2010/12/being-bad-to-your-customers-is-bad-for.html)
[19] "Top 500" (http://www.alexa.com/site/ds/top_sites?ts_mode=global&lang=none). Alexa. . Retrieved 2008-04-15.
[20] (http://www.frag.co.uk/blog/2011/10/googles-changes-the-operators/), Google changes the operators.
[21] Google.com (http://www.google.com/support/websearch/bin/answer.py?answer=136861)
[22] "Google:Stemming" (http://www.google.com/support/bin/answer.py?answer=35889#stemming). Google. .
[23] "I'm feeling lucky(button costs Google $110 million per year" (http://valleywag.com/tech/google/ im-feeling-lucky-button-costs-google-110-million-per-year-324927.php). Valleywag. 2007. . Retrieved 2008-01-19.
[24] "Google's New Homepage Motto: 'This Space Intentionally Left Blank'" (http://digitaldaily.allthingsd.com/20091030/goog-page/). WallStreetJournal. 2009. . Retrieved 2009-11-17.
[25] Google.com (http://www.google.com/pacman)
[26] Goel, Kavi; Ramanathan V. Guha, Othar Hansson (2009-05-12). "Introducing Rich Snippets" (http://googlewebmastercentral.blogspot. com/2009/05/introducing-rich-snippets.html). *Google Webmaster Central Blog*. Google. . Retrieved 2009-05-25.
[27] "Google and Search Engines"(http://www.law.emory.edu/law-library/research/advanced-legal-research-class/ finding-aids-and-searching/google.html). Emory University Law School. 2006. .
[28] Google.com (http://www.google.com/search?output=&sitesearch=&hl=en&q=2+2&submit=Search+the+Web#hl=en& expIds=17259,22104,25907,26637,26659,26741,26817,26992,27095&sugexp=ldymls&xhr=t&q=0^0&cp=3&pf=p&sclient=psy&aq=f& aqi=g4g-o1&aql=&oq=0^0&gs_rfai=&pbx=1&fp=433548e9226de17c)
[29] "Google Help Center – Alternate query types", 2009, webpage: G-help (http://www.google.com/help/operators.html).
[30] "Google error page" (http://www.google.com/support/bin/answer.py?answer=15661). . Retrieved 2008-12-31.
[31] Krebs, Brian (2009-01-31). "Google: This Internet May Harm Your Computer" (http://voices.washingtonpost.com/securityfix/2009/01/ google_this_internet_will_harm.html?hpid=news-col-blog). The Washington Post. . Retrieved 2009-01-31.
[32] Mayer, Marissa (2009-01-31). "This site may harm your computer on every search result?!?!" (http://googleblog.blogspot.com/2009/01/ this-site-may-harm-your-computer-on.html). *The Official Google Blog*. Google. . Retrieved 2009-01-31.
[33] Weinstein, Maxim (2009-1-31). "Google glitch causes confusion" (http://blog.stopbadware.org/2009/01/31/ google-glitch-causes-confusion). StopBadware.org. . Retrieved 2010-5-10.
[34] Cooper, Russ (January 31, 2009). "Serious problems with Google search" (http://securityblog.verizonbusiness.com/2009/01/31/ serious-problems-with-google-search/). Verizon Business Security Blog. . Retrieved 2010-5-10.
[35] Hwang, Dennis (June 8, 2004). "Oodles of Doodles" (http://googleblog.blogspot.com/2004/06/oodles-of-doodles.html). Google (corporate blog). . Retrieved July 19, 2006.
[36] "Doodle History" (http://www.google.com/doodle4google/history.html). Google, Inc.. . Retrieved 5-10-2010.
[37] "Google logos: *Valentine's Day logo*" (http://www.google.com/logos/valentine07.gif). February 14, 2007. . Retrieved April 6, 2007.
[38] Harvey, Mike (11 August 2009). "Google unveils new "Caffeine" search engine" (http://technology.timesonline.co.uk/tol/news/ tech_and_web/personal_tech/article6792403.ece). London: The Times. . Retrieved 14 August 2009.
[39] "What Does Google "Caffeine" Mean for My Website?" (http://www.siivo.com/blog/2010/07/ what-does-google-caffeine-mean-for-my-website). New York: Siivo Corp. 21 July 2010. . Retrieved 21 July 2010.
[40] Culp, Katie (12 August 2009). "Google introduces new "Caffeine" search system" (http://www.foxbusiness.com/story/markets/ industries/technology/google-introduces-new-caffeine-search/). Fox News. . Retrieved 14 August 2009.
[41] Martin, Paul (31 July 2009). "Bing - The new Search Engine from Microsoft and Yahoo" (http://blog.cube3marketing.com/2009/07/31/ bing-the-new-search-engine-from-microsoft-and-yahoo/). Cube3 Marketing. . Retrieved 12 January 2010.
[42] Martin, Paul (27 August 2009). "Caffeine - The New Google Update" (http://blog.cube3marketing.com/2009/08/27/ caffeine-the-new-google-update/). Cube3 Marketing. . Retrieved 12 January 2010.
[43] Barnett, Emma (11 August 2009). "Google reveals caffeine: a new faster search engine" (http://www.telegraph.co.uk/technology/ google/6009176/Google-reveals-caffeine-a-new-faster-search-engine.html). The Telegraph. . Retrieved 14 August 2009.
[44] Grimes, Carrie (8 June 2010). "Our new search index: Caffeine" (http://googleblog.blogspot.com/2010/06/ our-new-search-index-caffeine.html). *The Official Google Blog*. . Retrieved 18 June 2010.
[45] Google search index splits with MapReduce (http://www.theregister.co.uk/2010/09/09/google_caffeine_explained/) – The Register
[46] Google Caffeine: What it really is (http://www.theregister.co.uk/2009/08/14/google_caffeine_truth/) – The Register
[47] Google File System II: Dawn of the Multiplying Master Nodes (http://www.theregister.co.uk/2009/08/12/ google_file_system_part_deux/) – The Register
[48] Search Engine Land: Once Again, A Google Murder Case, 29 Jan 2008(http://searchengineland.com/ once-again-a-google-murder-case-13241)
[49] Search Engine Land: Google Anonymizing Search Records To Protect Privacy, 14 March 2007 (http://searchengineland.com/ google-anonymizing-search-records-to-protect-privacy-10736)

[50] "SSL Search: Features - Web Search Help" (http://www.google.com/support/websearch/bin/answer.py?answer=173733&hl=en). *Web Search Help*. Google. May 2010. . Retrieved 2010-07-07.
[51] Encrypted.google.com (http://encrypted.google.com)
[52] Peter Nowak (2010). *Tech Bytes: Google Instant* (Television production). United States: ABC News.
[53] "How Google Saved $100 Million By Launching Google Instant" (http://searchengineland.com/ how-google-saved-100-million-by-launching-google-instant-51270). . Retrieved 20 September 2010.
[54] Google Web Search Help Forum (http://www.google.com/support/forum/p/Web+Search/thread?tid=5a69f1094357f31b&hl=en) (WebCite archive (http://www.webcitation.org/5sy2ImYdO))
[55] 2600.com: Google Blacklist - Words That Google Instant Doesn't Like (http://www.2600.com/googleblacklist/)
[56] CNN: Which words does Google Instant blacklist? (http://www.cnn.com/2010/TECH/web/09/29/google.instant.blacklist.mashable/ index.html?eref=mrss_igoogle_cnn)
[57] The Huffington Post: Google Instant Censorship: The Strangest Terms Blacklisted By Google (http://www.huffingtonpost.com/2010/09/ 29/google-instant-censorship_n_743203.html)
[58] Boulton, Clint. "Google Redesign Backs Social Effort" (http://www.eweekeurope.co.uk/comment/ google-redesign-backs-social-effort-32954). *eWeek Europe*. eWeek Europe. . Retrieved 1 July 2011.
[59] Google redesigns its homepage [[Los Angeles Times (http://latimesblogs.latimes.com/technology/2011/06/ google-redesigns-its-homepage-with-new-black-bar-up-top-google-social-network.html)]]
[60] Google support forum, one of many threads on being unable to switch off the black navigation bar (http://www.google.com/support/ forum/p/Web+Search/thread?tid=7ddbf7a4c8fa04a9&hl=en)
[61] Language Tools (http://www.google.com/language_tools?hl=en)
[62] Google, Web Crawling and Distributed Synchronization (http://www.seas.upenn.edu/~zives/cis555/slides/I-Crawlers-Sync.ppt) p. 11.
[63] Blogspot.com (http://googleblog.blogspot.com/2009/01/powering-google-search.html), Powering a Google search

Further reading

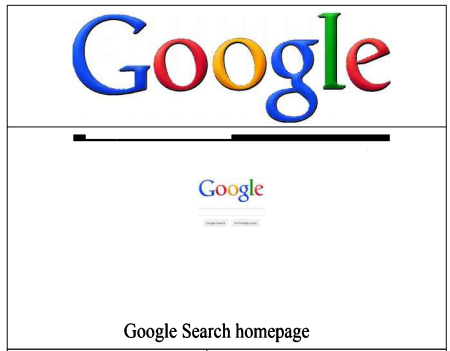
- *Google Hacks* from O'Reilly is a book containing tips about using Google effectively. Now in its third edition. ISBN 0-596-52706-3.
- *Google: The Missing Manual* by Sarah Milstein and Rael Dornfest (O'Reilly, 2004). ISBN 0-596-00613-6
- *How to Do Everything with Google* by Fritz Schneider, Nancy Blachman, and Eric Fredricksen (McGraw-Hill Osborne Media, 2003). ISBN 0-07-223174-2
- *Google Power* by Chris Sherman (McGraw-Hill Osborne Media, 2005). ISBN 0-07-225787-3
- Barroso, Luiz Andre; Dean, Jeffrey; Hölzle, Urs (2003). "Web Search for a Planet: The Google Cluster Architecture". *IEEE Micro* **23** (2): 22–28. doi:10.1109/MM.2003.1196112.

External links

- Google.com (http://www.google.com)
- The Original Google! (http://web.archive.org/web/19981111183552/google.stanford.edu/)

Google Search

Google Search



Google Search homepage

URL	Google.com [1]
Commercial?	Yes
Type of site	Web search engine
Registration	Optional
Available language(s)	Multilingual (124)
Owner	Google
Created by	SergeyBrinandLarryPage
Launched	September 15, 1997[1]
Alexa rank	1 (March 2012)[2]
Revenue	From AdWords
Current status	Active

Google Search (or **Google Web Search**) is a web search engine owned by Google Inc. Google Search is the most-used search engine on the World Wide Web,[3] receiving several hundred million queries each day through its various services.[4]

The order of search results on Google's search-results pages is based, in part, on a priority rank called a "PageRank". Google Search provides many options for customized search, using Boolean operators such as: implied "AND" (if several concatenated search terms separated by spaces are given, only pages containing all of them should be returned), exclusion ("-xx"), alternatives ("xx OR yy"), and wildcard ("x * x").[5]

The main purpose of Google Search is to hunt for text in Web pages, as opposed to other data, such as with Google Image Search. Google Search was originally developed by Larry Page and Sergey Brin in 1997.[6] Google Search provides at least 22 special features beyond the original word-search capability.[7] These include synonyms, weather forecasts, time zones, stock quotes, maps, earthquake data, movie showtimes, airports, home listings, and sports scores. There are special features for numbers, including ranges (70..73),[8] prices, temperatures, money/unit conversions ("10.5 cm in inches"), calculations ("3*4+sqrt(6)-pi/2"), package tracking, patents, area codes,[7] and language translation of displayed pages. In June 2011, Google introduced "Google Voice Search" and "Search by Image" features for allowing the users to search words by speaking and by giving images.[9]

The frequency of use of many search terms have reached a volume that they may indicate broader economic, social and health trends.[10] Data about the frequency of use of search terms on Google (available through Google Adwords, Google Trends, and Google Insights for Search) have been shown to correlate with flu outbreaks and

unemployment levels and provide the information faster than traditional reporting methods and government surveys.

Search engine

PageRank

Google's rise to success was in large part due to a patented algorithm called PageRank that helps rank web pages that match a given search string.[11] When Google was a Stanford research project, it was nicknamed BackRub because the technology checks backlinks to determine a site's importance. Previous keyword-based methods of ranking search results, used by many search engines that were once more popular than Google, would rank pages by how often the search terms occurred in the page, or how strongly associated the search terms were within each resulting page. The PageRank algorithm instead analyzes human-generated links assuming that web pages linked from many important pages are themselves likely to be important. The algorithm computes a recursive score for pages, based on the weighted sum of the PageRanks of the pages linking to them. PageRank is thought to correlate well with human concepts of importance. In addition to PageRank, Google, over the years, has added many other secret criteria for determining the ranking of pages on result lists, reported to be over 200 different indicators.[12] The specifics of which are kept secret to keep spammers at bay and help Google maintain an edge over its competitors globally.

Search results

The exact percentage of the total of web pages that Google indexes are not known, as it is very difficult to accurately calculate. Google not only indexes and caches web pages, but also takes "snapshots" of other file types, which include PDF, Word documents, Excel spreadsheets, Flash SWF, plain text files, and so on.[13] Except in the case of text and SWF files, the cached version is a conversion to (X)HTML, allowing those without the corresponding viewer application to read the file. Users can customize the search engine, by setting a default language, using the "SafeSearch" filtering technology and set the number of results shown on each page. Google has been criticized for placing long-term cookies on users' machines to store these preferences, a tactic which also enables them to track a user's search terms and retain the data for more than a year. For any query, up to the first 1000 results can be shown with a maximum of 100 displayed per page. The ability to specify the number of results is available only if "Instant Search" is not enabled. If "Instant Search" is enabled, only 10 results are displayed, regardless of this setting.

Non-indexable data

Despite its immense index, there is also a considerable amount of data available in online databases which are accessible by means of queries but not by links. This so-called invisible or deep Web is minimally covered by Google and other search engines.[14] The deep Web contains library catalogs, official legislative documents of governments, phone books, and other content which is dynamically prepared to respond to a query.

Google optimization

Since Google is the most popular search engine, many webmasters have become eager to influence their website's Google rankings. An industry of consultants has arisen to help websites increase their rankings on Google and on other search engines. This field, called search engine optimization, attempts to discern patterns in search engine listings, and then develop a methodology for improving rankings to draw more searchers to their client's sites. Search engine optimization encompasses both "on page" factors (like body copy, title elements, H1 heading elements and image alt attribute values) and Off Page Optimization factors (like anchor text and PageRank). The general idea is to affect Google's relevance algorithm by incorporating the keywords being targeted in various places "on page", in particular the title element and the body copy (note: the higher up in the page, presumably the better its keyword prominence and thus the ranking). Too many occurrences of the keyword, however, cause the page to look suspect to Google's spam checking algorithms. Google has published guidelines for website owners who would like to raise their rankings when using legitimate optimization consultants.[15] It has been hypothesized, and, allegedly, is the opinion of the owner of one business about which there has been numerous complaints, that negative publicity, for example, numerous consumer complaints, may serve as well to elevate page rank on Google Search as favorable comments.[16] The particular problem addressed in *The New York Times* article, which involved DecorMyEyes, was addressed shortly thereafter by an undisclosed fix in the Google algorithm. According to Google, it was not the frequently published consumer complaints about DecorMyEyes which resulted in the high ranking but mentions on news websites of events which affected the firm such as legal actions against it.[17]

Functionality

Google search consists of a series of localized websites. The largest of those, the google.com site, is the top most-visited website in the world.[18] Some of its features include a definition link for most searches including dictionary words, the number of results you got on your search, links to other searches (e.g. for words that Google believes to be misspelled, it provides a link to the search results using its proposed spelling), and many more.

Search syntax

Google's search engine normally accepts queries as a simple text, and breaks up the user's text into a sequence of search terms, which will usually be words that are to occur in the results, but one can also use Boolean operators, such as: quotations marks (") for a phrase, a prefix such as "+", "-" for qualified terms (no longer valid, the '+' was removed from google on 10/19/11[19]), or one of several advanced operators, such as "site:". The webpages of "Google Search Basics"[20] describe each of these additional queries and options (*see below:* Search options). Google's Advanced Search web form gives several additional fields which may be used to qualify searches by such criteria as date of first retrieval. All advanced queries transform to regular queries, usually with additional qualified term.

Query expansion

Google applies query expansion to the submitted search query, transforming it into the query that will actually be used to retrieve results. As with page ranking, the exact details of the algorithm Google uses are deliberately obscure, but certainly the following transformations are among those that occur:

- Term reordering: in information retrieval this is a standard technique to reduce the work involved in retrieving results. This transformation is invisible to the user, since the results ordering uses the original query order to determine relevance.
- Stemming is used to increase search quality by keeping small syntactic variants of search terms.[21]
- There is a limited facility to fix possible misspellings in queries.

"I'm Feeling Lucky"

Google's homepage includes a button labeled "I'm Feeling Lucky". When a user types in a search and clicks on the button the user will be taken directly to the first search result, bypassing the search engine results page. The thought is that if a user is "feeling lucky", the search engine will return the perfect match the first time without having to page through the search results. However, with the introduction of Google Instant, it is not possible to use the button properly unless the Google Instant function is switched off. According to a study by Tom Chavez of "Rapt", this feature costs Google $110 million a year as 1% of all searches use this feature and bypass all advertising.[22]

On October 30, 2009, for some users, the "I'm Feeling Lucky" button was removed from Google's main page, along with the regular search button. Both buttons were replaced with a field that reads, "This space intentionally left blank." This text faded out when the mouse was moved on the page, and normal search functionality is achieved by filling in the search field with the desired terms and pressing enter. A Google spokesperson explains, "This is just a test, and a way for us to gauge whether our users will like an even simpler search interface."[23] Personalized Google homepages retained both buttons and their normal functions.

On May 21, 2010, the 30th anniversary of Pac-Man, the "I'm Feeling Lucky" button was replaced with a button reading the words "Insert Coin". After pressing the button, the user would begin a Google-themed game of Pac-Man in the area where the Google logo would normally be. Pressing the button a second time would begin a two-player version of the same game that includes Ms. Pacman for player 2. This version can be accessed at `www.google.com/pacman/`[24] as a permanent link to the page.

Rich Snippets

On 12 May 2009, Google announced that they would be parsing the hCard, hReview, and hProduct microformats and using them to populate search result pages with what they called "Rich Snippets".[25]

Special features

Besides the main search-engine feature of searching for text, Google Search has more than 22 "special features" (activated by entering any of dozens of *trigger words*) when searching:[7][8][26]

- weather – The weather conditions, temperature, wind, humidity, and forecast,[7] for many cities, can be viewed by typing "weather" along with a city for larger cities or city and state, U.S. zip code, or city and country for smaller cities (such as: weather Lawrence, Kansas; weather Paris; weather Bremen, Germany).
- stock quotes – The market data[7] for a specific company or fund can be viewed, by typing the ticker symbol (or include "stock"), such as: CSCO; MSFT; IBM stock; F stock (lists Ford Motor Co.); or AIVSX (fund). Results show inter-day changes, or 5-year graph, etc. This does not work for stock names which are one letter long, such as Citigroup (C) or Macy's (M) (Ford being an exception), or are common words, such as Diamond Offshore (DO) or Majesco (COOL).

- time – The current time in many cities (worldwide),[7] can be viewed by typing "time" and the name of the city (such as: time Cairo; time Pratt, KS).
- sports scores – The scores and schedules, for sports teams,[7] can be displayed by typing the team name or league name into the search box.
- unit conversion – Measurements can be converted,[7] by entering each phrase, such as: 10.5 cm in inches; or 90 km in miles
- currency conversion – A money or currency converter can be selected,[7] by typing the names or currency codes (listed by ISO 4217): 6789 Euro in USD; 150 GBP in USD; 5000 Yen in USD; 5000 Yuan in lira (the U.S. dollar can be *USD* or "US$" or "$", while Canadian is *CAD*, etc.).
- calculator – Calculation results can be determined,[7] as calculated live, by entering a formula in numbers or words, such as: 6*77+pi+sqrt(e^3)/888 plus 0.45. The user is given the option to search for the formula, after calculation. The calculator also uses the unit and currency conversion functions to allow unit-aware calculations. For example, "(3 EUR/liter)/(40 miles/gallon) in USD/mile" calculates the dollar cost per mile for a 40 mpg car with gas costing 3 euros a liter. The caret "^" raises a number to an exponent power, and percentages are allowed ("40% of 300").[8] There is also some debate as to Google's calculation of 0^0. Many mathematicians believe that 0^0 is undefined but Google's calculator shows the result as 1.[27]
- numeric ranges – A set of numbers can be matched by using a double-dot between range numbers (70..73 or 90..100) to match any positive number in the range, inclusive.[8] Negative numbers are treated as using exclusion-dash to not match the number.
- dictionary lookup – A definition for a word or phrase can be found,[7] by entering "define" followed by a colon and the word(s) to lookup (such as, "define:philosophy")
- maps – Some related maps can be displayed,[7] by typing in the name or U.S. ZIP code of a location and the word "map" (such as: New York map; Kansas map; or Paris map).
- movie showtimes – Reviews or film showtimes can be listed for any movies playing nearby,[7] by typing "movies" or the name of any current film into the search box. If a specific location was saved on a previous search, the top search result will display showtimes for nearby theaters for that movie.
- public data – Trends for population (or unemployment rates)[7] can be found for U.S. states & counties, by typing "population" or "unemployment rate" followed by a state or county name.
- real estate and housing – Home listings in a given area can be displayed,[7] using the trigger words "housing", "home", or "real estate" followed by the name of a city or U.S. zip code.
- travel data/airports – The flight status for arriving or departing U.S. flights can be displayed,[7] by typing in the name of the airline and the flight number into the search box (such as: American airlines 18). Delays at a specific airport can also be viewed (by typing the name of the city or three-letter airport code plus word "airport").
- package tracking – Package mail can be tracked[7] by typing the tracking number of a Royal Mail, UPS, FedEx or USPS package directly into the search box. Results will include quick links to track the status of each shipment.
- patent numbers – U.S. patents can be searched[7][26] by entering the word "patent" followed by the patent number into the search box (such as: Patent 5123123).
- area code – The geographical location (for any U.S. telephone area code)[7] can be displayed by typing a 3-digit area code (such as: 650).
- synonym search – A search can match words similar to those specified,[7] by placing the tilde sign (~) immediately in front of a search term, such as: ~fast food.

Search options

The webpages maintained by the Google Help Center have text describing more than 15 various search options.[28] The Google operators:

- OR – Search for either one, such as "price high OR low" searches for *"price"* with *"high"* or *"low"*.
- "-" – Search while excluding a word, such as "apple -tree" searches where word *"tree"* is not used.
- "+" – (Removed on 10/19/11[19]) Force inclusion of a word, such as "Name+of+the Game" to require the words *"of"* & *"the"* to appear on a matching page.
- "*" – Wildcard operator to match any words between other specific words. Some of the

query options are as follows:

- define: – The query prefix "define:" will provide a definition[28] of the words listed after it.
- stocks: – After "stocks:" the query terms are treated as stock ticker symbols[28] for lookup.
- site: – Restrict the results to those websites in the given domain,[28] such as, site:www.acmeacme.com. The option "site:com" will search all domain URLs named with ".com" (no space after "site:").
- allintitle: – Only the page titles are searched[28] (not the remaining text on each webpage).
- intitle: – Prefix to search in a webpage title,[28] such as "intitle:googlesearch" will list pages with word "google" in title, and word "search" anywhere (no space after "intitle:").
- allinurl: – Only the page URL address lines are searched[28] (not the text inside each webpage).
- inurl: – Prefix for each word to be found in the URL;[28] others words are matched anywhere, such as "inurl:acme search" matches "acme" in a URL, but matches "search" anywhere (no space after "inurl:").

The page-display options (or query types) are:

- cache: – Highlights the search-words within the cached document, such as "cache:www.google.com xxx" shows cached content with word "xxx" highlighted.
- link: – The prefix "link:" will list webpages that have links to the specified webpage, such as "link:www.google.com" lists webpages linking to the Google homepage.
- related: – The prefix "related:" will list webpages that are "similar" to a specified web page.
- info: – The prefix "info:" will display some background information about one specified webpage, such as, info:www.google.com. Typically, the info is the first text (160 bytes, about 23 words) contained in the page, displayed in the style of a results entry (for just the 1 page as matching the search).
- filetype: – results will only show files of the desired type (ex filetype:pdf will return pdf files)

Error messages

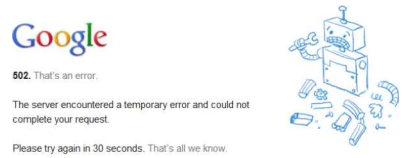
Some searches will give a 403 Forbidden error with the text "We're sorry...

```
... but your query looks similar to automated requests from a computer
virus or spyware application. To protect our users, we can't process
your request right now.
```

```
 the meantime, if you suspect that your computer or network has been
infected, you might want to run a virus checker or spyware remover to
make sure that your systems are free of viruses and other spurious
software.
```

We apologize for the inconvenience, and hope we'll see you again on Google." sometimes followed by a CAPTCHA prompt.[29]

The screen was first reported in 2005, and was a response to the heavy use of Google by search engine optimization companies to check on ranks of sites they were optimizing. The message is triggered by high volumes of requests from a single IP address. Google apparently uses the Google cookie as part of its determination of refusing service.[29]



Google's Server Error page

In June 2009, after the death of pop superstar Michael Jackson, this message appeared to many internet users who were searching Google for news stories related to the singer, and was assumed by Google to be a DDoS attack, although many queries were submitted by legitimate searchers.

January 2009 malware bug

Google flags search results with the message "This site may harm your computer" if the site is known to install malicious software in the background or otherwise surreptitiously. Google does this to protect users against visiting sites that could harm their computers. For approximately 40 minutes on January 31, 2009, *all* search results were mistakenly classified as malware and could therefore not be clicked; instead a warning message was displayed and the user was required to enter the requested URL manually. The bug was caused by human error.[30][31][32][33] The URL of "/" (which expands to all URLs) was mistakenly added to the malware patterns file.[31][32]

Google Doodles

On certain occasions, the logo on Google's webpage will change to a special version, known as a "Google Doodle". Clicking on the Doodle links to a string of Google search results about the topic. The first was a reference to the Burning Man Festival in 1998,[34][35] and others have been produced for the birthdays of notable people like Albert Einstein, historical events like the interlocking Lego block's 50th anniversary and holidays like Valentine's Day.[36] Some Google Doodles have interactivity beyond a simple search, such as the famous "Google Pacman" version that appeared on May 21,2010.

Google Caffeine

In August 2009, Google announced the rollout of a new search architecture, codenamed "Caffeine".[37] The new architecture was designed to return results faster and to better deal with rapidly updated information[38] from services including Facebook and Twitter.[37] Google developers noted that most users would notice little immediate change, but invited developers to test the new search in its sandbox.[39] Differences noted for their impact upon search engine optimization included heavier keyword weighting and the importance of the domain's age.[40][41] The move was interpreted in some quarters as a response to Microsoft's recent release of an upgraded version of its own search service, renamed Bing.[42] Google announced completion of Caffeine on 8 June 2010, claiming 50% fresher results due to continuous updating of its index.[43] With Caffeine, Google moved its back-end indexing system away from MapReduce and onto BigTable, the company's distributed database platform.[44] Caffeine is also based on Colossus, or GFS2,[45] an overhaul of the GFS distributed file system.[46]

Privacy

Searches made by search engines, including Google, leave traces, raising concerns about privacy but sometimes facilitating the administration of justice; murderers have been detected and convicted as a result of incriminating searches they made such as "tips with killing with a baseball bat".[47].

A search can be traced in several ways. When using a search engine through a browser program on a computer, search terms and other information will usually be stored on the computer by default, unless steps are taken to erase them. An Internet Service Provider may store records which relate search terms to an IP address and a time. The search engine provider (e.g., Google) may keep logs with the same information[48]. Whether such logs are kept, and access to them by law enforcement agencies, is subject to legislation and working practices; the law may mandate, prohibit, or say nothing about logging of various types of information.

The technically knowledgeable and forewarned user can avoid leaving traces.

Encrypted Search

In May 2010 Google rolled out SSL-encrypted web search.[49] The encrypted search can be accessed at `encrypted.google.com`[50]

Instant Search

Google Instant, a feature that displays suggested results while the user types, was introduced in the United States on September 8, 2010. In concert with the Google Instant launch, Google disabled the ability of users to choose to see more than 10 search results per page. At the time of the announcement Google expected Instant to save users 2 to 5 seconds in every search, collectively about 11 million seconds per hour.[51] Search engine marketing pundits speculate that Google Instant will have a great impact on local and paid search.[52]

Instant Search can be disabled via Google's "preferences" menu, but autocomplete-style search suggestions now cannot be disabled; Google confirm that this is intentional.[53]

The publication 2600: The Hacker Quarterly has compiled a list of words that are restricted by Google Instant.[54] These are terms the web giant's new instant search feature will not search.[55][56] Most terms are often vulgar and derogatory in nature, but some apparently irrelevant searches including "Myleak" are removed.[56]

Redesign

In late June 2011, Google introduced a new look to the Google home page in order to boost the use of the Google+ social tools.[57]

One of the major changes was replacing the classic navigation bar with a black one. Google's digital creative director Chris Wiggins explains: "We're working on a project to bring you a new and improved Google experience, and over the next few months, you'll continue to see more updates to our look and feel."[58] The new navigation bar has been negatively received by a vocal minority.[59]

International

Google is available in many languages and has been localized completely or partly for many countries.[60] The interface has also been made available in some languages for humorous purpose:

- Bork, bork, bork!
- Elmer Fudd
- Leetspeak
- Klingon
- Pig Latin
- Pirate

In addition to the main URL Google.com, Google Inc. owns 160 domain names for each of the countries/regions in which it has been localized.[60]

Search products

In addition to its tool for searching webpages, Google also provides services for searching images, Usenet newsgroups, news websites, videos, searching by locality, maps, and items for sale online. In 2006, Google has indexed over 25 billion web pages,[61] 400 million queries per day,[61] 1.3 billion images, and over one billion Usenet messages. It also caches much of the content that it indexes. Google operates other tools and services including Google News, Google Suggest, Google Product Search, Google Maps, Google Co-op, Google Earth, Google Docs, Picasa, Panoramio, YouTube, Google Translate, Google Blog Search and Google Desktop Search.

There are also products available from Google that are not directly search-related. Gmail, for example, is a webmail application, but still includes search features; Google Browser Sync does not offer any search facilities, although it aims to organize your browsing time.

Also Google starts many new beta products, like Google Social Search or Google Image Swirl.

Energy consumption

Google claims that a search query requires altogether about 1 kJ or 0.0003 kW·h.[62]

References

[1] "WHOIS - google.com" (http://reports.internic.net/cgi/whois?whois_nic=google.com&type=domain). . Retrieved 2009-01-27.

[2] "Google.com Site Info" (http://www.alexa.com/siteinfo/google.com). Alexa Internet. . Retrieved 2012-03-02.

[3] "Alexa Search Engine ranking" (http://www.alexa.com/siteinfo/google.com+yahoo.com+altavista.com). . Retrieved 2009-11-15.

[4] "Almost 12 Billion U.S. Searches Conducted in July" (http://searchenginewatch.com/showPage.html?page=3630718). SearchEngineWatch. 2008-09-02. .

[5] *...The *, or wildcard, is a little-known feature that can be very powerful...* (http://www.google.co.nz/support/websearch/bin/answer.py?answer=136861)

[6] "WHOIS - google.com" (http://reports.internic.net/cgi/whois?whois_nic=google.com&type=domain). . Retrieved 2009-01-27.

[7] "Search Features" (http://www.google.com/intl/en/help/features.html). Google.com. May 2009. .

[8] "Google Help : Cheat Sheet" (http://www.google.com/help/cheatsheet.html). Google. 2010. .

[9] *Voice Search for Google.com - Just click the mic and say your search. And, Search Google by giving Image* (http://qualitypoint.blogspot.com/2011/06/voice-search-for-googlecom.html)

[10] Hubbard, Douglas (2011). *Pulse: The New Science of Harnessing Internet Buzz to Track Threats and Opportunities*. John Wiley & Sons.

[11] Brin, S.; Page, L. (1998). "The anatomy of a large-scale hypertextual Web search engine" (http://infolab.stanford.edu/pub/papers/ google.pdf). *Computer Networks and ISDN Systems* **30**: 107–117. doi:10.1016/S0169-7552(98)00110-X. ISSN 0169-7552. .

[12] "Corporate Information: Technology Overview" (http://www.google.com/corporate/tech.html). Google. . Retrieved 2009-11-15. Wired.com (http://www.wired.com/magazine/2010/02/ff_google_algorithm/)

[13] "Google Frequently Asked Questions - File Types" (http://www.google.com/help/faq_filetypes.html#what). Google. . Retrieved 2011-09-12.

[14] Sherman, Chris and Price, Gary. "The Invisible Web: Uncovering Sources Search Engines Can't See, In: *Library Trends* 52 (2) 2003: Organizing the Internet:" (http://hdl.handle.net/2142/8528). pp. 282–298. .

[15] "Google Webmaster Guidelines" (http://www.google.com/webmasters/guidelines.html). Google. . Retrieved 2009-11-15.

[16] Segal, David (November 26, 2010). "A Bully Finds a Pulpit on the Web" (https://www.nytimes.com/2010/11/28/business/28borker. html). *The New York Times*. . Retrieved November 27, 2010.

[17] Blogspot.com (http://googleblog.blogspot.com/2010/12/being-bad-to-your-customers-is-bad-for.html)

[18] "Top 500" (http://www.alexa.com/site/ds/top_sites?ts_mode=global&lang=none). Alexa. . Retrieved 2008-04-15.

[19] (http://www.frag.co.uk/blog/2011/10/googles-changes-the-operators/), Google changes the operators.

[20] Google.com (http://www.google.com/support/websearch/bin/answer.py?answer=136861)

[21] "Google:Stemming" (http://www.google.com/support/bin/answer.py?answer=35889#stemming). Google. .

[22] "I'm feeling lucky(button costs Google $110 million per year" (http://valleywag.com/tech/google/ im-feeling-lucky-button-costs-google-110-million-per-year-324927.php). Valleywag. 2007. . Retrieved 2008-01-19.

[23] "Google's New Homepage Motto: 'This Space Intentionally Left Blank'" (http://digitaldaily.allthingsd.com/20091030/goog-page/). WallStreetJournal. 2009. . Retrieved 2009-11-17.

[24] Google.com (http://www.google.com/pacman)

[25] Goel, Kavi; Ramanathan V. Guha, Othar Hansson (2009-05-12). "Introducing Rich Snippets" (http://googlewebmastercentral.blogspot. com/2009/05/introducing-rich-snippets.html). *Google Webmaster Central Blog*. Google. . Retrieved 2009-05-25.

[26] "Google and Search Engines"(http://www.law.emory.edu/law-library/research/advanced-legal-research-class/ finding-aids-and-searching/google.html). Emory University Law School. 2006. .

[27] Google.com (http://www.google.com/search?output=&sitesearch=&hl=en&q=2+2&submit=Search+the+Web#hl=en& expIds=17259,22104,25907,26637,26659,26741,26817,26992,27095&sugexp=ldymls&xhr=t&q=0^0&cp=3&pf=p&sclient=psy&aq=f& aqi=g4g-o1&aql=&oq=0^0&gs_rfai=&pbx=1&fp=433548e9226de17c)

[28] "Google Help Center – Alternate query types", 2009, webpage: G-help (http://www.google.com/help/operators.html).

[29] "Google error page" (http://www.google.com/support/bin/answer.py?answer=15661). . Retrieved 2008-12-31.

[30] Krebs, Brian (2009-01-31). "Google: This Internet May Harm Your Computer" (http://voices.washingtonpost.com/securityfix/2009/01/ google_this_internet_will_harm.html?hpid=news-col-blog). The Washington Post. . Retrieved 2009-01-31.

[31] Mayer, Marissa (2009-01-31). "This site may harm your computer on every search result?!?!" (http://googleblog.blogspot.com/2009/01/ this-site-may-harm-your-computer-on.html). *The Official Google Blog*. Google. . Retrieved 2009-01-31.

[32] Weinstein, Maxim (2009-1-31). "Google glitch causes confusion" (http://blog.stopbadware.org/2009/01/31/ google-glitch-causes-confusion). StopBadware.org. . Retrieved 2010-5-10.

[33] Cooper, Russ (January 31, 2009). "Serious problems with Google search" (http://securityblog.verizonbusiness.com/2009/01/31/ serious-problems-with-google-search/). Verizon Business Security Blog. . Retrieved 2010-5-10.

[34] Hwang, Dennis (June 8, 2004). "Oodles of Doodles" (http://googleblog.blogspot.com/2004/06/oodles-of-doodles.html). Google (corporate blog). . Retrieved July 19, 2006.

[35] "Doodle History" (http://www.google.com/doodle4google/history.html). Google, Inc.. . Retrieved 5-10-2010.

[36] "Google logos: *Valentine's Day logo*" (http://www.google.com/logos/valentine07.gif). February 14, 2007. . Retrieved April 6, 2007.

[37] Harvey, Mike (11 August 2009). "Google unveils new "Caffeine" search engine" (http://technology.timesonline.co.uk/tol/news/ tech_and_web/personal_tech/article6792403.ece). London: The Times. . Retrieved 14 August 2009.

[38] "What Does Google "Caffeine" Mean for My Website?" (http://www.siivo.com/blog/2010/07/ what-does-google-caffeine-mean-for-my-website). New York: Siivo Corp. 21 July 2010. . Retrieved 21 July 2010.

[39] Culp, Katie (12 August 2009). "Google introduces new "Caffeine" search system" (http://www.foxbusiness.com/story/markets/ industries/technology/google-introduces-new-caffeine-search/). Fox News. . Retrieved 14 August 2009.

[40] Martin, Paul (31 July 2009). "Bing - The new Search Engine from Microsoft and Yahoo" (http://blog.cube3marketing.com/2009/07/31/ bing-the-new-search-engine-from-microsoft-and-yahoo/). Cube3 Marketing. . Retrieved 12 January 2010.

[41] Martin, Paul (27 August 2009). "Caffeine - The New Google Update" (http://blog.cube3marketing.com/2009/08/27/ caffeine-the-new-google-update/). Cube3 Marketing. . Retrieved 12 January 2010.

[42] Barnett, Emma (11 August 2009). "Google reveals caffeine: a new faster search engine" (http://www.telegraph.co.uk/technology/ google/6009176/Google-reveals-caffeine-a-new-faster-search-engine.html). The Telegraph. . Retrieved 14 August 2009.

[43] Grimes, Carrie (8 June 2010). "Our new search index: Caffeine" (http://googleblog.blogspot.com/2010/06/ our-new-search-index-caffeine.html). *The Official Google Blog*. . Retrieved 18 June 2010.

[44] Google search index splits with MapReduce (http://www.theregister.co.uk/2010/09/09/google_caffeine_explained/) – The Register

[45] Google Caffeine: What it really is (http://www.theregister.co.uk/2009/08/14/google_caffeine_truth/) – The Register

[46] Google File System II: Dawn of the Multiplying Master Nodes (http://www.theregister.co.uk/2009/08/12/ google_file_system_part_deux/) – The Register

[47] Search Engine Land: Once Again, A Google Murder Case, 29 Jan 2008(http://searchengineland.com/ once-again-a-google-murder-case-13241)

[48] Search Engine Land: Google Anonymizing Search Records To Protect Privacy, 14 March 2007 (http://searchengineland.com/ google-anonymizing-search-records-to-protect-privacy-10736)

[49] "SSL Search: Features - Web Search Help" (http://www.google.com/support/websearch/bin/answer.py?answer=173733&hl=en). *Web Search Help*. Google. May 2010. . Retrieved 2010-07-07.
[50] Encrypted.google.com (http://encrypted.google.com)
[51] Peter Nowak (2010). *Tech Bytes: Google Instant* (Television production). United States: ABC News.
[52] "How Google Saved $100 Million By Launching Google Instant" (http://searchengineland.com/ how-google-saved-100-million-by-launching-google-instant-51270). . Retrieved 20 September 2010.
[53] Google Web Search Help Forum (http://www.google.com/support/forum/p/Web+Search/thread?tid=5a69f1094357f31b&hl=en) (WebCite archive (http://www.webcitation.org/5sy2ImYdO))
[54] 2600.com: Google Blacklist - Words That Google Instant Doesn't Like (http://www.2600.com/googleblacklist/)
[55] CNN: Which words does Google Instant blacklist? (http://www.cnn.com/2010/TECH/web/09/29/google.instant.blacklist.mashable/ index.html?eref=mrss_igoogle_cnn)
[56] The Huffington Post: Google Instant Censorship: The Strangest Terms Blacklisted By Google (http://www.huffingtonpost.com/2010/09/ 29/google-instant-censorship_n_743203.html)
[57] Boulton, Clint. "Google Redesign Backs Social Effort" (http://www.eweekeurope.co.uk/comment/ google-redesign-backs-social-effort-32954). *eWeek Europe*. eWeek Europe. . Retrieved 1 July 2011.
[58] Google redesigns its homepage [[Los Angeles Times (http://latimesblogs.latimes.com/technology/2011/06/ google-redesigns-its-homepage-with-new-black-bar-up-top-google-social-network.html)]]
[59] Google support forum, one of many threads on being unable to switch off the black navigation bar (http://www.google.com/support/ forum/p/Web+Search/thread?tid=7ddbf7a4c8fa04a9&hl=en)
[60] Language Tools (http://www.google.com/language_tools?hl=en)
[61] Google, Web Crawling and Distributed Synchronization (http://www.seas.upenn.edu/~zives/cis555/slides/I-Crawlers-Sync.ppt) p. 11.
[62] Blogspot.com (http://googleblog.blogspot.com/2009/01/powering-google-search.html), Powering a Google search

Further reading

- *Google Hacks* from O'Reilly is a book containing tips about using Google effectively. Now in its third edition. ISBN 0-596-52706-3.
- *Google: The Missing Manual* by Sarah Milstein and Rael Dornfest (O'Reilly, 2004). ISBN 0-596-00613-6
- *How to Do Everything with Google* by Fritz Schneider, Nancy Blachman, and Eric Fredricksen (McGraw-Hill Osborne Media, 2003). ISBN 0-07-223174-2
- *Google Power* by Chris Sherman (McGraw-Hill Osborne Media, 2005). ISBN 0-07-225787-3
- Barroso, Luiz Andre; Dean, Jeffrey; Hölzle, Urs (2003). "Web Search for a Planet: The Google Cluster Architecture". *IEEE Micro* **23** (2): 22–28. doi:10.1109/MM.2003.1196112.

External links

- Google.com (http://www.google.com)
- The Original Google! (http://web.archive.org/web/19981111183552/google.stanford.edu/)

Cloaking

Cloaking is a search engine optimization (SEO) technique in which the content presented to the search engine spider is different from that presented to the user's browser. This is done by delivering content based on the IP addresses or the User-Agent HTTP header of the user requesting the page. When a user is identified as a search engine spider, a server-side script delivers a different version of the web page, one that contains content not present on the visible page, or that is present but not searchable. The purpose of cloaking is sometimes to deceive search engines so they display the page when it would not otherwise be displayed (black hat SEO). However, it can also be a functional (though antiquated) technique for informing search engines of content they would not otherwise be able to locate because it is embedded in non-textual containers such as video or certain Adobe Flash components.

As of 2006, better methods of accessibility, including progressive enhancement are available, so cloaking is not considered necessary by proponents of that method. Cloaking is often used as a spamdexing technique, to try to trick search engines into giving the relevant site a higher ranking; it can also be used to trick search engine users into visiting a site based on the search engine description which site turns out to have substantially different, or even pornographic content. For this reason, major search engines consider cloaking for deception to be a violation of their guidelines, and therefore, they delist sites when deceptive cloaking is reported.[1][2][3][4]

Cloaking is a form of the doorway page technique.

A similar technique is also used on the Open Directory Project web directory. It differs in several ways from search engine cloaking:

- It is intended to fool human editors, rather than computer search engine spiders.
- The decision to cloak or not is often based upon the HTTP referrer, the user agent or the visitor's IP; but more advanced techniques can be also based upon the client's behaviour analysis after a few page requests: the raw quantity, the sorting of, and latency between subsequent HTTP requests sent to a website's pages, plus the presence of a check for robots.txt file, are some of the parameters in which search engines spiders differ heavily from a natural user behaviour. The referrer tells the URL of the page on which a user clicked a link to get to the page. Some cloakers will give the fake page to anyone who comes from a web directory website, since directory editors will usually examine sites by clicking on links that appear on a directory web page. Other cloakers give the fake page to everyone *except* those coming from a major search engine; this makes it harder to detect cloaking, while not costing them many visitors, since most people find websites by using a search engine.

Black hat perspective

Increasingly, for a page without natural popularity due to compelling or rewarding content to rank well in the search engines, webmasters may be tempted to design pages solely for the search engines. This results in pages with too many keywords and other factors that might be search engine "friendly", but make the pages difficult for actual visitors to consume. As such, black hat SEO practitioners consider cloaking to be an important technique to allow webmasters to split their efforts and separately target the search engine spiders and human visitors.

In September 2007, Ralph Tegtmeier and Ed Purkiss coined the term "mosaic cloaking" whereby dynamic pages are constructed as tiles of content and only portions of the pages, javascript and CSS are changed, simultaneously decreasing the contrast between the cloaked page and the "friendly" page while increasing the capability for targeted delivery of content to various spiders and human visitors.

Cloaking versus IP delivery

IP delivery can be considered a more benign variation of cloaking, where different content is served based upon the requester's IP address. With cloaking, search engines and people never see the other's pages, whereas, with other uses of IP delivery, both search engines and people can see the same pages. This technique is sometimes used by graphics-heavy sites that have little textual content for spiders to analyze.

One use of IP delivery is to determine the requestor's location, and deliver content specifically written for that country. This isn't necessarily cloaking. For instance, Google uses IP delivery for AdWords and AdSense advertising programs to target users in different geographic locations.

IP delivery is a crude and unreliable method of determining the language in which to provide content. Many countries and regions are multi-lingual, or the requestor may be a foreign national. A better method of content negotiation is to examine the client's `Accept-Language` HTTP header.

As of 2006, many sites have taken up IP delivery to personalise content for their regular customers. Many of the top 1000 sites, including sites like Amazon (amazon.com), actively use IP delivery. None of these have been banned from search engines as their intent is not deceptive.

Notes

[1] "Ask.com Editorial Guidelines" (http://about.ask.com/en/docs/about/editorial_guidelines.shtml). About.ask.com. . Retrieved 2012-02-20.
[2] "Google's Guidelines on SEOs" (http://www.google.com/webmasters/seo.html). Google.com. 2012-01-24. . Retrieved 2012-02-20.
[3] "Google's Guidelines on Site Design" (http://www.google.com/webmasters/guidelines.html). Google.com. . Retrieved 2012-02-20.
[4] Yahoo! Search Content Quality Guidelines (http://help.yahoo.com/help/us/ysearch/deletions/deletions-05.html)

References

- Baoning Wu and Brian D. Davison: " Cloaking and Redirection: A Preliminary Study (http://airweb.cse.lehigh. edu/2005/wu.pdf)". Workshop on Adversarial Information Retrieval on the Web, Chiba, Japan, 2005.

Web search engine

A **web search engine** is designed to search for information on the World Wide Web and FTP servers. The search results are generally presented in a list of results often referred to as SERPS, or "search engine results pages". The information may consist of web pages, images, information and other types of files. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain real-time information by running an algorithm on a web crawler.

History

Timeline (full list)		
Year	**Engine**	**Current status**
1993	W3Catalog	Inactive
	Aliweb	Inactive
1994	WebCrawler	Active, Aggregator
	Go.com	Active, Yahoo Search
	Lycos	Active
1995	AltaVista	Inactive(URLredirectedtoYahoo!)
	Daum	Active
	Magellan	Inactive
	Excite	Active
	SAPO	Active
	Yahoo!	Active, Launched as a directory
1996	Dogpile	Active, Aggregator
	Inktomi	Acquired by Yahoo!
	HotBot	Active (lycos.com)
	Ask Jeeves	Active(ask.com,Jeeveswentaway)
1997	Northern Light	Inactive
	Yandex	Active
1998	Google	Active
	MSN Search	Active as Bing
1999	AlltheWeb	Inactive(URLredirectedtoYahoo!)
	GenieKnows	Active, rebranded Yellowee.com
	Naver	Active
	Teoma	Active
	Vivisimo	Inactive
2000	Baidu	Active
	Exalead	Acquired by Dassault Systèmes
2002	Inktomi	Acquired by Yahoo!
2003	Info.com	Active

2004	Yahoo! Search	Active, Launched own web search (see Yahoo! Directory, 1995)
	A9.com	Inactive
	Sogou	Active
2005	AOL Search	Active
	Ask.com	Active
	GoodSearch	Active
	SearchMe	Closed
2006	wikiseek	Inactive
	Quaero	Active
	Ask.com	Active
	Live Search	Active as Bing, Launched as rebranded MSN Search
	ChaCha	Active
	Guruji.com	Active
2007	wikiseek	Inactive
	Sproose	Inactive
	Wikia Search	Inactive
	Blackle.com	Active
2008	Powerset	Inactive (redirects to Bing)
	Picollator	Inactive
	Viewzi	Inactive
	Boogami	Inactive
	LeapFish	Inactive
	Forestle	Inactive (redirects to Ecosia)
	VADLO	Active
	Duck Duck Go	Active, Aggregator
2009	Bing	Active, Launched as rebranded Live Search
	Yebol	Active
	Megafore	Active
	Mugurdy	Inactive due to a lack of funding
	Goby	Active
2010	Black Google Mobile	Active
	Blekko	Active
	Cuil	Inactive
	Yandex	Active, Launched global (English) search
	Yummly	Active
2011	Interred	Active
2012	Volunia	Active , only *Power User*

During the early development of the web, there was a list of webservers edited by Tim Berners-Lee and hosted on the CERN webserver. One historical snapshot from 1992 remains.[1] As more webservers went online the central list could not keep up. On the NCSA site new servers were announced under the title "What's New!"[2]

The very first tool used for searching on the Internet was Archie.[3] The name stands for "archive" without the "v". It was created in 1990 by Alan Emtage, Bill Heelan and J. Peter Deutsch, computer science students at McGill University in Montreal. The program downloaded the directory listings of all the files located on public anonymous FTP (File Transfer Protocol) sites, creating a searchable database of file names; however, Archie did not index the contents of these sites since the amount of data was so limited it could be readily searched manually.

The rise of Gopher (created in 1991 by Mark McCahill at the University of Minnesota) led to two new search programs, Veronica and Jughead. Like Archie, they searched the file names and titles stored in Gopher index systems. Veronica (*V*ery *E*asy *R*odent-*O*riented *N*et-wide *I*ndex to *C*omputerized *A*rchives) provided a keyword search of most Gopher menu titles in the entire Gopher listings. Jughead (*J*onzy's *U*niversal *G*opher *H*ierarchy *E*xcavation *A*nd *D*isplay) was a tool for obtaining menu information from specific Gopher servers. While the name of the search engine "Archie" was not a reference to the Archie comic book series, "Veronica" and "Jughead" are characters in the series, thus referencing their predecessor.

In the summer of 1993, no search engine existed yet for the web, though numerous specialized catalogues were maintained by hand. Oscar Nierstrasz at the University of Geneva wrote a series of Perl scripts that would periodically mirror these pages and rewrite them into a standard format which formed the basis for W3Catalog, the web's first primitive search engine, released on September 2, 1993.[4]

In June 1993, Matthew Gray, then at MIT, produced what was probably the first web robot, the Perl-based World Wide Web Wanderer, and used it to generate an index called 'Wandex'. The purpose of the Wanderer was to measure the size of the World Wide Web, which it did until late 1995. The web's second search engine Aliweb appeared in November 1993. Aliweb did not use a web robot, but instead depended on being notified by website administrators of the existence at each site of an index file in a particular format.

JumpStation (released in December 1993[5]) used a web robot to find web pages and to build its index, and used a web form as the interface to its query program. It was thus the first WWW resource-discovery tool to combine the three essential features of a web search engine (crawling, indexing, and searching) as described below. Because of the limited resources available on the platform on which it ran, its indexing and hence searching were limited to the titles and headings found in the web pages the crawler encountered.

One of the first "full text" crawler-based search engines was WebCrawler, which came out in 1994. Unlike its predecessors, it let users search for any word in any webpage, which has become the standard for all major search engines since. It was also the first one to be widely known by the public. Also in 1994, Lycos (which started at Carnegie Mellon University) was launched and became a major commercial endeavor.

Soon after, many search engines appeared and vied for popularity. These included Magellan, Excite, Infoseek, Inktomi, Northern Light, and AltaVista. Yahoo! was among the most popular ways for people to find web pages of interest, but its search function operated on its web directory, rather than full-text copies of web pages. Information seekers could also browse the directory instead of doing a keyword-based search.

In 1996, Netscape was looking to give a single search engine an exclusive deal to be the featured search engine on Netscape's web browser. There was so much interest that instead a deal was struck with Netscape by five of the major search engines, where for $5 million per year each search engine would be in rotation on the Netscape search engine page. The five engines were Yahoo!, Magellan, Lycos, Infoseek, and Excite.[6][7]

Search engines were also known as some of the brightest stars in the Internet investing frenzy that occurred in the late 1990s.[8] Several companies entered the market spectacularly, receiving record gains during their initial public offerings. Some have taken down their public search engine, and are marketing enterprise-only editions, such as Northern Light. Many search engine companies were caught up in the dot-com bubble, a speculation-driven market

boom that peaked in 1999 and ended in 2001.

Around 2000, Google's search engine rose to prominence. The company achieved better results for many searches with an innovation called PageRank. This iterative algorithm ranks web pages based on the number and PageRank of other web sites and pages that link there, on the premise that good or desirable pages are linked to more than others. Google also maintained a minimalist interface to its search engine. In contrast, many of its competitors embedded a search engine in a web portal.

By 2000, Yahoo! was providing search services based on Inktomi's search engine. Yahoo! acquired Inktomi in 2002, and Overture (which owned AlltheWeb and AltaVista) in 2003. Yahoo! switched to Google's search engine until 2004, when it launched its own search engine based on the combined technologies of its acquisitions.

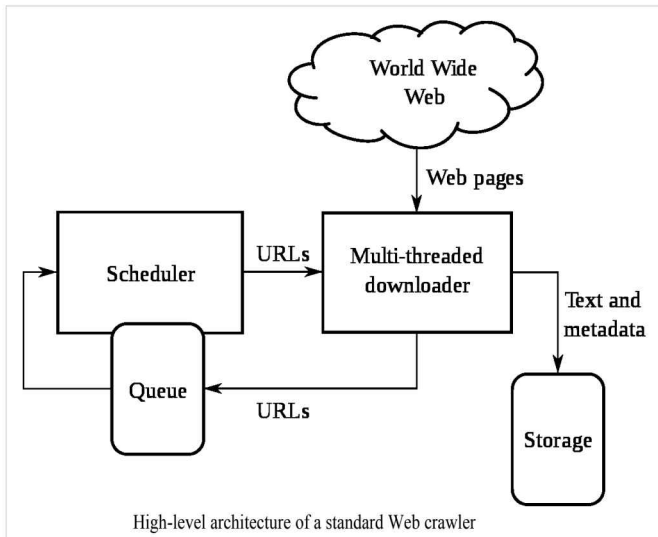
Microsoft first launched MSN Search in the fall of 1998 using search results from Inktomi. In early 1999 the site began to display listings from Looksmart blended with results from Inktomi except for a short time in 1999 when results from AltaVista were used instead. In 2004, Microsoft began a transition to its own search technology, powered by its own web crawler (called msnbot).

Microsoft's rebranded search engine, Bing, was launched on June 1, 2009. On July 29, 2009, Yahoo! and Microsoft finalized a deal in which Yahoo! Search would be powered by Microsoft Bing technology.

How web search engines work

A search engine operates in the following order:

1. Web crawling
2. Indexing
3. Searching



High-level architecture of a standard Web crawler

Web search engines work by storing information about many web pages, which they retrieve from the HTML itself. These pages are retrieved by a Web crawler (sometimes also known as aspider) — an automated Web browser which follows every link on the site. Exclusions can be made by the use of robots.txt. The contents of each page are then analyzed to determine how it should be indexed (for example, words are extracted from the titles, headings, or special fields called meta tags). Data about web pages are stored in an index database for use in later queries. A query can be a single word. The purpose of an index is to allow information to be found as quickly as possible. Some search engines, such as Google, store all or part of the source page (referred to as a cache) as well as information about the web pages, whereas others, such as AltaVista, store every word of every page they find. This cached page always holds the actual search text since it is the one that was actually indexed, so it can be very useful when the content of the current page has been updated and the search terms are no longer in it. This problem might be considered to be a mild form of linkrot, and Google's handling of it increases usability by satisfying user expectations that the search terms will be on the returned webpage. This satisfies the principle of least astonishment since the user normally expects the search terms to be on the returned pages. Increased search relevance makes these cached pages very useful, even beyond the fact that they may contain data that may no longer be available elsewhere.

When a user enters a query into a search engine (typically by using keywords), the engine examines its index and provides a listing of best-matching web pages according to its criteria, usually with a short summary containing the document's title and sometimes parts of the text. The index is built from the information stored with the data and the

method by which the information is indexed. Unfortunately, there are currently no known public search engines that allow documents to be searched by date. Most search engines support the use of the boolean operators AND, OR and NOT to further specify the search query. Boolean operators are for literal searches that allow the user to refine and extend the terms of the search. The engine looks for the words or phrases exactly as entered. Some search engines provide an advanced feature called proximity search which allows users to define the distance between keywords. There is also concept-based searching where the research involves using statistical analysis on pages containing the words or phrases you search for. As well, natural language queries allow the user to type a question in the same form one would ask it to a human. A site like this would be ask.com.

The usefulness of a search engine depends on the relevance of the **result set** it gives back. While there may be millions of web pages that include a particular word or phrase, some pages may be more relevant, popular, or authoritative than others. Most search engines employ methods to rank the results to provide the "best" results first. How a search engine decides which pages are the best matches, and what order the results should be shown in, varies widely from one engine to another. The methods also change over time as Internet usage changes and new techniques evolve. There are two main types of search engine that have evolved: one is a system of predefined and hierarchically ordered keywords that humans have programmed extensively. The other is a system that generates an "inverted index" by analyzing texts it locates. This second form relies much more heavily on the computer itself to do the bulk of the work.

Most Web search engines are commercial ventures supported by advertising revenue and, as a result, some employ the practice of allowing advertisers to pay money to have their listings ranked higher in search results. Those search engines which do not accept money for their search engine results make money by running search related ads alongside the regular search engine results. The search engines make money every time someone clicks on one of these ads.

Market share

Search engine	Market share in May 2011		Market share in December 2010[9]	
Google	82.80%		84.65%	
Yahoo!	6.42%		6.69%	
Baidu	4.89%		3.39%	
Bing	3.91%		3.29%	
Ask	0.52%		0.56%	
AOL	0.36%		0.42%	

Google's worldwide market share peaked at 86.3% in April 2010.[10] Yahoo!, Bing and other search engines are more popular in the US than in Europe.

According to Hitwise, market share in the U.S. for October 2011 was Google 65.38%, Bing-powered (Bing and Yahoo!) 28.62%, and the remaining 66 search engines 6%. However, an Experian Hit wise report released in August 2011 gave the "success rate" of searches sampled in July. Over 80 percent of Yahoo! and Bing searches resulted in the users visiting a web site, while Google's rate was just under 68 percent.[11] [12]

In the People's Republic of China, Baidu held a 61.6% market share for web search in July 2009.[13]

Search engine bias

Although search engines are programmed to rank websites based on their popularity and relevancy, empirical studies indicate various political, economic, and social biases in the information they provide.[14][15] These biases could be a direct result of economic and commercial processes (e.g., companies that advertise with a search engine can become also more popular in its organic search results), and political processes (e.g., the removal of search results in order to comply with local laws).[16] Google Bombing is one example of an attempt to manipulate search results for political, social or commercial reasons.

References

[1] World-Wide Web Servers (http://www.w3.org/History/19921103-hypertext/hypertext/DataSources/WWW/Servers.html)

[2] What's New! February 1994 (http://home.mcom.com/home/whatsnew/whats_new_0294.html)

[3] "Internet History - Search Engines" (from Search Engine Watch), Universiteit Leiden, Netherlands, September 2001, web: LeidenU-Archie (http://www.internethistory.leidenuniv.nl/index.php3?c=7).

[4] Oscar Nierstrasz (2 September 1993). "Searchable Catalog of WWW Resources (experimental)" (http://groups.google.com/group/comp.infosystems.www/browse_thread/thread/2176526a36dc8bd3/2718fd17812937ac?hl=en&lnk=gst&q=Oscar+ Nierstrasz#2718fd17812937ac). .

[5] Archive of NCSA what's new in December 1993 page(http://web.archive.org/web/20010620073530/http://archive.ncsa.uiuc.edu/SDG/Software/Mosaic/Docs/old-whats-new/whats-new-1293.html)

[6] "Yahoo! And Netscape Ink International Distribution Deal" (http://files.shareholder.com/downloads/YHOO/701084386x0x27155/ 9a3b5ed8-9e84-4cba-a1e5-77a3dc606566/YHOO_News_1997_7_8_General.pdf).

[7] *Browser Deals Push Netscape Stock Up 7.8%* (http://articles.latimes.com/1996-04-01/business/fi-53780_1_netscape-home). Los Angeles Times. 1 April 1996.

[8] Gandal, Neil (2001). "The dynamics of competition in the internet search engine market". *International Journal of Industrial Organization* **19** (7): 1103–1117. doi:10.1016/S0167-7187(01)00065-0.

[9] Net Marketshare - World (http://marketshare.hitslink.com/search-engine-market-share.aspx?qprid=4)

[10] Net Market share - Google (http://marketshare.hitslink.com/report.aspx?qprid=5&qpcustom=Google - Global&qptimeframe=M& qpsp=120&qpnp=25)

[11] "Google Remains Ahead of Bing, But Relevance Drops" (http://news.yahoo.com/ google-remains-ahead-bing-relevance-drops-210457139.html). August 12, 2011..

[12] Experian Hitwise reports Bing-powered share of searches at 29 percent in October 2011 (http://www.hitwise.com/us/about-us/ press-center/press-releases/bing-powered-share-of-searches-at-29-percent), Experian Hitwise, November 16, 2011

[13] Search Engine Market Share July 2009 | Rise to the Top Blog (http://risetothetop.techwyse.com/internet-marketing/ search-engine-market-share-july-2009/)

[14] Segev, Elad (2010). Google and the Digital Divide: The Biases of Online Knowledge, Oxford: Chandos Publishing.

[15] Vaughan, L. & Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes, Information Processing & Management, 40(4), 693-707.

[16] Berkman Center for Internet & Society (2002), "Replacement of Google with Alternative Search Systems in China: Documentation and Screen Shots" (http://cyber.law.harvard.edu/filtering/china/google-replacements/), Harvard Law School.

- GBMW: Reports of 30-day punishment, re: Carmaker BMW had its German website bmw.de delisted from Google, such as: Slashdot-BMW (http://slashdot.org/article.pl?sid=06/02/05/235218) (05-Feb-2006).
- INSIZ: Maximum size of webpages indexed by MSN/Google/Yahoo! ("100-kb limit"): Max Page-size (http://www.sitepoint.com/article/indexing-limits-where-bots-stop) (28-Apr-2006).

Further reading

- For a more detailed history of early search engines, see Search Engine Birthdays (http://searchenginewatch.com/showPage.html?page=3071951) (from Search Engine Watch), Chris Sherman, September 2003.
- Steve Lawrence; C. Lee Giles (1999). "Accessibility of information on the web". *Nature* **400** (6740): 107–9. doi:10.1038/21987. PMID 10428673.
- Bing Liu (2007), *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data* (http://www.cs.uic.edu/~liub/WebMiningBook.html). Springer, ISBN 3540378812
- Bar-Ilan, J. (2004). The use of Web search engines in information science research. ARIST, 38, 231-288.
- Levene, Mark (2005). *An Introduction to Search Engines and Web Navigation*. Pearson.
- Hock, Randolph (2007). *The Extreme Searcher's Handbook*. ISBN 978-0-910965-76-7
- Javed Mostafa (February 2005). "Seeking Better Web Searches" (http://www.sciam.com/article.cfm?articleID=0006304A-37F4-11E8-B7F483414B7F0000). *Scientific American Magazine*.
- Ross, Nancy; Wolfram, Dietmar (2000). "End user searching on the Internet: An analysis of term pair topics submitted to the Excite search engine". *Journal of the American Society for Information Science* **51** (10): 949–958. doi:10.1002/1097-4571(2000)51:10<949::AID-ASI70>3.0.CO;2-5.
- Xie, M. et al (1998). "Quality dimensions of Internet search engines". *Journal of Information Science* **24** (5): 365–372. doi:10.1177/016555159802400509.
- *Information Retrieval: Implementing and Evaluating Search Engines* (http://www.ir.uwaterloo.ca/book/). MIT Press. 2010.

External links

- Search Engines (http://www.dmoz.org/Computers/Internet/Searching/Search_Engines//) at the Open Directory Project

Bing

Bing

The Bing homepage features an image or video that changes daily.

URL	www.bing.com [1]
Slogan	Bing and decide
Commercial?	Yes
Type of site	Web search engine
Registration	Optional
Available language(s)	40 languages
Owner	Microsoft
Created by	Microsoft
Launched	June 3, 2009
Alexa rank	26 (March 2012)[2]
Current status	Active

Bing (formerly **Live Search**, **Windows Live Search**, and **MSN Search**) is a web search engine (advertised as a "decision engine"[3]) from Microsoft. Bing was unveiled by Microsoft CEO Steve Ballmer on May 28, 2009 at the *All Things Digital* conference in San Diego. It went fully online on June 3, 2009,[4] with a preview version released on June 1, 2009.

Notable changes include the listing of search suggestions as queries are entered and a list of related searches (called "Explore pane") based on[5] semantic technology from Powerset that Microsoft purchased in 2008.[6]

On July 29, 2009, Microsoft and Yahoo! announced a deal in which Bing would power Yahoo! Search.[7] All Yahoo! Search global customers and partners are expected to have made the transition by early 2012.[8]

In October 2011, Bing announced it is working on new back-end search infrastructure, with the goal of delivering faster and slightly more relevant search results for users. Known as "Tiger," the new index-serving technology has being incorporated into Bing globally, since August 2011.[9]

History

MSN Search

MSN Search was a search engine by Microsoft that consisted of a search engine, index, and web crawler. MSN Search first launched in the third quarter of 1998 and used search results from Inktomi. In early 1999, MSN Search launched a version which displayed listings from Looksmart blended with results from Inktomi except for a short time in 1999 when results from AltaVista were used instead. Since then Microsoft upgraded MSN Search to provide its own self-built search engine results, the index of which was updated weekly and sometimes daily. The upgrade started as a beta program in November 2004, and came out of beta in February 2005. Image search was powered by a

third party, Picsearch. The service also started providing its search results to other search engine portals in an effort to better compete in the market.

Windows Live Search

The first public beta of Windows Live Search was unveiled on March 8, 2006, with the final release on September 11, 2006 replacing MSN Search. The new search engine used search tabs that include Web, news, images, music, desktop, local, and Microsoft Encarta.

In the roll-over from MSN Search to Windows Live Search, Microsoft stopped using Picsearch as their image search provider and started performing their own image search, fueled by their own internal image search algorithms.[10]

Live Search

On March 21, 2007, Microsoft announced that it would separate its search developments from the Windows Live services family, rebranding the service as Live Search. Live Search was integrated into the *Live Search and Ad Platform* headed by Satya Nadella, part of Microsoft's Platform and Systems division. As part of this change, Live Search was merged with Microsoft adCenter.[11]

A series of reorganisations and consolidations of Microsoft's search offerings were made under the Live Search branding. On May 23, 2008, Microsoft announced the discontinuation of Live Search Books and Live Search Academic and integrated all academic and book search results into regular search, and as a result this also included the closure of Live Search Books Publisher Program. Soon after, Windows Live Expo was discontinued on July 31, 2008. Live Search Macros, a service for users to create their own custom search engines or use macros created by other users, was also discontinued shortly after. On May 15, 2009, Live Product Upload, a service which allowed merchants to upload products information onto Live Search Products, was discontinued. The final reorganisation came as Live Search QnA was rebranded as MSN QnA on February 18, 2009, however, it was subsequently discontinued on May 21, 2009.[12]

Microsoft recognised that there would be a brand issue as long as the word "Live" remained in the name.[13] As an effort to create a new identity for Microsoft's search services, Live Search was officially replaced by Bing on June 3, 2009.[14]

Yahoo! search deal

On July 29, 2009, Microsoft and Yahoo! announced that they had made a 10-year deal in which the Yahoo! search engine would be replaced by Bing. Yahoo! will get to keep 88% of the revenue from all search ad sales on its site for the first five years of the deal, and have the right to sell adverts on some Microsoft sites. Yahoo! Search will still maintain its own user interface, but will eventually feature "Powered by Bing™" branding.[15][16] All Yahoo! Search global customers and partners are expected to be transitioned by early 2012.[17]

Market share

Before the launch of Bing, the marketshare of Microsoft web search pages (MSN and Live search) had been small but steady. By January 2011, Experian Hitwise show that Bing's market share had increased to 12.8% at the expense of Yahoo and Google. Bing powered searches also continued to have a higher "success rate" compared to Google, with more users clicking on the resulting links.[18] In the same period, comScore's "2010 U.S. Digital Year in Review" report showed that "Bing was the big gainer in year-over-year search activity, picking up 29% more searches in 2010 than it did in 2009."[19] The Wall Street Journal notes the 1% jump in share "appeared to come at the expense of rival Google Inc".[20] In February 2011 Bing beat out Yahoo! for the first time ever in terms of search marketshare. Bing received 4.37% search share while Yahoo! received 3.93% according to StatCounter.[21]

In March 2011, Bing-powered search accounts for over 30% of US searches, up 5% over February. In the same period, Google fell 3%.[22]

Counting core searches only, i.e. those where the user has an intent to interact with the search result, Bing achieved a market share of 14.54% in the second quarter of 2011 in the US.[23][24]

In December of 2011, Bing search queries overtook Yahoo for the first time ever. Bing's share of searches was 15.1% in December 2011, as compared to Yahoo's, which fell to 14.5%.[25]

Features

Interface features

- Daily changing of background image. The images are mostly of noteworthy places in the world, though it sometimes displays animals, people, and sports. The background image also contains information about the element(s) shown in the image.
- Video homepage for HTML-5 enabled browsers on occasional events, similar to the daily background images.
- Images page shows the main picture from that day and four searches that refers to that image with three preview pictures per search term.
- Left side navigation pane. Includes navigation and, on results pages, related searches and prior searches.
- Right side extended preview which shows a bigger view of the page and gives URLs to links inside of the page.
- Sublinks. On certain search results, the search result page also shows section links within the article (this is also done on other search engines, including Google)
- *Enhanced view* where third party site information can be viewed inside Bing.
- On certain sites, search from within the website on the results page.
- On certain sites, Bing will display the Customer Service number on the results page.

Media features

- Video thumbnail Preview where, by hovering over a video thumbnail, the video automatically starts playing
- Image search with continuous scrolling images results page that has adjustable settings for size, layout, color, style and people.[26]
- Advanced filters - allow users to refine search results based on properties such as image size, aspect ratio, color or black and white, photo or illustration, and facial features recognition
- Video search with adjustable setting for length, screen size, resolution and source

Instant answers

- Sports. Bing can directly display scores from a specific day, recent scores from a league or scores and statistics on teams or players.
- Finance. When entering a company name or stock symbol **and** either *stock* or *quote* in the search box Bing will show direct stock information like a stock chart, price, volume, and p/e ratio[27] in a webslice that users can subscribe to.
- Math calculations (e.g., 2*pi*24).[28] Users can enter math expressions in the search box using a variety of math operators and trigonometric functions[29] and Bing will provide a direct calculation of the expression.
- Advanced computations. Using the WolframAlpha computational engine, Bing can also give results to advanced math problems (e.g. lim x/2xasx->2) and other WolframAlpha related queries (e.g. calories in pizza).
- Package tracking and tracing. When a user types the name of the shipping company and the tracking number, Bing will provide direct tracking information.
- Dictionary. When "define", "definition" or "what is" followed by a word is entered in the search box Bing will show a direct answer from the Encarta dictionary.

- Spell check. Will change frequently misspelled search terms to the more commonly spelled alternative.
- Best match (plus similar sites)
- Product shopping and "Bing cashback"
- Health information
- Flight tracking
- Translate. Auto translation of certain search phrases, often with phrases including "translate" or "in English." For example, to translate "me llamo" from Spanish to English you would simply type *"translate" me llamo in english* and you will be redirected to a search results page with Bing Translator with your translation from Spanish to English.

Local info

- Current traffic information
- Business listing
- People listing
- Collections
- Localized searching for restaurants and services
- Restaurant reviews
- Movies played in an area. When a current movie title is entered in the search box Bing will provide listings of local theaters showing the movie. When a city is added to the search box, Bing provides the movie listings localized for that city.
- City hotel listings. When 'hotels' and a city name is entered in the search box Bing can provide hotel listings with a map. The listing leads to a detail search page with the hotels listed that holds extended information on the hotels and contains links to reviews, directions reservations and bird eye view of the hotel. On the page with the listings the list can be refined by settings on ratings, pricing, amenities, payment and parking

Integration with Hotmail

- With Hotmail's "Quick Add" feature, users can insert derivatives of Bing search results such as restaurant reviews, movie times, images, videos, and maps directly into their e-mail messages.[30]

Integration with Facebook

- Bing's search results can display one's Facebook friends when a Facebook account is linked with Bing via Facebook Connect.
- Users have the option to send messages to their friends in the search results.

International

Bing is available in many languages and has been localized for many countries.[31]

Languages in which Bing can find results

- Albanian
- Arabic
- Bulgarian
- Catalan
- Chinese (Simplified and Traditional scripts)
- Croatian
- Czech
- Danish
- Dutch
- English (American and British)
- Estonian
- Finnish
- French
- German
- Greek
- Hebrew
- Hungarian
- Icelandic
- Indonesian
- Italian
- Japanese
- Korean
- Latvian
- Lithuanian
- Malay
- Norwegian
- Persian
- Polish
- Portuguese (Brazil)
- Portuguese (Portugal)
- Romanian
- Russian
- Serbian (Cyrillic)
- Slovak
- Slovenian
- Spanish
- Swedish
- Tamil
- Thai
- Turkish
- Ukrainian
- Vietnamese

Languages in which Bing can be displayed

- Basque
- Bulgarian
- Catalan
- Chinese (Simplified and Traditional scripts)
- Croatian
- Czech
- Danish
- Dutch
- English (American and British)
- Estonian
- Finnish
- French
- Galician
- German
- Greek
- Hindi
- Hungarian
- Icelandic
- Italian
- Japanese
- Korean
- Latin
- Latvian
- Lithuanian
- Malay
- Norwegian
- Polish
- Portuguese (Brazil)
- Portuguese (Portugal)
- Romanian
- Russian
- Serbian (Cyrillic)
- Slovak
- Slovenian
- Spanish
- Swedish
- Tamil
- Thai
- Turkish
- Ukrainian
- Vietnamese

Search products

In addition to its tool for searching web pages, Bing also provides the following search offerings:[32]

Service	Description
Dictionary	*Bing Dictionary* enables users to quickly search for definitions of English words. Bing Dictionary results are based on Microsoft Encarta World English Dictionary. In addition, Bing Dictionary also provides an audio player for users to hear the pronunciation of the dictionary words.
Entertainment	*Bing Entertainment* allow users to view and search for detailed information and reviews for music, movies, television shows, and video games. Bing Entertainment partners with Microsoft Games to allow users to directly play online games within Bing Online Games.
Events	*Bing Events* allow users to search for upcoming events from Zvents, and displays the date and time, venue details, brief description, as well as method to purchase tickets for the events listed. Users can also filter the search results by date and categories.

Finance	*Bing Finance* enables users to search for exchange listed stocks and displays the relevant stock information, company profile and statistics, financial statements, stock ratings, analyst recommendations, as well as news related to the particular stock or company. Bing Finance also allow users to view the historical data of the particular stock, and allows comparison of the stock to major indices. In addition, Bing Finance also features a javascript-based *Stock screener*, allowing investors to quickly filter for value, contrarian, high-yield, and bargain investment strategies.
Health	*Bing Health* refines health searches using related medical concepts to get relevant health information and to allow users to navigate complex medical topics with inline article results from experts. This feature is based on the Medstory acquisition.
Images	*Bing Images* enables the user to quickly search and display most relevant photos and images of interest. The infinite scroll feature allows browsing a large number of images quickly. The advance filters allows refining search results in terms of properties such as image size, aspect ratio, color or black and white, photo or illustration, and facial features recognition.
Local	*Bing Local* searches local business listings with business details and reviews, allowing users to make more informed decisions.
Maps	*Bing Maps* enables the user to search for businesses, addresses, landmarks and street names worldwide, and can select from a road-map style view, a satellite view or a hybrid of the two. Also available are "bird's-eye" images for many cities worldwide, and 3D maps which include virtual 3D navigation and to-scale terrain and 3D buildings. For business users it will be available as "Bing Maps For Enterprise".
News	*Bing News* is a news aggregator and provides news results relevant to the search query from a wide range of online news and information services.
Recipe	*Bing Recipe* allow users to search for cooking recipes sourced from Delish.com, MyRecipes.com, and Epicurious.com, and allow users to filter recipe results based on their ratings, cuisine, convenience, occasion, ingredient, course, cooking method, and recipe provider.
Reference	*Bing Reference* semantically indexes Wikipedia content and displays them in an enhanced view within Bing. It also allow users to input search queries that resembles full questions and highlights the answer within search results. This feature is based on the Powerset acquisition.
Social	*Bing Social* allow users to search for and retrieve real-time information from Twitter and Facebook services. Bing Social search also provides "best match" and "social captions" functionalities that prioritises results based on relevance and contexts. Only public feeds from the past 7 days will be displayed in Bing Social search results.
Shopping	*Bing Shopping* lets users search from a wide range of online suppliers and marketer's merchandise for all types of products and goods. This service also integrates with Bing cashback offering money back for certain purchases made through the site. This feature is based on the Jellyfish.com acquisition, but will be discontinued July 30, 2010.
Translator	*Bing Translator* lets users translate texts or entire web pages into different languages.
Travel	*Bing Travel* searches for airfare and hotel reservations online and predicts the best time to purchase them. This feature is based on the Farecast acquisition.
University	*Bing University* allow users to search for and view detailed information about United States universities, including information such as admissions, cost, financial aid, student body, and graduation rate.
Videos	*Bing Videos* enables the user to quickly search and view videos online from various websites. The Smart Preview feature allows the user to instantly watch a short preview of an original video. Bing Videos also allow users to access editorial video contents from MSN Video.
Visual Search	*Bing Visual Search* (now deprecated) allowed users to refine their search queries for structured results through data-grouping image galleries that resembles "large online catalogues", powered by Silverlight[33]
Weather	*Bing Weather* allow users to search for the local weather for cities around the world, displaying the current weather information and also extended weather forecasts for the next 10 days. Weather information are provided by Intellicast and Foreca.
Wolfram Alpha	*Bing Wolfram Alpha* allow users to directly enter factual queries within Bing and provides answers and relevant visualizations from a core knowledge base of curated, structured data provided by Wolfram Alpha. Bing Wolfram Alpha can also answer mathematical and algebraic questions.
xRank	*Bing xRank* allowed users to search for celebrities, musicians, politicians and bloggers, read short biographies and news about them, and track their trends or popularity rankings. As of October 2010, this feature was shut down.

Webmaster services

Bing allows webmasters to manage the web crawling status of their own websites through Bing Webmaster Center. Additionally, users may also submit contents to Bing via the following methods:

- **Bing Local Listing Center** allow businesses to add business listings onto Bing Maps and Bing Local
- **Soapbox on MSN Video** allow users to upload videos for searching via Bing Videos

Mobile services

Bing Mobile allow users to conduct search queries on their mobile devices, either via the mobile browser or a downloadable mobile application. In the United States, Microsoft also operates a toll-free number for directory assistance called Bing411.[32]

Developer services

Bing Application Programming Interface enables developers to programmatically submit queries and retrieve results from the Bing Engine. http://www.bing.com/developers

To use the Bing API developers have to obtain an Application ID, http://www.bing.com/developers/createapp. aspx

Bing API can be used with following protocols:

- XML
- JSON
- SOAP

Query examples:

- http://api.bing.net/xml.aspx?AppId=YOUR_APPID&Version=2.2&Market=en-US&Query=YOUR_QUERY&Sources=web+spe
- http://api.bing.net/json.aspx?AppId=YOUR_APPID&Version=2.2&Market=en-US&Query=YOUR_QUERY&Sources=web+spe
- http://api.bing.net/search.wsdl?AppID=YourAppId&Version=2.2

Other services

BingTweets is a service that combines Twitter trends with Bing search results, enabling users to see real-time information about the hottest topics on Twitter. The BingTweets service was initiated on July 14, 2009 in a partnership between Microsoft, Twitter and Federated Media.[34]

Bing Rewards is a service that allows users to earn points for searching with Bing. These points can be redeemed for various products such as electronics, multimedia and gift cards.

Toolbars, gadgets and plug-ins

Toolbars

Both Windows Live Toolbar and MSN Toolbar will be powered by Bing and aim to offer users a way to access Bing search results. Together with the launch of Bing, MSN Toolbar 4.0 will be released with inclusion of new Bing-related features such as Bing cashback offer alerts.[32] (See "**Bing Rewards**")

Gadgets

The Bing Search gadget is a Windows Sidebar Gadget that uses Bing Search to fetch the user's search results and render them directly in the gadget. Another gadget, the Bing Maps gadget, displays real-time traffic conditions using Bing Maps.[35] The gadget provides shortcuts to driving directions, local search and full-screen traffic view of major US and Canadian cities, including Atlanta, Boston, Chicago, Denver, Detroit, Houston, Los Angeles, Milwaukee,

Montreal, New York City, Oklahoma City, Ottawa, Philadelphia, Phoenix, Pittsburgh, Portland, Providence, Sacramento, Salt Lake City, San Diego, San Francisco, Seattle, St. Louis, Tampa, Toronto, Vancouver, and Washington, D.C.

Prior to October 30, 2007, the gadgets were known as *Live Search gadget* and *Live Search Maps gadget*; both gadgets were removed from Windows Live Gallery due to possible security concerns.[36] The Live Search Maps gadget was made available for download again on January 24, 2008 with the security concern addressed.[37] However around the introduction of Bing in June 2009 both gadgets have been removed again for download from Windows Live Gallery.

Accelerators

Accelerators allow users to access Bing features directly from selected text in a webpage. Accelerators provided by the Bing team include:

- Bing Translator
- Bing Maps
- Bing Shopping

Web Slices

Web Slices can be used to monitor information gathered by Bing. Web Slices provided by the Bing team include:

- Weather from Bing
- Finance from Bing
- Traffic from Bing

Plug-ins

The Bing team provides an official Bing Firefox add-on, which adds search suggestions to the Firefox search box from Bing.[38]

Marketing and advertisements

Live Search

Since 2006, Microsoft had conducted a number of tie-ins and promotions for promoting Microsoft's search offerings. These include:

- Amazon's A9 search service and the experimental Ms. Dewey interactive search site syndicated all search results from Microsoft's then search engine, Live Search. This tie-in started on May 1, 2006.
- Search and Give - a promotional website launched on 17 January 2007 where all searches done from a special portal site would lead to a donation to the UNHCR's organization for refugee children, ninemillion.org. Reuters AlertNet reported in 2007 that the amount to be donated would be $0.01 per search, with a minimum of $100,000 and a maximum of $250,000 (equivalent to 25 million searches).[39] According to the website the service was decommissioned on June 1, 2009, having donated over $500,000 to charity and schools.[40]
- Club Bing - a promotional website where users can win prizes by playing word games that generate search queries on Microsoft's then search service Live Search. This website began in April 2007 as Live Search Club.
- Big Snap Search - a promotional website similar to Live Search Club. This website began in February 2008, but was discontinued shortly after.[41]
- Live Search SearchPerks! - a promotional website which allowed users to redeem tickets for prizes while using Microsoft's search engine. This website began on October 1, 2008 and was decommissioned on April 15, 2009.

Debut

Bing's debut featured an $80 to $100 million online, TV, print, and radio advertising campaign in the US. The advertisements do not mention other search engine competitors, such as Google and Yahoo, directly by name; rather, they attempt to convince users to switch to Bing by focusing on Bing's search features and functionality.[42] The ads claim that Bing does a better job countering "search overload".[43]

Bing Rewards

Launched by Microsoft in September 2010, Bing Rewards provides credits to users through regular Bing searches and special promotions.[44] These credits are then redeemed for various products including electronics, gift cards and charitable donations.[45] Initially, participants in the program were required to download and use the Bing Bar for Internet Explorer in order to earn credits; however, this is no longer the case, and the service now works with all major browsers.[46] The Bing Rewards program is similar to two earlier services, SearchPerks! and Bing Cashback, which have now been discontinued.

The Colbert Report

During the episode of *The Colbert Report* that aired on June 8, 2010, Stephen Colbert stated that Microsoft would donate $2,500 to help clean up the Gulf oil spill each time he mentioned the word "Bing" on air. Colbert mostly mentioned Bing in out-of-context situations, such as Bing Crosby and Bing cherries. By the end of the show, Colbert had said the word 40 times, for a total donation of $100,000. Colbert poked fun at their rivalry with Google, stating "Bing is a great website for doing Internet searches. I know that, because I Googled it."[47][48]

Los Links Son Malos

An advertising campaign during 2010, *Los Links Son Malos* (English: The Links are Bad), took the form of a Mexican telenovela, with people conversing in Spanish, subtitled in English. In it, somebody rides in on a horse and takes a woman away when he shows her how easy Bing is to use in order to get movie tickets or travel.

Search deals

As of Opera 10.6, Bing has been incorporated into the Opera browser, but Google is still the default search engine. Bing will also be incorporated into all future versions of Opera.[49] Mozilla Firefox has made a deal with Microsoft to jointly release "Firefox with Bing"[50], an edition of Firefox where Bing has replaced Google as the default search engine.[51][52] However, the default edition of Firefox still has Google as its default search engine, but has included Bing in its default list of search providers since Firefox version 4.0.[53]

In addition, Microsoft also paid Verizon Wireless $550 million USD[54] to use Bing as the default search provider on Verizon's Blackberry, and in turn, have Verizon "turn off" (via Blackberry service books) the other search providers available. Though users can still access other search engines via the mobile browser.[55]

Name origin

Through focus groups, Microsoft decided that the name Bing was memorable, short, easy to spell, and that it would function well as a URL around the world. The word would remind people of the sound made during "the moment of discovery and decision making."[56] Microsoft was assisted by branding consultancy Interbrand in their search for the best name for the new search engine.[57] The name also has strong similarity to the word 'bingo', which is used to mean that something sought has been found or realized, as is interjected when winning the game Bingo. Microsoft advertising strategist David Webster originally proposed the name "Bang" for the same reasons the name Bing was ultimately chosen (easy to spell, one syllable, and easy to remember). He noted, "It's there, it's an exclamation point [...] It's the opposite of a question mark." This name was ultimately not chosen because it could not be properly used

as a verb in the context of an internet search.[58]

According to the *Guardian* "[Microsoft] hasn't confirmed that it stands recursively for Bing Is Not Google, but that's the sort of joke software engineers enjoy."[59] Qi Lu, president of Microsoft Online Services, also announced that Bing's official Chinese name is *bì yìng* (simplified Chinese: 必 应 ; traditional Chinese: 必 應), which literally means "very certain to respond" or "very certain to answer" in Chinese.[60]

While being tested internally by Microsoft employees, Bing's codename was *Kumo* (くも),[61] which came from the Japanese word for *spider* (蜘蛛 ; くも , *kumo*) as well as *cloud* (雲 ; くも , *kumo*), referring to the manner in which search engines "spider" Internet resources to add them to their database, as well as cloud computing.

Legal challenges

On July 31, 2009, The Laptop Company, Inc. released a press release stating that it is challenging Bing's trademark application, alleging that Bing may cause confusion in the marketplace as Bing and their product BongoBing both do online product search.[62] Software company TeraByte Unlimited, which has a product called BootIt Next Generation (abbreviated to BING), also contended the trademark application on similar grounds, as did a Missouri-based design company called Bing! Information Design.[63]

Microsoft contends that claims challenging its trademark are without merit because these companies filed for U.S. federal trademark applications only after Microsoft filed for the Bing trademark in March 2009.[64]

Adult content

Video content

Bing's video search tool has a preview mode that could potentially be used to preview pornographic videos.[65] By simply turning off safe search, users can search for and view pornographic videos by hovering the cursor over a thumbnail, since the video and audio, in some cases, are cached on Microsoft's Server.

Since the videos are playing within Bing instead of the site where they are hosted, the videos are not necessarily blocked by parental control filters. Monitoring programs designed to tell parents what sites their children have visited are likely to simply report "Bing.com" instead of the site that actually hosts the video. The same situation can be said about corporate filters, many of which have been fooled by this feature.[66] Users do not need to leave Bing's site to view these videos.[67][68]

Microsoft responded in a blog post on June 4, 2009, with a short term work-around.[69] By adding "&adlt=strict" to the end of a query, no matter what the settings are for that session it will return results as if safe search were set to strict. The query would look like this: http://www.bing.com/videos/search?q=adulttermgoeshere&adlt=strict (case sensitive).

On June 12, 2009, Microsoft announced two changes regarding Bing's Smart Motion Preview and SafeSearch features. All potentially explicit content will be coming from a separate single domain, explicit.bing.net. Additionally, Bing will also return source URL information in the query string for image and video contents. Both changes allow both home users and corporate users to filter content by domain regardless of what the SafeSearch settings might be.[70]

Regional censorship

Bing censors results for adult search terms for some of the regions including India, People's Republic of China, Germany and Arab countries.[71] This censoring is done based on the local laws of those countries.[72] However, Bing allows users to simply change their country/region preference to somewhere without restrictions – such as the United States, United Kingdom or Republic of Ireland – to sidestep this censorship.

Criticism

Censorship

Microsoft has been criticized for censoring Bing search results to queries made in simplified Chinese characters, used in mainland China. This is done to comply with the censorship requirements of the government in China.[73] Microsoft has not indicated a willingness to stop censoring search results in simplified Chinese characters in the wake of Google's decision to do so.[74] All simplified Chinese searches in Bing are censored regardless of the user's country.[75]

Performance issues

Bing has been criticized for being slower to index websites than Google. It has also been criticized for not indexing some websites at all.[76][77][78]

Copying Google's results

Bing has been criticized by competitor Google, for utilizing user input via Internet Explorer, the Bing Toolbar, or Suggested Sites, to add results to Bing. After discovering in October 2010 that Bing appeared to be imitating Google's auto-correct results for a misspelling, despite not actually fixing the spelling of the term, Google set up a honeypot, configuring the Google search engine to return specific unrelated results for 100 nonsensical queries such as *hiybbprqag*.[79] Over the next couple of weeks, Google engineers entered the search term into Google, while using Microsoft Internet Explorer, with the Bing Toolbar installed and the optional Suggested Sites enabled. In 9 out of the 100 queries, Bing later started returning the same results as Google, despite the only apparent connection between the result and search term being that Google's results connected the two.[80][81]

Microsoft's response to this issue, coming from a company's spokesperson, was clear: "We do not copy Google's results." Bing's Vice President, Harry Shum, later reiterated that the search result data Google claimed that Bing copied had in fact come from Bing's very own users. Shum further wrote that "we use over 1,000 different signals and features in our ranking algorithm. A small piece of that is clickstream data we get from some of our customers, who opt-in to sharing anonymous data as they navigate the web in order to help us improve the experience for all users."[82] Microsoft commented that clickstream data from customers who had opted in was collected, but said that it was just a small piece of over 1000 signals used in their ranking algorithm, and that their intention was to learn from their collective customers. They stated that Bing was not intended to be a duplicate of any existing search engines.[83] Representatives for Google have said the company simply wants the practice to stop.[80]

References

[1] http://www.bing.com

[2] "Bing.com Site Info" (http://www.alexa.com/siteinfo/bing.com). Alexa Internet. . Retrieved 2012-03-02.

[3] "Welcome to Discover Bing" (http://discoverbing.com/). Discover Bing. . Retrieved 2010-01-16.

[4] "Microsoft's New Search at Bing.com Helps People Make Better Decisions: Decision Engine goes beyond search to help customers deal with information overload" (http://www.microsoft.com/presspass/press/2009/may09/05-28NewSearchPR.mspx?rss_fdn=Press Releases). Microsoft. . Retrieved 2009-05-29.

[5] "Microsoft Bing rides open source to semantic search" (http://www.theregister.co.uk/2009/06/04/bing_and_powerset/). The Register. . Retrieved 2010-01-01.

[6] "Microsoft to Acquire Powerset" (http://www.bing.com/community/blogs/powerset/archive/2008/07/01/ microsoft-to-acquire-powerset.aspx). Bing. . Retrieved 2010-01-01.

[7] Microsoft and Yahoo seal web deal (http://news.bbc.co.uk/1/hi/business/8174763.stm).

[8] "When will the change happen? How long will the transition take?" (http://help.yahoo.com/l/us/yahoo/search/alliance/alliance-2.html;_ylt=AvrC8b99B5.r4JmW33gA5ChaMnlG), help.yahoo.com, March 23, 2011

[9] "Bing Unleashing Tiger to Speed Search Results" (http://searchenginewatch.com/article/2113363/ Bing-Unleashing-Tiger-to-Speed-Search-Results), Search Engine Watch, October 3, 2011

[10] Chris Sherman, September 12, 2006, Microsoft Upgrades Live Search Offerings (http://searchenginewatch.com/showPage. html?page=3623401).

[11] Mary Jo Foley: Microsoft severs Live Search from the rest of the Windows Live family (http://blogs.zdnet.com/microsoft/?p=339).

[12] Live QnA team blog announcement (http://liveqna.spaces.live.com/blog/cns!2933A3E375F68349!2125.entry).

[13] Keynote with Kevin Johnson at Microsoft (http://www.seroundtable.com/archives/017296.html).

[14] *Wired*, 28 May 2009, Hands On With Microsoft's New Search Engine: Bing, But No Boom (http://www.wired.com/epicenter/2009/05/ microsofts-bing-hides-its-best-features/).

[15] "Microsoft and Yahoo seal web deal" (http://news.bbc.co.uk/2/hi/business/8174763.stm). BBC News. Wednesday, 29 July 2009 13:58 UK. . Retrieved 2009-07-29.

[16] Tiffany Wu; Derek Caney (Wed Jul 29, 2009 8:27am EDT). "REFILE-UPDATE 1-Microsoft, Yahoo in 10-year Web search deal" (http://www.reuters.com/article/CMPSRV/idUSN2921665320090729). Thomson Reuters. . Retrieved 2009-07-29.

[17] When will the change happen? How long will the transition take?(http://help.yahoo.com/l/us/yahoo/search/alliance/alliance-2.html;_ylt=AvrC8b99B5.r4JmW33gA5ChaMnlG) Yahoo! Search Help

[18] "Experian Hitwise reports Bing searches increase 21 percent in January 2011" (http://www.hitwise.com/us/press-center/press-releases/ bing-searches-increase-twenty-one-percent/). .

[19] "Bing Search Volume Up 29% In 2010, Google Up 13%, comScore Says" (http://searchengineland.com/ bing-search-volume-up-29-in-2010-google-up-13-comscore-says-64075). .

[20] Wingfield, Nick (February 10, 2011). "Microsoft's Bing Gains Share" (http://online.wsj.com/article/SB10001424052748703745704576136742065119876.html?mod=wsjcrmain). *The Wall Street Journal*. .

[21] "StatCounter: Bing Just Beat Yahoo Worldwide" (http://www.readwriteweb.com/archives/statcounter_bing_just_beat_yahoo_for_first_time.php). *Read, Write, Web*. March 1, 2011. .

[22] "Experian Hitwise reports Bing-powered share of searches reaches 30 percent in March 2011" (http://www.winrumors.com/ bing-powered-u-s-searches-rise-to-over-30-market-share/). .

[23] Jay Yarow, Kamelia Angelova (2011-07-13). "CHART OF THE DAY: This Is What Microsoft Is Getting For Its Big Bing Investment" (http://www.businessinsider.com/chart-of-the-day-us-search-market-share-2011-7?op=1). Business Insider. .

[24] Stephanie Lyn Flosi (2011-07-13). "comScore Releases June 2011 U.S. Search Engine Rankings" (http://www.comscore.com/layout/set/popup/Press_Events/Press_Releases/2011/7/comScore_Releases_June_2011_U.S._Search_Engine_Rankings). comScore. .

[25] Rao, Leena. January 11, 2012. "Microsoft Bing Search Queries Overtake Yahoo For The First Time In December." http://techcrunch.com/ 2012/01/11/microsoft-bing-search-queries-overtake-yahoo-for-the-first-time-in-december/

[26] Limit Image Results to Color or Black and White Images (http://malektips.com/bing-images-color-black-white.html).

[27] Display Stock Quotes (http://malektips.com/bing-stock-quote.html).

[28] Use Bing to Calculate, Perform Trigonometry, or Solve Your Algebra Homework (http://malektips.com/ bing-math-trigonometry-equations.html).

[29] Mathematical notations for use with Math Answers (http://help.live.com/Help.aspx?market=en-US&project=WL_Searchv1&querytype=topic&query=WL_SEARCH_REF_MathNotations.htm).

[30] "Bing! Instantly find answers and add them to your e-mail" (http://windowslivewire.spaces.live.com/blog/cns!2F7EB29B42641D59!41224.entry). Windows Live team. 2009-07-09. .

[31] Language Tools (http://www.bing.com/settings.aspx?sh=2&FORM=WIWA).

[32] Explore Bing (http://www.bing.com/explore).

[33] Shiels, Maggie (07:39 GMT, Tuesday, 15 September 2009 08:39 UK). "Microsoft Bing adds visual search" (http://news.bbc.co.uk/2/hi/ technology/8256046.stm). *Technology* (BBC News). . Retrieved 2009-09-15.

[34] Bing Community: BingTweets Debuts (http://www.bing.com/community/blogs/search/archive/2009/07/14/bingtweets-debuts.aspx) Retrieved on 2009-07-20.
[35] Traffic by Live Search Maps Vista Gadget Returns (http://www.bing.com/community/blogs/maps/archive/2008/02/11/ traffic-by-live-search-maps-vista-gadget-returns.aspx).
[36] LiveSide.net: Yes, the Live Search and Live Search Traffic gadgets are gone: security concerns cited (http://www.liveside.net/blogs/ main/archive/2007/10/30/yes-the-live-search-and-live-search-traffic-gadgets-are-gone-security-concerns-cited.aspx).
[37] LiveSide.net: The Traffic Gadget is Back! (http://www.liveside.net/blogs/developer/archive/2008/01/23/the-traffic-gadget-is-back. aspx).
[38] Bing Firefox addon(https://addons.mozilla.org/en-US/firefox/addon/bing/).
[39] *Reuters AlertNet*, 17 January 2007, Microsoft launches "Click for Cause" initiative to support UNHCR Net campaign (http://www.alertnet.org/thenews/newsdesk/UNHCR/329ac7cacd8c9f683e9f270d84fc78e9.htm).
[40] searchandgive.com (http://www.searchandgive.com/). Retrieved 1 June 2009.
[41] "Microsoft challenges search users to game of snap" (http://www.brandrepublic.com/Digital/News/785817/ Microsoft-challenges-search-users-game-snap/?DCMP=EMC-Digital Bulletin).
[42] Microsoft Aims Big Guns at Google, Asks Consumers to Rethink Search (http://adage.com/digital/article?article_id=136847).
[43] "Microsoft's Bing Ad Claims to Terminate 'Search Overload'" (http://www.pcworld.com/article/166067/microsofts_bing_ad_claims_to_terminate_search_overload.html). PC World. 2009-06-03. . Retrieved 2010-01-16.
[44] Sterling, Greg (22 September 2010). "Microsoft Launches A New Loyalty Program: Bing Rewards" (http://searchengineland.com/ microsoft-launches-a-new-loyalty-program-bing-rewards-51374). *Search Engine Land*. . Retrieved 11 May 2011.
[45] "Bing Rewards Shop" (https://ssl.bing.com/rewards/redeem). . Retrieved 11 May 2011.
[46] "FAQ - Bing Rewards Preview" (http://www.bing.com/rewards/faq/Questions#WilltheBingbarworkonmycomputerandbrowser). . Retrieved 11 May 2011.
[47] "Charity Begins at 11:30 - The Colbert Report - 2010-07-06 - Video Clip | Comedy Central" (http://www.colbertnation.com/ the-colbert-report-videos/311926/june-07-2010/charity-begins-at-11-30). Colbertnation.com. . Retrieved 2011-12-16.
[48] Eaton, Nick (2010-06-08). "Stephen Colbert makes Bing donate $100K for oil spill | The Microsoft Blog - seattlepi.com" (http://blog.seattlepi.com/microsoft/archives/210083.asp). Blog.seattlepi.com. . Retrieved 2011-12-16.
[49] "Microsoft hits search deal with Opera Software" (http://topnews.co.uk/26953-microsoft-hits-search-deal-opera-software). .
[50] http://www.firefoxwithbing.com/
[51] "Bing Introducing Firefox with Bing - Search Blog - Site Blogs - Bing Community" (http://www.bing.com/community/site_blogs/b/search/archive/2011/10/26/bff.aspx). Bing.com. . Retrieved 2011-12-16.
[52] Mozilla. "Offering a Customized Firefox Experience for Bing Users | The Mozilla Blog" (http://blog.mozilla.com/blog/2011/10/26/ offering-a-customized-firefox-experience-for-bing-users/). Blog.mozilla.com. . Retrieved 2011-12-16.
[53] jsullivan. "Refreshing the Firefox Search Bar | The Mozilla Blog" (http://blog.mozilla.com/blog/2010/10/06/ refreshing-the-firefox-search-bar/). Blog.mozilla.com. . Retrieved 2011-12-16.
[54] See, Dianne (2009-01-07). "Microsoft Beats Out Google To Win Verizon Search Deal" (http://moconews.net/article/ 419-microsoft-beats-out-google-to-win-verizon-search-deal/). mocoNews. . Retrieved 2011-12-16.
[55] "As Verizon Implements Bing Default Search Deal, Company Sees User Backlash" (http://searchengineland.com/as-verizon-implements-bing-default-search-deal-company-sees-user-backlash-32650). Searchengineland.com. . Retrieved 2011-12-16.
[56] "The sound of found: Bing!" (http://blogs.msdn.com/livesearch/archive/2009/05/28/the-sound-of-found-bing.aspx). *Neowin.net*. May 28, 2009. . Retrieved May 29, 2009.
[57] "Interbrand Blog | Interbrand names Microsoft's new search engine Bing!" (http://www.interbrand.com/blog/post/2009/06/05/ Interbrand-names-Microsofts-new-search-engine-Bing!.aspx). Interbrand.com. . Retrieved 2010-01-16.
[58] Fried, Ina (2010-03-29). "conversation with Microsoft's marketing strategist" (http://news.cnet.com/8301-13860_3-20001293-56.html?tag=newsLeadStoriesArea.1A). News.cnet.com. . Retrieved 2011-12-16.
[59] Schofield, Jack (June 8, 2009). "Bing Is Not Google, but it might be better than you think" (http://www.guardian.co.uk/technology/ 2009/jun/08/netbytes-microsoft-bing). *The Guardian* (London). .
[60] Binging on search by design (http://news.ninemsn.com.au/technology/819478/binging-on-search-by-design).
[61] "First screenshot of Microsoft's Kumo emerges" (http://www.neowin.net/index.php?act=view&id=53126). *Neowin.net*. March 3, 2009. . Retrieved May 29, 2009.
[62] Wauters, Robin (2009-07-31). "BongoBing Opposes Microsoft Trademark Application For "Bing"" (http://www.techcrunch.com/2009/ 07/31/bongobing-opposes-microsoft-trademark-application-for-bing/). Techcrunch.com. . Retrieved 2010-01-16.
[63] Johnson, Bobbie (Monday 21 December 2009). "Microsoft sued over Bing trademark" (http://www.guardian.co.uk/technology/2009/ dec/21/microsoft-bing-trademark). *The Guardian* (London: Guardian News and Media Limited). . Retrieved 5 March 2010.
[64] Johnson, Bobbie (December 21, 2009). "Microsoft sued over Bing trademark | Technology | guardian.co.uk" (http://www.guardian.co.uk/technology/2009/dec/21/microsoft-bing-trademark). London: Guardian. . Retrieved 2010-01-16.
[65] Magid, Larry (June 2, 2009). "Parents beware: Bing previews video porn" (http://news.cnet.com/8301-19518_3-10255043-238. html?tag=mncol;txt). bing.com. . Retrieved 2009-06-08.
[66] Krazit, Tom (June 4, 2009). "Microsoft gives Bing stronger search filter option" (http://news.cnet.com/8301-17939_109-10257397-2. html). . Retrieved 2009-06-10.

[67] Magid, Larry (June 5, 2009). "Microsoft offers unworkable solution to Bing porn" (http://news.cnet.com/8301-19518_3-10258458-238. html?tag=rtcol;pop). . Retrieved 2009-06-08.

[68] McDougall, Paul (June 8, 2009). "Bing Porn Draws Flak" (http://www.informationweek.com/news/internet/search/showArticle. jhtml?articleID=217800024). . Retrieved2009-06-08.

[69] Nichols, Mike (June 4, 2009). "Bing Community: smart motion preview and safesearch" (http://www.bing.com/community/blogs/ search/archive/2009/06/04/smart-motion-preview-and-safesearch.aspx). bing.com. . Retrieved2009-06-08.

[70] Nichols, Mike (June 12, 2009). "Bing Community: Safe Search Update" (http://www.bing.com/community/blogs/search/archive/2009/ 06/12/safe-search-update.aspx). Bing.com. . Retrieved 2009-06-14.

[71] "No sex for Indians on Microsoft Bing" (http://webcache.googleusercontent.com/search?q=cache:8fJ0g1riPNwJ:infotech.indiatimes. com/articleshow/msid-4612759,flstry-1.cms). .

[72] "Why You Can't Search The Word 'Sex' On Bing" (http://in.reuters.com/article/paidmediaAtoms/idIN196193078720090604). *Reuters*. .

[73] Kristof, Nicholas (November 20, 2009). "Boycott Microsoft Bing" (http://kristof.blogs.nytimes.com/2009/11/20/ boycott-microsoft-bing/). *The New York Times*. . Retrieved March 31, 2010.

[74] "Activists applaud Google's censorship move, China grumbles" (http://www.itpro.co.uk/621706/ activists-applaud-googles-censorship-move-china-grumbles). IT PRO. 2010-03-23. . Retrieved 2012-01-30.

[75] "Boycott Microsoft Bing" (http://kristof.blogs.nytimes.com/2009/11/20/boycott-microsoft-bing/). *The New York Times*. November 20, 2009. .

[76] Protalinski, Emil (2010-01-17). "Microsoft has a plan to improve Bing's poor indexing" (http://arstechnica.com/microsoft/news/2010/ 01/microsoft-outlines-plan-to-improve-bings-slow-indexing.ars). Arstechnica.com. . Retrieved2011-12-16.

[77] "Bing Re: Website Not Listed In Bing Search - Indexing and Ranking Discussion - Webmaster - Bing Community" (http://www.bing. com/community/forums/p/653570/9582219.aspx). Bing.com. . Retrieved 2012-01-30.

[78] 12/09/2011 10:50 am (2010-01-07). "Microsoft Bing Says They Are "Fairly Slow"" (http://www.seroundtable.com/archives/021475. html). Seroundtable.com. . Retrieved 2011-12-16.

[79] "Google accuses Bing of 'copying' its search results" (http://www.bbc.co.uk/news/technology-12343597). BBC Online. February 2, 2011. .

[80] Singhal, Amit (2 February 2011). "Microsoft's Bing uses Google search results—and denies it" (http://googleblog.blogspot.com/2011/ 02/microsofts-bing-uses-google-search.html). Google Blog. . Retrieved 2 February 2011.

[81] Sullivan, Danny (1 February 2011). "Google: Bing Is Cheating, Copying Our Search Results" (http://searchengineland.com/ google-bing-is-cheating-copying-our-search-results-62914). Search Engine Land. . Retrieved 2 February 2011.

[82] "Google: Bing's Search Results Are a "Cheap Imitation"" (http://mashable.com/2011/02/02/google-bing-copying/). 25 October 2011. . Retrieved 25 October 2011.

[83] Shum, Harry (2 February 2011). "Thoughts on search quality" (http://www.bing.com/community/site_blogs/b/search/archive/2011/ 02/01/thoughts-on-search-quality.aspx?PageIndex=2). Bing Search Blog. . Retrieved 2 February 2011.

External links

- Official website (http://www.bing.com/) (Mobile (http://m.bing.com/))
- Discover Bing (http://www.discoverbing.com/)
- Decision Engine (http://www.decisionengine.com/)
- Bing Community (http://www.bing.com/community)
- Bing Newsroom (http://www.microsoft.com/presspass/presskits/bing/)
- Bing Toolbox (http://www.bing.com/toolbox) - for developers and webmasters
- Bing API (http://www.bing.com/developers) - for developers

Ask.com

Ask.com

Type	Search Engine
Founded	1996
Headquarters	Oakland, California, US
Key people	Garrett Gruener David Warthen (Founders) Doug Leeds (CEO)
Industry	Internet
Revenue	▲ US$227 million
Parent	InterActiveCorp
Slogan	
Website	Ask Jeeves [1]
Alexa rank	▼51 (March 2012)[2]
Registration	Optional
Screenshot	

Ask (sometimes known as **Ask Jeeves**) is a Q&A focused search engine founded in 1996 by Garrett Gruener and David Warthen in Berkeley, California. The original software was implemented by Gary Chevsky from his own design. Warthen, Chevsky, Justin Grant, and others built the early AskJeeves.com website around that core engine.

Three venture capital firms, Highland Capital Partners, Institutional Venture Partners, and The RODA Group were early investors.[3] Ask.com is currently owned by InterActiveCorp under the NASDAQ symbol IACI. In late 2010, facing insurmountable competition from Google, the company outsourced its web search technology to an unspecified third party and returned to its roots as a question and answer site.[4] Doug Leeds was appointed from president to CEO in January 2011.[5]

History

Ask.com was originally known as **Ask Jeeves**, where "Jeeves" is the name of the "gentleman's personal gentleman", or valet, fetching answers to any question asked. The character was based on Jeeves, Bertie Wooster's fictional valet from the works of P. G. Wodehouse.

The original idea behind Ask Jeeves was to allow users to get answers to questions posed in everyday, natural language, as well as traditional keyword searching. The current Ask.com still supports this, with added support for math, dictionary, and conversion questions.

Ask.com headquarters in Oakland, California

In 2005, the company announced plans to phase out Jeeves. On February 27, 2006, the character disappeared from Ask.com, and was stated to be "going in to retirement." The U.K./Ireland edition of the website, at uk.ask.com [6], prominently brought the character back in 2009.

InterActiveCorp owns a variety of sites including country-specific sites for UK, Germany, Italy, Japan, the Netherlands, and Spain along with Ask Kids [7], Teoma (now ExpertRank[8]) and several others (see this page for a complete list). On June 5, 2007 Ask.com relaunched with a 3D look.[9]

On May 16, 2006, Ask implemented a "Binoculars Site Preview" into its search results. On search results pages, the "Binoculars" let searchers capture a sneak peek of the page they could visit with a mouse-over activating screenshot pop-up.[10]

In December 2007, Ask released the AskEraser feature,[11] allowing users to opt-out from tracking of search queries and IP and cookie values. They also vowed to erase this data after 18 months if the AskEraser option is not set. HTTP cookies must be enabled for AskEraser to function.[12][13]

On July 4, 2008 InterActiveCorp announced the acquisition of Lexico Publishing Group, which owns Dictionary.com, Thesaurus.com, and Reference.com.[14][15]

On July 26, 2010, Ask.com released a closed-beta Q&A service. The service was released to the public on July 29, 2010.[16] Ask.com launched its mobile Q&A app for the iPhone in late 2010.[17]

Corporate details

Ask Jeeves, Inc. stock traded on the NASDAQ stock exchange from July 1999 to July 2005, under the ticker symbol ASKJ. In July 2005, the ASKJ ticker was retired upon the acquisition by InterActiveCorp, valuing ASKJ at US$1.85 billion.

Ask Sponsored Listings

Ask Sponsored Listings is the search engine marketing tool offered to advertisers to increase the visibility of their websites (and subsequent businesses, services, and products) by producing more prominent and frequent search engine listing.

Ask Toolbar

The Ask Toolbar is a web-browser add-on that can appear as an extra bar added to the browser's window and/or menu. It is often installed during the process of another installation; Ask.com has entered into partnerships with some software security vendors, whereby they are paid to distribute the toolbar alongside their software.

Marketing and promotion

Information-revolution.org campaign

In early 2007, a number of advertisements appeared on London Underground trains warning commuters that 75% of all the information on the web flowed through one site (implied to be Google), with a URL for www.information-revolution.org.[18]

Advertising

Apostolos Gerasoulis, the co-creator of Ask's Teoma algorithmic search technology, starred in four television advertisements in 2007, extolling the virtues of Ask.com's usefulness for information relevance.[19] There was a Jeeves balloon in the 2001 Macy's Thanksgiving Day Parade.

NASCAR sponsorship

On a January 14, 2009, Ask.com became the official sponsor of NASCAR driver Bobby Labonte's No.96 car. Ask would become the official search engine of NASCAR.[20] Ask.com will be the primary sponsor for the No. 96 for 18 of the first 21 races and has rights to increase this to a total of 29 races this season.[21] The Ask.com car debuted in the 2009 Bud Shootout where it failed to finish the race but subsequently has come back strong placing as high as 5th in the March 1, 2009 Shelby 427 race at Las Vegas Motor Speedway.[22] Ask.com's foray into NASCAR is the first instance of its venture into what it calls Super Verticals.[23]

The Simpsons

On January 15th 2012, The Simpsons episode "The D'oh-cial Network" referenced Ask Jeeves as an "Internet failure" at Lisa Simpson's trial (due to her website called "SpringFace" causing 35 deaths due to citizens of Springfield being too preoccupied with SpringFace to pay attention to anything else, causing various accidents around the city).

References

[1] http://www.ask.co.uk/

[2] "Ask.com Site Info" (http://www.alexa.com/siteinfo/Ask.com). Alexa Internet. . Retrieved 2012-03-02.

[3] "Ask Jeeves, Inc. initial public offering prospectus" (http://www.sec.gov/Archives/edgar/data/1054298/0000950149-99-001225.txt). . Retrieved July 12, 2011.

[4] Kopytoff, Verne G. (November 9, 2010). "Ask.com Giving Up Search to Return to Q-and-A Service" (http://www.nytimes.com/2010/11/10/technology/internet/10ask.html?src=busln). *The New York Times*..

[5] "IAC Management" (http://iac.mediaroom.com/index.php?s=20&item=2491). *IAC*. .

[6] http://www.uk.ask.com

[7] http://www.askkids.com/

[8] Ask.com Search Technology (http://about.ask.com/en/docs/about/webmasters.shtml). Retrieved on May 11, 2009.

[9] Major Relaunch For Ask: Ask3D (http://www.techcrunch.com/2007/06/04/major-relaunch-for-ask-ask3d/), Techcrunch, 2007-06-04. Retrieved on June 5, 2007

[10] United States Patent Database (http://patft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=/netahtml/PTO/srchnum.htm&r=1&f=G&l=50&s1=7,047,502.PN.&OS=PN/7,047,502&RS=PN/7,047,502), 2006-06-16. Retrieved on May 16, 2006

[11] Ask.com Takes the Lead on Log Retention; Microsoft and Yahoo! Follow (http://www.eff.org/deeplinks/2007/07/ ask-com-takes-lead-log-retention-microsoft-and-yahoo-follow), eff.org, Retrieved on January 3, 2008

[12] "Does AskEraser Really Erase?" (http://epic.org/privacy/ask/default.html). Electronic Privacy Information Center. . Retrieved March 10, 2008.

[13] "Letter to U.S. Federal Trade Commission" (http://www.cdt.org/privacy/20080123_FTC_Ask.pdf) (PDF). Center for Democracy and Technology. January 23, 2008. . Retrieved March 10, 2008.

[14] Auchard, Eric (July 3, 2008). "Ask.com closes acquisition of Dictionary.com" (http://www.reuters.com/article/internetNews/idUSN0337985120080703?feedType=RSS&feedName=internetNews). Reuters..

[15] "Ask.com closes Dictionary.com deal" (http://news.cnet.com/8300-10784_3-7-0.html?keyword=Dictionary.com). CNet. July 4, 2008..

[16] "Ask.com Q&A Service Drops July 29th" (http://news.softpedia.com/news/Ask-com-Q-A-Service-Drops-July-29th-149176.shtml). Softpedia. July 27, 2010..

[17] Christian, Zibreg (September 24, 2010). "Ask.com has an iPhone app that lets you ask and get local answers" (http://www.geek.com/ articles/mobile/ask-com-has-an-iphone-app-that-lets-you-ask-and-get-local-answers-20100924/). *Geek.com*. .

[18] "– Information Revolution" (http://web.archive.org/web/20070313223519/http://information-revolution.org/). Web.archive.org. March 13, 2007. . Retrieved July 12, 2011.

[19] "About Ask.com: TV Spots" (http://about.ask.com/docs/about/televisionads.shtml). . Retrieved April 25, 2007.

[20] Official Release (January 14, 2009). "– Ask.com enters NASCAR with multi-faceted program – Jan 14, 2009" (http://www.nascar.com/2009/news/headlines/cup/01/14/ask.com.partnerships/index.html). Nascar.com. . Retrieved July 12, 2011.

[21] Duane Cross. "NASCAR.COM – Labonte will drive No. 96 for Hall of Fame in 2009 – Jan 14, 2009" (http://bbs.cid.cn.nascar.com/2009/news/headlines/cup/01/13/blabonte.hof.racing/index.html). Bbs.cid.cn.nascar.com. . Retrieved July 12, 2011.

[22] http://www.ask.com/nascar/2009-Shelby-427-race#results

[23] "Ask.com Partners With NASCAR, Says "Super Verticals" Will Put It Back In Search Race" (http://searchengineland.com/askcom-partners-with-nascar-says-super-vertical-will-put-it-back-in-search-race-16143). Searchengineland.com. January 13, 2009. . Retrieved July 12, 2011.

External links

- Official website (http://www.ask.co.uk)

Yahoo! Search

Yahoo! Search

YAHOO! SEARCH	
Yahoo! Search	
URL	search.yahoo.com [1]
Commercial?	Yes
Type of site	Search Engine
Registration	Optional
Available language(s)	Multilingual (40)
Owner	Yahoo!
Created by	Yahoo!
Launched	March 1, 1995
Alexa rank	4(November2011)[2]
Current status	Active

Yahoo! Search is a web search directory, owned by Yahoo! Inc. and was as of December 2009, the 2nd largest search directory on the web by query volume, at 6.42%, after its competitor Google at 85.35% and before Baidu at 3.67%, according to NetApplications.[3]

Yahoo! Search, originally referred to as *Yahoo!* provided Search interface, would send queries to a searchable index of pages supplemented with its directory of sites. The results were presented to the user under the Yahoo! brand. Originally, none of the actual web crawling and storage/retrieval of data was done by Yahoo! itself. In 2001 the searchable index was powered by Inktomi and later was powered by Google until 2004, when Yahoo! Search became independent.

On July 29, 2009, Microsoft and Yahoo! announced a deal in which Bing would power Yahoo! Search.[] All Yahoo! Search global customers and partners are expected to be transitioned by early 2012.[4]

Search technology acquisition

Seeking to provide its own search engine results, Yahoo! acquired their own search technology.

In 2002, they bought Inktomi, a "behind the scenes" or OEM search engine provider, whose results are shown on other companies' websites and powered Yahoo! in its earlier days. In 2003, they purchased Overture Services, Inc., which owned the AlltheWeb and AltaVista search engines. Initially, even though Yahoo! owned multiple search engines, they didn't use them on the main yahoo.com website, but kept using Google's search engine for its results.

Starting in 2003, Yahoo! Search became its own web crawler-based search engine, with a reinvented crawler called Yahoo! Slurp. Yahoo! Search combined the capabilities of all the search engine companies they had acquired, with its existing research, and put them into a single search engine. The new search engine results were included in all of Yahoo!'s sites that had a web search function. Yahoo! also started to sell the search engine results to other companies, to show on their own web sites. Their relationship with Google was terminated at that time, with the former partners becoming each other's main competitors.

In October 2007, Yahoo! Search was updated with a more modern appearance in line with the redesigned Yahoo! home page. In addition, *Search Assist* was added; which provides real-time query suggestions and related concepts as they are typed.

In July 2008, Yahoo! Search announced the introduction of a new service called "Build Your Own Search Service," or BOSS. This service opens the doors for developers to use Yahoo!'s system for indexing information and images and create their own custom search engine.[5]

In July 2009, Yahoo! signed a deal with Microsoft, the result of which was that Yahoo! Search would be powered by Bing. This is now in effect.[]

Yahoo! Search blog and announcements

The team at Yahoo! Search frequently blogged about search announcements, features, updates and enhancements. The Yahoo! Search Blog, as stated provided *A look inside the world of search from the people at Yahoo!*.[6] This included index updates named *Weather Updates* and their *Yahoo! Search Assist* feature.

International presence

Yahoo! Search also provided their search interface in at least 38 international markets and a variety of available languages.[7] Yahoo! has a presence in Europe, Asia and across the Emerging Markets.

Yahoo UK homepage

Languages

- Arabic
- Bulgarian
- Catalan
- Chinese (Simplified)
- Chinese (Traditional)
- Croatian
- Czech
- Danish
- Dutch
- English
- Estonian
- Finnish
- French
- German
- Greek
- Hebrew
- Hungarian
- Icelandic
- Indonesian
- Italian
- Japanese
- Korean
- Latvian
- Lithuanian
- Malay
- Norwegian
- Persian
- Polish
- Portuguese
- Romanian
- Russian
- Serbian
- Slovak
- Slovenian
- Spanish
- Swedish
- Tagalog
- Thai
- Turkish
- Vietnamese

Search results

Yahoo! Search indexed and cached the common HTML page formats, as well as several of the more popular file-types, such as PDF, Excel spreadsheets, PowerPoint, Word documents, RSS/XML and plain text files. For some of these supported file-types, Yahoo! Search provided *cached* links on their search results allowing for viewing of these file-types in standard HTML.

Using the Advanced Search interface or Preferences settings, Yahoo! Search allowed the customization of search results and enabling of certain settings such as: SafeSearch, Language Selection, Number of results, Domain restrictions, etc.[8]

For a Basic and starter guide to Yahoo! Search, they also provided a *Search Basics* tutorial.[9]

In 2005, Yahoo! began to provide links to previous versions of pages archived on the Wayback Machine.[10]

In the first week of May 2008, Yahoo! launched a new search mash up called *Yahoo! Glue*, which is in beta testing.

Selection-based search

On June 20, 2007, Yahoo! introduced a selection-based search feature called Yahoo! Shortcuts. When activated this selection-based search feature enabled users to invoke search using only their mouse and receive search suggestions in floating windows while remaining on Yahoo! properties such as Yahoo! Mail. This feature was only active on Yahoo web pages or pages within the Yahoo! Publisher Network. Yahoo! Shortcuts required the content-owner to modify the underlying HTML of his or her webpage to call out the specific keywords to be enhanced. The technology for context-aware selection-based search on Yahoo pages was first developed by Reiner Kraft.[11]

SearchScan

On May 11, 2008, Yahoo! introduced SearchScan. If enabled this add-on/feature enhanced Yahoo! Search by automatically alerting users of viruses, spyware and spam websites.[12]

Search verticals

Yahoo! Search provided the ability to search across numerous vertical properties outside just the Web at large. These included Images, Videos, Local, Shopping, Yahoo! Answers, Audio, Directory, Jobs, News, Mobile, Travel and various other services as listed on their *About Yahoo! Search* page.[13]

References

[1] http://search.yahoo.com/
[2] "Alexa yahoo traffic results" (http://www.alexa.com/siteinfo/yahoo.com). Alexa. . Retrieved 2011-02-07.
[3] "Search Engine Market Share as reported by Compete, comScore, Hitwise, Net applications, Nielsen Online & StatCounter" (http:// marketshare.hitslink.com/search-engine-market-share.aspx?qprid=4). antezeta.com. December 20, 2009. . Retrieved 2010-12-16.
[4] When will the change happen? How long will the transition take?(http://help.yahoo.com/l/us/yahoo/search/alliance/alliance-2.html;_ylt=AvrC8b99B5.r4JmW33gA5ChaMnlG) Yahoo! Search Help
[5] "Yahoo! Opens Up Search Technology Infrastructure for Innovative, New Search Experiences, Providing Third Parties with Unprecedented Access, Re-Ranking and Presentation Control of Web Search Results" (http://yhoo.client.shareholder.com/press/releasedetail. cfm?ReleaseID=320623). Yahoo!. July 10, 2008. . Retrieved 2008-07-25.
[6] Yahoo! Search Blog (http://www.ysearchblog.com/)
[7] Yahoo! international presence (http://world.yahoo.com/)
[8] Advanced Search (http://search.yahoo.com/web/advanced?ei=UTF-8)
[9] Search Basics tutorial (http://help.yahoo.com/l/us/yahoo/search/basics/)
[10] Yahoo Cache Now Offers Direct Links to Wayback Machine (http://blog.searchenginewatch.com/050918-143500) Search Engine Watch, September 18 2005
[11] Yahoo! Shortcuts (http://www.ysearchblog.com/archives/000462.html)
[12] Yahoo! SearchScan information page (http://tools.search.yahoo.com/newsearch/searchscan)

[13] About Yahoo! Search(http://tools.search.yahoo.com/about/forsearchers.html)

External links

- Yahoo! Search (http://search.yahoo.com)
- Search Engines: costs vs. benefits (http://www.johnsankey.ca/searchbots.html)

Tim Berners-Lee

Sir
Tim Berners-Lee
OM, KBE, FRS, FREng, FRSA

Berners-Lee in 2010

Born	Tim Berners-Lee 8 June 1955[1] London, England U.K.
Residence	Massachusetts, U.S.[1]
Nationality	English
Alma mater	Queen's College, Oxford
Occupation	Computer scientist
Employer	• World Wide WebConsortium • University of Southampton
Known for	• Inventing the World WideWeb • Holder of the 3Com Founders Chair at MIT's Computer Science and Artificial Intelligence Laboratory
Title	Professor
Religion	Unitarian Universalism
Spouse	Nancy Carlson
Parents	• Conway Berners-Lee • Mary Lee Woods
Website	
www.w3.org/People/Berners-Lee/ [2]	

Sir Timothy John "Tim" Berners-Lee, OM, KBE, FRS, FREng, FRSA (born 8 June 1955[1]), also known as "**TimBL**", is an English computer scientist, MIT professor and the inventor of the World Wide Web. He made a proposal for an information management system in March 1989[3] and on 25 December 1990, with the help of Robert Cailliau and a young student at CERN, he implemented the first successful communication between a Hypertext Transfer Protocol (HTTP) client and server via the Internet.[4]

Berners-Lee is the director of the World Wide Web Consortium (W3C), which oversees the Web's continued development. He is also the founder of the World Wide Web Foundation, and is a senior researcher and holder of the 3Com Founders Chair at the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL).[5] He is a director of The Web Science Research Initiative (WSRI),[6] and a member of the advisory board of the MIT Center for Collective Intelligence.[7][8]

In 2004, Berners-Lee was knighted by Queen Elizabeth II for his pioneering work.[9] In April 2009, he was elected a foreign associate of the United States National Academy of Sciences, based in Washington, D.C.[10][11]

Early life

Tim Berners-Lee was born in southwest London, England, on 8 June 1955, the son of Conway Berners-Lee and Mary Lee Woods. His parents worked on the first commercially built computer, the Ferranti Mark 1. One of four children, he attended Sheen Mount Primary School, and then went on to Emanuel School in London, from 1969 to 1973.[9] He studied at The Queen's College, Oxford, from 1973 to 1976, where he received a first-class degree in Physics.[1]

Career

Berners-Lee, 2005

While being an independent contractor at CERN from June to December 1980, Berners-Lee proposed a project based on the concept of hypertext, to facilitate sharing and updating information among researchers.[12] While there, he built a prototype system named ENQUIRE.[13]

After leaving CERN in 1980, he went to work at John Poole's Image Computer Systems, Ltd, in Bournemouth, England.[14] The project he worked on was a *real-time remote procedure call* which gave him experience in computer networking.[14] In 1984 he returned to CERN as a fellow.[13]

In 1989, CERN was the largest Internet node in Europe, and Berners-Lee saw an opportunity to join hypertext with the Internet: "I just had to take the hypertext idea and connect it to the Transmission Control Protocol and domain name system ideas and—ta-da!—the World Wide Web."[15] "Creating the web was really an act of desperation, because the situation without it was very difficult when I was working at CERN later. Most of the technology involved in the web, like the hypertext, like the Internet, multifont text objects, had all been designed already. I just had to put them together. It was a step of generalising, going to a higher level of abstraction, thinking about all the documentation systems out there as being possibly part of a larger imaginary documentation system."[16] He wrote his initial proposal in March 1989, and in 1990, with the help of Robert Cailliau (with whom he shared the 1995 ACM Software System Award), produced a revision which was accepted by his manager, Mike Sendall.[17] He used similar ideas to those underlying the ENQUIRE system to create the World Wide Web, for which he designed and built the first Web browser. This also functioned as an editor (WorldWideWeb, running on the NeXTSTEP operating system), and the first Web server, CERN HTTPd (short for Hypertext Transfer Protocol daemon).

> " Mike Sendall buys a NeXT cube for evaluation, and gives it to Tim [Berners-Lee]. Tim's prototype implementation on NeXTStep is made in the space of a few months, thanks to the qualities of the NeXTStep software development system. This prototype offers WYSIWYG browsing/authoring! Current Web browsers used in "surfing the Internet" are mere passive windows, depriving the user of the possibility to contribute. During some sessions in the CERN cafeteria, Tim and I try to find a catching name for the system. I was

> determined that the name should not yet again be taken from Greek mythology. Tim proposes "World-Wide Web". I like this very much, except that it is difficult to pronounce in French..." by Robert Cailliau, 2 November 1995.[18]

The first web site built was at CERN within the border of France[19], and was first put online on 6 August 1991:

> "Info.cern.ch was the address of the world's first-ever web site and web server, running on a NeXT computer at CERN. The first web page address was http://info.cern.ch/hypertext/WWW/TheProject. html, which centred on information regarding the WWW project. Visitors could learn more about hypertext, technical details for creating their own webpage, and even an explanation on how to search the Web for information. There are no screenshots of this original page and, in any case, changes were made daily to the information available on the page as the WWW project developed. You may find a later copy (1992) on the World Wide Web Consortium website." -CERN

It provided an explanation of what the World Wide Web was, and how one could use a browser and set up a web server.[20][21][22][23]

In 1994, Berners-Lee founded the W3C at MIT. It comprised various companies that were willing to create standards and recommendations to improve the quality of the Web. Berners-Lee made his idea available freely, with no patent and no royalties due. The World Wide Web Consortium decided that its standards should be based on royalty-free technology, so that they could easily be adopted by anyone.[24]

In 2001, Berners-Lee became a patron of the East Dorset Heritage Trust, having previously lived in Colehill in Wimborne, East Dorset, England.[25]

In December 2004, he accepted a chair in Computer Science at the School of Electronics and Computer Science, University of Southampton, England, to work on his new project, the Semantic Web.[26][27]

Current work

In June 2009 then British Prime Minister Gordon Brown announced Berners-Lee would work with the UK Government to help make data more open and accessible on the Web, building on the work of the Power of Information Task Force.[28] Berners-Lee and Professor Nigel Shadbolt are the two key figures behind data.gov.uk, a UK Government project to open up almost all data acquired for official purposes for free re-use. Commenting on the opening up of Ordnance Survey data in April 2010 Berners-Lee said that: "The changes signal a wider cultural change in Government based on an assumption that information should be in the public domain unless there is a good reason not to—not the other way around." He went on to say "Greater openness, accountability and transparency in Government will give people greater choice and make it easier for individuals to get more directly involved in issues that matter to them."[29]

Tim Berners-Lee at the Home Office, London, on 11 March 2010

In November 2009, Berners-Lee launched the World Wide Web Foundation in order to "Advance the Web to empower humanity by launching transformative programs that build local capacity to leverage the Web as a medium for positive change."[30]

Berners-Lee is one of the pioneer voices in favour of Net Neutrality,[31] and has expressed the view that ISPs should supply "connectivity with no strings attached," and should neither control nor monitor customers' browsing activities without their expressed consent.[32][33] He advocates the idea that net neutrality is a kind of human network right: "Threats to the Internet, such as companies or governments that interfere with or snoop on Internet traffic,

compromise basic human network rights."[34]

In a *Times* article in October 2009, Berners-Lee admitted that the forward slashes ("//") in a web address were actually "unnecessary". He told the newspaper that he could easily have designed URLs not to have the forward slashes. "There you go, it seemed like a good idea at the time," he said in his lighthearted apology.[35]

Recognition

This NeXT Computer was used by Berners-Lee at CERN and became the world's first web server.

- In 1994 he became one of only six members of the World Wide Web Hall of Fame.[36]
- In 1995 he won the Kilby Foundation's "Young Innovator of the Year" Award.[1]
- In 1995 he received also the Software System Award from the Association for Computing Machinery (ACM).[37]
- In 1998 he was awarded with an honorary doctorate from the University of Essex.[38]
- In 1999, *Time Magazine* named Berners-Lee one of the 100 Most Important People of the 20th century.[4]
- In March 2000 he was awarded an honorary degree from The Open University as Doctor of the University.[39]
- In 2001, he was elected a Fellow of the American Academy of Arts and Sciences.[40]
- In 2003, he received the Computer History Museum's Fellow Award, for his seminal contributions to the development of the World Wide Web.[41]
- On 15 April 2004, he was named as the first recipient of Finland's Millennium Technology Prize, for inventing the World Wide Web. The cash prize, worth one million euros (about £892,000, or US$1.3 million, as of Sept 2011), was awarded on 15 June, in Helsinki, Finland, by the President of the Republic of Finland, Tarja Halonen.[42]
- He was appointed to the rank of Knight Commander of the Most Excellent Order of the British Empire (the second-highest class within this Order that entails a knighthood) by Queen Elizabeth II, in the 2004 New Year's Honours List, and was formally invested on 16 July 2004.[9][43]
- On 21 July 2004, he was presented with an honorary Doctor of Science degree from Lancaster University.[44]
- On 27 January 2005, he was named Greatest Briton of 2004, both for his achievements and for displaying the key British characteristics of "diffidence, determination, a sharp sense of humour and adaptability", as put by David Hempleman-Adams, a panel member.[45]
- In 2007, Berners-Lee received the Academy of Achievement's Golden Plate Award.
- In 2007, he was ranked Joint First, alongside Albert Hofmann, in *The Telegraph*'s list of 100 greatest living geniuses.[46]
- On 13 June 2007, he received the Order of Merit, becoming one of only 24 living members entitled to hold the honour, and to use the post-nominals 'O.M.' after their name.[47] (The Order of Merit is within the personal bestowal of The Queen, and does not require recommendation by ministers or the Prime Minister)
- He was awarded the 2008 IEEE/RSE Wolfson James Clerk Maxwell Award, for "conceiving and further developing the World Wide Web".[48]
- On 2 December 2008, Berners-Lee was awarded an honorary doctorate from the University of Manchester. His parents worked on the Manchester Mark 1 in the 1940s and 50s.[49]
- On 21 April 2009, he was awarded an honorary doctorate by the Universidad Politécnica de Madrid.[50]
- On 28 April 2009, he was elected member of the United States National Academy of Sciences.
- On 8 June 2009, he received the Webby Award for Lifetime Achievement, at the awards ceremony held in New York City.[51]

- In October 2009, he was awarded an honorary doctorate by the Vrije Universiteit Amsterdam[52][53]
- On 30 March 2011, he was one of the first three recipients of the Mikhail Gorbachev award for "The Man Who Changed the World", at the inuagural awards ceremony held in London. The other recipients were Evans Wadongo for solar power development and anti-poverty work in Africa, and media mogul Ted Turner.
- On 26 May 2011, Berners-Lee was awarded with an honorary Doctor of Science degree from Harvard University.[54]
- In 2011, he was inducted into IEEE Intelligent Systems' AI's Hall of Fame for the "significant contributions to the field of AI and intelligent systems".[55][56]

Personal life

Berners-Lee had a religious upbringing, but left the Church of England as a teenager, just after being confirmed and "told how essential it was to believe in all kinds of unbelievable things". He and his family eventually joined a Unitarian Universalist church while they were living in Boston. They now live in Lexington, Massachusetts.[57]

Publications

- Berners-Lee, Tim; Mark Fischetti (1999). *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its inventor*. Britain: Orion Business. ISBN 0-7528-2090-7.
- Berners-Lee, T. (2010). "Long Live the Web". *Scientific American* **303** (6): 80–85. doi:10.1038/scientificamerican1210-80. PMID 21141362.
- Berners-Lee, T. (2010). "Long Live the Web". *Scientific American* **303** (6): 80–85. doi:10.1038/scientificamerican1210-80. PMID 21141362.
- Shadbolt, N.; Berners-Lee, T. (2008). "Web science emerges". *Scientific American* **299** (4): 76–81. doi:10.1038/scientificamerican1008-76. PMID 18847088.
- Berners-Lee, T.; Hall, W.; Hendler, J.; Shadbolt, N.; Weitzner, D. (2006). "COMPUTER SCIENCE: Enhanced: Creating a Science of the Web". *Science* **313** (5788): 769–771. doi:10.1126/science.1126902. PMID 16902115.

Notes

[1] "Berners-Lee Longer Biography" (http://www.w3.org/People/Berners-Lee/Longer.html). World Wide Web Consortium. . Retrieved 18 January 2011.

[2] http://www.w3.org/People/Berners-Lee/

[3] "cern.info.ch - Tim Berners-Lee's proposal" (http://info.cern.ch/Proposal.html). Info.cern.ch. . Retrieved 2011-12-21.

[4] Quittner, Joshua (29 March 1999). "Tim Berners Lee—Time 100 People of the Century" (http://www.time.com/time/magazine/article/ 0,9171,990627,00.html). *Time Magazine*. . "He wove the World Wide Web and created a mass medium for the 21st century. The World Wide Web is Berners-Lee's alone. He designed it. He loosed it on the world. And he more than anyone else has fought to keep it open, nonproprietary and free."

[5] "Draper Prize" (http://web.mit.edu/newsoffice/2007/draper-prize.html). Massachusetts Institute of Technology. . Retrieved 25 May 2008.

[6] "People" (http://web.archive.org/web/20080628052526/http://webscience.org/about/people/). The Web Science Research Initiative. Archived from the original (http://webscience.org/about/people/) on 28 June 2008. . Retrieved 17 January 2011.

[7] "MIT Center for Collective Intelligence (homepage)" (http://cci.mit.edu). Cci.mit.edu. . Retrieved 15 August 2010.

[8] "MIT Center for Collective Intelligence (people)" (http://cci.mit.edu/people/index.html). Cci.mit.edu. . Retrieved 15 August 2010.

[9] "Web's inventor gets a knighthood" (http://news.bbc.co.uk/1/hi/technology/3357073.stm). BBC. 31 December 2003. . Retrieved 25 May 2008.

[10] "Timothy Berners-Lee Elected to National Academy of Sciences" (http://www.ddj.com/217200450). Dr. Dobb's Journal. . Retrieved 9 June 2009.

[11] "72 New Members Chosen By Academy" (http://www8.nationalacademies.org/onpinews/newsitem.aspx?RecordID=04282009) (Press release). United States National Academy of Sciences. 28 April 2009. . Retrieved 17 January 2011.

[12] "Berners-Lee's original proposal to CERN" (http://www.w3.org/History/1989/proposal.html). World Wide Web Consortium. March 1989. . Retrieved 25 May 2008.

[13] Stewart, Bill. "Tim Berners-Lee, Robert Cailliau, and the World Wide Web" (http://www.livinginternet.com/w/wi_lee.htm). . Retrieved 22 July 2010.
[14] Tim Berners-Lee. "Frequently asked questions" (http://www.w3.org/People/Berners-Lee/FAQ.html). World Wide Web Consortium. . Retrieved 22 July 2010.
[15] Tim Berners-Lee. "Answers for Young People" (http://www.w3.org/People/Berners-Lee/Kids). World Wide Web Consortium. . Retrieved 25 May 2008.
[16] "Biography and Video Interview of Timothy Berners-Lee at Academy of Achievement" (http://www.achievement.org/autodoc/page/ ber1int-1). Achievement.org. . Retrieved 2011-12-21.
[17] "Ten Years Public Domain for the Original Web Software" (http://tenyears-www.web.cern.ch/tenyears-www/Story/WelcomeStory. html). CERN. . Retrieved 21 July 2010.
[18] Roads and Crossroads of Internet History (http://www.netvalley.com/cgi-bin/intval/net_history.pl?chapter=4) Chapter 4: Birth of the Web
[19] "Tim Berners-Lee. Confirming The Exact Location Where the Web Was Invented" (http://davidgalbraith.org/uncategorized/ the-exact-location-where-the-web-was-invented/2343/). .
[20] "Welcome to info.cern.ch, the website of the world's first-ever web server" (http://info.cern.ch/). CERN. . Retrieved 25 May 2008.
[21] "World Wide Web—Archive of world's first website" (http://www.w3.org/History/19921103-hypertext/hypertext/WWW/TheProject. html). World Wide Web Consortium. . Retrieved 25 May 2008.
[22] "World Wide Web—First mentioned on USENET" (http://groups.google.co.uk/group/alt.hypertext/msg/06dad279804cb3ba?dmode=source&hl=en). Google. 6 August 1991. . Retrieved 25 May 2008.
[23] "The original post to alt.hypertalk describing the WorldWideWeb Project" (http://groups.google.com/group/comp.archives/browse_thread/thread/9fb079523583d42/37bb6783d03a3b0d?lnk=st&q=&rnum=2&hl=en#37bb6783d03a3b0d). *Google Groups*. Google. 9 August 1991. . Retrieved 25 May 2008.
[24] "Patent Policy—5 February 2004" (http://www.w3.org/Consortium/Patent-Policy-20040205/). World Wide Web Consortium. 5 February 2004. . Retrieved 25 May 2008.
[25] John W. Klooster (2009) Icons of invention: the makers of the modern world from Gutenberg to Gates (http://books.google.com/ books?id=WKuG-VIwID8C&pg=PA611&dq=tim+berners+lee++-+east+dorset+heritage+trust&hl=en&ei=0cE9TtyhIcyo8QPQx_zgDA&sa=X&oi=book_result&ct=result&resnum=1&ved=0CCoQ6AEwAA#v=onepage&q=tim berners lee - east dorset heritage trust&f=false) p.611. ABC-CLIO, 2009
[26] Berners-Lee, T.; Hendler, J.; Lassila, O. (2001). "The Semantic Web" (http://www.scientificamerican.com/article. cfm?id=the-semantic-web). *Scientific American* **284** (5): 34. doi:10.1038/scientificamerican0501-34. .
[27] "Tim Berners-Lee, World Wide Web inventor, to join ECS" (http://www.ecs.soton.ac.uk/news/658). World Wide Web Consortium. 2 December 2004. . Retrieved 25 May 2008.
[28] "Tim Berners-Lee" (http://www.w3.org/News/2009#item98). World Wide Web Consortium. 10 June 2009. . Retrieved 10 July 2009.
[29] "Ordnance Survey offers free data access" (http://news.bbc.co.uk/1/hi/technology/8597779.stm). BBC News. 1 April 2010. . Retrieved 3 April 2009.
[30] FAQ—World Wide Web Foundation (http://www.webfoundation.org/about/faq/) Retrieved 18 January 2011
[31] Ghosh, Pallab (15 September 2008). "Web creator rejects net tracking" (http://news.bbc.co.uk/2/hi/technology/7613201.stm). BBC. . Retrieved 15 September 2008. "Warning sounded on web's future."
[32] Cellan-Jones, Rory (March 2008). "Web creator rejects net tracking" (http://news.bbc.co.uk/1/hi/technology/7299875.stm). BBC. . Retrieved 25 May 2008. "Sir Tim rejects net tracking like Phorm."
[33] Adams, Stephen (March 2008). "Web inventor's warning on spy software" (http://www.telegraph.co.uk/news/uknews/1581938/Web-inventor's-warning-on-spy-software.html). *The Daily Telegraph* (London). . Retrieved 25 May 2008. "Sir Tim rejects net tracking like Phorm."
[34] Berners, Tim (2011-05-04). "Tim Berners-Lee, Long Live the Web: A Call for Continued Open Standards and Neutrality, Scientific American Magazine, December 2010" (http://www.scientificamerican.com/article.cfm?id=long-live-the-web). Scientificamerican.com. . Retrieved 2011-12-21.
[35] "Berners-Lee 'sorry' for slashes" (http://news.bbc.co.uk/1/hi/technology/8306631.stm). BBC. 14 October 2009. . Retrieved 14 October 2009.
[36] "The World-Wide Web Hall of Fame" (http://botw.org/1994/awards/fame.html). Best of the Web Directory. .
[37] "Software System Award" (http://awards.acm.org/homepage.cfm?srt=all&awd=149). *ACM Awards*. Association for Computing Machinery. . Retrieved 25 October 2011.
[38] "Honorary Graduates of University of Essex" (http://www.essex.ac.uk/honorary_graduates/hg/profiles/1998/t-berners-lee.aspx).. Retrieved December 15, 2011.
[39] "Open University's online graduation" (http://news.bbc.co.uk/1/hi/education/696176.stm). *BBC NEWS*. 31. March 2000. . Retrieved 22 September 2010.
[40] "Book of Members, 1780–2010: Chapter B" (http://www.amacad.org/publications/BookofMembers/ChapterB.pdf). American Academy of Arts and Sciences. . Retrieved 24 June 2011.
[41] "Fellow Awards | Fellows Home" (http://www.computerhistory.org/fellowawards/index.php?id=88). Computerhistory.org. 11 January 2010. . Retrieved 15 August 2010.

[42] "Millennium Technology Prize 2004 awarded to inventor of World Wide Web" (http://web.archive.org/web/20070830111145/http://www.technologyawards.org/index.php?m=2&s=1&id=16&sm=4). Millennium Technology Prize. Archived from the original (http://www.technologyawards.org/index.php?m=2&s=1&id=16&sm=4) on 30 August 2007. . Retrieved 25 May 2008.

[43] "Creator of the web turns knight" (http://news.bbc.co.uk/1/hi/technology/3899723.stm). BBC. 16 July 2004. . Retrieved 25 May 2008.

[44] "Lancaster University Honorary Degrees, July 2004" (http://domino.lancs.ac.uk/info/lunews.nsf/l/2768F56EB38B32F780256ECC00404E69). Lancaster University. . Retrieved 25 May 2008.

[45] "Three loud cheers for the father of the web" (http://www.telegraph.co.uk/news/uknews/1482211/Three-loud-cheers-for-the-father-of-the-web.html). *The Daily Telegraph* (London). 28 January 2005. . Retrieved 25 May 2008.

[46] 8:01AM GMT 30 Oct 2007 (2007-10-30). ""Top 100 living geniuses" "The Daily Telegraph" 28 October 2007" (http://www.telegraph.co. uk/news/uknews/1567544/Top-100-living-geniuses.html). Telegraph.co.uk. . Retrieved 2011-12-21.

[47] "Web inventor gets Queen's honour" (http://news.bbc.co.uk/1/hi/technology/6750395.stm). BBC. 13 June 2007. . Retrieved 25 May 2008.

[48] "IEEE/RSE Wolfson James Clerk Maxwell Award Recipients" (http://www.ieee.org/documents/maxwell_rl.pdf). IEEE. . Retrieved 4 October 2011.

[49] "Scientific pioneers honoured by The University of Manchester" (http://www.manchester.ac.uk/aboutus/news/display/?id=4216). *manchester.ac.uk*. 2 December 2008. . Retrieved 10 October 2011.

[50] "Universidad Politécnica de Madrid: Berners-Lee y Vinton G. Cerf—Doctores Honoris Causa por la UPM" (http://www2.upm.es/portal/site/institucional/menuitem.fa77d63875fa4490b99bfa04dffb46a8/?vgnextoid=c5d0492bf33c0210VgnVCM10000009c7648aRCRD). . Retrieved 15 August 2010.

[51] Press Release: Sir Tim Berners Lee, Inventor of the World Wide Web, to receive Webby Lifetime Award At the 13th Annual Webby Awards (http://www.webbyawards.com/press/press-release.php?id=187) *Webby Awards.com* Retrieved 21 January 2011

[52] Vrije Universiteit Amsterdam (22 July 2008). "Uitvinder World Wide Web krijgt eredoctoraat Vrije Universiteit" (http://www.vu.nl/nl/ Images/pb 09.082 Eredoctoraat_tcm9-94528.pdf) (in Dutch). . Retrieved 22 July 2009.

[53] NU.nl (22 July 2008). "'Bedenker' wereldwijd web krijgt eredoctoraat VU" (http://www.nu.nl/internet/2046688/ bedenker-wereldwijd-web-krijgt-eredoctoraat-vu.html) (in Dutch). . Retrieved 22 July 2009.

[54] Harvard awards 9 honorary degrees (http://news.harvard.edu/gazette/story/2011/05/harvard-to-award-nine-honorary-degrees/ #berners-lee) *news.harvard.edu* Retrieved 11 June 2011

[55] "AI's Hall of Fame" (http://www.computer.org/cms/Computer.org/ComputingNow/homepage/2011/0811/rW_IS_AIsHallofFame. pdf). *IEEE Intelligent Systems* (IEEE Computer Society) **26** (4): 5–15. 2011. doi:10.1109/MIS.2011.64. .

[56] "IEEE Computer Society Magazine Honors Artificial Intelligence Leaders" (http://www.digitaljournal.com/pr/399442). *DigitalJournal.com*. 24 August 2011. . Retrieved 18 September 2011. Press release source: *PRWeb* (Vocus).

[57] Berners-Lee, Timothy (1998). "The World Wide Web and the "Web of Life"" (http://www.w3.org/People/Berners-Lee/UU.html"). World Wide Web Consortium. . Retrieved 25 May 2008.

Further reading

- *Tim Berners-Lee and the Development of the World Wide Web (Unlocking the Secrets of Science)*, Ann Gaines (Mitchell Lane Publishers, 2001) ISBN 1-58415-096-3
- *Tim Berners-Lee: Inventor of the World Wide Web (Ferguson's Career Biographies)*, Melissa Stewart (Ferguson Publishing Company, 2001) ISBN 0-89434-367-X children's biography
- *Weaving the Web* Berners-Lee, Tim, with Fischetti, Mark (Harper Collins Publishers, 1999) ISBN 0-06-251586-1 (cloth) ISBN 0-06-251587-X (paper)
- *How the Web was Born: The Story of the World Wide Web* Robert Cailliau, James Gillies, R. Cailliau (Oxford University Press, 2000) ISBN 0-19-286207-3
- Tim Berners-Lee Gives the Web a New Definition (http://computemagazine.com/ man-who-invented-world-wide-web-gives-new-definition/)
- BBC2 Newsnight – Transcript of video interview of Berners-Lee on the read/write Web (http://news.bbc.co.uk/2/hi/technology/4132752.stm)
- Technology Review interview (http://www.technologyreview.com/Infotech/13784/)

External links

- Tim Berners-Lee (https://twitter.com/timberners_lee) on Twitter
- timbl (http://identi.ca/timbl) on identi.ca
- Tim Berners-Lee (http://www.ted.com/speakers/tim_berners_lee.html/) at TED Conferences
- Tim Berners-Lee (http://www.imdb.com/name/nm3805083/) at the Internet Movie Database
- Tim Berners-Lee (http://www.nndb.com/people/573/000023504) at the Notable Names Database
- Works by or about Tim Berners-Lee (http://worldcat.org/identities/lccn-no99-10609) in libraries (WorldCat catalog)
- Tim Berners-Lee (http://www.w3.org/People/Berners-Lee/) on the W3C site
- First World Wide Web page (http://www.w3.org/History/19921103-hypertext/hypertext/WWW/TheProject. html)

Web search query

A **web search query** is a query that a user enters into web search engine to satisfy his or her information needs. Web search queries are distinctive in that they are unstructured and often ambiguous; they vary greatly from standard query languages which are governed by strict syntax rules.

Types

There are four broad categories that cover most web search queries[1]:

- **Informational queries** – Queries that cover a broad topic (e.g., *colorado* or *trucks*) for which there may be thousands of relevant results.
- **Navigational queries** – Queries that seek a single website or web page of a single entity (e.g., *youtube* or *delta air lines*).
- **Transactional queries** – Queries that reflect the intent of the user to perform a particular action, like purchasing a car or downloading a screen saver.

Search engines often support a fourth type of query that is used far less frequently:

- **Connectivity queries** – Queries that report on the connectivity of the indexed web graph (e.g., Which links point to this URL?, and How many pages are indexed from this domain name?).

Characteristics

Most commercial web search engines do not disclose their search logs, so information about what users are searching for on the Web is difficult to come by.[2] Nevertheless, a study in 2001[3] analyzed the queries from the Excite search engine showed some interesting characteristics of web search:

- The average length of a search query was 2.4 terms.
- About half of the users entered a single query while a little less than a third of users entered three or more unique queries.
- Close to half of the users examined only the first one or two pages of results (10 results per page).
- Less than 5% of users used advanced search features (e.g., boolean operators like AND, OR, and NOT).
- The top four most frequently used terms were , *(empty search), and, of, and sex.*

A study of the same Excite query logs revealed that 19% of the queries contained a geographic term (e.g., place names, zip codes, geographic features, etc.).[4]

A 2005 study of Yahoo's query logs revealed 33% of the queries from the same user were repeat queries and that 87% of the time the user would click on the same result.[5] This suggests that many users use repeat queries to revisit or re-find information. This analysis is confirmed by a Bing search engine blog post telling about 30% queries are navigational queries [6]

In addition, much research has shown that query term frequency distributions conform to the power law, or *long tail* distribution curves. That is, a small portion of the terms observed in a large query log (e.g. >100 million queries) are used most often, while the remaining terms are used less often individually.[7] This example of the Pareto principle (or *80–20 rule*) allows search engines to employ optimization techniques such as index or database partitioning, caching and pre-fetching.

But in a recent study in 2011 it was found that the average length of queries has grown steadily over time and average length of non-English languages queries had increased more than English queries.[8]

Structured queries

With search engines that support Boolean operators and parentheses, a technique traditionally used by librarians can be applied. A user who is looking for documents that cover several topics or *facets* may want to describe each of them by a disjunction of characteristic words, such as `vehicles OR cars OR automobiles`. A *faceted query* is a conjunction of such facets; e.g. a query such as `(electronic OR computerized OR DRE) AND (voting OR elections OR election OR balloting OR electoral)` is likely to find documents about electronic voting even if they omit one of the words "electronic" and "voting", or even both.[9]

References

[1] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze (2007), Introduction to Information Retrieval (http://nlp.stanford.edu/ IR-book/pdf/19web.pdf), Ch. 19

[2] Dawn Kawamoto and Elinor Mills (2006), AOL apologizes for release of user search data (http://news.com.com/2100-1030_3-6102793. html)

[3] Amanda Spink, Dietmar Wolfram, Major B. J. Jansen, Tefko Saracevic (2001). "Searching the web: The public and their queries". *Journal of the American Society for Information Science and Technology* **52** (3): 226–234. doi:10.1002/1097-4571(2000)9999:9999<::AID-ASI1591>3.3.CO;2-I.

[4] Mark Sanderson and Janet Kohler (2004). "Analyzing geographic queries" (http://www.geo.unizh.ch/~rsp/gir/abstracts/sanderson.pdf). *Proceedings of the Workshop on Geographic Information (SIGIR '04)*. .

[5] Jaime Teevan, Eytan Adar, Rosie Jones, Michael Potts (2005). "History repeats itself: Repeat Queries in Yahoo's query logs" (http://www.csail.mit.edu/~teevan/work/publications/posters/sigir06.pdf). *Proceedings of the 29th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR '06)*. pp. 703–704. doi:10.1145/1148170.1148326. .

[6] http://www.bing.com/community/site_blogs/b/search/archive/2011/02/10/making-search-yours.aspx

[7] Ricardo Baeza-Yates (2005). *Applications of Web Query Mining* (http://www.springerlink.com/content/kpphaktugag5mbv0/). **3408**. Springer Berlin / Heidelberg. pp. 7–22. ISBN 978-3-540-25295-5. .

[8] Mona Taghavi, Ahmed Patel, Nikita Schmidt, Christopher Wills, Yiqi Tew (2011). *An analysis of web proxy logs with query distribution pattern approach for search engines* (http://www.sciencedirect.com/science/article/pii/S0920548911000808). **34**. Elsevier. pp. 162–170. .

[9] Vojkan Mihajlović, Djoerd Hiemstra, Henk Ernst Blok, Peter M.G. Apers. "Exploiting Query Structure and Document Structure to Improve Document Retrieval Effectiveness" (http://eprints.eemcs.utwente.nl/6918/01/TR-CTIT-06-57.pdf).

Web crawling

A **Web crawler** is a computer program that browses the World Wide Web in a methodical, automated manner or in an orderly fashion. Other terms for Web crawlers are *ants*, *automatic indexers*, *bots*,[1] *Web spiders*,[2] *Web robots*,[2] or—especially in the FOAF community—*Web scutters*.[3]

This process is called *Web crawling* or *spidering*. Many sites, in particular search engines, use spidering as a means of providing up-to-date data. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches. Crawlers can also be used for automating maintenance tasks on a Web site, such as checking links or validating HTML code. Also, crawlers can be used to gather specific types of information from Web pages, such as harvesting e-mail addresses (usually for sending spam).

A Web crawler is one type of bot, or software agent. In general, it starts with a list of URLs to visit, called the *seeds*. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the *crawl frontier*. URLs from the frontier are recursively visited according to a set of policies.

The large volume implies that the crawler can only download a fraction of the Web pages within a given time, so it needs to prioritize its downloads. The high rate of change implies that the pages might have already been updated or even deleted.

The number of possible crawlable URLs being generated by server-side software has also made it difficult for web crawlers to avoid retrieving duplicate content. Endless combinations of HTTP GET (URL-based) parameters exist, of which only a small selection will actually return unique content. For example, a simple online photo gallery may offer three options to users, as specified through HTTP GET parameters in the URL. If there exist four ways to sort images, three choices of thumbnail size, two file formats, and an option to disable user-provided content, then the same set of content can be accessed with 48 different URLs, all of which may be linked on the site. This mathematical combination creates a problem for crawlers, as they must sort through endless combinations of relatively minor scripted changes in order to retrieve unique content.

As Edwards *et al.* noted, "Given that the bandwidth for conducting crawls is neither infinite nor free, it is becoming essential to crawl the Web in not only a scalable, but efficient way, if some reasonable measure of quality or freshness is to be maintained."[4] A crawler must carefully choose at each step which pages to visit next.

The behavior of a Web crawler is the outcome of a combination of policies:[5]

- a *selection policy* that states which pages to download,
- a *re-visit policy* that states when to check for changes to the pages,
- a *politeness policy* that states how to avoid overloading Web sites, and
- a *parallelization policy* that states how to coordinate distributed Web crawlers.

Selection policy

Given the current size of the Web, even large search engines cover only a portion of the publicly-available part. A 2005 study showed that large-scale search engines index no more than 40%-70% of the indexable Web;[6] a previous study by Dr. Steve Lawrence and Lee Giles showed that no search engine indexed more than 16% of the Web in 1999.[7] As a crawler always downloads just a fraction of the Web pages, it is highly desirable that the downloaded fraction contains the most relevant pages and not just a random sample of the Web.

This requires a metric of importance for prioritizing Web pages. The importance of a page is a function of its intrinsic quality, its popularity in terms of links or visits, and even of its URL (the latter is the case of vertical search engines restricted to a single top-level domain, or search engines restricted to a fixed Web site). Designing a good selection policy has an added difficulty: it must work with partial information, as the complete set of Web pages is not known during crawling.

Cho *et al.* made the first study on policies for crawling scheduling. Their data set was a 180,000-pages crawl from the `stanford.edu` domain, in which a crawling simulation was done with different strategies.[8] The ordering metrics tested were breadth-first, backlink-count and partial Pagerank calculations. One of the conclusions was that if the crawler wants to download pages with high Pagerank early during the crawling process, then the partial Pagerank strategy is the better, followed by breadth-first and backlink-count. However, these results are for just a single domain. Cho also wrote his Ph.D. dissertation at Stanford on web crawling.[9]

Najork and Wiener performed an actual crawl on 328 million pages, using breadth-first ordering.[10] They found that a breadth-first crawl captures pages with high Pagerank early in the crawl (but they did not compare this strategy against other strategies). The explanation given by the authors for this result is that "the most important pages have many links to them from numerous hosts, and those links will be found early, regardless of on which host or page the crawl originates."

Abiteboul designed a crawling strategy based on an algorithm called OPIC (On-line Page Importance Computation).[11] In OPIC, each page is given an initial sum of "cash" that is distributed equally among the pages it points to. It is similar to a Pagerank computation, but it is faster and is only done in one step. An OPIC-driven crawler downloads first the pages in the crawling frontier with higher amounts of "cash". Experiments were carried in a 100,000-pages synthetic graph with a power-law distribution of in-links. However, there was no comparison with other strategies nor experiments in the real Web.

Boldi *et al.* used simulation on subsets of the Web of 40 million pages from the `.it` domain and 100 million pages from the WebBase crawl, testing breadth-first against depth-first, random ordering and an omniscient strategy. The comparison was based on how well PageRank computed on a partial crawl approximates the true PageRank value. Surprisingly, some visits that accumulate PageRank very quickly (most notably, breadth-first and the omniscent visit) provide very poor progressive approximations.[12][13]

Baeza-Yates *et al.* used simulation on two subsets of the Web of 3 million pages from the `.gr` and `.cl` domain, testing several crawling strategies.[14] They showed that both the OPIC strategy and a strategy that uses the length of the per-site queues are better than breadth-first crawling, and that it is also very effective to use a previous crawl, when it is available, to guide the current one.

Daneshpajouh *et al.* designed a community based algorithm for discovering good seeds.[15] Their method crawls web pages with high PageRank from different communities in less iteration in comparison with crawl starting from random seeds. One can extract good seed from a previously-crawled-Web graph using this new method. Using these seeds a new crawl can be very effective.

Focused crawling

The importance of a page for a crawler can also be expressed as a function of the similarity of a page to a given query. Web crawlers that attempt to download pages that are similar to each other are called **focused crawler** or **topical crawlers**. The concepts of topical and focused crawling were first introduced by Menczer[16][17] and by Chakrabarti *et al.*[18]

The main problem in focused crawling is that in the context of a Web crawler, we would like to be able to predict the similarity of the text of a given page to the query before actually downloading the page. A possible predictor is the anchor text of links; this was the approach taken by Pinkerton[19] in the first web crawler of the early days of the Web. Diligenti *et al.* [20] propose using the complete content of the pages already visited to infer the similarity between the driving query and the pages that have not been visited yet. The performance of a focused crawling depends mostly on the richness of links in the specific topic being searched, and a focused crawling usually relies on a general Web search engine for providing starting points..

Restricting followed links

A crawler may only want to seek out HTML pages and avoid all other MIME types. In order to request only HTML resources, a crawler may make an HTTP HEAD request to determine a Web resource's MIME type before requesting the entire resource with a GET request. To avoid making numerous HEAD requests, a crawler may examine the URL and only request a resource if the URL ends with certain characters such as .html, .htm, .asp, .aspx, .php, .jsp,
.jspx or a slash. This strategy may cause numerous HTML Web resources to be unintentionally skipped.

Some crawlers may also avoid requesting any resources that have a "?" in them (are dynamically produced) in order to avoid spider traps that may cause the crawler to download an infinite number of URLs from a Web site. This strategy is unreliable if the site uses URL rewriting to simplify its URLs.

URL normalization

Crawlers usually perform some type of URL normalization in order to avoid crawling the same resource more than once. The term *URL normalization*, also called *URL canonicalization*, refers to the process of modifying and standardizing a URL in a consistent manner. There are several types of normalization that may be performed including conversion of URLs to lowercase, removal of "." and ".." segments, and adding trailing slashes to the non-empty path component.[21]

Path-ascending crawling

Some crawlers intend to download as many resources as possible from a particular web site. So *path-ascending crawler* was introduced that would ascend to every path in each URL that it intends to crawl.[22] For example, when given a seed URL of http://llama.org/hamster/monkey/page.html, it will attempt to crawl /hamster/monkey/,
/hamster/, and /. Cothey found that a path-ascending crawler was very effective in finding isolated resources, or resources for which no inbound link would have been found in regular crawling.

Many path-ascending crawlers are also known as Web harvesting software, because they're used to "harvest" or collect all the content—perhaps the collection of photos in a gallery—from a specific page or host.

Re-visit policy

The Web has a very dynamic nature, and crawling a fraction of the Web can take weeks or months. By the time a Web crawler has finished its crawl, many events could have happened, including creations, updates and deletions.

From the search engine's point of view, there is a cost associated with not detecting an event, and thus having an outdated copy of a resource. The most-used cost functions are freshness and age.[23]

Freshness: This is a binary measure that indicates whether the local copy is accurate or not. The freshness of a page p in the repository at time t is defined as:

$$F_p(t) = \begin{cases} 1 & \text{if } p \text{ is equal to the local copy at time } t \\ 0 & \text{otherwise} \end{cases}$$

Age: This is a measure that indicates how outdated the local copy is. The age of a page p in the repository, at time t is defined as:

$$A_p(t) = \begin{cases} 0 & \text{if } p \text{ is not modified at time } t \\ t - \text{modification time of } p & \text{otherwise} \end{cases}$$

Coffman *et al.* worked with a definition of the objective of a Web crawler that is equivalent to freshness, but use a different wording: they propose that a crawler must minimize the fraction of time pages remain outdated. They also noted that the problem of Web crawling can be modeled as a multiple-queue, single-server polling system, on which the Web crawler is the server and the Web sites are the queues. Page modifications are the arrival of the customers, and switch-over times are the interval between page accesses to a single Web site. Under this model, mean waiting

time for a customer in the polling system is equivalent to the average age for the Web crawler.[24]

The objective of the crawler is to keep the average freshness of pages in its collection as high as possible, or to keep the average age of pages as low as possible. These objectives are not equivalent: in the first case, the crawler is just concerned with how many pages are out-dated, while in the second case, the crawler is concerned with how old the local copies of pages are.

Two simple re-visiting policies were studied by Cho and Garcia-Molina:[25]

Uniform policy: This involves re-visiting all pages in the collection with the same frequency, regardless of their rates of change.

Proportional policy: This involves re-visiting more often the pages that change more frequently. The visiting frequency is directly proportional to the (estimated) change frequency.

(In both cases, the repeated crawling order of pages can be done either in a random or a fixed order.)

Cho and Garcia-Molina proved the surprising result that, in terms of average freshness, the uniform policy outperforms the proportional policy in both a simulated Web and a real Web crawl. Intuitively, the reasoning is that, as web crawlers have a limit to how many pages they can crawl in a given time frame, (1) they will allocate too many new crawls to rapidly changing pages at the expense of less frequently updating pages, and (2) the freshness of rapidly changing pages lasts for shorter period than that of less frequently changing pages. In other words, a proportional policy allocates more resources to crawling frequently updating pages, but experiences less overall freshness time from them.

To improve freshness, the crawler should penalize the elements that change too often.[26] The optimal re-visiting policy is neither the uniform policy nor the proportional policy. The optimal method for keeping average freshness high includes ignoring the pages that change too often, and the optimal for keeping average age low is to use access frequencies that monotonically (and sub-linearly) increase with the rate of change of each page. In both cases, the optimal is closer to the uniform policy than to the proportional policy: as Coffman *et al.* note, "in order to minimize the expected obsolescence time, the accesses to any particular page should be kept as evenly spaced as possible".[24] Explicit formulas for the re-visit policy are not attainable in general, but they are obtained numerically, as they depend on the distribution of page changes. Cho and Garcia-Molina show that the exponential distribution is a good fit for describing page changes,[26] while Ipeirotis *et al.* show how to use statistical tools to discover parameters that affect this distribution.[27] Note that the re-visiting policies considered here regard all pages as homogeneous in terms of quality ("all pages on the Web are worth the same"), something that is not a realistic scenario, so further information about the Web page quality should be included to achieve a better crawling policy.

Politeness policy

Crawlers can retrieve data much quicker and in greater depth than human searchers, so they can have a crippling impact on the performance of a site. Needless to say, if a single crawler is performing multiple requests per second and/or downloading large files, a server would have a hard time keeping up with requests from multiple crawlers.

As noted by Koster, the use of Web crawlers is useful for a number of tasks, but comes with a price for the general community.[28] The costs of using Web crawlers include:

- network resources, as crawlers require considerable bandwidth and operate with a high degree of parallelism during a long period of time;
- server overload, especially if the frequency of accesses to a given server is too high;
- poorly-written crawlers, which can crash servers or routers, or which download pages they cannot handle; and
- personal crawlers that, if deployed by too many users, can disrupt networks and Web servers.

A partial solution to these problems is the robots exclusion protocol, also known as the robots.txt protocol that is a standard for administrators to indicate which parts of their Web servers should not be accessed by crawlers.[29] This standard does not include a suggestion for the interval of visits to the same server, even though this interval is the

most effective way of avoiding server overload. Recently commercial search engines like Ask Jeeves, MSN and Yahoo are able to use an extra "Crawl-delay:" parameter in the robots.txt file to indicate the number of seconds to delay between requests.

The first proposed interval between connections was 60 seconds.[30] However, if pages were downloaded at this rate from a website with more than 100,000 pages over a perfect connection with zero latency and infinite bandwidth, it would take more than 2 months to download only that entire Website; also, only a fraction of the resources from that Web server would be used. This does not seem acceptable.

Cho uses 10 seconds as an interval for accesses,[25] and the WIRE crawler uses 15 seconds as the default.[31] The MercatorWeb crawler follows an adaptive politeness policy: if it took t seconds to download a document from a given server, the crawler waits for $10t$ seconds before downloading the next page.[32] Dill *et al.* use 1 second.[33]

For those using Web crawlers for research purposes, a more detailed cost-benefit analysis is needed and ethical considerations should be taken into account when deciding where to crawl and how fast to crawl.[34]

Anecdotal evidence from access logs shows that access intervals from known crawlers vary between 20 seconds and 3–4 minutes. It is worth noticing that even when being very polite, and taking all the safeguards to avoid overloading Web servers, some complaints from Web server administrators are received. Brin and Page note that: "... running a crawler which connects to more than half a million servers (...) generates a fair amount of e-mail and phone calls. Because of the vast number of people coming on line, there are always those who do not know what a crawler is, because this is the first one they have seen."[35]

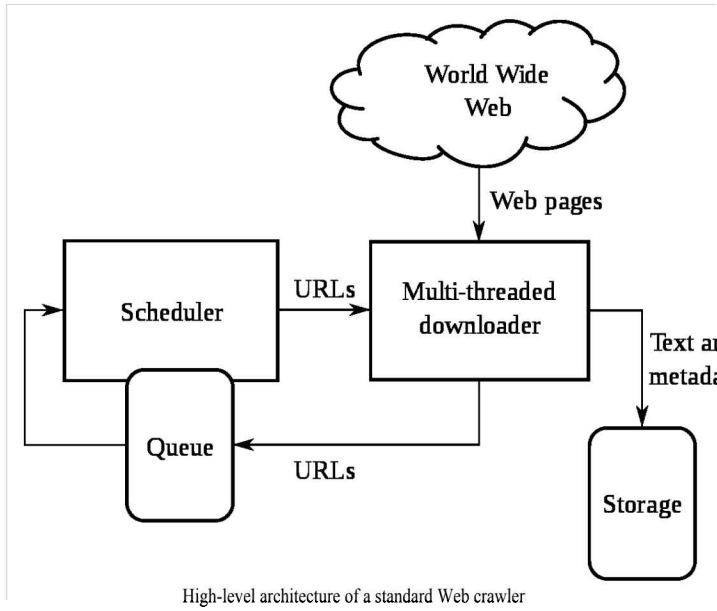
Parallelization policy

A parallel crawler is a crawler that runs multiple processes in parallel. The goal is to maximize the download rate while minimizing the overhead from parallelization and to avoid repeated downloads of the same page. To avoid downloading the same page more than once, the crawling system requires a policy for assigning the new URLs discovered during the crawling process, as the same URL can be found by two different crawling processes.

Architectures

A crawler must not only have a good crawling strategy, as noted in the previous sections, but it should also have a highly optimized architecture. Shkapenyuk and Suel noted that:[36]

> While it is fairly easy to build a slow crawler that downloads a few pages per second for a short period of time, building a high-performance system that can download hundreds of millions of pages over several weeks presents a number of challenges in system design, I/O and network efficiency, and robustness and manageability.



High-level architecture of a standard Web crawler

Web crawlers are a central part of search engines, and details on their algorithms and architecture are kept as business secrets. When crawler designs are published, there is often an important lack of detail that prevents others from reproducing the work. There are also emerging concerns about "search engine spamming", which prevent major search engines from publishing their ranking algorithms.

Crawler identification

Web crawlers typically identify themselves to a Web server by using the User-agent field of an HTTP request. Web site administrators typically examine their Web servers' log and use the user agent field to determine which crawlers have visited the web server and how often. The user agent field may include a URL where the Web site administrator may find out more information about the crawler. Spambots and other malicious Web crawlers are unlikely to place identifying information in the user agent field, or they may mask their identity as a browser or other well-known crawler.

It is important for Web crawlers to identify themselves so that Web site administrators can contact the owner if needed. In some cases, crawlers may be accidentally trapped in a crawler trap or they may be overloading a Web server with requests, and the owner needs to stop the crawler. Identification is also useful for administrators that are interested in knowing when they may expect their Web pages to be indexed by a particular search engine.

Examples

The following is a list of published crawler architectures for general-purpose crawlers (excluding focused web crawlers), with a brief description that includes the names given to the different components and outstanding features:

- **Yahoo! Slurp** is the name of the Yahoo Search crawler.
- **Bingbot** is the name of Microsoft's Bing webcrawler. It replaced **Msnbot**.
- **FAST Crawler**[37] is a distributed crawler, used by Fast Search & Transfer, and a general description of its architecture is available.
- **Googlebot**[35] is described in some detail, but the reference is only about an early version of its architecture, which was based in C++ and Python. The crawler was integrated with the indexing process, because text parsing was done for full-text indexing and also for URL extraction. There is a URL server that sends lists of URLs to be fetched by several crawling processes. During parsing, the URLs found were passed to a URL server that checked if the URL have been previously seen. If not, the URL was added to the queue of the URL server.
- **PolyBot**[36] is a distributed crawler written in C++ and Python, which is composed of a "crawl manager", one or more "downloaders" and one or more "DNS resolvers". Collected URLs are added to a queue on disk, and processed later to search for seen URLs in batch mode. The politeness policy considers both third and second level domains (e.g.: www.example.com and www2.example.com are third level domains) because third level domains are usually hosted by the same Web server.
- **RBSE**[38] was the first published web crawler. It was based on two programs: the first program, "`spider`" maintains a queue in a relational database, and the second program "`mite`", is a modified `www` ASCII browser that downloads the pages from the Web.
- **WebCrawler**[19] was used to build the first publicly-available full-text index of a subset of the Web. It was based on lib-WWW to download pages, and another program to parse and order URLs for breadth-first exploration of the Web graph. It also included a real-time crawler that followed links based on the similarity of the anchor text with the provided query.
- **World Wide Web Worm**[39] was a crawler used to build a simple index of document titles and URLs. The index could be searched by using the `grep` Unix command.
- **WebFountain**[4] is a distributed, modular crawler similar to Mercator but written in C++. It features a "controller" machine that coordinates a series of "ant" machines. After repeatedly downloading pages, a change

rate is inferred for each page and a non-linear programming method must be used to solve the equation system for maximizing freshness. The authors recommend to use this crawling order in the early stages of the crawl, and then switch to a uniform crawling order, in which all pages are being visited with the same frequency.

- **WebRACE**[40] is a crawling and caching module implemented in Java, and used as a part of a more generic system called eRACE. The system receives requests from users for downloading web pages, so the crawler acts in part as a smart proxy server. The system also handles requests for "subscriptions" to Web pages that must be monitored: when the pages change, they must be downloaded by the crawler and the subscriber must be notified. The most outstanding feature of WebRACE is that, while most crawlers start with a set of "seed" URLs, WebRACE is continuously receiving new starting URLs to crawl from.

In addition to the specific crawler architectures listed above, there are general crawler architectures published by Cho[41] and Chakrabarti.[42]

Open-source crawlers

- **Aspseek** is a crawler, indexer and a search engine written in C++ and licensed under the GPL
- **DataparkSearch** is a crawler and search engine released under the GNU General Public License.
- **GNU Wget** is a command-line-operated crawler written in C and released under the GPL. It is typically used to mirror Web and FTP sites.
- **GRUB** is an open source distributed search crawler that Wikia Search used to crawl the web.
- **Heritrix** is the Internet Archive's archival-quality crawler, designed for archiving periodic snapshots of a large portion of the Web. It was written in Java.
- **ht://Dig** includes a Web crawler in its indexing engine.
- **HTTrack** uses a Web crawler to create a mirror of a web site for off-line viewing. It is written in C and released under the GPL.
- **ICDL Crawler** is a cross-platform web crawler written in C++ and intended to crawl Web sites based on Web-site Parse Templates using computer's free CPU resources only.
- **mnoGoSearch** is a crawler, indexer and a search engine written in C and licensed under the GPL (Linux machines only)
- **Nutch** is a crawler written in Java and released under an Apache License. It can be used in conjunction with the Lucene text-indexing package.
- **Open Search Server** is a search engine and web crawler software release under the GPL.
- **Pavuk** is a command-line Web mirror tool with optional X11 GUI crawler and released under the GPL. It has bunch of advanced features compared to wget and httrack, e.g., regular expression based filtering and file creation rules.
- **PHP-Crawler** is a simple PHP and MySQL based crawler released under the BSD. Easy to install it became popular for small MySQL-driven websites on shared hosting.
- the **tkWWW Robot**, a crawler based on the tkWWW web browser (licensed under GPL).
- **YaCy**, a free distributed search engine, built on principles of peer-to-peer networks (licensed under GPL).
- **Seeks**, a free distributed search engine (licensed under Affero General Public License).

Crawling the Deep Web

A vast amount of Web pages lie in the deep or invisible Web.[43] These pages are typically only accessible by submitting queries to a database, and regular crawlers are unable to find these pages if there are no links that point to them. Google's Sitemap Protocol and mod oai[44] are intended to allow discovery of these deep-Web resources.

Deep Web crawling also multiplies the number of Web links to be crawled. Some crawlers only take some of the `<a href="URL">`-shaped URLs. In some cases, such as the Googlebot, Web crawling is done on all text contained inside the hypertext content, tags, or text.

Crawling Web 2.0 Applications

- Sheeraj Shah provides insight into Crawling Ajax-driven Web 2.0 Applications[45].
- Interested readers might wish to read AJAXSearch: Crawling, Indexing and Searching Web 2.0 Applications[46].
- Making AJAX Applications Crawlable[47], from Google Code. It defines an agreement between web servers and search engine crawlers that allows for dynamically created content to be visible to crawlers. Google currently supports this agreement.[45]

References

[1] Kobayashi, M. and Takeda, K. (2000). "Information retrieval on the web" (http://doi.acm.org/10.1145/358923.358934). *ACM Computing Surveys* (ACM Press) **32** (2): 144–173. doi:10.1145/358923.358934. .

[2] Spetka, Scott. "The TkWWW Robot: Beyond Browsing" (http://web.archive.org/web/20040903174942/archive.ncsa.uiuc.edu/SDG/IT94/Proceedings/Agents/spetka/spetka.html). NCSA. Archived from the original (http://archive.ncsa.uiuc.edu/SDG/IT94/Proceedings/Agents/spetka/spetka.html) on 3 September 2004. . Retrieved 21 November 2010.

[3] See definition of scutter on FOAF Project's wiki (http://wiki.foaf-project.org/w/Scutter)

[4] Edwards, J., McCurley, K. S., and Tomlin, J. A. (2001). "An adaptive model for optimizing performance of an incremental web crawler" (http://www10.org/cdrom/papers/210/index.html). *In Proceedings of the Tenth Conference on World Wide Web* (Hong Kong: Elsevier Science): 106–113. doi:10.1145/371920.371960. .

[5] Castillo, Carlos (2004). *Effective Web Crawling* (http://chato.cl/research/crawling_thesis) (Ph.D. thesis). University of Chile. . Retrieved 2010-08-03.

[6] Gulli, A.; Signorini, A. (2005). "The indexable web is more than 11.5 billion pages" (http://doi.acm.org/10.1145/1062745.1062789). *Special interest tracks and posters of the 14th international conference on World Wide Web*. ACM Press.. pp. 902–903. doi:10.1145/1062745.1062789. .

[7] Lawrence, Steve; C. Lee Giles (1999-07-08). "Accessibility of information on the web". *Nature* **400** (6740): 107. doi:10.1038/21987. PMID 10428673.

[8] Cho, J.; Garcia-Molina, H.; Page, L. (1998-04). "Efficient Crawling Through URL Ordering" (http://ilpubs.stanford.edu:8090/347/). *Seventh International World-Wide Web Conference*. Brisbane, Australia. . Retrieved 2009-03-23.

[9] Cho, Junghoo, "Crawling the Web: Discovery and Maintenance of a Large-Scale Web Data" (http://oak.cs.ucla.edu/~cho/papers/ cho-thesis.pdf), Ph.D. dissertation, Department of Computer Science, Stanford University, November 2001

[10] Marc Najork and Janet L. Wiener. Breadth-first crawling yields high-quality pages (http://www10.org/cdrom/papers/pdf/p208.pdf). In Proceedings of the Tenth Conference on World Wide Web, pages 114–118, Hong Kong, May 2001. Elsevier Science.

[11] Abiteboul, Serge; Mihai Preda, Gregory Cobena (2003). "Adaptive on-line page importance computation" (http://www2003.org/cdrom/ papers/refereed/p007/p7-abiteboul.html). *Proceedings of the 12th international conference on World Wide Web*. Budapest, Hungary: ACM. pp. 280–290. doi:10.1145/775152.775192. ISBN 1-58113-680-3. . Retrieved 2009-03-22.

[12] Boldi, Paolo; Bruno Codenotti, Massimo Santini, Sebastiano Vigna (2004). "UbiCrawler: a scalable fully distributed Web crawler" (http://vigna.dsi.unimi.it/ftp/papers/UbiCrawler.pdf). *Software: Practice and Experience* **34** (8): 711–726. doi:10.1002/spe.587. . Retrieved 2009-03-23.

[13] Boldi, Paolo; Massimo Santini, Sebastiano Vigna (2004). "Do Your Worst to Make the Best: Paradoxical Effects in PageRank Incremental Computations" (http://vigna.dsi.unimi.it/ftp/papers/ParadoxicalPageRank.pdf). *Algorithms and Models for the Web-Graph* (http://springerlink.com/content/g10m122f9hb6). pp. 168–180. . Retrieved 2009-03-23.

[14] Baeza-Yates, R., Castillo, C., Marin, M. and Rodriguez, A. (2005). Crawling a Country: Better Strategies than Breadth-First for Web Page Ordering (http://www.dcc.uchile.cl/~ccastill/papers/baeza05_crawling_country_better_breadth_first_web_page_ordering.pdf). In Proceedings of the Industrial and Practical Experience track of the 14th conference on World Wide Web, pages 864–872, Chiba, Japan. ACM Press.

[15] Shervin Daneshpajouh, Mojtaba Mohammadi Nasiri, Mohammad Ghodsi, A Fast Community Based Algorithm for Generating Crawler Seeds Set (http://ce.sharif.edu/~daneshpajouh/publications/A Fast Community Based Algorithm for Generating Crawler Seeds Set.pdf),

In proceeding of 4th International Conference on Web Information Systems and Technologies (WEBIST-2008 (http://www.webist.org/)), Funchal, Portugal, May 2008.

[16] Menczer, F. (1997). ARACHNID: Adaptive Retrieval Agents Choosing Heuristic Neighborhoods for Information Discovery (http://informatics.indiana.edu/fil/Papers/ICML.ps). In D. Fisher, ed., Machine Learning: Proceedings of the 14th International Conference (ICML97). Morgan Kaufmann

[17] Menczer, F. and Belew, R.K. (1998). Adaptive Information Agents in Distributed Textual Environments (http://informatics.indiana.edu/ fil/Papers/AA98.ps). In K. Sycara and M. Wooldridge (eds.) Proc. 2nd Intl. Conf. on Autonomous Agents (Agents '98). ACM Press

[18] Chakrabarti, S., van den Berg, M., and Dom, B. (1999). Focused crawling: a new approach to topic-specific web resource discovery (http://web.archive.org/web/20040317210216/http://www.fxpal.com/people/vdberg/pubs/www8/www1999f.pdf). Computer Networks, 31(11–16):1623–1640.

[19] Pinkerton, B. (1994). Finding what people want: Experiences with the WebCrawler (http://web.archive.org/web/20010904075500/http://archive.ncsa.uiuc.edu/SDG/IT94/Proceedings/Searching/pinkerton/WebCrawler.html). In Proceedings of the First World Wide Web Conference, Geneva, Switzerland.

[20] Diligenti, M., Coetzee, F., Lawrence, S., Giles, C. L., and Gori, M. (2000). Focused crawling using context graphs (http://nautilus.dii.unisi.it/pubblicazioni/files/conference/2000-Diligenti-VLDB.pdf). In Proceedings of 26th International Conference on Very Large Databases (VLDB), pages 527-534, Cairo, Egypt.

[21] Pant, Gautam; Srinivasan, Padmini; Menczer, Filippo (2004). "Crawling the Web" (http://dollar.biz.uiowa.edu/~pant/Papers/crawling. pdf). In Levene, Mark; Poulovassilis, Alexandra. *Web Dynamics: Adapting to Change in Content, Size, Topology and Use*. Springer. pp. 153–178. ISBN 9783540406761. . Retrieved 2009-03-22.

[22] Cothey, Viv (2004). "Web-crawling reliability". *Journal of the American Society for Information Science and Technology* **55** (14): 1228–1238. doi:10.1002/asi.20078.

[23] Cho, Junghoo; Hector Garcia-Molina (2000). "Synchronizing a database to improve freshness" (http://www.cs.brown.edu/courses/ cs227/2002/cache/Cho.pdf). *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. Dallas, Texas, United States: ACM. pp. 117–128. doi:10.1145/342009.335391. ISBN 1-58113-217-4. . Retrieved 2009-03-23.

[24] Jr, E. G. Coffman; Zhen Liu, Richard R. Weber (1998). "Optimal robot scheduling for Web search engines". *Journal of Scheduling* **1** (1): 15–29. doi:10.1002/(SICI)1099-1425(199806)1:1<15::AID-JOS3>3.0.CO;2-K.

[25] Cho, J. and Garcia-Molina, H. (2003). Effective page refresh policies for web crawlers (http://portal.acm.org/citation.cfm?doid=958942. 958945). ACM Transactions on Database Systems, 28(4).

[26] Cho, Junghoo; Hector Garcia-Molina (2003). "Estimating frequency of change" (http://portal.acm.org/citation.cfm?doid=857166. 857170). *ACM Trans. Interet Technol.* **3** (3): 256–290. doi:10.1145/857166.857170. . Retrieved 2009-03-22.

[27] Ipeirotis, P., Ntoulas, A., Cho, J., Gravano, L. (2005) Modeling and managing content changes in text databases (http://pages.stern.nyu.edu/~panos/publications/icde2005.pdf). In Proceedings of the 21st IEEE International Conference on Data Engineering, pages 606-617, April 2005, Tokyo.

[28] Koster, M. (1995). Robots in the web: threat or treat? ConneXions, 9(4).

[29] Koster, M. (1996). A standard for robot exclusion (http://www.robotstxt.org/wc/exclusion.html).

[30] Koster, M. (1993). Guidelines for robots writers (http://www.robotstxt.org/wc/guidelines.html).

[31] Baeza-Yates, R. and Castillo, C. (2002). Balancing volume, quality and freshness in Web crawling (http://www.chato.cl/papers/ baeza02balancing.pdf). In Soft Computing Systems – Design, Management and Applications, pages 565–572, Santiago, Chile. IOS Press Amsterdam.

[32] Heydon, Allan; Najork, Marc (1999-06-26) (PDF). *Mercator: A Scalable, Extensible Web Crawler* (http://www.cindoc.csic.es/cybermetrics/pdf/68.pdf). . Retrieved 2009-03-22.

[33] Dill, S., Kumar, R., Mccurley, K. S., Rajagopalan, S., Sivakumar, D., and Tomkins, A. (2002). Self-similarity in the web (http://www.mccurley.org/papers/fractal.pdf). ACM Trans. Inter. Tech., 2(3):205–223.

[34] "Web crawling ethics revisited: Cost, privacy and denial of service" (http://www.scit.wlv.ac.uk/~cm1993/papers/Web_Crawling_Ethics_preprint.doc). *Journal of the American Society for Information Science and Technology*. 2006. .

[35] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine (http://infolab.stanford.edu/~backrub/google. html). Computer Networks and ISDN Systems, 30(1-7):107–117.

[36] Shkapenyuk, V. and Suel, T. (2002). Design and implementation of a high performance distributed web crawler (http://cis.poly.edu/tr/tr-cis-2001-03.pdf). In Proceedings of the 18th International Conference on Data Engineering (ICDE), pages 357-368, San Jose, California. IEEE CS Press.

[37] Risvik, K. M. and Michelsen, R. (2002). Search Engines and Web Dynamics (http://citeseer.ist.psu.edu/rd/1549722,509701,1,0.25,Download/http://citeseer.ist.psu.edu/cache/papers/cs/26004/http:zSzzSzwww.idi.ntnu.nozSz~algkonzSzgenereltzSzse-dynamicweb1.pdf/risvik02search.pdf). Computer Networks, vol. 39, pp. 289–302, June 2002.

[38] Eichmann, D. (1994). The RBSE spider: balancing effective search against Web load (http://mingo.info-science.uiowa.edu/eichmann/ www94/Spider.ps). In Proceedings of the First World Wide Web Conference, Geneva, Switzerland.

[39] McBryan, O. A. (1994). GENVL and WWWW: Tools for taming the web. In Proceedings of the First World Wide Web Conference, Geneva, Switzerland.

[40] Zeinalipour-Yazti, D. and Dikaiakos, M. D. (2002). Design and implementation of a distributed crawler and filtering processor (http://www.cs.ucr.edu/~csyiazti/downloads/papers/ngits02/ngits02.pdf). In Proceedings of the Fifth Next Generation Information Technologies

and Systems (NGITS), volume 2382 of Lecture Notes in Computer Science, pages 58–74, Caesarea, Israel. Springer.

[41] Cho, Junghoo; Hector Garcia-Molina (2002). "Parallel crawlers" (http://portal.acm.org/citation.cfm?id=511464). *Proceedings of the 11th international conference on World Wide Web*. Honolulu, Hawaii, USA: ACM. pp. 124–135. doi:10.1145/511446.511464. ISBN 1-58113-449-5. . Retrieved 2009-03-23.

[42] Chakrabarti, S. (2003). Mining the Web (http://www.cs.berkeley.edu/~soumen/mining-the-web/). Morgan Kaufmann Publishers. ISBN 1-55860-754-4

[43] Shestakov, Denis (2008). Search Interfaces on the Web: Querying and Characterizing (https://oa.doria.fi/handle/10024/38506). *TUCS Doctoral Dissertations 104, University of Turku*

[44] Nelson, Michael L; Herbert Van de Sompel, Xiaoming Liu, Terry L Harrison, Nathan McFarland (2005-03-24). "mod_oai: An Apache Module for Metadata Harvesting". *Eprint arXiv:cs/0503069*: 3069. arXiv:cs/0503069. Bibcode 2005cs.....3069N.

[45] Making AJAX Applications Crawlable: Full Specification (http://code.google.com/web/ajaxcrawling/docs/specification.html)

Further reading

- Cho, Junghoo, "Web Crawling Project" (http://oak.cs.ucla.edu/~cho/research/crawl.html), UCLA Computer Science Department.

Social search

Social search or a **social search engine** is a type of web search that takes into account the Social Graph of the person initiating the search query. When applied to web search this Social-Graph approach to relevance is in contrast to established algorithmic or machine-based approaches where relevance is determined by analyzing the text of each document or the link structure of the documents.[1] Search results produced by **social search engine** give more visibility to content created or touched by users in the Social Graph.

Social search takes many forms, ranging from simple shared bookmarks or tagging of content with descriptive labels to more sophisticated approaches that combine human intelligence with computer algorithms.[2][3]

The search experience takes into account varying sources of metadata, such as collaborative discovery of web pages, tags, social ranking, commenting on bookmarks, news, images, videos, knowledge sharing, podcasts and other web pages. Example forms of user input include social bookmarking or direct interaction with the search results such as promoting or demoting results the user feels are more or less relevant to their query.[4]

History

The term social search began to emerge between 2004 and 2005. The concept of social ranking can be considered to derive from Google's PageRank algorithm, which assigns importance to web pages based on analysis of the link structure of the web, because PageRank is relying on the collective judgment of webmasters linking to other content on the web. Links, in essence, are positive votes by the webmaster community for their favorite sites.

In 2008, there were a few startup companies that focused on ranking search results according to one's social graph on social networks.[5][6] Companies in the social search space include Wajam, folkd, Slangwho, Sproose, Mahalo, Jumper 2.0, Qitera, Scour, Wink, Eurekster, Baynote, Delver, OneRiot, and SideStripe. Former efforts include Wikia Search. In 2008, a story on *TechCrunch* showed Google potentially adding in a voting mechanism to search results similar to Digg's methodology.[7] This suggests growing interest in how social groups can influence and potentially enhance the ability of algorithms to find meaningful data for end users. There are also other services like Sentimnt that turn search personal by searching within the users' social circles.

The term 'Lazyweb' has been used to describe the act of out-sourcing your questions to your friends, usually by broadcasting them on Twitter or Facebook (as opposed to posting them on Q&A websites such as Yahoo Answers). The company Aardvark, acquired by Google in February 2010, has created a more targeted version of this, which directs your questions to people in your social networks, based on relating the content of the question to the content of their social network pages. Aardvark users primarily use the Aardvark IM buddy, also integrated into Google

Gmail, to ask and answer their questions. The company Cofacio released a beta platform in August 2009 in the UK which marks a return to the open, broadcast method of social search for the Twitter/Facebook generation.

In October 2009, Google rolled out its "Social Search" feature; after a time in beta, the feature was expanded to multiple languages in May 2011. However, after a search deal with Twitter ended without renewal, Google began to retool its Social Search. In January 2012, Google released "Search plus Your World", a further development of Social Search. The feature, which is integrated into Google's regular search as an opt-out feature, pulls references to results from Google+ profiles. The company was subsequently criticized by Twitter for the perceived potential impact of "Search plus Your World" upon web publishers, describing the feature's release to the public as a "bad day for the web", while Google replied that Twitter refused to allow deep search crawling by Google of Twitter's content[8].

Benefits

To date social search engines have not demonstrated measurably improved search results over algorithmic search engines. However, there are potential benefits deriving from the human input qualities of social search.

- Reduced impact of link spam by relying less on link structure of web pages.
- Increased relevance because each result has been selected by users.
- Leverage a network of trusted individuals by providing an indication of whether they thought a particular result was good or bad.
- The introduction of 'human judgement' suggests that each web page has been viewed and endorsed by one or more people, and they have concluded it is relevant and worthy of being shared with others using human techniques that go beyond the computer's current ability to analyze a web page.
- Web pages are considered to be relevant from the reader's perspective, rather than the author who desires their content to be viewed, or the web master as they create links.
- More current results. Because a social search engine is constantly getting feedback it is potentially able to display results that are more current or in context with changing information.

Concerns

- Risk of spam. Because users can directly add results to a social search engine there is a risk that some users could insert search spam directly into the search engine. Elimination or prevention of this spam would require the ability to detect the validity of a user's' contribution, such as whether it agrees with other trusted users.
- "The Long Tail" of search is a concept that there are so many unique searches conducted that most searches, while valid, are performed very infrequently. A search engine that relied on users filling in all the searches would be at a disadvantage to one that used machines to crawl and index the entire web.

References

[1] What's the Big Deal With Social Search? (http://searchenginewatch.com/showPage.html?page=3623153), SearchEngineWatch, Aug 15, 2006
[2] Chi, Ed H. Information Seeking Can Be Social, Computer, vol. 42, no. 3, pp. 42-46, Mar. 2009, doi:10.1109/MC.2009.87 (http://www2.computer.org/portal/web/csdl/doi/10.1109/MC.2009.87)
[3] A Taxonomy of Social Search Approaches (http://blog.delver.com/index.php/2008/07/31/taxonomy-of-social-search-approaches/), Delver company blog, Jul 31, 2008
[4] Google's Marissa Mayer: Social search is the future (http://venturebeat.com/2008/01/31/ googles-marissa-mayer-social-search-is-the-future), VentureBeat, Jan 31, 2008
[5] New Sites Make It Easier To Spy on Your Friends (http://online.wsj.com/public/article/SB121063460767286631.html), Wall Street Journal, May 13. 2008
[6] Social Search Guide: 40+ Social Search Engines (http://mashable.com/2007/08/27/social-search/), Mashable, Aug 27. 2007
[7] Is This The Future Of Search? (http://www.techcrunch.com/2008/07/16/is-this-the-future-of-search/), TechCrunch, July 16, 2008

[8] "Twitter unhappy about Google's social search changes" (http://www.bbc.co.uk/news/technology-16511794). BBC News. 11 January 2012. . Retrieved 11 January 2012.

Vertical search

A **vertical search** engine, as distinct from a general web search engine, focuses on a specific segment of online content. The vertical content area may be based on topicality, media type, or genre of content. Common verticals include shopping, the automotive industry, legal information, medical information, and travel. In contrast to general Web search engines, which attempt to index large portions of the World Wide Web using a web crawler, vertical search engines typically use a focused crawler that attempts to index only Web pages that are relevant to a pre-defined topic or set of topics.

Some vertical search sites focus on individual verticals, while other sites include multiple vertical searches within one search engine.

Vertical search offers several potential benefits over general search engines:

- Greater precision due to limited scope
- Leverage domain knowledge including taxonomies and ontologies
- Support specific unique user tasks

Domain-specific search

Domain-specific verticals focus on a specific topic. John Battelle describes this in his book, *The Search*:

> Domain-specific search solutions focus on one area of knowledge, creating customized search experiences, that because of the domain's limited corpus and clear relationships between concepts, provide extremely relevant results for searchers.[1]

References

[1] Battelle, John (2005). *The Search: How Google and its Rivals Rewrote the Rules of Business and Transformed Our Culture*. New York: Portfolio.

Web analytics

Web analytics is the measurement, collection, analysis and reporting of internet data for purposes of understanding and optimizing web usage.[1]

Web analytics is not just a tool for measuring web traffic but can be used as a tool for business research and market research, and to assess and improve the effectiveness of a web site. Web analytics applications can also help companies measure the results of traditional print advertising campaigns. It helps one to estimate how traffic to a website changes after the launch of a new advertising campaign. Web analytics provides information about the number of visitors to a website and the number of page views. It helps gauge traffic and popularity trends which is useful for market research.

There are two categories of web analytics; *off-site* and *on-site* web analytics.

Off-site web analytics refers to web measurement and analysis regardless of whether you own or maintain a website. It includes the measurement of a website's *potential* audience (opportunity), share of voice (visibility), and buzz (comments) that is happening on the Internet as a whole.

On-site web analytics measure a visitor's journey once *on your website*. This includes its drivers and conversions; for example, which landing pages encourage people to make a purchase. On-site web analytics measures the performance of your website in a commercial context. This data is typically compared against key performance indicators for performance, and used to improve a website or marketing campaign's audience response.

Historically, web analytics has referred to on-site visitor measurement. However in recent years this has blurred, mainly because vendors are producing tools that span both categories.

On-site web analytics technologies

Many different vendors provide on-site web analytics software and services. There are two main technological approaches to collecting the data. The first method, *log file analysis*, reads the logfiles in which the web server records all its transactions. The second method, *page tagging*, uses JavaScript or images on each page to notify a third-party server when a page is rendered by a web browser. Both collect data that can be processed to produce web traffic reports.

In addition other data sources may also be added to augment the data. For example; e-mail response rates, direct mail campaign data, sales and lead information, user performance data such as click heat mapping, or other custom metrics as needed.

Web server logfile analysis

Web servers record some of their transactions in a logfile. It was soon realized that these logfiles could be read by a program to provide data on the popularity of the website. Thus arose web log analysis software.

In the early 1990s, web site statistics consisted primarily of counting the number of client requests (or *hits*) made to the web server. This was a reasonable method initially, since each web site often consisted of a single HTML file. However, with the introduction of images in HTML, and web sites that spanned multiple HTML files, this count became less useful. The first true commercial Log Analyzer was released by IPRO in 1994.[2]

Two units of measure were introduced in the mid 1990s to gauge more accurately the amount of human activity on web servers. These were *page views* and *visits* (or *sessions*). A *page view* was defined as a request made to the web server for a page, as opposed to a graphic, while a *visit* was defined as a sequence of requests from a uniquely identified client that expired after a certain amount of inactivity, usually 30 minutes. The page views and visits are still commonly displayed metrics, but are now considered rather rudimentary.

The emergence of search engine spiders and robots in the late 1990s, along with web proxies and dynamically assigned IP addresses for large companies and ISPs, made it more difficult to identify unique human visitors to a website. Log analyzers responded by tracking visits by cookies, and by ignoring requests from known spiders.

The extensive use of web caches also presented a problem for logfile analysis. If a person revisits a page, the second request will often be retrieved from the browser's cache, and so no request will be received by the web server. This means that the person's path through the site is lost. Caching can be defeated by configuring the web server, but this can result in degraded performance for the visitor and bigger load on the servers.

Page tagging

Concerns about the accuracy of logfile analysis in the presence of caching, and the desire to be able to perform web analytics as an outsourced service, led to the second data collection method, page tagging or 'Web bugs'.

In the mid 1990s, Web counters were commonly seen — these were images included in a web page that showed the number of times the image had been requested, which was an estimate of the number of visits to that page. In the late 1990s this concept evolved to include a small invisible image instead of a visible one, and, by using JavaScript, to pass along with the image request certain information about the page and the visitor. This information can then be processed remotely by a web analytics company, and extensive statistics generated.

The web analytics service also manages the process of assigning a cookie to the user, which can uniquely identify them during their visit and in subsequent visits. Cookie acceptance rates vary significantly between web sites and may affect the quality of data collected and reported.

Collecting web site data using a third-party data collection server (or even an in-house data collection server) requires an additional DNS look-up by the user's computer to determine the IP address of the collection server. On occasion, delays in completing a successful or failed DNS look-ups may result in data not being collected.

With the increasing popularity of Ajax-based solutions, an alternative to the use of an invisible image, is to implement a call back to the server from the rendered page. In this case, when the page is rendered on the web browser, a piece of Ajax code would call back to the server and pass information about the client that can then be aggregated by a web analytics company. This is in some ways flawed by browser restrictions on the servers which can be contacted with XmlHttpRequest objects. Also, this method can lead to slightly lower reported traffic levels, since the visitor may stop the page from loading in mid-response before the Ajax call is made.

Logfile analysis vs page tagging

Both logfile analysis programs and page tagging solutions are readily available to companies that wish to perform web analytics. In some cases, the same web analytics company will offer both approaches. The question then arises of which method a company should choose. There are advantages and disadvantages to each approach.[3]

Advantages of logfile analysis

The main advantages of logfile analysis over page tagging are as follows:

- The web server normally already produces logfiles, so the raw data is already available. No changes to the website are required.
- The data is on the company's own servers, and is in a standard, rather than a proprietary, format. This makes it easy for a company to switch programs later, use several different programs, and analyze historical data with a new program.
- Logfiles contain information on visits from search engine spiders, which generally do not execute JavaScript on a page and are therefore not recorded by page tagging. Although these should not be reported as part of the human activity, it is useful information for search engine optimization.
- Logfiles require no additional DNS Lookups. Thus there are no external server calls which can slow page load speeds, or result in uncounted page views.

- The web server reliably records every transaction it makes, including e.g. serving PDF documents and content generated by scripts, and does not rely on the visitors' browsers co-operating

Advantages of page tagging

The main advantages of page tagging over logfile analysis are as follows:

- Counting is activated by opening the page (given that the web client runs the tag scripts), not requesting it from the server. If a page is cached, it will not be counted by the server. Cached pages can account for up to one-third of all pageviews. Not counting cached pages seriously skews many site metrics. It is for this reason server-based log analysis is not considered suitable for analysis of human activity on websites.
- Data is gathered via a component ("tag") in the page, usually written in JavaScript, though Java can be used, and increasingly Flash is used. JQuery and AJAX can also be used in conjunction with a server-side scripting language (such as PHP) to manipulate and (usually) store it in a database, basically enabling complete control over how the data is represented.
- The script may have access to additional information on the web client or on the user, not sent in the query, such as visitors' screen sizes and the price of the goods they purchased.
- Page tagging can report on events which do not involve a request to the web server, such as interactions within Flash movies, partial form completion, mouse events such as onClick, onMouseOver, onFocus, onBlur etc.
- The page tagging service manages the process of assigning cookies to visitors; with logfile analysis, the server has to be configured to do this.
- Page tagging is available to companies who do not have access to their own web servers.
- Lately page tagging has become a standard in web analytics.[4]

Economic factors

Logfile analysis is almost always performed in-house. Page tagging can be performed in-house, but it is more often provided as a third-party service. The economic difference between these two models can also be a consideration for a company deciding which to purchase.

- Logfile analysis typically involves a one-off software purchase; however, some vendors are introducing maximum annual page views with additional costs to process additional information. In addition to commercial offerings, several open-source logfile analysis tools are available free of charge.
- For Logfile analysis you have to store and archive your own data, which often grows very large quickly. Although the cost of hardware to do this is minimal, the overhead for an IT department can be considerable.
- For Logfile analysis you need to maintain the software, including updates and security patches.
- Complex page tagging vendors charge a monthly fee based on volume i.e. number of pageviews per month collected.

Which solution is cheaper to implement depends on the amount of technical expertise within the company, the vendor chosen, the amount of activity seen on the web sites, the depth and type of information sought, and the number of distinct web sites needing statistics.

Regardless of the vendor solution or data collection method employed, the cost of web visitor analysis and interpretation should also be included. That is, the cost of turning raw data into actionable information. This can be from the use of third party consultants, the hiring of an experienced web analyst, or the training of a suitable in-house person. A cost-benefit analysis can then be performed. For example, what revenue increase or cost savings can be gained by analysing the web visitor data?

Hybrid methods

Some companies are now producing programs that collect data through both logfiles and page tagging. By using a hybrid method, they aim to produce more accurate statistics than either method on its own. The first Hybrid solution was produced in 1998 by Rufus Evison, who then spun the product out to create a company based upon the increased accuracy of hybrid methods.

Geolocation of visitors

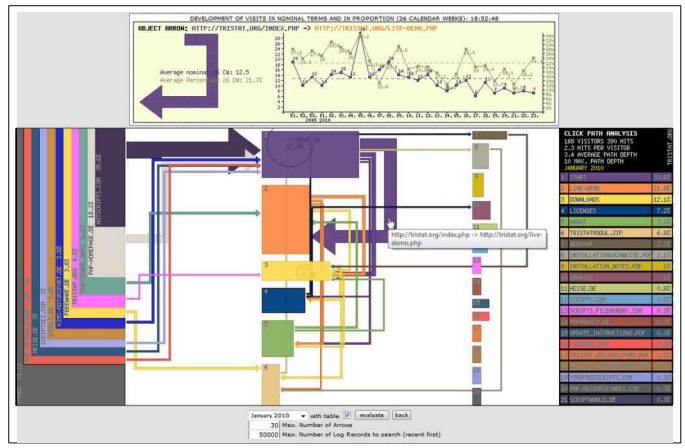
With IP geolocation, it is possible to track visitors location. Using IP geolocation database or API, visitors can be geolocated to city, region or country level.[5]

IP Intelligence, or Internet Protocol (IP) Intelligence, is a technology that maps the Internet and catalogues IP addresses by parameters such as geographic location (country, region, state, city and postcode), connection type, Internet Service Provider (ISP), proxy information, and more. The first generation of IP Intelligence was referred to as geotargeting or geolocation technology. This information is used by businesses for online audience segmentation in applications such online advertising, behavioral targeting, content localization (or website localization), digital rights management, personalization, online fraud detection, geographic rights management, localized search, enhanced analytics, global traffic management, and content distribution.

Click analytics

Click analytics is a special type of web analytics that gives special attention to clicks.

Commonly, click analytics focuses on on-site analytics. An editor of a web site uses click analytics to determine the performance of his or her particular site, with regards to where the users of the site are clicking.



Clickpath Analysis with referring pages on the left and arrows and rectangles differing in thickness and expanse to symbolize movement quantity.

Also, click analytics may happen real-time or "unreal"-time, depending on the type of information sought. Typically, front-page editors on high-traffic news media sites will want to monitor their pages in real-time, to optimize the content. Editors, designers or other types of stakeholders may analyze clicks on a wider time frame to aid them assess performance of writers, design elements or advertisements etc.

Data about clicks may be gathered in at least two ways. Ideally, a click is "logged" when it occurs, and this method requires some functionality that picks up relevant information when the event occurs. Alternatively, one may institute the assumption that a page view is a result of a click, and therefore log a simulated click that led to that page view.

Customer lifecycle analytics

Customer lifecycle analytics is a visitor-centric approach to measuring that falls under the umbrella of lifecycle marketing. Page views, clicks and other events (such as API calls, access to third-party services, etc.) are all tied to an individual visitor instead of being stored as separate data points. Customer lifecycle analytics attempts to connect all the data points into a marketing funnel that can offer insights into visitor behavior and website optimization.

Other methods

Other methods of data collection are sometimes used. Packet sniffing collects data by sniffing the network traffic passing between the web server and the outside world. Packet sniffing involves no changes to the web pages or web servers. Integrating web analytics into the web server software itself is also possible.[6] Both these methods claim to provide better real-time data than other methods.

Key definitions

There are no globally agreed definitions within web analytics as the industry bodies have been trying to agree definitions that are useful and definitive for some time. The main bodies who have had input in this area have been JICWEBS (The Joint Industry Committee for Web Standards in the UK and Ireland) [7], ABCe (Audit Bureau of Circulations electronic, UK and Europe) [8], The WAA (Web Analytics Association, US) and to a lesser extent the IAB (Interactive Advertising Bureau). This does not prevent the following list from being a useful guide, suffering only slightly from ambiguity. Both the WAA and the ABCe provide more definitive lists for those who are declaring their statistics using the metrics defined by either.

- **Hit** - A request for a file from the web server. Available only in log analysis. The number of hits received by a website is frequently cited to assert its popularity, but this number is extremely misleading and dramatically over-estimates popularity. A single web-page typically consists of multiple (often dozens) of discrete files, each of which is counted as a hit as the page is downloaded, so the number of hits is really an arbitrary number more reflective of the complexity of individual pages on the website than the website's actual popularity. The total number of visitors or page views provides a more realistic and accurate assessment of popularity.
- **Page view** - A request for a file whose type is defined as a page in log analysis. An occurrence of the script being run in page tagging. In log analysis, a single page view may generate multiple hits as all the resources required to view the page (images, .js and .css files) are also requested from the web server.
- **Visit / Session** - A visit is defined as a series of page requests from the same uniquely identified client with a time of no more than 30 minutes between each page request. A session is defined as a series of page requests from the same uniquely identified client with a time of no more than 30 minutes and no requests for pages from other domains intervening between page requests. In other words, a session ends when someone goes to another site, or 30 minutes elapse between page views, whichever comes first. A visit ends only after a 30 minute time delay. If someone leaves a site, then returns within 30 minutes, this will count as one visit but two sessions. In practice, most systems ignore sessions and many analysts use both terms for visits. Because time between page views is critical to the definition of visits and sessions, a single page view does not constitute a visit or a session (it is a "bounce").
- **First Visit / First Session** - (also known as 'Absolute Unique Visitor) A visit from a visitor who has not made any previous visits.
- **Visitor / Unique Visitor / Unique User** - The uniquely identified client generating requests on the web server (log analysis) or viewing pages (page tagging) within a defined time period (i.e. day, week or month). A Unique Visitor counts once within the timescale. A visitor can make multiple visits. Identification is made to the visitor's computer, not the person, usually via cookie and/or IP+User Agent. Thus the same person visiting from two different computers or with two different browsers will count as two Unique Visitors. Increasingly visitors are uniquely identified by Flash LSO's (Local Shared Object), which are less susceptible to privacy enforcement.

- **Repeat Visitor** - A visitor that has made at least one previous visit. The period between the last and current visit is called visitor recency and is measured in days.
- **New Visitor** - A visitor that has not made any previous visits. This definition creates a certain amount of confusion (see common confusions below), and is sometimes substituted with analysis of first visits.
- **Impression** - An impression is each time an advertisement loads on a user's screen. Anytime you see a banner, that is an impression.
- **Singletons** - The number of visits where only a single page is viewed (a 'bounce'). While not a useful metric in and of itself the number of singletons is indicative of various forms of Click fraud as well as being used to calculate bounce rate and in some cases to identify automatons bots.
- **Bounce Rate** - The percentage of visits where the visitor enters and exits at the same page without visiting any other pages on the site in between.
- **% Exit** - The percentage of users who exit from a page.
- **Visibility time** - The time a single page (or a blog, Ad Banner...) is viewed.
- **Session Duration** - Average amount of time that visitors spend on the site each time they visit. This metric can be complicated by the fact that analytics programs can not measure the length of the final page view.[9]
- **Page View Duration / Time on Page** - Average amount of time that visitors spend on each page of the site. As with Session Duration, this metric is complicated by the fact that analytics programs can not measure the length of the final page view unless they record a page close event, such as onUnload().
- **Active Time / Engagement Time** - Average amount of time that visitors spend actually interacting with content on a web page, based on mouse moves, clicks, hovers and scrolls. Unlike Session Duration and Page View Duration / Time on Page, this metric **can** accurately measure the length of engagement in the final page view.
- **Page Depth / Page Views per Session** - Page Depth is the average number of page views a visitor consumes before ending their session. It is calculated by dividing total number of page views by total number of sessions and is also called Page Views per Session or PV/Session.
- **Frequency / Session per Unique** - Frequency measures how often visitors come to a website. It is calculated by dividing the total number of sessions (or visits) by the total number of unique visitors. Sometimes it is used to measure the loyalty of your audience.
- **Click path** - the sequence of hyperlinks one or more website visitors follows on a given site.
- **Click** - "refers to a single instance of a user following a hyperlink from one page in a site to another".[10] Some use click analytics to analyze their web sites.
- **Site Overlay** is a technique in which graphical statistics are shown beside each link on the web page. These statistics represent the percentage of clicks on each link.

Common sources of confusion in web analytics

The hotel problem

The hotel problem is generally the first problem encountered by a user of web analytics. The problem is that the unique visitors for each day in a month do not add up to the same total as the unique visitors for that month. This appears to an inexperienced user to be a problem in whatever analytics software they are using. In fact it is a simple property of the metric definitions.

The way to picture the situation is by imagining a hotel. The hotel has two rooms (Room A and Room B).

	Day 1	Day 2	Day 3	Total
Room A	John	John	Jane	2 Unique Users
Room B	Mark	Jane	Mark	2 Unique Users
Total	2	2	2	?

As the table shows, the hotel has two unique users each day over three days. The sum of the totals with respect to the days is therefore six.

During the period each room has had two unique users. The sum of the totals with respect to the rooms is therefore four.

Actually only three visitors have been in the hotel over this period. The problem is that a person who stays in a room for two nights will get counted twice if you count them once on each day, but is only counted once if you are looking at the total for the period. Any software for web analytics will sum these correctly for whatever time period, thus leading to the problem when a user tries to compare the totals.

New visitors + Repeat visitors unequal to total visitors

Another common misconception in web analytics is that the sum of the new visitors and the repeat visitors ought to be the total number of visitors. Again this becomes clear if the visitors are viewed as individuals on a small scale, but still causes a large number of complaints that analytics software cannot be working because of a failure to understand the metrics.

Here the culprit is the metric of a new visitor. There is really no such thing as a new visitor when you are considering a web site from an ongoing perspective. If a visitor makes their first visit on a given day and then returns to the web site on the same day they are both a new visitor and a repeat visitor for that day. So if we look at them as an individual which are they? The answer has to be both, so the definition of the metric is at fault.

A new visitor is not an individual; it is a fact of the web measurement. For this reason it is easiest to conceptualize the same facet as a first visit (or first session). This resolves the conflict and so removes the confusion. Nobody expects the number of first visits to add to the number of repeat visitors to give the total number of visitors. The metric will have the same number as the new visitors, but it is clearer that it will not add in this fashion.

On the day in question there was a first visit made by our chosen individual. There was also a repeat visit made by the same individual. The number of first visits and the number of repeat visits will add up to the total number of visits for that day.

Web analytics methods

Problems with cookies

Historically, vendors of page-tagging analytics solutions have used third-party cookies sent from the vendor's domain instead of the domain of the website being browsed. Third-party cookies can handle visitors who cross multiple unrelated domains within the company's site, since the cookie is always handled by the vendor's servers.

However, third-party cookies in principle allow tracking an individual user across the sites of different companies, allowing the analytics vendor to collate the user's activity on sites where he provided personal information with his activity on other sites where he thought he was anonymous. Although web analytics companies deny doing this, other companies such as companies supplying banner ads have done so. Privacy concerns about cookies have therefore led a noticeable minority of users to block or delete third-party cookies. In 2005, some reports showed that about 28% of Internet users blocked third-party cookies and 22% deleted them at least once a month.[11]

Most vendors of page tagging solutions have now moved to provide at least the option of using first-party cookies (cookies assigned from the client subdomain).

Another problem is cookie deletion. When web analytics depend on cookies to identify unique visitors, the statistics are dependent on a persistent cookie to hold a unique visitor ID. When users delete cookies, they usually delete both first- and third-party cookies. If this is done between interactions with the site, the user will appear as a first-time visitor at their next interaction point. Without a persistent and unique visitor id, conversions, click-stream analysis, and other metrics dependent on the activities of a unique visitor over time, cannot be accurate.

Cookies are used because IP addresses are not always unique to users and may be shared by large groups or proxies. In some cases, the IP address is combined with the user agent in order to more accurately identify a visitor if cookies are not available. However, this only partially solves the problem because often users behind a proxy server have the same user agent. Other methods of uniquely identifying a user are technically challenging and would limit the trackable audience or would be considered suspicious. Cookies are the selected option because they reach the lowest common denominator without using technologies regarded as spyware.

Secure analytics (metering) methods

All the methods described above (and some other methods not mentioned here, like sampling) have the central problem of being vulnerable to manipulation (both inflation and deflation). This means these methods are imprecise and insecure (in any reasonable model of security). This issue has been addressed in a number of papers[12][13][14]
,[15] but to-date the solutions suggested in these papers remain theoretic, possibly due to lack of interest from the engineering community, or because of financial gain the current situation provides to the owners of big websites. For more details, consult the aforementioned papers.

References

[1] The Official WAA Definition of Web Analytics (http://www.webanalyticsassociation.org/?page=aboutus)

[2] Web Traffic Data Sources and Vendor Comparison (http://www.advanced-web-metrics.com/docs/web-data-sources.pdf) by Brian Clifton and Omega Digital Media Ltd

[3] Increasing Accuracy for Online Business Growth (http://www.advanced-web-metrics.com/blog/2008/02/16/accuracy-whitepaper/) - a web analytics accuracy whitepaper

[4] Revisiting log file analysis versus page tagging (http://web.analyticsblog.ca/2010/02/revisiting-log-file-analysis-versus-page-tagging/) - McGill University Web Analytics blog article (CMIS 530)

[5] IPInfoDB (2009-07-10). "IP geolocation database" (http://ipinfodb.com/ip_database.php). IPInfoDB. . Retrieved 2009-07-19.

[6] Web analytics integrated into web software itself (http://portal.acm.org/citation.cfm?id=1064677.1064679&coll=GUIDE&dl=GUIDE&CFID=66492168&CFTOKEN=93187844)

[7] http://www.jicwebs.org/

[8] http://www.abc.org.uk/

[9] ClickTale Blog » Blog Archive » What Google Analytics Can't Tell You, Part 1 (http://blog.clicktale.com/2009/10/14/ what-google-analytics-cant-tell-you-part-1/)

[10] Clicks - Analytics Help (http://www.google.com/support/googleanalytics/bin/answer.py?hl=en&answer=32981)

[11] clickz report (http://www.clickz.com/showPage.html?page=3489636)

[12] Naor, M.; Pinkas, B. (1998). "Secure and efficient metering". *Advances in Cryptology — EUROCRYPT'98*. Lecture Notes in Computer Science. **1403**. pp. 576. doi:10.1007/BFb0054155. ISBN 3-540-64518-7.

[13] Naor, M.; Pinkas, B. (1998). "Secure accounting and auditing on the Web". *Computer Networks and ISDN Systems* **30**: 541. doi:10.1016/S0169-7552(98)00116-0.

[14] Franklin, M. K.; Malkhi, D. (1997). "Auditable metering with lightweight security". *Financial Cryptography*. Lecture Notes in Computer Science. **1318**. pp. 151. doi:10.1007/3-540-63594-7_75. ISBN 978-3-540-63594-9.

[15] Johnson, R.; Staddon, J. (2007). "Deflation-secure web metering". *International Journal of Information and Computer Security* **1**: 39. doi:10.1504/IJICS.2007.012244.

Bibliography

- Clifton, Brian (2010) Advanced Web Metrics with Google Analytics, 2nd edition, Sybex (Paperback.)
- Kaushik, Avinash (2009) Web Analytics 2.0 - The Art of Online Accountability and Science of Customer Centricity. Sybex, Wiley.
- Mortensen, Dennis R. (2009) Yahoo! Web Analytics. Sybex.
- Farris, P., Bendle, N.T., Pfeifer, P.E. Reibstein, D.J. (2009) Key Marketing Metrics The 50+ Metrics Every Manager needs to know, Prentice Hall, London.
- Plaza, B (2009) Monitoring web traffic source effectiveness with Google Analytics: An experiment with time series. *Aslib Proceedings*, 61(5):474–482.
- Arikan, Akin (2008) Multichannel Marketing. Metrics and Methods for On and Offline Success. Sybex.
- Tullis, Tom & Albert, Bill (2008) Measuring the User Experience. Collecting, Analyzing and Presenting Usability Metrics. Morgan Kaufmann, Elsevier, Burlington MA.
- Kaushik, Avinash (2007) Web Analytics: An Hour a Day, Sybex, Wiley.
- Bradley N (2007) Marketing Research. Tools and Techniques. Oxford University Press, Oxford.
- Burby, Jason and Atchison, Shane (2007) Actionable Web Analytics: Using Data to Make Smart Business Decisions.
- Davis, J. (2006) 'Marketing Metrics: How to create Accountable Marketing plans that really work' John Wiley & Sons (Asia).
- Peterson Eric T (2005) Web Site Measurement Hacks. O'Reilly ebook.
- Peterson Eric T (2004) Web Analytics Demystified: A Marketer's Guide to Understanding How Your Web Site Affects Your Business. Celilo Group Media
- Lenskold, J. (2003) 'Marketing ROI: how to plan, Measure and Optimise strategies for Profit' London: McGraw Hill Contemporary
- Sterne, J. (2002) Web metrics, Proven Methods for Measuring Web Site Success, London: John Wiley & Sons.
- Srinivasan , J .(2001) E commerce Metrics, Models and Examples, London: Prentice Hall.

External links

- Technology enablers and business goals for web analytics initiatives (http://www.joelichtenberg.com/2011/ 02/02/web-analytics-â□"-overview-options-and-technology-enablers/)
- ABCe (Audit Bureau of Circulations electronic, UK and Europe), (http://www.abc.org.uk/)
- JICWEBS (The Joint Industry Committee for Web Standards in the UK and Ireland) (http://www.jicwebs.org/)

Pay per click

Pay per click (**PPC**) (also called **Cost per click**) is an Internet advertising model used to direct traffic to websites, where advertisers pay the publisher (typically a website owner) when the ad is clicked. With search engines, advertisers typically bid on keyword phrases relevant to their target market. Content sites commonly charge a fixed price per click rather than use a bidding system. PPC "display" advertisements are shown on web sites or search engine results with related content that have agreed to show ads. This approach differs from the "pay per impression" methods used in television and newspaper advertising.

In contrast to the generalized portal, which seeks to drive a high volume of traffic to one site, PPC implements the so-called affiliate model, that provides purchase opportunities wherever people may be surfing. It does this by offering financial incentives (in the form of a percentage of revenue) to affiliated partner sites. The affiliates provide purchase-point click-through to the merchant. It is a pay-for-performance model: If an affiliate does not generate sales, it represents no cost to the merchant. Variations include banner exchange, pay-per-click, and revenue sharing programs.

Websites that utilize PPC ads will display an advertisement when a keyword query matches an advertiser's keyword list, or when a content site displays relevant content. Such advertisements are called *sponsored links* or *sponsored ads*, and appear adjacent to or above organic results on search engine results pages, or anywhere a web developer chooses on a content site.[1]

Among PPC providers, Google AdWords, Yahoo! Search Marketing, and Microsoft adCenter are the three largest network operators, and all three operate under a bid-based model. [1]

The PPC advertising model is open to abuse through click fraud, although Google and others have implemented automated systems[2] to guard against abusive clicks by competitors or corrupt web developers.[3]

Determining cost per click

There are two primary models for determining cost per click: flat-rate and bid-based. In both cases the advertiser must consider the potential value of a click from a given source. This value is based on the type of individual the advertiser is expecting to receive as a visitor to his or her website, and what the advertiser can gain from that visit, usually revenue, both in the short term as well as in the long term. As with other forms of advertising targeting is key, and factors that often play into PPC campaigns include the target's interest (often defined by a search term they have entered into a search engine, or the content of a page that they are browsing), intent (e.g., to purchase or not), location (for geo targeting), and the day and time that they are browsing.

Flat-rate PPC

In the flat-rate model, the advertiser and publisher agree upon a fixed amount that will be paid for each click. In many cases the publisher has a rate card that lists the Cost Per Click (CPC) within different areas of their website or network. These various amounts are often related to the content on pages, with content that generally attracts more valuable visitors having a higher CPC than content that attracts less valuable visitors. However, in many cases advertisers can negotiate lower rates, especially when committing to a long-term or high-value contract.

The flat-rate model is particularly common to comparison shopping engines, which typically publish rate cards.[4] However, these rates are sometimes minimal, and advertisers can pay more for greater visibility. These sites are usually neatly compartmentalized into product or service categories, allowing a high degree of targeting by advertisers. In many cases, the entire core content of these sites is paid ads.

Bid-based PPC

In the bid-based model, the advertiser signs a contract that allows them to compete against other advertisers in a private auction hosted by a publisher or, more commonly, an advertising network. Each advertiser informs the host of the maximum amount that he or she is willing to pay for a given ad spot (often based on a keyword), usually using online tools to do so. The auction plays out in an automated fashion every time a visitor triggers the ad spot.

When the ad spot is part of a search engine results page (SERP), the automated auction takes place whenever a search for the keyword that is being bid upon occurs. All bids for the keyword that target the searcher's geo-location, the day and time of the search, etc. are then compared and the winner determined. In situations where there are multiple ad spots, a common occurrence on SERPs, there can be multiple winners whose positions on the page are influenced by the amount each has bid. The ad with the highest bid generally shows up first, though additional factors such as ad quality and relevance can sometimes come into play (see Quality Score).

In addition to ad spots on SERPs, the major advertising networks allow for contextual ads to be placed on the properties of 3rd-parties with whom they have partnered. These publishers sign up to host ads on behalf of the network. In return, they receive a portion of the ad revenue that the network generates, which can be anywhere from 50% to over 80% of the gross revenue paid by advertisers. These properties are often referred to as a *content network* and the ads on them as *contextual ads* because the ad spots are associated with keywords based on the context of the page on which they are found. In general, ads on content networks have a much lower click-through rate (CTR) and conversion rate (CR) than ads found on SERPs and consequently are less highly valued. Content network properties can include websites, newsletters, and e-mails.[5]

Advertisers pay for each click they receive, with the actual amount paid based on the amount bid. It is common practice amongst auction hosts to charge a winning bidder just slightly more (e.g. one penny) than the next highest bidder or the actual amount bid, whichever is lower.[6] This avoids situations where bidders are constantly adjusting their bids by very small amounts to see if they can still win the auction while paying just a little bit less per click.

To maximize success and achieve scale, automated bid management systems can be deployed. These systems can be used directly by the advertiser, though they are more commonly used by advertising agencies that offer PPC bid management as a service. These tools generally allow for bid management at scale, with thousands or even millions of PPC bids controlled by a highly automated system. The system generally sets each bid based on the goal that has been set for it, such as maximize profit, maximize traffic at breakeven, and so forth. The system is usually tied into the advertiser's website and fed the results of each click, which then allows it to set bids. The effectiveness of these systems is directly related to the quality and quantity of the performance data that they have to work with - low-traffic ads can lead to a scarcity of data problem that renders many bid management tools useless at worst, or inefficient at best.

History

In February 1998 Jeffrey Brewer of Goto.com, a 25-employee startup company (later Overture, now part of Yahoo!), presented a pay per click search engine proof-of-concept to the TED conference in California.[7] This presentation and the events that followed created the PPC advertising system. Credit for the concept of the PPC model is generally given to Idealab and Goto.com founder Bill Gross.

Google started search engine advertising in December 1999. It was not until October 2000 that the AdWords system was introduced, allowing advertisers to create text ads for placement on the Google search engine. However, PPC was only introduced in 2002; until then, advertisements were charged at cost-per-thousand impressions. Overture has filed a patent infringement lawsuit against Google, saying the rival search service overstepped its bounds with its ad-placement tools. [8]

Although GoTo.com started PPC in 1998, Yahoo! did not start syndicating GoTo.com (later Overture) advertisers until November 2001.[9] Prior to this, Yahoo's primary source of SERPS advertising included contextual IAB

advertising units (mainly 468x60 display ads). When the syndication contract with Yahoo! was up for renewal in July 2003, Yahoo! announced intent to acquire Overture for $1.63 billion.[10]

References

[1] "Customers Now", David Szetela, 2009.

[2] Shuman Ghosemajumder (March 18, 2008). "Using data to help prevent fraud" (http://googleblog.blogspot.com/2008/03/ using-data-to-help-prevent-fraud.html). Google Blog. . Retrieved May 18, 2010.

[3] How do you prevent invalid clicks and impressions? (https://www.google.com/adsense/support/bin/answer.py?answer=9718&ctx=en:search&query=invalid+click&topic=&type=f) Google AdSense Help Center, Accessed January 9, 2008

[4] Shopping.com Merchant Enrollment (https://merchant.shopping.com/enroll/app?service=page/RateCard) Shopping.com, Accessed June 12, 2007

[5] Yahoo! Search Marketing (May 18, 2010). "Sponsored Search" (http://advertising.yahoo.com/smallbusiness/whatsincluded). Website Traffic Yahoo! Search Marketing (formerly Overture). . Retrieved May 18, 2010.

[6] AdWords Discounter (http://adwords.google.com/support/bin/answer.py?hl=en&answer=6302) Google AdWords Help, Accessed February 23, 2009

[7] Overture and Google: Internet Pay Per Click (PPC) Advertising Auctions (http://faculty.london.edu/mottaviani/PPCA.pdf), London Business School, Accessed June 12, 2007

[8] Stefanie Olsen and Gwendolyn Mariano (April 5, 2002). "Overture sues Google over search patent" (http://news.cnet.com/ 2100-1023-876861.html). CNET. . Retrieved Jan 28, 2011.

[9] Yahoo! Inc. (2002). "Yahoo! and Overture Extend Pay-for-Performance Search Agreement" (http://docs.yahoo.com/docs/pr/release975. html). Yahoo! Press Release. . Retrieved May 18, 2010.

[10] Stefanie Olsen (July 14, 2003). "Yahoo to buy Overture for $1.63 billion" (http://news.cnet.com/2100-1030_3-1025394.html). CNET. . Retrieved May 18, 2010.

External links

- Paid listings confuse web searchers (http://www.pcworld.com/article/112132/study_paid_listings_still_confuse_web_searchers.html), PCWorld

Social media marketing

Social media marketing refers to the process of gaining website traffic or attention through social media sites.[1]

Social media marketing programs usually center on efforts to create content that attracts attention and encourages readers to share it with their social networks. A corporate message spreads from user to user and presumably resonates because it appears to come from a trusted, third-party source, as opposed to the brand or company itself. Hence, this form of marketing is driven by word-of-mouth, meaning it results in earned media rather than paid media.

Social media has become a platform that is easily accessible to anyone with internet access. Increased communication for organizations fosters brand awareness and often, improved customer service. Additionally, social media serves as a relatively inexpensive platform for organizations to implement marketing campaigns.

Social media outlets/platforms

Twitter, Facebook, Google+, YouTube, blogs

Social networking websites allow individuals to interact with one another and build relationships. When products or companies join those sites, people can interact with the product or company. That interaction feels personal to users because of their previous experiences with social networking site interactions.

Social networking sites like Twitter, Facebook, Google Plus, YouTube, Pinterest, blogs and Bclicky allow individual followers to "retweet" or "repost" comments made by the product being promoted. By repeating the message, all of the users connections are able to see the message, therefore reaching more people. Social networking sites act as word of mouth.[2] Because the information about the product is being put out there and is getting repeated, more traffic is brought to the product/company.[2]

Through social networking sites, products/companies can have conversations and interactions with individual followers. This personal interaction can instill a feeling of loyalty into followers and potential customers.[2] Also, by choosing whom to follow on these sites, products can reach a very narrow target audience.[2]

Cell phones

Cell phone usage has also become a benefit for social media marketing. Today, many cell phones have social networking capabilities: individuals are notified of any happenings on social networking sites through their cell phones, in real-time. This constant connection to social networking sites means products and companies can constantly remind and update followers about their capabilities, uses, importance, etc. Because cell phones are connected to social networking sites, advertisements are always in sight. Also many companies are now putting QR codes along with products for individuals to access the companies website or online services with their smart-phones.

Engagement

In the context of the social web, **engagement** means that customers and stakeholders are participants rather than viewers. Social media in business allows anyone and everyone to express and share an opinion or idea somewhere along the business's path to market. Each participating customer becomes part of the marketing department, as other customers read their comments or reviews. The engagement process is then fundamental to successful social media marketing.[3]

Campaigns

Adidas

In 2007, Adidas, and their agency Carat, created a social media experience for soccer players. Adidas pitted two different cleat types against one another and asked people to "choose your side." The content focused on fostering an environment of friendly discussion and debate of Adidas' two models of elite soccer cleats/boots, Predator and F50 TUNIT. Visitors to the community had the opportunity to align themselves with one product "team" and offer comments in support of their preferred model. The community included content about professional Adidas soccer players on each "team," rotational product views, downloadable graphics, forum discussions, a link to additional product information, and a link to the adidas Mexico Fútbol profile page.

Betty White

Social networking sites can have a large impact on the outcome of events. In 2010, a Facebook campaign surfaced in the form of a petition. Users virtually signed a petition asking NBC Universal to have actress Betty White host Saturday Night Live.[4][5] Once signed, users forwarded the petition to all of their followers. The petition went viral and on May 8, 2010, Betty White hosted SNL.

2008 Presidential Election

The 2008 presidential campaign had a huge presence on social networking sites. Barack Obama, a Democratic candidate for US President, used Twitter and Facebook to differentiate his campaign.[6] His social networking site profile pages were constantly being updated and interacting with followers. The use of social networking sites gave Barack Obama's campaign access to e-mail addresses, as posted on social networking site profile pages. This allowed the Democratic Party to launch e-mail campaigns asking for votes and campaign donations.[6]

Local businesses

Small businesses also use social networking sites as a promotional technique. Businesses can follow individuals social networking site uses in the local area and advertise specials and deals.[6] These can be exclusive and in the form of "get a free drink with a copy of this tweet".[6] This type of message encourages other locals to follow the business on the sites in order to obtain the promotional deal. In the process, the business is getting seen and promoting itself.

Tactics

Twitter

Twitter allows companies to promote products on an individual level. The use of a product can be explained in short messages that followers are more likely to read. These messages appear on followers' home pages. Messages can link to the product's website, Facebook profile, photos, videos, etc. This link provides followers the opportunity to spend more time interacting with the product online. This interaction can create a loyal connection between product and individual and can also lead to larger advertising opportunities. Twitter promotes a product in real-time and brings customers in.

Facebook

Facebook profiles are more detailed than Twitter. They allow a product to provide videos, photos, and longer descriptions. Videos can show when a product can be used as well as how to use it. These also can include testimonials as other followers can comment on the product pages for others to see. Facebook can link back to the product's Twitter page as well as send out event reminders. Facebook promotes a product in real-time and brings customers in.

As marketers see more value in social media marketing, advertisers continue to increase sequential ad spend in social by 25%. Strategies to extend the reach with Sponsored Stories and acquire new fans with Facebook ads continue to an uptick in spend across the site. The study attributes 84% of "engagement" or clicks to Likes that link back to Facebook advertising. Today, brands increase fan counts on average of 9% monthly, increasing their fan base by two-times the amount annually.[7]

Blogs

Blogs allow a product or company to provide longer descriptions of products or services. The longer description can include reasoning and uses. It can include testimonials and can link to and from Facebook, Twitter and many social network and blog pages. Blogs can be updated frequently and are promotional techniques for keeping customers. Other promotional uses are acquiring followers and subscribers and direct them to your social network pages.

Social media marketing tools

Besides research tools,[8] there are many companies providing specialized platforms/tools for social media marketing, such as tools for:

- Social Media Monitoring
- Social Aggregation
- Social Book Marking and Tagging
- Social Analytics and Reporting
- Automation
- Social Media
- Blog Marketing
- Validation

Implications on traditional advertising

Minimizing use

Traditional advertising techniques include print and television advertising. The Internet had already overtaken television as the largest advertising market.[2] Websites often include banner or pop-up ads. Social networking sites don't always have ads. In exchange, products have entire pages and are able to interact with users. Television commercials often end with a spokesperson asking viewers to check out the product website for more information. Print ads are also starting to include barcodes on them. These barcodes can be scanned by cell phones and computers, sending viewers to the product website. Advertising is beginning to move viewers from the traditional outlets to the electronic ones.

Leaks

Internet and social networking leaks are one of the issues facing traditional advertising. Video and print ads are often leaked to the world via the Internet earlier than they are scheduled to premiere. Social networking sites allow those leaks to go viral, and be seen by many users more quickly. Time difference is also a problem facing traditional advertisers. When social events occur and are broadcast on television, there is often a time delay between airings on the east coast and west coast of the United States. Social networking sites have become a hub of comment and interaction concerning the event. This allows individuals watching the event on the west coast (time-delayed) to know the outcome before it airs. The 2011 Grammy Awards highlighted this problem. Viewers on the west coast learned who won different awards based on comments made on social networking sites by individuals watching live on the east coast.[9] Since viewers knew who won already, many tuned out and ratings were lower. All the advertisement and promotion put into the event was lost because viewers didn't have a reason to watch.

References

[1] "What is Social Media Marketing" (http://searchengineland.com/guide/what-is-social-media-marketing). Search Engine Land. . Retrieved 11 January 2012.

[2] "NU Libraries" (http://0-ehis.ebscohost.com.ilsprod.lib.neu.edu/eds/detail?hid=22&sid=e8949099-0cb3-410a-91ac-0b9bf1986d28@sessionmgr15&vid=6&bdata=JnNpdGU9ZWRzLWxpdmU=#db=psyh& AN=2010-06092-007). 0-ehis.ebscohost.com.ilsprod.lib.neu.edu. . Retrieved 2011-11-17.

[3] *Social Media Marketing: The Next... - Dave Evans, Jake McKee - Google Books* (http://books.google.com/books?hl=en&lr=&id=7l2OR6giC6AC&oi=fnd&pg=PT15&dq=social+media+promoter&ots=jLG3wU2N5U&sig=X-D-jOgcbNEUxMgqNOhs0XjJkaU#v=onepage&q=social media promoter&f=false). Books.google.com. . Retrieved 2011-11-17.

[4] Itzkoff, Dave (2010-05-10). "Betty White Helps Boost Ratings of 'SNL'" (http://www.nytimes.com/2010/05/10/arts/television/ 10arts-BETTYWHITEHE_BRF.html). *The New York Times*. .

[5] Levin, Gary (2010-03-12). "Live, from New York, it's ... Betty White hosting 'SNL'" (http://www.usatoday.com/life/television/news/ 2010-03-11-bettywhite11_ST_N.htm). *USA Today*. .

[6] "NU Libraries" (http://0-ehis.ebscohost.com.ilsprod.lib.neu.edu/eds/detail?hid=22&sid=e8949099-0cb3-410a-91ac-0b9bf1986d28@sessionmgr15&vid=8&bdata=JnNpdGU9ZWRzLWxpdmU=#db=bth&AN=55584217). 0-ehis.ebscohost.com.ilsprod.lib.neu.edu. . Retrieved 2011-11-17.

[7] "Marketers Spend More" (http://www.mediapost.com/publications/article/160225/marketers-spend-more-on-mobile-search.html). Mediapost.com. . Retrieved 2011-12-21.

[8] Erik Cambria; Marco Grassi, Amir Hussain and Catherine Havasi (2011). "Sentic Computing for Social Media Marketing" (http://springerlink.com/content/q1vq625w2x27x4r7). *In press: Multimedia Tools and Applications Journal. Springer-Verlag, Berlin Heidelberg (DOI 10.1007/s11042-011-0815-0)*. .

[9] "Hey Grammys, you can't tape-delay social media" (http://www.lostremote.com/2011/02/13/hey-grammys-you-cant-tape-delay-social-media/). Lostremote.com. 2011-02-13. . Retrieved 2011-11-17.

Affiliate marketing

Affiliate marketing is a marketing practice in which a business rewards one or more affiliates for each visitor or customer brought about by the affiliate's own marketing efforts. The industry has four core players: the merchant (also known as 'retailer' or 'brand'), the network (that contains offers for the affiliate to choose from and also takes care of the payments), the publisher (also known as 'the affiliate'), and the customer. The market has grown in complexity to warrant a secondary tier of players, including affiliate management agencies, super-affiliates and specialized third party vendors.

Affiliate marketing overlaps with other Internet marketing methods to some degree, because affiliates often use regular advertising methods. Those methods include organic search engine optimization (SEO), paid search engine marketing (PPC - Pay Per Click), e-mail marketing, and in some sense display advertising. On the other hand, affiliates sometimes use less orthodox techniques, such as publishing fake reviews of products or services offered by a partner.

Affiliate marketing is commonly confused with referral marketing, as both forms of marketing use third parties to drive sales to the retailer.[1] However, both are distinct forms of marketing and the main difference between them is that affiliate marketing relies purely on financial motivations to drive sales while referral marketing relies on trust and personal relationships to drive sales.[1]

Affiliate marketing is frequently overlooked by advertisers.[2] While search engines, e-mail, and websitesyndication capture much of the attention of online retailers, affiliate marketing carries a much lower profile. Still, affiliates continue to play a significant role in e-retailers' marketing strategies.

History

Origin

The concept of affiliate marketing on the Internet was conceived of, put into practice and patented by William J. Tobin, the founder of PC Flowers & Gifts. Launched on the Prodigy Network in 1989, PC Flowers & Gifts remained on the service until 1996. By 1993, PC Flowers & Gifts generated sales in excess of $6 million dollars per year on the Prodigy service. In 1998, PC Flowers and Gifts developed the business model of paying a commission on sales to The Prodigy network (Reference-Chicago Tribune-Oct, 4, 1995).(Ref The Sunsentinal 1991 and www.dankawaski.com).

In 1994, Mr. Tobin launched a beta version of PC Flowers & Gifts on the Internet in cooperation with IBM who owned half of Prodigy (Reference-PC Week Article Jan 9, 1995). By 1995 PC Flowers & Gifts had launched a commercial version of the website and had 2,600 affiliate marketing partners on the World Wide Web. Mr. Tobin applied for a patent on tracking and affiliate marketing on January 22, 1996 and was issued U.S. Patent number 6,141,666 on Oct 31, 2000. Mr. Tobin also received Japanese Patent number 4021941 on Oct 5, 2007 and U.S. Patent number 7,505,913 on Mar 17, 2009 for affiliate marketing and tracking (Reference-Business Wire-Jan, 24, 2000). In July 1998 PC Flowers and Gifts merged with Fingerhut and Federated Department Stores (Reference- Business Wire- March 31,1999).

On March 9, 2009 Mr. Tobin assigned his patents to the Tobin Family Education and Health Foundation. The Foundation licenses the patents to many of the largest affiliate marketing companies in the US and Japan. Mr. Tobin discusses the P.C Flowers & Gifts service on the Internet as well as the other nine companies he has founded in his book entitled "Confessions of an Obsessive Entrepreneur".

The concept of revenue sharing—paying commission for referred business—predates affiliate marketing and the Internet. The translation of the revenue share principles to mainstream e-commerce happened in November 1994,[3] almost four years after the origination of the World Wide Web.

Cybererotica was among the early innovators in affiliate marketing with a cost per click program.[4]

During November 1994, CDNOW launched its BuyWeb program. CDNOW had the idea that music-oriented websites could review or list albums on their pages that their visitors may be interested in purchasing. These websites could also offer a link that would take the visitor directly to CDNOW to purchase the albums. The idea for remote purchasing originally arose because of conversations with music label Geffen Records in the fall of 1994. The management at Geffen wanted to sell its artists' CDs directly from its website, but did not want to implement this capability itself. Geffen asked CDNOW if it could design a program where CDNOW would handle the order fulfillment. Geffen realized that CDNOW could link directly from the artist on its website to Geffen's website, bypassing the CDNOW home page and going directly to an artist's music page.[5]

Amazon.com (Amazon) launched its associate program in July 1996: Amazon associates could place banner or text links on their site for individual books, or link directly to the Amazon home page.[6]

When visitors clicked from the associate's website through to Amazon and purchased a book, the associate received a commission. Amazon was not the first merchant to offer an affiliate program, but its program was the first to become widely known and serve as a model for subsequent programs.[7][8]

In February 2000, Amazon announced that it had been granted a patent[9] on components of an affiliate program. The patent application was submitted in June 1997, which predates most affiliate programs, but not PC Flowers & Gifts.com (October 1994), AutoWeb.com (October 1995), Kbkids.com/BrainPlay.com (January 1996), EPage (April 1996), and several others.[4]

Historic development

Affiliate marketing has grown quickly since its inception. The e-commerce website, viewed as a marketing toy in the early days of the Internet, became an integrated part of the overall business plan and in some cases grew to a bigger business than the existing offline business. According to one report, the total sales amount generated through affiliate networks in 2006 was £2.16 billion in the United Kingdom alone. The estimates were £1.35 billion in sales in 2005.[10] MarketingSherpa's research team estimated that, in 2006, affiliates worldwide earned US$6.5 billion in bounty and commissions from a variety of sources in retail, personal finance, gaming and gambling, travel, telecom, education, publishing, and forms of lead generation other than contextual advertising programs.[11]

Currently the most active sectors for affiliate marketing are the adult, gambling, retail industries and file-sharing services.[12] The three sectors expected to experience the greatest growth are the mobile phone, finance, and travel sectors.[12] Soon after these sectors came the entertainment (particularly gaming) and Internet-related services (particularly broadband) sectors. Also several of the affiliate solution providers expect to see increased interest from business-to-business marketers and advertisers in using affiliate marketing as part of their mix.[12]

Web 2.0

Websites and services based on Web 2.0 concepts—blogging and interactive online communities, for example—have impacted the affiliate marketing world as well. The new media allowed merchants to become closer to their affiliates and improved the communication between them.

Web 2.0 platforms have also opened affiliate marketing channels to personal bloggers, writers, and independent website owners. Regardless of web traffic, size, or business age, programs through Google, LinkShare, and Amazon allow publishers at all levels of web traffic to place contextual ads in blog posts.

Forms of new media have also diversified how companies, brands, and ad networks serve ads to visitors. For instance, YouTube allows video-makers to embed advertisements through Google's affiliate network.
[13] [14]

Compensation methods

Predominant compensation methods

Eighty percent of affiliate programs today use revenue sharing or pay per sale (PPS) as a compensation method, nineteen percent use cost per action (CPA), and the remaining programs use other methods such as cost per click (CPC) or cost per mille (CPM).

Diminished compensation methods

Within more mature markets, less than one percent of traditional affiliate marketing programs today use cost per click and cost per mille. However, these compensation methods are used heavily in display advertising and paid search.

Cost per mille requires only that the publisher make the advertising available on his website and display it to his visitors in order to receive a commission. Pay per click requires one additional step in the conversion process to generate revenue for the publisher: A visitor must not only be made aware of the advertisement, but must also click on the advertisement to visit the advertiser's website.

Cost per click was more common in the early days of affiliate marketing, but has diminished in use over time due to click fraud issues very similar to the click fraud issues modern search engines are facing today. Contextual advertising programs are not considered in the statistic pertaining to diminished use of cost per click, as it is uncertain if contextual advertising can be considered affiliate marketing.

While these models have diminished in mature e-commerce and online advertising markets they are still prevalent in some more nascent industries. China is one example where Affiliate Marketing does not overtly resemble the same model in the West. With many affiliates being paid a flat "Cost Per Day" with some networks offering Cost Per Click or CPM.

Performance marketing

In the case of cost per mille/click, the publisher is not concerned about a visitor being a member of the audience that the advertiser tries to attract and is able to convert, because at this point the publisher has already earned his commission. This leaves the greater, and, in case of cost per mille, the full risk and loss (if the visitor can not be converted) to the advertiser.

Cost per action/sale methods require that referred visitors do more than visit the advertiser's website before the affiliate receives commission. The advertiser must convert that visitor first. It is in the best interest for the affiliate to send the most closely targeted traffic to the advertiser as possible to increase the chance of a conversion. The risk and loss is shared between the affiliate and the advertiser.

Affiliate marketing is also called "performance marketing", in reference to how sales employees are typically being compensated. Such employees are typically paid a commission for each sale they close, and sometimes are paid performance incentives for exceeding targeted baselines.[15] Affiliates are not employed by the advertiser whose products or services they promote, but the compensation models applied to affiliate marketing are very similar to the ones used for people in the advertisers' internal sales department.

The phrase, "Affiliates are an extended sales force for your business", which is often used to explain affiliate marketing, is not completely accurate. The primary difference between the two is that affiliate marketers provide little if any influence on a possible prospect in the conversion process once that prospect is directed to the advertiser's website. The sales team of the advertiser, however, does have the control and influence up to the point where the prospect signs the contract or completes the purchase.

Multi-tier programs

Some advertisers offer multi-tier programs that distribute commission into a hierarchical referral network of sign-ups and sub-partners. In practical terms, publisher "A" signs up to the program with an advertiser and gets rewarded for the agreed activity conducted by a referred visitor. If publisher "A" attracts publishers "B" and "C" to sign up for the same program using his sign-up code, all future activities performed by publishers "B" and "C" will result in additional commission (at a lower rate) for publisher "A".

Two-tier programs exist in the minority of affiliate programs; most are simply one-tier. Referral programs beyond two-tier resemble multi-level marketing (MLM) or network marketing but are different: Multi-level marketing (MLM) or network marketing associations tend to have more complex commission requirements/qualifications than standard affiliate programs.

From the advertiser's perspective

Pros and cons

Merchants favor affiliate marketing because in most cases it uses a "pay for performance" model, meaning that the merchant does not incur a marketing expense unless results are accrued (excluding any initial setup cost).[16] Some businesses owe much of their success to this marketing technique, a notable example being Amazon.com. Unlike display advertising, however, affiliate marketing is not easily scalable.[17]

Implementation options

Some merchants run their own (in-house) affiliate programs using popular software while others use third-party services provided by intermediaries to track traffic or sales that are referred from affiliates (*see* outsourced program management). Merchants can choose from two different types of affiliate management solutions: standalone software or hosted services, typically called affiliate networks. Payouts to affiliates or publishers are either made by the networks on behalf of the merchant, by the network, consolidated across all merchants where the publisher has a relationship with and earned commissions or directly by the merchant itself.

Affiliate management and program management outsourcing

Successful affiliate programs require significant work and maintenance. Having a successful affiliate program is more difficult than when such programs were just emerging. With the exception of some vertical markets, it is rare for an affiliate program to generate considerable revenue with poor management or no management ("auto-drive"). Uncontrolled affiliate programs did still do aid rogue affiliates, who use spamming,[18] trademark infringement, false advertising, "cookie cutting", typosquatting,[19] and other unethical methods that have given affiliate marketing a negative reputation.

The increased number of Internet businesses and the increased number of people that trust the current technology enough to shop and do business online allows further maturation of affiliate marketing.

The opportunity to generate a considerable amount of profit combined with a crowded marketplace filled with competitors of equal quality and size makes it more difficult for merchants to be noticed. In this environment, however, being noticed can yield greater rewards.

Recently, the Internet marketing industry has become more advanced. In some areas online media has been rising to the sophistication of offline media, in which advertising has been largely professional and competitive. There are significantly more requirements that merchants must meet to be successful, and those requirements are becoming too burdensome for the merchant to manage successfully in-house.

An increasing number of merchants are seeking alternative options found in relatively new outsourced (affiliate) program management (OPM) companies, which are often founded by veteran affiliate managers and network

program managers.[20] OPM companies perform affiliate program management for the merchants as a service, similar to advertising agencies promoting a brand or product as done in offline marketing.

Types of affiliate websites

Affiliate websites are often categorized by merchants (advertisers) and affiliate networks. There are currently no industry-wide standards for the categorization. The following types of websites are generic, yet are commonly understood and used by affiliate marketers.

- Search affiliates that utilize pay per click search engines to promote the advertisers' offers (i.e., search arbitrage)
- Comparison shopping websites and directories
- Loyalty websites, typically characterized by providing a reward system for purchases via points back, cash back
- CRM sites that offer charitable donations
- Coupon and rebate websites that focus on sales promotions
- Content and niche market websites, including product review sites
- Personal websites
- Weblogs and website syndication feeds
- E-mail list affiliates (i.e., owners of large opt-in-mail lists that typically employe-mail drip marketing) and newsletter list affiliates, which are typically more content-heavy
- Registration path or co-registration affiliates who include offers from other merchants during the registration process on their own website
- Shopping directories that list merchants by categories without providing coupons, price comparisons, or other features based on information that changes frequently, thus requiring continual updates
- Cost per action networks (i.e., top-tier affiliates) that expose offers from the advertiser with which they are affiliated to their own network of affiliates
- Websites using adbars (e.g. Adsense) to display context-sensitive, highly relevant ads for products on the site
- Virtual Currency: a new type of publisher that utilizes the social media space to couple an advertiser's offer with a handout of "virtual currency" in a game or virtual platform.
- Video Blog: Video content that allows viewers to click on and purchase products related to the video's subject.
- File-Sharing: Web sites that host directories of music, movies, games and other software. Users upload content (usually in violation of copyright) to file-hosting sites, and then post descriptions of the material and their download links on directory sites. Uploaders are paid by the file-hosting sites based on the number of times their files are downloaded. The file-hosting sites sell premium download access to the files to the general public. The web sites that host the directory services sell advertising and do not host the files themselves.

Publisher recruitment

Affiliate networks that already have several advertisers typically also have a large pool of publishers. These publishers could be potentially recruited, and there is also an increased chance that publishers in the network apply to the program on their own, without the need for recruitment efforts by the advertiser.

Relevant websites that attract the same target audiences as the advertiser but without competing with it are potential affiliate partners as well. Vendors or existing customers can also become recruits if doing so makes sense and does not violate any laws or regulations.

Almost any website could be recruited as an affiliate publisher, but high-traffic websites are more likely interested in (for their own sake) low-risk cost per mille or medium-risk cost per click deals rather than higher-risk cost per action or revenue share deals.[21]

Locating affiliate programs

There are three primary ways to locate affiliate programs for a target website:

1. Affiliate program directories,
2. Large affiliate networks that provide the platform for dozens or even hundreds of advertisers, and
3. The target website itself. (Websites that offer an affiliate program often have a link titled "affiliate program", "affiliates","referral program",or"webmasters"—usuallyinthefooteror"About"sectionofthewebsite.)

If the above locations do not yield information pertaining to affiliates, it may be the case that there exists a non-public affiliate program. Utilizing one of the common website correlation methods may provide clues about the affiliate network. The most definitive method for finding this information is to contact the website owner directly, if a contact method can be located.

Past and current issues

Since the emergence of affiliate marketing, there has been little control over affiliate activity. Unscrupulous affiliates have used spam, false advertising, forced clicks (to get tracking cookies set on users' computers), adware, and other methods to drive traffic to their sponsors. Although many affiliate programs have terms of service that contain rules against spam, this marketing method has historically proven to attract abuse from spammers.

E-mail spam

In the infancy of affiliate marketing, many Internet users held negative opinions due to the tendency of affiliates to use spam to promote the programs in which they were enrolled.[22] As affiliate marketing matured, many affiliate merchants have refined their terms and conditions to prohibit affiliates from spamming.

Search engine spam

As search engines have become more prominent, some affiliate marketers have shifted from sending e-mail spam to creating automatically generated webpages that often contain product data feeds provided by merchants. The goal of such webpages is to manipulate the relevancy or prominence of resources indexed by a search engine, also known as *spamdexing*. Each page can be targeted to a different niche market through the use of specific keywords, with the result being a skewed form of search engine optimization.

Spam is the biggest threat to organic search engines, whose goal is to provide quality search results for keywords or phrases entered by their users. Google's PageRank algorithm update ("BigDaddy") in February 2006—the final stage of Google's major update ("Jagger") that began in mid-summer 2005—specifically targeted spamdexing with great success. This update thus enabled Google to remove a large amount of mostly computer-generated duplicate content from its index.[23]

Websites consisting mostly of affiliate links have previously held a negative reputation for underdelivering quality content. In 2005 there were active changes made by Google, where certain websites were labeled as "thin affiliates".[24] Such websites were either removed from Google's index or were relocated within the results page (i.e., moved from the top-most results to a lower position). To avoid this categorization, affiliate marketer webmasters must create quality content on their websites that distinguishes their work from the work of spammers or banner farms, which only contain links leading to merchant sites.

Some commentators originally suggested that affiliate links work best in the context of the information contained within the website itself. For instance, if a website contains information pertaining to publishing a website, an affiliate link leading to a merchant's internet service provider (ISP) within that website's content would be appropriate. If a website contains information pertaining to sports, an affiliate link leading to a sporting goods website may work well within the context of the articles and information about sports. The goal in this case is to publish quality information within the website and provide context-oriented links to related merchant's websites.

However, more recent examples exist of "thin" affiliate sites that are using the affiliate marketing model to create value for Consumers by offering them a service. These thin content service Affiliate fall into three categories:

- Price comparison
- Cause related marketing
- Time saving

Virus and Trojan distribution through advertising networks

Server farms hosting advertising content are periodically infected by hackers who alter the behavior of these servers such that the content they serve to end-users includes hidden I-frames and other exploits that leverage vulnerabilities in various web-browsers and operating systems for the purpose of infecting those systems with malware. End users frequently confuse the source of their computer infection with a particular website they were viewing at the time, and not the advertising network that was linked to, by the website (commonly users themselves do not understand or appreciate there is adistinction).

Consumer countermeasures

The implementation of affiliate marketing on the internet relies heavily on various techniques built into the design of many web-pages and web-sites, and the use of calls to external domains to track user actions (click tracking, Ad Sense) and to serve up content (advertising) to the user. Most of this activity adds time and is generally a nuisance to the casual web-surfer and is seen as visual clutter. Various countermeasures have evolved over time to prevent or eliminate the appearance of advertising when a web-page is rendered. Third party programs (Ad Aware, SpyBot, pop-up blockers, etc.) and particularly, the use of a comprehensive HOSTS file can effectively eliminate the visual clutter and the extra time and bandwidth needed to render many web pages. The use of specific entries in the HOSTS file to block these well-known and persistent marketing and click-tracking domains can also aid in reducing a system's exposure to malware by preventing the content of infected advertising or tracking servers to reach a user's web-browser.

Adware

Although it differs from spyware, adware often uses the same methods and technologies. Merchants initially were uninformed about adware, what impact it had, and how it could damage their brands. Affiliate marketers became aware of the issue much more quickly, especially because they noticed that adware often overwrites tracking cookies, thus resulting in a decline of commissions. Affiliates not employing adware felt that it was stealing commission from them. Adware often has no valuable purpose and rarely provides any useful content to the user, who is typically unaware that such software is installed on his/her computer.

Affiliates discussed the issues in Internet forums and began to organize their efforts. They believed that the best way to address the problem was to discourage merchants from advertising via adware. Merchants that were either indifferent to or supportive of adware were exposed by affiliates, thus damaging those merchants' reputations and tarnishing their affiliate marketing efforts. Many affiliates either terminated the use of such merchants or switched to a competitor's affiliate program. Eventually, affiliate networks were also forced by merchants and affiliates to take a stand and ban certain adware publishers from their network. The result was Code of Conduct by Commission Junction/beFree and Performics,[25] LinkShare's Anti-Predatory Advertising Addendum,[26] and ShareASale's complete ban of software applications as a medium for affiliates to promote advertiser offers.[27] Regardless of the progress made, adware continues to be an issue, as demonstrated by the class action lawsuit against ValueClick and its daughter company Commission Junction filed on April 20, 2007.[28]

Trademark bidding

Affiliates were among the earliest adopters of pay per click advertising when the first pay-per-click search engines emerged during the end of the 1990s. Later in 2000 Google launched its pay per click service, Google AdWords, which is responsible for the widespread use and acceptance of pay per click as an advertising channel. An increasing number of merchants engaged in pay per click advertising, either directly or via a search marketing agency, and realized that this space was already well-occupied by their affiliates. Although this situation alone created advertising channel conflicts and debates between advertisers and affiliates, the largest issue concerned affiliates bidding on advertisers names, brands, and trademarks.[29] Several advertisers began to adjust their affiliate program terms to prohibit their affiliates from bidding on those type of keywords. Some advertisers, however, did and still do embrace this behavior, going so far as to allow, or even encourage, affiliates to bid on any term, including the advertiser's trademarks. And some affiliates abuse it by bidding on those terms by excluding the location of the advertiser alone in many Search engines.

Lack of self-regulation and collaboration

Affiliate marketing is driven by entrepreneurs who are working at the edge of Internet marketing. Affiliates are often the first to take advantage of emerging trends and technologies. The "trial and error" approach is probably the best way to describe the operation methods for affiliate marketers. This risky approach is one of the reasons why most affiliates fail or give up before they become successful "super affiliates", capable of generating US$10,000 or more per month in commission. This "frontier" life combined with the attitude found in such communities is likely the main reason why the affiliate marketing industry is unable to self-regulate beyond individual contracts between advertisers and affiliates. Affiliate marketing has experienced numerous failed attempts to create an industry organization or association of some kind that could be the initiator of regulations, standards, and guidelines for the industry.[30] Some examples of failed regulation efforts are the Affiliate Union and iAfma.

Online forums and industry trade shows are the only means for the different members from the industry—affiliates/publishers, merchants/advertisers, affiliate networks, third-party vendors, and service providers such as outsourced program managers—to congregate at one location. Online forums are free, enable small affiliates to have a larger say, and provide anonymity. Trade shows are cost-prohibitive to small affiliates because of the high price for event passes. Larger affiliates may even be sponsored by an advertiser they promote.

Because of the anonymity of online forums, the quantitative majority of industry members are unable to create any form of legally binding rule or regulation that must be followed throughout the industry. Online forums have had very few successes as representing the majority of the affiliate marketing industry. The most recent example of such a success was the halt of the "Commission Junction Link Management Initiative" (CJ LMI) in June/July 2006, when a single network tried to impose the use of a Javascript tracking code as a replacement for common HTML links on its affiliates.[31]

Compensation Disclosure

Bloggers and other publishers may not be aware of disclosure guidelines set forth by the FTC. Guidelines affect celebrity endorsements, advertising language, and blogger compensation.[32]

Lack of industry standards

Certification and training

Affiliate marketing currently lacks industry standards for training and certification. There are some training courses and seminars that result in certifications; however, the acceptance of such certifications is mostly due to the reputation of the individual or company issuing the certification. Affiliate marketing is not commonly taught in universities, and only a few college instructors work with Internet marketers to introduce the subject to students

majoring in marketing.[33]

Education occurs most often in "real life" by becoming involved and learning the details as time progresses. Although there are several books on the topic, some so-called "how-to" or "silver bullet" books instruct readers to manipulate holes in the Google algorithm, which can quickly become out of date,[33] or suggest strategies no longer endorsed or permitted by advertisers.[34]

Outsourced Program Management companies typically combine formal and informal training, providing much of their training through group collaboration and brainstorming. Such companies also try to send each marketing employee to the industry conference of their choice.[35]

Other training resources used include online forums, weblogs, podcasts, video seminars, and specialty websites.

Affiliate Summit is the largest conference in the industry, and many other affiliate networks host their own annual events.

Code of conduct

A code of conduct was released by affiliate networks Commission Junction/beFree and Performics in December 2002 to guide practices and adherence to ethical standards for online advertising.

Marketing term

Members of the marketing industry are recommending that "affiliate marketing" be substituted with an alternative name.[36] Affiliate marketing is often confused with either network marketing or multi-level marketing. *Performance marketing* is a common alternative, but other recommendations have been made as well.

Sales tax vulnerability

In April 2008 the State of New York inserted an item in the state budget asserting sales tax jurisdiction over Amazon.com sales to residents of New York, based on the existence of affiliate links from New York–based websites to Amazon.[37] The state asserts that even one such affiliate constitutes Amazon having a business presence in the state, and is sufficient to allow New York to tax all Amazon sales to state residents. Amazon challenged the amendment and lost at the trial level in January, 2009. The case is currently making its way through the New York appeals courts.

Cookie stuffing

Cookie stuffing involves placing an affiliate tracking cookie on a website visitor's computer without their knowledge, which will then generate revenue for the person doing the cookie stuffing. This not only generates fraudulent affiliate sales, but also has the potential to overwrite other affiliates' cookies, essentially stealing their legitimately earned commissions.

Click to reveal

Many voucher code web sites use a click-to-reveal format, which requires the web site user to click to reveal the voucher code. The action of clicking places the cookie on the website visitor's computer. The IAB[38] have stated that "Affiliates must not use a mechanism whereby users are encouraged to click to interact with content where it is unclear or confusing what the outcome will be."

Affiliate services

- Affiliate tracking software
- Affiliate programs directories
- Affiliate networks (see also Category:Internet advertising services and affiliate networks)
- Affiliate manager and Outsourced Program Management (OPM or APM) (manages affiliates)
- Category:Internet marketing trade shows

References

[1] "Referral and Affiliate Marketing – What's the Difference?" (http://blog.referralcandy.com/2010/11/05/ referral-and-affiliate-marketing-whats-the-difference/). ReferralCandy. . Retrieved 10 January2012.

[2] Prussakov, Evgenii (2007). "A Practical Guide to Affiliate Marketing" (pp.16-17), 2007. ISBN 0-9791927-0-6.

[3] ShashankSHEKHAR(2009-06-29)."OnlineMarketingSystem:Affiliatemarketing"(http://feedmoney.com/archives/2009/06/29/ online-marketing-system-affiliate-marketing). Feed Money.com. . Retrieved 2011-04-20. "During November 1994, CDNOW released its BuyWeb program. With this program CDNOW was the first non-adult website to launch the concept of an affiliate or associate program with its idea of click-through purchasing."

[4] Collins, Shawn (2000-11-10). History of Affiliate Marketing. *ClickZ Network*, 10 November 2000. Retrieved on 2007-10-15 from http://www.clickz.com/showPage.html?page=832131.

[5] Olim, Jason; Olim, Matthew; and Kent, Peter (1999-01). "The Cdnow Story: Rags to Riches on the Internet", *Top Floor Publishing*, January 1999. ISBN 0-9661032-6-2.

[6] "What is the Amazon Associates program?" (https://affiliate-program.amazon.com/gp/associates/join/getstarted). https://affiliate-program.amazon.com/: amazon associates. . Retrieved 2011-04-20. "Amazon Associates is one of the first online affiliate marketing programs and was launched in 1996."

[7] Frank Fiore and Shawn Collins, "Successful Affiliate Marketing for Merchants", from pages 12,13 and 14. *QUE Publishing*, April 2001 ISBN 0-7897-2525-8

[8] Gray, Daniel (1999-11-30). "The Complete Guide to Associate and Affiliate Programs on the Net". *McGraw-Hill Trade*, 30 November 1999. ISBN 0-07-135310-0.

[9] US 6029141 (http://worldwide.espacenet.com/textdoc?DB=EPODOC&IDX=US6029141)

[10] October 2006, Affiliate Marketing Networks Buyer's Guide (2006) (http://www.e-consultancy.com/publications/ affiliate-marketing-networks-buyers-guide/), Page 6, *e-Consultancy.com*, retrieved June 25, 2007

[11] Anne Holland, publisher (January 11, 2006), Affiliate Summit 2006 Wrap-Up Report--Commissions to Reach $6.5 Billion in 2006 (http://www.marketingsherpa.com/barrier.cfm?contentID=3157), *MarketingSherpa*, retrieved on May 17, 2007

[12] February 2007, Internet Statistics Compendium 2007 (http://www.e-consultancy.com/publications/internet-stats-compendium/), Pages 149–150, *e-Consultancy*, retrieved June 25, 2007

[13] Dion, Hinchcliffe. "Web 2.0's Real Secret Sauce: Network Effects" (http://dionhinchcliffe.com/2006/07/15/ web-2-0s-real-secret-sauce-network-effects/). . Retrieved 10 January 2012.

[14] Dion, Hinchcliffe. "Social Media Goes Mainstream" (http://dionhinchcliffe.com/2007/01/29/social-media-goes-mainstream/). . Retrieved 10 January 2012.

[15] CellarStone Inc. (2006), Sales Commission (http://www.qcommission.com/salescommission_details.htm), *QCommission.com*, retrieved June 25, 2007

[16] Tom Taulli (9 November 2005), Creating A Virtual Sales Force (http://www.forbes.com/business/2005/11/08/ marketing-ecommerce-internet-cx_tt_1109straightup.html), *Forbes.com Business*. Retrieved 14 May 2007.

[17] Jeff Molander (June 22, 2006), Google's Content Referral Network: A Grab for Advertisers (http://www.thoughtshapers.com/index.php/ weblog/google-cpa-content-referral-network-goog-vclk-valueclick-david-jackson/), *Thought Shapers*, retrieved on December 16th, 2007

[18] Danny Sullivan (June 27, 2006), The Daily SearchCast News from June 27, 2006 (http://www.webmasterradio.fm/episodes/index. php?showId=30), *WebmasterRadio.fm*, retrieved May 17, 2007

[19] Wayne Porter (September 6, 2006), NEW FIRST: LinkShare- Lands' End Versus The Affiliate on Typosquatting (http://www.revenews.com/wayneporter/archives/002263.html), *ReveNews*, retrieved on May 17, 2007

[20] Jennifer D. Meacham (July/August 2006), Going Out Is In (http://www.revenuetoday.com/story/Going+Out+Is+In), *Revenue Magazine*, published by Montgomery Research Inc, Issue 12., Page 36

[21] Marios Alexandrou (February 4th, 2007), CPM vs. CPC vs. CPA (http://www.allthingssem.com/cpm-cpc-cpa/), *All Things SEM*, retrieved November 11, 2007

[22] Ryan Singel (October 2, 2005), Shady Web of Affiliate Marketing (http://www.wired.com/politics/security/news/2005/02/66556), *Wired.com*, retrieved May 17, 2007

[23] Jim Hedger (September 6, 2006), Being a Bigdaddy Jagger Meister (http://www.webpronews.com/expertarticles/2006/06/09/ being-a-bigdaddy-jagger-meister), *WebProNews.com*, retrieved on December 16, 2007

[24] SpamRecognitionGuideforRaters(http://www.searchbistro.com/spamguide.doc)(Worddocument)supposedlyleakedoutfrom Google (http://www.threadwatch.org/node/2709) in 2005. The authenticity of the document was neither acknowledged nor challenged by Google.
[25] December 10, 2002, Online Marketing Service Providers Announce Web Publisher Code of Conduct (http://www.cj.com/news/press_releases0102/press_021210.html) (contains original CoC text), *CJ.com*, retrieved June 26, 2007
[26] December 12, 2002, LinkShare's Anti-Predatory Advertising Addendum (http://www.linkshare.com/press/addendum.html), *LinkShare.com*, retrieved June 26, 2007
[27] ShareASale Affiliate Service Agreement (http://www.shareasale.com/agreement.cfm), *ShareASale.com*, retrieved June 26, 2007
[28] April 20, 2007, AdWare Class Action Lawsuit against - ValueClick, Commission Junction and beFree (http://www.cjclassaction.com/), *Law Firms of Nassiri & Jung LLP and Hagens Berman*, retrieved from CJClassAction.com on June 26, 2007
[29] Rosso, Mark; Jansen, Bernard (Jim) (August 2010), "Brand Names as Keywords in Sponsored Search Advertising" (http://aisel.aisnet.org/ cais/vol27/iss1/6), *Communications of the Association for Information Systems* **27** (1): 81-98,
[30] Carsten Cumbrowski (November 4, 2006), Affiliate Marketing Organization Initiative Vol.2 - We are back to Step 0 (http://www.revenews.com/carstencumbrowski/2006/11/affiliate_marketing_organizati_1.html), *Reve News*, retrieved May 17, 2007
[31] May 2006, New Javascript Links? (http://forum.abestweb.com/showthread.php?t=73747) main discussion thread to CJ's LMI, *ABestWeb*, retrieved on May 17, 2007
[32] http://www.ftc.gov/opa/2009/10/endortest.shtm
[33] Alexandra Wharton (March/April 2007), Learning Outside the Box (http://www.revenuetoday.com/story/Learning+Outside+the+Box&readpage=1), *Revenue Magazine*, Issue: March/April 2007, Page 58, link to online version retrieved June 26, 2007
[34] Shawn Collins (June 9, 2007), Affiliate Millions - Book Report (http://blog.affiliatetip.com/archives/affiliate-millions-book-report/), *AffiliateTip Blog*, retrieved June 26, 2007
[35] March/April 2007, How Do Companies Train Affiliate Managers? (http://www.revenuetoday.com/story/webextra-issue16-2) (Web Extra), *RevenueToday.com*, retrieved June 26, 2007
[36] Vinny Lingham (11.October, 2005), Profit Sharing - The Performance Marketing Model of the Future (http://www.vinnylingham.com/ 2006/10/special-report-profit-sharing-the-performance-marketing-model-of-the-future.html), *Vinny Lingham's Blog*, retrieved on 14.May, 2007
[37] Linda Rosencrance, 15.April, 2008), N.Y. to tax goods bought on Amazon (http://www.computerworld.com/action/article.do?command=viewArticleBasic&taxonomyName=government&articleId=9077963&taxonomyId=13&intsrc=kc_top), *Computerworld*, retrieved on 16.April, 2008
[38] IAB, Friday, 27 March 2009 IAB affiliate council strengthens voucher code guidelines (http://www.iabuk.net/en/1/iabaffiliatemarketingcouncilstrengthensonlinevouchercodebestpracticeguidelines270309.mxs)

External links

- Affiliatemarketing(http://www.dmoz.org/Business/Opportunities/Online_Opportunities/Affiliate_Programs/) at the Open Directory Project
- Website Affiliate Programs (http://dir.yahoo.com/Business_and_Economy/Business_to_Business/Business_Opportunities/Directories/Web_Site_Affiliate_Programs/) at the Yahoo! Directory
- Affiliate Programs (http://botw.org/top/Computers/Internet/Web_Design_and_Development/Authoring/Webmaster_Resources/Affiliate_Programs/) at the BOTW Directory

Article marketing

Article marketing is a type of advertising in which businesses write short articles about themselves as a marketing strategy. A primary style for the articles includes a *bio box* and *byline* (collectively known as the resource box) about the business.

Traditional Article Marketing

Article marketing has been used by professionals for nearly as long as mass print has been available. A business provides content to a newspaper, possibly on a timely topic such as an article on tax audits during tax season, and the newspaper may use the article and include the business's name and contact information. Newspapers and other traditional media have limited budgets for gathering content and these articles may be used in the business section of the newspaper.

Internet Article Marketing

Internet article marketing is used to promote products and services online via article directories. Article directories with good web page ranks receive a lot of site visitors and are may be considered authority sites by search engines,leading to high traffic. These directories then go on PageRank to the author's website and in addition send traffic from readers.

Internet marketers attempt to maximize the results of an article advertising campaign by submitting their articles to a number of article directories. However, most of the major search engines filter duplicate content to stop the identical content material from being returned multiple times in searches. Some marketers attempt to circumvent this filter by creating a number of variations of an article, known as article spinning. By doing this, one article can theoretically acquire site visitors from a number of article directories.

Most forms of search engine optimization and internet marketing require a domain, internet hosting plan, and promoting budget. However, article advertising makes use of article directories as a free host and receives traffic by way of organic searches due to the listing's search engine authority.

The primary goal behind article marketing is to get search engine traffic and authors generally incorporate relevant keywords or keyphrases in their articles.

References

Digital marketing

Digital marketing is the use of digital sources based on electronic signal like Internet, digital display advertising and other digital media such as television, radio, and mobile phones in the promotion of brands and products to consumers. Digital marketing may cover the more traditional marketing areas such as Direct Marketing by providing the same method of communicating with an audience but in a digital fashion.

Digital marketing – Pull versus Push

Two different forms of digital marketing exist.

Pull digital marketing in which the consumer must actively seek the marketing content, often via web searches, and push digital marketing where the marketer sends the content to the consumer, as in email. Websites, blogs and streaming media (audio and video) are examples of pull digital marketing. In each of these users have to link to the website to view the content. Only current web browser technology is required to maintain static content. However, additional internet marketing technologies (search engine optimization) may be required to attract the desired consumer demographic.

Push digital marketing technologies involve both the marketer as well as the recipients. Email, text messaging and web feeds are examples of push digital marketing. In each of these, the marketer has to send the messages to the subscribers. In the case of web feeds, content is pulled on a periodic basis (polling), thus simulating a push. Push technologies can deliver content immediately as it becomes available and is better targeted to its consumer demographic, although audiences are often smaller, and the cost for creation and distribution is higher.

Digital Marketing and Multi-Channel Communications

Push and pull message technologies can be used in conjunction with each other. For example, an email campaign can include a banner ad or link to a content download. This enables a marketer to benefit from both types of digital marketing.

References

Hilltop algorithm

The **Hilltop algorithm** is an algorithm used to find documents relevant to a particular keyword topic. Created by Krishna Bharat while he was at Compaq Systems Research Center and George A. Mihăilă, then at the University of Toronto, it was acquired by Google in February 2003. Whenever you enter a query or keyword in Search engine hilltop algorithm helps to find relevant keywords matched results. Which are more informative about the query or keyword. The algorithm operates on a special index of *expert documents*. These are pages that are about a specific topic and have links to many non-affiliated pages on that topic. Pages are defined as non-affiliated if they are authored by people from non-affiliated organizations. Results are ranked based on the match between the query and relevant descriptive text for hyperlinks on expert pages pointing to a given result page. Websites which have backlinks from many of the best expert pages are *authorities* and are ranked well. Basically, it looks at the relationship between the "Expert" and "Authority" pages. An "Expert" is a page that links to lots of other relevant documents. An "Authority" is a page that has links pointing to it from the "Expert" pages. Here they mean pages about a specific topic and having links to many non-affiliated pages on that topic. Pages are defined as non-affiliated if they are authored by authors from non-affiliated organizations. So, if your website has backlinks from many of the best expert pages it will be an "Authority".

In theory, Google finds "Expert" pages and then the pages that they link to would rank well. Pages on sites like Yahoo!, DMOZ, college sites and library sites can be considered experts.

External links

- Hilltop: A Search Engine based on Expert Documents [1] by K. Bharat and G.A. Mihaila
  - At archive.org: [2]
  - When Experts Agree: Using Non-Affiliated Experts to Rank Popular Topics [3] by K. Bharat and G. A. Mihaila is substantially the same, but under a different title.
- The Hilltop algorithm [4]

References

[1] ftp://ftp.cs.toronto.edu/pub/reports/csri/405/hilltop.html
[2] http://web.archive.org/web/20070401224626/http://www.cs.toronto.edu/~georgem/hilltop/
[3] http://citeseer.ist.psu.edu/bharat01when.html
[4] http://pagerank.suchmaschinen-doktor.de/hilltop.html

TrustRank

TrustRank is a link analysis technique described in a paper by Stanford University and Yahoo! researchers for semi-automatically separating useful webpages from spam.[1]

Many Web spam pages are created only with the intention of misleading search engines. These pages, chiefly created for commercial reasons, use various techniques to achieve higher-than-deserved rankings on the search engines' result pages. While human experts can easily identify spam, it is too expensive to manually evaluate a large number of pages.

One popular method for improving rankings is to increase artificially the perceived importance of a document through complex linking schemes. Google's PageRank and similar methods for determining the relative importance of Web documents have been subjected to manipulation.

TrustRank method calls for selecting a small set of seed pages to be evaluated by an expert. Once the reputable seed pages are manually identified, a crawl extending outward from the seed set seeks out similarly reliable and trustworthy pages. TrustRank's reliability diminishes with increased distance between documents and the seed set.

The researchers who proposed the TrustRank methodology have continued to refine their work by evaluating related topics, such as measuring spam mass.

References

[1] Gyöngyi, Zoltán; Hector Garcia-Molina, Jan Pedersen (2004). "Combating Web Spam with TrustRank" (http://www.vldb.org/conf/2004/ RS15P3.PDF). *Proceedings of the International Conference on Very Large Data Bases* **30**: 576. . Retrieved 2007-10-26.

External links

- Z. Gyöngyi, H. Garcia-Molina, J. Pedersen: *Combating Web Spam with TrustRank* (http://www.vldb.org/conf/2004/RS15P3.PDF)
- Link-based spam detection (http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PG01&p=1&u=/netahtml/PTO/srchnum.html&r=1&f=G&l=50&s1="20060095416".PGNR.&OS=DN/20060095416&RS=DN/20060095416) Yahoo! assigned patent application using Trustrank
- TrustRank algorithm explained (http://pagerank.suchmaschinen-doktor.de/trustrank.html)

Latent semantic indexing

Latent Semantic Indexing (LSI) is an indexing and retrieval method that uses a mathematical technique called Singular value decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text. LSI is based on the principle that words that are used in the same contexts tend to have similar meanings. A key feature of LSI is its ability to extract the conceptual content of a body of text by establishing associations between those terms that occur in similar contexts.[1]

LSI is also an application of correspondence analysis, a multivariate statistical technique developed by Jean-Paul Benzécri[2] in the early 1970s, to a contingency table built from word counts in documents.

Called Latent Semantic Indexing because of its ability to correlate semantically related terms that are latent in a collection of text, it was first applied to text at Bell Laboratories in the late 1980s. The method, also called latent semantic analysis (LSA), uncovers the underlying latent semantic structure in the usage of words in a body of text and how it can be used to extract the meaning of the text in response to user queries, commonly referred to as concept searches. Queries, or concept searches, against a set of documents that have undergone LSI will return results that are conceptually similar in meaning to the search criteria even if the results don't share a specific word or words with the search criteria.

Benefits of LSI

LSI overcomes two of the most problematic constraints of Boolean keyword queries: multiple words that have similar meanings (synonymy) and words that have more than one meaning (polysemy). Synonymy and polysemy are often the cause of mismatches in the vocabulary used by the authors of documents and the users of information retrieval systems.[3] As a result, Boolean keyword queries often return irrelevant results and miss information that is relevant.

LSI is also used to perform automated document categorization. In fact, several experiments have demonstrated that there are a number of correlations between the way LSI and humans process and categorize text.[4] Document categorization is the assignment of documents to one or more predefined categories based on their similarity to the conceptual content of the categories.[5] LSI uses *example* documents to establish the conceptual basis for each category. During categorization processing, the concepts contained in the documents being categorized are compared to the concepts contained in the example items, and a category (or categories) is assigned to the documents based on the similarities between the concepts they contain and the concepts that are contained in the example documents.

Dynamic clustering based on the conceptual content of documents can also be accomplished using LSI. Clustering is a way to group documents based on their conceptual similarity to each other without using example documents to establish the conceptual basis for each cluster. This is very useful when dealing with an unknown collection of unstructured text.

Because it uses a strictly mathematical approach, LSI is inherently independent of language. This enables LSI to elicit the semantic content of information written in any language without requiring the use of auxiliary structures, such as dictionaries and thesauri. LSI can also perform cross-linguistic concept searching and example-based categorization. For example, queries can be made in one language, such as English, and conceptually similar results will be returned even if they are composed of an entirely different language or of multiple languages.

LSI is not restricted to working only with words. It can also process arbitrary character strings. Any object that can be expressed as text can be represented in an LSI vector space.[6] For example, tests with MEDLINE abstracts have shown that LSI is able to effectively classify genes based on conceptual modeling of the biological information contained in the titles and abstracts of the MEDLINE citations.[7]

LSI automatically adapts to new and changing terminology, and has been shown to be very tolerant of noise (i.e., misspelled words, typographical errors, unreadable characters, etc.).[8] This is especially important for applications

using text derived from Optical Character Recognition (OCR) and speech-to-text conversion. LSI also deals effectively with sparse, ambiguous, and contradictory data.

Text does not need to be in sentence form for LSI to be effective. It can work with lists, free-form notes, email, Web-based content, etc. As long as a collection of text contains multiple terms, LSI can be used to identify patterns in the relationships between the important terms and concepts contained in the text.

LSI has proven to be a useful solution to a number of conceptual matching problems.[9][10] The technique has been shown to capture key relationship information, including causal, goal-oriented, and taxonomic information.[11]

LSI Timeline

Mid-1960s – Factor analysis technique first described and tested (H. Borko and M. Bernick)

1988 – Seminal paper on LSI technique published (Deerwester et al.)

1989 – Original patent granted (Deerwester et al.)

1992 – First use of LSI to assign articles to reviewers[12] (Dumais and Nielsen) **1994 –** Patent granted for the cross-lingual application of LSI (Landauer et al.) **1995 –** First use of LSI for grading essays (Foltz, et al., Landauer et al.)

1999 – First implementation of LSI technology for intelligence community for analyzing unstructured text (SAIC).

2002 – LSI-based product offering to intelligence-based government agencies (SAIC)

2005 – First vertical-specific application – publishing – EDB (EBSCO, Content Analyst Company)

Mathematics of LSI

LSI uses common linear algebra techniques to learn the conceptual correlations in a collection of text. In general, the process involves constructing a weighted term-document matrix, performing a **Singular Value Decomposition** on the matrix, and using the matrix to identify the concepts contained in the text.

Term Document Matrix

LSI begins by constructing a term-document matrix, A, to identify the occurrences of the unique terms within a collection of documents. In a term-document matrix, each term is represented by a row, and each document is represented by a column, with each matrix cell, a_{ij}, initially representing the number of times the associated term appears in the indicated document, tf_{ij}. This matrix is usually very large and very sparse.

Once a term-document matrix is constructed, local and global weighting functions can be applied to it to condition the data. The weighting functions transform each cell, a_{ij} of A, to be the product of a local term weight, l_{ij}, which describes the relative frequency of a term in a document, and a global weight, g_i, which describes the relative frequency of the term within the entire collection of documents.

Some common local weighting functions [13] are defined in the following table.

Binary	if the term exists in the document, or else 0
TermFrequency	, the number of occurrences of term i in document j
Log	$l_{ij} = \log(\mathrm{tf}_{ij} + 1)$
Augnorm	$l_{ij} = \frac{\left(\frac{\mathrm{tf}_{ij}}{\max_i(\mathrm{tf}_{ij})}\right) + 1}{2}$

Some common global weighting functions are defined in the following table.

Binary	$g_i = 1$
Normal	$g_i = \frac{1}{\sqrt{\sum_j \mathrm{tf}_{ij}^2}}$
GfIdf	, where gf_i is the total number of times term i occurs in the whole collection, and df_i is the number of documents in which term i occurs.
Idf	$g_i = \log_2 \frac{n}{1 + \mathrm{df}_i}$
Entropy	$g_i = 1 + \sum_j \frac{p_{ij} \log p_{ij}}{\log n}$, where $p_{ij} = \frac{\mathrm{tf}_{ij}}{\mathrm{gf}_i}$

Empirical studies with LSI report that the Log Entropy weighting functions work well, in practice, with many data sets.[14] In other words, each entry a_{ij} of A is computed as:

$$g_i = 1 + \sum_j \frac{p_{ij} \log p_{ij}}{\log n}$$

$$a_{ij} = g_i \ \log(\mathrm{tf}_{ij} + 1)$$

Rank-Reduced Singular Value Decomposition

A rank-reduced, Singular Value Decomposition is performed on the matrix to determine patterns in the relationships between the terms and concepts contained in the text. The SVD forms the foundation for LSI.[15] It computes the term and document vector spaces by transforming the single term-frequency matrix, A, into three other matrices—
an ***m*** by ***r*** term-concept vector matrix T, an ***r*** by ***r*** singular values matrix S, and a ***n*** by ***r*** concept-document vector matrix, D, which satisfy the following relations:

$$A = TSD^T$$
$$T^T T = I_r \quad D^T D = I_r$$
$$S_{1,1} \geq S_{2,2} \geq \ldots \geq S_{r,r} > 0 \quad S_{i,j} = 0 \text{ where } i \neq j$$

In the formula, **A**, is the supplied ***m*** by ***n*** weighted matrix of term frequencies in a collection of text where ***m*** is the number of unique terms, and ***n*** is the number of documents. **T** is a computed ***m*** by ***r*** matrix of term vectors where ***r*** is the rank of **A**—a measure of its unique dimensions **≤ min(*m*,*n*)**. **S** is a computed ***r*** by ***r*** diagonal matrix of decreasing singular values, and **D** is a computed ***n*** by ***r*** matrix of document vectors.

The LSI modification to a standard SVD is to reduce the rank or truncate the singular value matrix **S** to size ***k*** « ***r***, typically on the order of a ***k*** in the range of 100 to 300 dimensions, effectively reducing the term and document vector matrix sizes to ***m*** by ***k*** and ***n*** by ***k*** respectively. The SVD operation, along with this reduction, has the effect of preserving the most important semantic information in the text while reducing noise and other undesirable artifacts of the original space of **A**. This reduced set of matrices is often denoted with a modified formula such as:

$$\mathbf{A} \approx \mathbf{A}_k = \mathbf{T}_k \mathbf{S}_k \mathbf{D}_k^{\mathbf{T}}$$

Efficient LSI algorithms only compute the first k singular values and term and document vectors as opposed to computing a full SVD and then truncating it.

Note that this rank reduction is essentially the same as doing Principal Component Analysis (PCA) on the matrix **A**, except that PCA subtracts off the means. PCA provides cleaner mathematics, but loses the sparseness of the **A** matrix, which can make it infeasible for large lexicons.

Querying and Augmenting LSI Vector Spaces

The computed $\mathbf{T}_k$ and $\mathbf{D}_k$ matrices define the term and document vector spaces, which with the computed singular values, $\mathbf{S}_k$, embody the conceptual information derived from the document collection. The similarity of terms or documents within these spaces is a factor of how close they are to each other in these spaces, typically computed as a function of the angle between the corresponding vectors.

The same steps are used to locate the vectors representing the text of queries and new documents within the document space of an existing LSI index. By a simple transformation of the $\mathbf{A} = \mathbf{T}\,\mathbf{S}\,\mathbf{D}^{\mathbf{T}}$ equation into the equivalent $\mathbf{D} = \mathbf{A}^{\mathbf{T}}\,\mathbf{T}\,\mathbf{S}^{-1}$ equation, a new vector, $\boldsymbol{d}$, for a query or for a new document can be created by computing a new column in **A** and then multiplying the new column by $\mathbf{T}\,\mathbf{S}^{-1}$. The new column in **A** is computed using the originally derived global term weights and applying the same local weighting function to the terms in the query or in the new document.

A drawback to computing vectors in this way, when adding new searchable documents, is that terms that were not known during the SVD phase for the original index are ignored. These terms will have no impact on the global weights and learned correlations derived from the original collection of text. However, the computed vectors for the new text are still very relevant for similarity comparisons with all other document vectors.

The process of augmenting the document vector spaces for an LSI index with new documents in this manner is called *folding-in*. Although the folding-in process does not account for the new semantic content of the new text, adding a substantial number of documents in this way will still provide good results for queries as long as the terms and concepts they contain are well represented within the LSI index to which they are being added. When the terms and concepts of a new set of documents need to be included in an LSI index, the term-document matrix, and the SVD, must either be recomputed or an incremental update method (such as the one described in [16]) be used.

Additional Uses of LSI

It is generally acknowledged that the ability to work with text on a semantic basis is essential to modern information retrieval systems. As a result, the use of LSI has significantly expanded in recent years as earlier challenges in scalability and performance have been overcome.

LSI is being used in a variety of information retrieval and text processing applications, although its primary application has been for concept searching and automated document categorization.[17] Below are some otherways in which LSI is being used:

- Information discovery[18] (eDiscovery, Government/Intelligence community, Publishing)
- Automated document classification (eDiscovery, Government/Intelligence community, Publishing)[19]
- Text summarization[20] (eDiscovery, Publishing)
- Relationship discovery[21] (Government, Intelligence community, Social Networking)
- Automatic generation of link charts of individuals and organizations[22] (Government, Intelligence community)
- Matching technical papers and grants with reviewers[23] (Government)
- Online customer support[24] (Customer Management)
- Determining document authorship[25] (Education)
- Automatic keyword annotation of images[26]
- Understanding software source code[27] (Software Engineering)

- Filtering spam[28] (System Administration)
- Information visualization[29]
- Essay scoring[30] (Education)
- Literature-based discovery[31]

LSI is increasingly being used for electronic document discovery (eDiscovery) to help enterprises prepare for litigation. In eDiscovery, the ability to cluster, categorize, and search large collections of unstructured text on a conceptual basis is essential. Concept-based searching using LSI has been applied to the eDiscovery process by leading providers as early as 2003.[32]

Challenges to LSI

Early challenges to LSI focused on scalability and performance. LSI requires relatively high computational performance and memory in comparison to other information retrieval techniques.[33] However, with the implementation of modern high-speed processors and the availability of inexpensive memory, these considerations have been largely overcome. Real-world applications involving more than 30 million documents that were fully processed through the matrix and SVD computations are not uncommon in some LSI applications.

Another challenge to LSI has been the alleged difficulty in determining the optimal number of dimensions to use for performing the SVD. As a general rule, fewer dimensions allow for broader comparisons of the concepts contained in a collection of text, while a higher number of dimensions enable more specific (or more relevant) comparisons of concepts. The actual number of dimensions that can be used is limited by the number of documents in the collection. Research has demonstrated that around 300 dimensions will usually provide the best results with moderate-sized document collections (hundreds of thousands of documents) and perhaps 400 dimensions for larger document collections (millions of documents).[34] However, recent studies indicate that 50-1000 dimensions are suitable depending on the size and nature of the document collection.[35]

Checking the amount of variance in the data after computing the SVD can be used to determine the optimal number of dimensions to retain. The variance contained in the data can be viewed by plotting the singular values (S) in a scree plot. Some LSI practitioners select the dimensionality associated with the knee of the curve as the cut-off point for the number of dimensions to retain. Others argue that some quantity of the variance must be retained, and the amount of variance in the data should dictate the proper dimensionality to retain. Seventy percent is often mentioned as the amount of variance in the data that should be used to select the optimal dimensionality for recomputing the SVD.[36][37][38]

References

[1] Deerwester, S., et al, Improving Information Retrieval with Latent Semantic Indexing, Proceedings of the 51st Annual Meeting of the American Society for Information Science 25, 1988, pp. 36–40.

[2] Benzécri, J.-P. (1973). *L'Analyse des Données. Volume II. L'Analyse des Correspondences*. Paris, France: Dunod.

[3] Furnas, G., et al, The Vocabulary Problem in Human-System Communication, Communications of the ACM, 1987, 30(11), pp. 964971.

[4] Landauer, T., et al., Learning Human-like Knowledge by Singular Value Decomposition: A Progress Report, M. I. Jordan, M. J. Kearns & S. A. Solla (Eds.), Advances in Neural Information Processing Systems 10, Cambridge: MIT Press, 1998, pp. 45–51.

[5] Dumais, S., Platt J., Heckerman D., and Sahami M., Inductive Learning Algorithms and Representations For Text Categorization, Proceedings of ACM-CIKM'98, 1998.

[6] Zukas, Anthony, Price, Robert J., Document Categorization Using Latent Semantic Indexing, White Paper, Content Analyst Company, LLC

[7] Homayouni, Ramin, Heinrich, Kevin, Wei, Lai, Berry, Michael W., Gene Clustering by Latent Semantic Indexing of MEDLINE Abstracts, August 2004, pp. 104–115.

[8] Price, R., and Zukas, A., Application of Latent Semantic Indexing to Processing of Noisy Text, Intelligence and Security Informatics, Lecture Notes in Computer Science, Volume 3495, Springer Publishing, 2005, pp. 602–603.

[9] Ding, C., A Similarity-based Probability Model for Latent Semantic Indexing, Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999, pp. 59–65.

[10] Bartell, B., Cottrell, G., and Belew, R., Latent Semantic Indexing is an Optimal Special Case of Multidimensional Scaling, Proceedings, ACM SIGIR Conference on Research and Development in Information Retrieval, 1992, pp. 161–167.

[11] Graesser, A., and Karnavat, A., Latent Semantic Analysis Captures Causal, Goal-oriented, and Taxonomic Structures, Proceedings of CogSci 2000, pp. 184–189.

[12] Dumais, S., and Nielsen, J., Automating the Assignment of Submitted Manuscripts to Reviewers, Proceedings of the Fifteenth Annual International Conference on Research and Development in Information Retrieval, 1992, pp. 233–244.

[13] Berry, M. W., and Browne, M., Understanding Search Engines: Mathematical Modeling and Text Retrieval, Society for Industrial and Applied Mathematics, Philadelphia, (2005).

[14] Landauer, T., et al., Handbook of Latent Semantic Analysis, Lawrence Erlbaum Associates, 2007.

[15] Berry, Michael W., Dumais, Susan T., O'Brien, Gavin W., Using Linear Algebra for Intelligent Information Retrieval, December 1994, SIAM Review 37:4 (1995), pp. 573–595.

[16] Matthew Brand (2006). "Fast Low-Rank Modifications of the Thin Singular Value Decomposition" (http://www.merl.com/reports/docs/ TR2006-059.pdf) (PDF). *Linear Algebra and Its Applications* **415**: 20–30. doi:10.1016/j.laa.2005.07.021. .

[17] Dumais, S., Latent Semantic Analysis, ARIST Review of Information Science and Technology, vol. 38, 2004, Chapter 4.

[18] Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery, the Sedona Conference, 2007, pp. 189–223.

[19] Foltz, P. W. and Dumais, S. T. Personalized Information Delivery: An analysis of information filtering methods, Communications of the ACM, 1992, 34(12), 51-60.

[20] Gong, Y., and Liu, X., Creating Generic Text Summaries, Proceedings, Sixth International Conference on Document Analysis and Recognition, 2001, pp. 903–907.

[21] Bradford, R., Efficient Discovery of New Information in Large Text Databases, Proceedings, IEEE International Conference on Intelligence and Security Informatics, Atlanta, Georgia, LNCS Vol. 3495, Springer, 2005, pp. 374–380.

[22] Bradford, R., Application of Latent Semantic Indexing in Generating Graphs of Terrorist Networks, in: Proceedings, IEEE International Conference on Intelligence and Security Informatics, ISI 2006, San Diego, CA, USA, May 23–24, 2006, Springer, LNCS vol. 3975, pp. 674–675.

[23] Yarowsky, D., and Florian, R., Taking the Load off the Conference Chairs: Towards a Digital Paper-routing Assistant, Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very-Large Corpora, 1999, pp. 220–230.

[24] Caron, J., Applying LSA to Online Customer Support: A Trial Study, Unpublished Master's Thesis, May 2000.

[25] Soboroff, I., et al, Visualizing Document Authorship Using N-grams and Latent Semantic Indexing, Workshop on New Paradigms in Information Visualization and Manipulation, 1997, pp. 43–48.

[26] Monay, F., and Gatica-Perez, D., On Image Auto-annotation with Latent Space Models, Proceedings of the 11th ACM international conference on Multimedia, Berkeley, CA, 2003, pp. 275–278.

[27] Maletic, J., and Marcus, A., Using Latent Semantic Analysis to Identify Similarities in Source Code to Support Program Understanding, Proceedings of 12th IEEE International Conference on Tools with Artificial Intelligence, Vancouver, British Columbia, November 13–15, 2000, pp. 46–53.

[28] Gee, K., Using Latent Semantic Indexing to Filter Spam, in: Proceedings, 2003 ACM Symposium on Applied Computing, Melbourne, Florida, pp. 460–464.

[29] Landauer, T., Laham, D., and Derr, M., From Paragraph to Graph: Latent Semantic Analysis for Information Visualization, Proceedings of the National Academy of Science, 101, 2004, pp. 5214–5219.

[30] Foltz, Peter W., Laham, Darrell, and Landauer, Thomas K., Automated Essay Scoring: Applications to Educational Technology, Proceedings of EdMedia, 1999.

[31] Gordon, M., and Dumais, S., Using Latent Semantic Indexing for Literature Based Discovery, Journal of the American Society for Information Science, 49(8), 1998, pp. 674–685.

[32] There Has to be a Better Way to Search, 2008, White Paper, Fios, Inc.

[33] Karypis, G., Han, E., Fast Supervised Dimensionality Reduction Algorithm with Applications to Document Categorization and Retrieval, Proceedings of CIKM-00, 9th ACM Conference on Information and Knowledge Management.

[34] Bradford, R., An Empirical Study of Required Dimensionality for Large-scale Latent Semantic Indexing Applications, Proceedings of the 17th ACM Conference on Information and Knowledge Management, Napa Valley, California, USA, 2008, pp. 153–162.

[35] Landauer, Thomas K., and Dumais, Susan T., Latent Semantic Analysis, Scholarpedia, 3(11):4356, 2008.

[36] Cangelosi, R., Goriely A., Component Retention In Principal Component Analysis With Application to Cdna Microarray Data, BMC Biology Direct 2(2) (2007).

[37] Jolliffe, L. T., Principal Component Analysis, Springer-Verlag, New York, (1986).

[38] Hu, X., Z. Cai, et al., LSA: First Dimension and Dimensional Weighting, 25th Annual Meeting of the Cognitive Science Society, Boston, MA.

External links

- Michael Berry's site (http://www.cs.utk.edu/~lsi/)
- Gensim (http://radimrehurek.com/gensim) contains a Python+NumPy implementation of LSI for matrices larger than the available RAM.
- Text to Matrix Generator (TMG) (http://scgroup.hpclab.ceid.upatras.gr/scgroup/Projects/TMG/) MATLAB toolbox that can be used for various tasks in text mining (TM) specifically i) indexing, ii) retrieval, iii) dimensionality reduction, iv) clustering, v) classification. Most of TMG is written in MATLAB and parts in Perl. It contains implementations of LSI, clustered LSI, NMF and other methods.

Further reading

Berry, M. W., Browne M., Understanding Search Engines: Mathematical Modeling and Text Retrieval, Philadelphia, Society for Industrial and Applied Mathematics, (2005).

Berry, M. W., (Editor), Survey of Text Mining: Clustering, Classification, and Retrieval, New York, Springer, (2004).

Landauer, T., et al., Handbook of Latent Semantic Analysis, Lawrence Erlbaum Associates, 2007.

Manning, C. D., Schutze H., Foundations of Statistical Natural Language Processing, Cambridge, MA, The MIT Press, (1999).

Semantic targeting

Semantic targeting is a technique enabling the delivery of targeted advertising for advertisements appearing on websites and is used by online publishers and advertisers to increase the effectiveness of their campaigns. The selection of advertisements are served by automated systems based on the content displayed to the user.

Origins

Semantic Targeting has originated from the developments arising from Semantic Web. The Semantic Web enables the representation of concepts expressed in human language to data in such a way that facilitates automatic processing, where software can programmatically understand and reason how different elements of data are related. The concept of semantic targeting utilises this capability to identify these concepts and the contexts in which they occur, enabling marketers to deliver highly targeted and specific ad campaigns to webpages.

The evolution of online advertising

The targeting of advertising to specific micro segments is a fundamental requirement for an effective ad campaign. The two methods of targeting of recent times have been behavioral targeting and contextual targeting. It is now generally accepted that these forms have pitfalls for both advertiser and consumer.

Behavioral targeting aggregates data based upon a user's viewing of pages from a website. Generally this is facilitated by the placing of a cookie upon the user's PC. The cookie then reports the user's viewing behavior allowing for the identification of patterns of viewing behavior. However, great concern is expressed about the treatment of the user's right to privacy amongst consumer groups and legislators.[1][2]

Contextual advertising scans the content of webpages, seeking to identify keywords, against which advertisers have bid to have their ad linked. If a match is made the ad is placed alongside the content, through an automated process. However, such systems are unable to identify the context of the entire page and therefore, a placement could be made against content that is inappropriate, derogatory or insensitive to the subject.[3][4][5][6] They are also unable to

identify the sense or meaning of words, leading to a misplacement of ads. For example, the word "orange" can be a color, a fruit, a telecommunications company, a mountain bike, and countless other variants.

How semantic targeting works

Semantic targeting aims to match the specific context of content on page within a website to an available advertising campaign. A key difference of semantic targeting to a contextual advertising system is that, instead of scanning a page for bided keywords, a semantic system examines all the words and identifies the senses of those words.[7] Because most words are polysemous, i.e. have more than one meaning, without having an understanding of the true context in which words occur, it is possible to incorrectly assign an advertisement where there is no contextual link. A semantic targeting system has to examine all the words before it can accurately identify the subject matter of the entire text and deliver an in context advertisement.[8] For example, if the user is viewing a website relating to golf, where that website uses semantic targeting, the user may see advertisements for golf related topics, such as golf equipment, golf holidays etc. Advertisers can locate their ads in given categories using an ontology (computer science) or taxonomy, ensuring that their ads will only appear in the context that they request.

Semantic targeting is also capable of identifying the sentiment of a webpage, through effective analysis of the language used on page. Sentiment analysis can determine whether content is talking about a subject in a positive or negative light. If the page was being detrimental about a particular subject, the semantic targeting system could deter the placement of a related ad alongside the story.

Other capabilities of a semantic targeting system include the availability of brand protection filtering. This can enable the blocking of an ad placed alongside content of a controversial nature. Such systems can deter placement against such subjects as Adult/Erotica, Alcohol, Nudity, Offensive language, Bad News and other such topics. This would then avoid the potentially brand damaging occurrence of, for example, and airline advertising alongside a story about an air disaster.[9]

References

[1] ReneeBoucherFerguson(2008-03-27)."ABattleIsBrewingOverOnlineBehavioralAdvertising"(http://www.eweek.com/c/a/ Enterprise-Applications/A-Battle-Is-Brewing-Over-Online-Behavioral-Advertising-Market/). E-week.com. . Retrieved 2008-10-10.

[2] "FTC Staff Proposes Online Behavioral Advertising Privacy Principles" (http://www.ftc.gov/opa/2007/12/principles.shtm). Federal Trade Commission. 2008-12-20. . Retrieved 2008-10-10.

[3] "Steve Irwin's Death : Contextual Advertising Gone Bad" (http://www.shmula.com/194/ steve-irwins-death-contextual-advertising-gone-bad). Shmula.com. 2006-09-05. . Retrieved 2008-10-10.

[4] "Contextual advertising gone bad" (http://www.etre.com/blog/2007/11/contextual_advertising_gone_bad/). etre.com. 2007.11.02. . Retrieved 2008-10-10.

[5] "When Contextual Advertising Goes Horribly Wrong" (http://mashable.com/2008/06/19/contextual-advertising/). Mashable.com. 2008.06.19. . Retrieved2008-10-10.

[6] "McCain campaign pulls ads from some anti-Obama Web sites" (http://www.cnn.com/2008/POLITICS/07/01/mccain.ads/). CNN.com. 2008-07-01. . Retrieved2008-10-10.

[7] Graham Charlton (2008-05-16). "Q & A Prof. David Crystal & Sacha Carton on semantic targeting" (http://www.e-consultancy.com/ news-blog/365601/q-a-prof-david-crystal-and-sacha-carton-on-semantic-targeting.html). e-consultancy. . Retrieved 2008-10-10.

[8] "Semantic Targeting" (http://web.archive.org/web/20080213195315/http://www.isense.net/index.php?id=68). isense.net (copy on web.archive.org). Archived from the original (http://www.isense.net/index.php?id=68) on 2008-02-13. . Retrieved2010-10-20.

[9] Scott Brinker, (2008-09-13). "Semantic advertising: 4 different kinds"(http://www.chiefmartec.com/2008/09/ semantic-advertising-of-4-different-kinds.html).iChiefmarketingtechnologist..Retrieved2008-10-10.

Canonical meta tag

A **canonical link element** is an HTML element that helps webmasters prevent duplicate content issues by specifying the "canonical", or "preferred", version of a web page[1][2][3] as part of search engine optimization.

Duplicate content issues occur when the same content is accessible from multiple URLs.[4] For example, `http://www.example.com/page.html` would be considered by search engines to be an entirely different page to `http://www.example.com/page.html?parameter=1`, even though both URLs return the same content. Another example is essentially the same (tabular) content, but sorted differently.

In February 2009, Google, Yahoo and Microsoft announced support for the canonical link element, which can be inserted into the <head> section of a web page, to allow webmasters to prevent these issues.[5] The canonical link element helps webmasters make clear to the search engines which page should be credited as the original.

According to Google, the canonical link element is not considered to be a directive, but a hint that the web crawler will "honor strongly".[1]

While the canonical link element has its benefits, Matt Cutts, who is the head of Google's webspam team, has claimed that the search engine prefers the use of 301 redirects. Cutts claims the preference for redirects is because Google's spiders can choose to ignore a canonical link element if they feel it is more beneficial to do so.[6]

Examples of the `canonical` link element

```
<link rel="canonical" href="http://www.example.com/" />
```

```
<link rel="canonical" href="http://www.example.com/page.html" />
```

```
<link rel="canonical" href="http://www.example.com/directory/page.html" />
```

References

[1] http://googlewebmastercentral.blogspot.com/2009/02/specify-your-canonical.html
[2] http://www.mattcutts.com/blog/canonical-link-tag/
[3] http://www.seomoz.org/blog/canonical-url-tag-the-most-important-advancement-in-seo-practices-since-sitemaps
[4] http://www.google.com/support/webmasters/bin/answer.py?answer=66359
[5] http://searchengineland.com/canonical-tag-16537
[6] http://www.seonewyorkcity.org/seo-blog/301-redirect-vs-relcanonical-tags/

Keyword research

Keyword research is a practice used by search engine optimization professionals to find and research actual search terms people enter into the search engines when conducting a search. Search engine optimization professionals research keywords in order to achieve better rankings in their desired keywords.[1]

Potential barriers

Existing brands

If a company decides to sell Nike trainers online, the market is pretty competitive, and the Nike brand itself is Predominant.

Sources of traditional research data

- Google AdWords Keyword Tool, traffic estimator, Webmaster Tools; Google Suggest and Google Trends
- MSN Keyword Forecast
- Hitwise

References

[1] Daniel Lofton (2010). "Importance of Keyword Research" (http://www.articlemarketinghq.com/keyword-research/ keyword-research-importance). Article Marketing HQ. . Retrieved November 9, 2010.

Latent Dirichlet allocation

In statistics, **latent Dirichlet allocation (LDA)** is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics. LDA is an example of a topic model and was first presented as a graphical model for topic discovery by David Blei, Andrew Ng, and Michael Jordan in 2002.[1]

Topics in LDA

In LDA, each document may be viewed as a mixture of various topics. This is similar to probabilistic latent semantic analysis (pLSA), except that in LDA the topic distribution is assumed to have a Dirichlet prior. In practice, this results in more reasonable mixtures of topics in a document. It has been noted, however, that the pLSA model is equivalent to the LDA model under a uniform Dirichlet prior distribution.[2]

For example, an LDA model might have topics that can be classified as **CAT** and **DOG**. However, the classification is arbitrary because the topic that encompasses these words cannot be named. Furthermore, a topic has probabilities of generating various words, such as *milk*, *meow*, and *kitten*, which can be classified and interpreted by the viewer as "CAT". Naturally, *cat* itself will have high probability given this topic. The **DOG** topic likewise has probabilities of generating each word: *puppy*, *bark*, and *bone* might have high probability. Words without special relevance, such as *the* (see function word), will have roughly even probability between classes (or can be placed into a separate category).

A document is given the topics. This is a standard bag of words model assumption, and makes the individual words exchangeable.

Model

With plate notation, the dependencies among the many variables can be captured concisely. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document. M denotes the number of documents, N the number of words in a document. Thus:

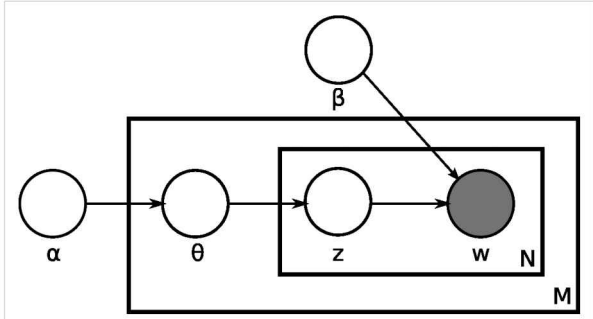


Plate notation representing the LDA model.

α is the parameter of the Dirichlet prior on the per-document topic distributions.

β is the parameter of the Dirichlet prior on the per-topic word distribution.

θ_i is the topic distribution for document i, is the

ϕ_k word distribution for topic k,

z_{ij} is the topic for the jth word in document i, and

w_{ij} is the specific word.

The w_{ij} are the only observable variables, and the other variables are latent variables. Mostly, the basic LDA model will be extended to a smoothed version to gain better results. The plate notation is shown on the right, where K denotes the number of topics considered in the model and:

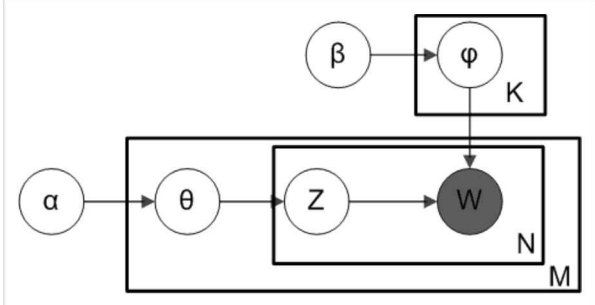


Plate notation for smoothed LDA

ϕ is a $K*V$ (V is the dimension of the vocabulary) Markov matrix each row of which denotes the word distribution of a topic.

The generative process behind is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. LDA assumes the following generative process for each document i in a corpus D :

1. Choose $\theta_i \sim \mathrm{Dir}(\alpha)$, where $i \in \{1, \ldots, M\}$ and $\mathrm{Dir}(\alpha)$ is the Dirichlet distribution for parameter α
2. Choose $\phi_k \sim \mathrm{Dir}(\beta)$, where $k \in \{1, \ldots, K\}$
3. For each of the words w_{ij}, where $j \in \{1, \ldots, N_i\}$
   (a) Choose a topic $z_{i,j} \sim \mathrm{Multinomial}(\theta_i)$.
   (b) Choose a word $w_{i,j} \sim \mathrm{Multinomial}(\phi_{z_{i,j}})$.

(Note that the Multinomial distribution here refers to the Multinomial with only one trial. It is formally equivalent to the categorical distribution.)

The lengths N_i are treated as independent of all the other data generating variables (q and z). The subscript is often dropped, as in the plate diagrams shown here.

Mathematical definition

A formal description of smoothed LDA is as follows:

Definition of variables in the model

Variable	Type	Meaning
K	integer	number of topics (e.g. 50)
V	integer	number of words in the vocabulary (e.g. 50,000 or 1,000,000)
M	integer	number of documents
$N_{d=1\ldots M}$	integer	number of words in document *d*
N	integer	total number of words in all documents; sum of all values, i.e. $N = \sum_{d=1}^{M} N_d$
$\alpha_{k=1\ldots K}$	positive real	prior weight of topic *k* in a document; usually the same for all topics; normally a number less than 1, e.g. 0.1, to prefer sparse topic distributions, i.e. few topics per document
$\boldsymbol{\alpha}$	*K*-dimension vector of positive reals	collection of all α_k values, viewed as a single vector
$\beta_{w=1\ldots V}$	positive real	prior weight of word *w* in a topic; usually the same for all words; normally a number much less than 1, e.g. 0.001, to strongly prefer sparse word distributions, i.e. few words per topic
$\boldsymbol{\beta}$	*V*-dimension vector of positive reals	collection of all β_w values, viewed as a single vector
$\phi_{k=1\ldots K,w=1\ldots V}$	probability (real number between 0 and 1)	probability of word *w* occurring in topic *k*
$\boldsymbol{\phi}_{k=1\ldots K}$	*V*-dimension vector of probabilities, which must sum to 1	distribution of words in topic *k*
$\theta_{d=1\ldots M,k=1\ldots K}$	probability (real number between 0 and 1)	probability of topic *k* occurring in document *d* for a given word
$\boldsymbol{\theta}_{d=1\ldots M}$	*K*-dimension vector of probabilities, which must sum to 1	distribution of topics in document *d*
$z_{d=1\ldots M,w=1\ldots N_d}$	integer between 1 and *K*	identity of topic of word *w* in document *d*
$\mathbf{Z}$	*N*-dimension vector of integers between 1 and *K*	identity of topic of all words in all documents
$w_{d=1\ldots M,w=1\ldots N_d}$	integer between 1 and *V*	identity of word *w* in document *d*
$\mathbf{W}$	*N*-dimension vector of integers between 1 and *V*	identity of all words in all documents

We can then mathematically describe the random variables as follows:

$$
\begin{aligned}
\boldsymbol{\phi}_{k=1\ldots K} &\sim \text{Dirichlet}_V(\boldsymbol{\beta}) \\
\boldsymbol{\theta}_{d=1\ldots M} &\sim \text{Dirichlet}_K(\boldsymbol{\alpha}) \\
z_{d=1\ldots M,w=1\ldots N_d} &\sim \text{Categorical}_K(\boldsymbol{\theta}_d) \\
w_{d=1\ldots M,w=1\ldots N_d} &\sim \text{Categorical}_V(\boldsymbol{\phi}_{z_{dw}})
\end{aligned}
$$

Inference

Learning the various distributions (the set of topics, their associated word probabilities, the topic of each word, and the particular topic mixture of each document) is a problem of Bayesian inference. The original paper used a variational Bayes approximation of the posterior distribution;[1] alternative inference techniques use Gibbs sampling[3] and expectation propagation.[4]

Following is the derivation of the equations for collapsed Gibbs sampling, which means φs and θs will be integrated out. For simplicity, in this derivation the documents are all assumed to have the same length N. The derivation is equally valid if the document lengths vary.

According to the model, the total probability of the model is:

$$P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi}; \alpha, \beta) = \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{t=1}^{N} P(Z_{j,t}|\theta_j) P(W_{j,t}|\varphi_{Z_{j,t}}),$$

where the bold-font variables denote the vector version of the variables. First of all, $\boldsymbol{\varphi}$ and $\boldsymbol{\theta}$ need to be integrated out.

$$P(\boldsymbol{Z}, \boldsymbol{W}; \alpha, \beta) = \int_{\boldsymbol{\theta}} \int_{\boldsymbol{\varphi}} P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi}; \alpha, \beta)\, d\boldsymbol{\varphi}\, d\boldsymbol{\theta}$$

$$= \int_{\boldsymbol{\varphi}} \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{j=1}^{M} \prod_{t=1}^{N} P(W_{j,t}|\varphi_{Z_{j,t}})\, d\boldsymbol{\varphi} \int_{\boldsymbol{\theta}} \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{t=1}^{N} P(Z_{j,t}|\theta_j)\, d\boldsymbol{\theta}.$$

Note that all the θs are independent to each other and the same to all the φs. So we can treat each θ and each φ separately. We now focus only on the θ part.

$$\int_{\boldsymbol{\theta}} \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{t=1}^{N} P(Z_{j,t}|\theta_j) d\boldsymbol{\theta} = \prod_{j=1}^{M} \int_{\theta_j} P(\theta_j; \alpha) \prod_{t=1}^{N} P(Z_{j,t}|\theta_j)\, d\theta_j.$$

We can further focus on only one θ as the following:

$$\int_{\theta_j} P(\theta_j; \alpha) \prod_{t=1}^{N} P(Z_{j,t}|\theta_j)\, d\theta_j.$$

Actually, it is the hidden part of the model for the j^{th} document. Now we replace the probabilities in the above equation by the true distribution expression to write out the explicit equation.

$$\int_{\theta_j} P(\theta_j; \alpha) \prod_{t=1}^{N} P(Z_{j,t}|\theta_j)\, d\theta_j$$

$$= \int_{\theta_j} \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} \theta_{j,i}^{\alpha_i - 1} \prod_{t=1}^{N} P(Z_{j,t}|\theta_j)\, d\theta_j.$$

Let $n_{j,r}^{i}$ be the number of word tokens in the j^{th} document with the same word symbol (the r^{th} word in the vocabulary) assigned to the i^{th} topic. So, $n_{j,r}^{i}$ is three dimensional. If any of the three dimensions is not limited to a specific value, we use a parenthesized point $(\cdot)$ to denote. For example, $n_{j,(\cdot)}^{i}$ denotes the number of word tokens in the j^{th} document assigned to the i^{th} topic. Thus, the right most part of the above equation can be rewritten as:

$$\prod_{t=1}^{N} P(Z_{j,t}|\theta_j) = \prod_{i=1}^{K} \theta_{j,i}^{n_{j,(\cdot)}^{i}}.$$

So the θ_j integration formula can be changed to:

$$\int_{\theta_j} \frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i)} \prod_{i=1}^{K}\theta_{j,i}^{\alpha_i-1} \prod_{i=1}^{K}\theta_{j,i}^{n_{j,(\cdot)}^{i}}\, d\theta_j$$
$$= \int_{\theta_j} \frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i)} \prod_{i=1}^{K}\theta_{j,i}^{n_{j,(\cdot)}^{i}+\alpha_i-1}\, d\theta_j.$$

Clearly, the equation inside the integration has the same form as the Dirichlet distribution. According to the Dirichlet distribution,

Thus,

$$\int_{\theta_j} P(\theta_j;\alpha)\prod_{t=1}^{N}P(Z_{j,t}|\theta_j)\,d\theta_j = \int_{\theta_j} \frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i)} \prod_{i=1}^{K}\theta_{j,i}^{n_{j,(\cdot)}^{i}+\alpha_i-1}\, d\theta_j$$
$$= \frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i)} \frac{\prod_{i=1}^{K}\Gamma(n_{j,(\cdot)}^{i}+\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}n_{j,(\cdot)}^{i}+\alpha_i\right)} \int_{\theta_j} \frac{\Gamma\left(\sum_{i=1}^{K}n_{j,(\cdot)}^{i}+\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(n_{j,(\cdot)}^{i}+\alpha_i)} \prod_{i=1}^{K}\theta_{j,i}^{n_{j,(\cdot)}^{i}+\alpha_i-1}\, d\theta_j$$
$$= \frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i)} \frac{\prod_{i=1}^{K}\Gamma(n_{j,(\cdot)}^{i}+\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}n_{j,(\cdot)}^{i}+\alpha_i\right)}.$$

Now we turn our attentions to the φ part. Actually, the derivation of the φ part is very similar to the θ part. Here we only list the steps of the derivation:

$$\int_{\varphi} \prod_{i=1}^{K}P(\varphi_i;\beta)\prod_{j=1}^{M}\prod_{t=1}^{N}P(W_{j,t}|\varphi_{Z_{j,t}})\,d\varphi$$
$$= \prod_{i=1}^{K}\int_{\varphi_i} P(\varphi_i;\beta)\prod_{j=1}^{M}\prod_{t=1}^{N}P(W_{j,t}|\varphi_{Z_{j,t}})\,d\varphi_i$$
$$= \prod_{i=1}^{K}\int_{\varphi_i} \frac{\Gamma\left(\sum_{r=1}^{V}\beta_r\right)}{\prod_{r=1}^{V}\Gamma(\beta_r)} \prod_{r=1}^{V}\varphi_{i,r}^{\beta_r-1} \prod_{r=1}^{V}\varphi_{i,r}^{n_{(\cdot),r}^{i}}\, d\varphi_i$$
$$= \prod_{i=1}^{K}\int_{\varphi_i} \frac{\Gamma\left(\sum_{r=1}^{V}\beta_r\right)}{\prod_{r=1}^{V}\Gamma(\beta_r)} \prod_{r=1}^{V}\varphi_{i,r}^{n_{(\cdot),r}^{i}+\beta_r-1}\, d\varphi_i$$
$$= \prod_{i=1}^{K} \frac{\Gamma\left(\sum_{r=1}^{V}\beta_r\right)}{\prod_{r=1}^{V}\Gamma(\beta_r)} \frac{\prod_{r=1}^{V}\Gamma(n_{(\cdot),r}^{i}+\beta_r)}{\Gamma\left(\sum_{r=1}^{V}n_{(\cdot),r}^{i}+\beta_r\right)}.$$

For clarity, here we write down the final equation with both ϕ and θ integrated out:

$$P(\boldsymbol{Z},\boldsymbol{W};\alpha,\beta)$$
$$= \prod_{j=1}^{M} \frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i)} \frac{\prod_{i=1}^{K}\Gamma(n_{j,(\cdot)}^{i}+\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}n_{j,(\cdot)}^{i}+\alpha_i\right)} \times \prod_{i=1}^{K} \frac{\Gamma\left(\sum_{r=1}^{V}\beta_r\right)}{\prod_{r=1}^{V}\Gamma(\beta_r)} \frac{\prod_{r=1}^{V}\Gamma(n_{(\cdot),r}^{i}+\beta_r)}{\Gamma\left(\sum_{r=1}^{V}n_{(\cdot),r}^{i}+\beta_r\right)}.$$

The goal of Gibbs Sampling here is to approximate the distribution of $P(\boldsymbol{Z}|\boldsymbol{W};\alpha,\beta)$. Since $P(\boldsymbol{W};\alpha,\beta)$ is invariable for any of Z, Gibbs Sampling equations can be derived from $P(\boldsymbol{Z},\boldsymbol{W};\alpha,\beta)$ directly. The key point is to derive the following conditional probability:

$$P(Z_{(m,n)}|\boldsymbol{Z}_{-(\boldsymbol{m},\boldsymbol{n})},\boldsymbol{W};\alpha,\beta) = \frac{P(Z_{(m,n)},\boldsymbol{Z}_{-(\boldsymbol{m},\boldsymbol{n})},\boldsymbol{W};\alpha,\beta)}{P(\boldsymbol{Z}_{-(\boldsymbol{m},\boldsymbol{n})},\boldsymbol{W};\alpha,\beta)},$$

where $Z_{(m,n)}$ denotes the Z hidden variable of the n^{th} word token in the m^{th} document. And further we assume that the word symbol of it is the v^{th} word in the vocabulary. $\boldsymbol{Z}_{-(m,n)}$ denotes all the Z s but $Z_{(m,n)}$. Note that Gibbs Sampling needs only to sample a value for $Z_{(m,n)}$, according to the above probability, we do not need the exact value of $P(Z_{m,n}|\boldsymbol{Z}_{-(m,n)}, \boldsymbol{W}; \alpha, \beta)$ but the ratios among the probabilities that $Z_{(m,n)}$ can take value. So, the above equation can be simplified as:

$$
\begin{aligned}
& P(Z_{(m,n)} = k|\boldsymbol{Z}_{-(m,n)}, \boldsymbol{W}; \alpha, \beta) \\
\propto & P(Z_{(m,n)} = k, \boldsymbol{Z}_{-(m,n)}, \boldsymbol{W}; \alpha, \beta) \\
= & \left(\frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\right)^{M}\prod_{j\neq m}\frac{\prod_{i=1}^{K}\Gamma(n_{j,(\cdot)}^{i}+\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}n_{j,(\cdot)}^{i}+\alpha_i\right)} \\
& \times\left(\frac{\Gamma\left(\sum_{r=1}^{V}\beta_r\right)}{\prod_{r=1}^{V}\Gamma(\beta_r)}\right)^{K}\prod_{i=1}^{K}\prod_{r\neq v}\Gamma(n_{(\cdot),r}^{i}+\beta_r) \\
& \times\frac{\prod_{i=1}^{K}\Gamma(n_{m,(\cdot)}^{i}+\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}n_{m,(\cdot)}^{i}+\alpha_i\right)}\prod_{i=1}^{K}\frac{\Gamma(n_{(\cdot),v}^{i}+\beta_v)}{\Gamma\left(\sum_{r=1}^{V}n_{(\cdot),r}^{i}+\beta_r\right)} \\
\propto & \frac{\prod_{i=1}^{K}\Gamma(n_{m,(\cdot)}^{i}+\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}n_{m,(\cdot)}^{i}+\alpha_i\right)}\prod_{i=1}^{K}\frac{\Gamma(n_{(\cdot),v}^{i}+\beta_v)}{\Gamma\left(\sum_{r=1}^{V}n_{(\cdot),r}^{i}+\beta_r\right)}.
\end{aligned}
$$

Finally, let $n_{m,(\cdot)}^{i,-(m,n)}$ be the same meaning as $n_{m,(\cdot)}^{i}$ but with the n excluded. The above equation can be further simplified by treating terms not dependent on $n_{j,r}^{i}$ as constants: $Z_{(m,n)} = k$

$$
\begin{aligned}
\propto & \frac{\prod_{i\neq k}\Gamma(n_{m,(\cdot)}^{i,-(m,n)}+\alpha_i)}{\Gamma\left(\left(\sum_{i=1}^{K}n_{m,(\cdot)}^{i,-(m,n)}+\alpha_i\right)+1\right)}\prod_{i\neq k}\frac{\Gamma(n_{(\cdot),v}^{i,-(m,n)}+\beta_v)}{\Gamma\left(\sum_{r=1}^{V}n_{(\cdot),r}^{i,-(m,n)}+\beta_r\right)} \\
& \times\Gamma(n_{m,(\cdot)}^{k,-(m,n)}+\alpha_k+1)\frac{\Gamma(n_{(\cdot),v}^{k,-(m,n)}+\beta_v+1)}{\Gamma\left(\left(\sum_{r=1}^{V}n_{(\cdot),r}^{k,-(m,n)}+\beta_r\right)+1\right)} \\
\propto & \frac{\Gamma(n_{m,(\cdot)}^{k,-(m,n)}+\alpha_k+1)}{\Gamma\left(\left(\sum_{i=1}^{K}n_{m,(\cdot)}^{i,-(m,n)}+\alpha_i\right)+1\right)}\frac{\Gamma(n_{(\cdot),v}^{k,-(m,n)}+\beta_v+1)}{\Gamma\left(\left(\sum_{r=1}^{V}n_{(\cdot),r}^{k,-(m,n)}+\beta_r\right)+1\right)} \\
= & \frac{\Gamma(n_{m,(\cdot)}^{k,-(m,n)}+\alpha_k)(n_{m,(\cdot)}^{k,-(m,n)}+\alpha_k)}{\Gamma\left(\sum_{i=1}^{K}n_{m,(\cdot)}^{i,-(m,n)}+\alpha_i\right)\left(\sum_{i=1}^{K}n_{m,(\cdot)}^{i,-(m,n)}+\alpha_i\right)}\frac{\Gamma(n_{(\cdot),v}^{k,-(m,n)}+\beta_v)(n_{(\cdot),v}^{k,-(m,n)}+\beta_v)}{\Gamma\left(\sum_{r=1}^{V}n_{(\cdot),r}^{k,-(m,n)}+\beta_r\right)\left(\sum_{r=1}^{V}n_{(\cdot),r}^{k,-(m,n)}+\beta_r\right)} \\
\propto & \frac{(n_{m,(\cdot)}^{k,-(m,n)}+\alpha_k)}{\left(\sum_{i=1}^{K}n_{m,(\cdot)}^{i,-(m,n)}+\alpha_i\right)}\frac{(n_{(\cdot),v}^{k,-(m,n)}+\beta_v)}{\left(\sum_{r=1}^{V}n_{(\cdot),r}^{k,-(m,n)}+\beta_r\right)} \\
\propto & (n_{m,(\cdot)}^{k,-(m,n)}+\alpha_k)\frac{(n_{(\cdot),v}^{k,-(m,n)}+\beta_v)}{\left(\sum_{r=1}^{V}n_{(\cdot),r}^{k,-(m,n)}+\beta_r\right)}.
\end{aligned}
$$

Note that the same formula is derived in the article on the Dirichlet compound multinomial distribution, as part of a more general discussion of integrating Dirichlet distribution priors out of a Bayesian network.

Applications, extensions and similar techniques

Topic modeling is a classic problem in information retrieval. Related models and techniques are, among others, latent semantic indexing, independent component analysis, probabilistic latent semantic indexing, non-negative matrix factorization, and Gamma-Poisson.

The LDA model is highly modular and can therefore be easily extended. The main field of interest is modeling relations between topics. This is achieved by using another distribution on the simplex instead of the Dirichlet. The Correlated Topic Model[5] follows this approach, inducing a correlation structure between topics by using the logistic normal distribution instead of the Dirichlet. Another extension is the hierarchical LDA (hLDA),[6] where topics are joined together in a hierarchy by using the nested Chinese restaurant process.

As noted earlier, PLSA is similar to LDA. The LDA model is essentially the Bayesian version of PLSA model. Bayesian formulation tends to perform better on small datasets because Bayesian methods can avoid overfitting the data. In a very large dataset, the results are probably the same. One difference is that PLSA uses a variable d to represent a document in the training set. So in PLSA, when presented with a document the model hasn't seen before, we fix $\Pr(w \mid z)$--the probability of words under topics—to be that learned from the training set and use the same EM algorithm to infer $\Pr(z \mid d)$--the topic distribution under d. Blei argues that this step is cheating because you are essentially refitting the model to the new data.

Notes

[1] Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). Lafferty, John. ed. "Latent Dirichlet allocation" (http://jmlr.csail.mit. edu/papers/v3/blei03a.html). *Journal of Machine Learning Research* **3** (4–5): *pp.* 993–1022. doi:10.1162/jmlr.2003.3.4-5.993. .

[2] Girolami, Mark; Kaban, A. (2003). "On an Equivalence between PLSI and LDA" (http://www.cs.bham.ac.uk/~axk/sigir2003_mgak. pdf). Proceedings of SIGIR 2003. New York: Association for Computing Machinery. ISBN 1-58113-646-3. .

[3] Griffiths, Thomas L.; Steyvers, Mark (April 6 2004). "Finding scientific topics". *Proceedings of the National Academy of Sciences* **101** (Suppl. 1): 5228–5235. doi:10.1073/pnas.0307752101. PMC 387300. PMID 14872004.

[4] Minka, Thomas; Lafferty, John (2002). "Expectation-propagation for the generative aspect model" (https://research.microsoft.com/~minka/papers/aspect/minka-aspect.pdf). Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence. San Francisco, CA: Morgan Kaufmann. ISBN 1-55860-897-4. .

[5] Blei, David M.; Lafferty, John D. (2006). "Correlated topic models" (http://www.cs.cmu.edu/~lafferty/pub/ctm.pdf). *Advances in Neural Information Processing Systems* **18**. .

[6] Blei, David M.; Jordan, Michael I.; Griffiths, Thomas L.; Tenenbaum; Joshua B (2004). "Hierarchical Topic Models and the Nested [[Chinese restaurant process|Chinese Restaurant Process (http://cocosci.berkeley.edu/tom/papers/ncrp.pdf)]"]. Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference. MIT Press. ISBN 0-262-20152-6. .

External links

- D. Mimno's LDA Bibliography (http://www.cs.princeton.edu/~mimno/topics.html) An exhaustive list of LDA-related resources (incl. papers and some implementations)
- Gensim (http://radimrehurek.com/gensim) Python+NumPy implementation of LDA for input larger than the available RAM.
- topicmodels (http://cran.r-project.org/web/packages/topicmodels/index.html) and lda (http://cran.r-project.org/web/packages/lda/index.html) are two R packages for LDA analysis.
- LDA and Topic Modelling Video Lecture by David Blei (http://videolectures.net/mlss09uk_blei_tm/)
- "Text Mining with R" including LDA methods (http://www.r-bloggers.com/RUG/2010/10/285/), video of Rob Zinkov's presentation to the October 2011 meeting of the Los Angeles R users group
- MALLET (http://mallet.cs.umass.edu/index.php) Open source Java-based package from the University of Massachusetts-Amherst for topic modeling with LDA, also has an independently developed GUI, the Topic Modeling Tool (http://code.google.com/p/topic-modeling-tool/)
- LDA in Mahout (https://cwiki.apache.org/confluence/display/MAHOUT/Latent+Dirichlet+Allocation) implementation of LDA using MapReduce on the Hadoop platform

- Perl implementation of LDA (http://www.people.fas.harvard.edu/~ptoulis/code/LDA_Perl.zip) A non-optimized implementation of the LDA model in Perl, with documentation
- The LDA Buffet is Now Open; or, Latent Dirichlet Allocation for English Majors (https://www.stanford.edu/~mjockers/cgi-bin/drupal/node/61) A non-technical introduction to LDA by Matthew Jocker

Vanessa Fox

Vanessa Fox

Website
http://ninebyblue.com

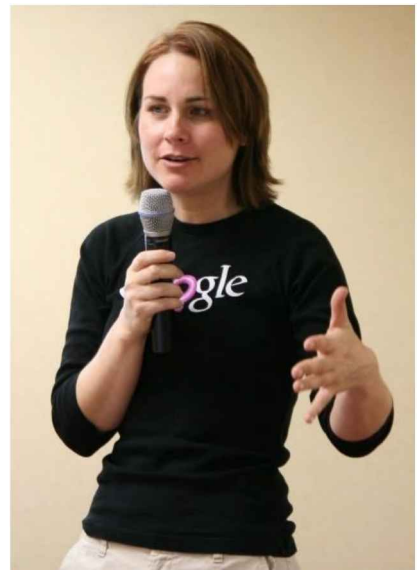
Vanessa Fox (born 1972) is a search engine optimization expert, writer and consultant best known for her work creating Google Webmaster Central and as a Google spokesperson. Google Webmaster Central is a set of tools, educational materials, and community to advise webmasters on how to have their sites listed in Google, and to help solve problems they might have with the way Google indexes their pages.[1] [2] She is a prominent technology blogger,[3] and a frequent speaker at industry events.[4][5][6]

Career

Fox joined Google in 2005 as a technical writer in its Kirkland, Washington office. She left Google in June 2007, and briefly worked for real-estate startup Zillow. She is an Entrepreneur In Residence for Ignition Partners, a Seattle, Washington-based venture capital firm. Additionally, she is the founder of Nine By Blue, a marketing consultancy with an emphasis on search.

Fox, originally from Southern California, worked at a Seattle web startup and AOL before joining Google in 2005. While at Google, she was based at Google's Kirkland, Washington office.[7][8]

She is an adviser to Thinglabs and she is on the University of Washington's MSIM/Informatics advisory board.[9]


Vanessa Fox 2007

Events

Vanessa Fox is a frequent speaker at conferences worldwide, including Search Engine Strategies, Search Marketing Expo, Web 2.0 Conference, BlogHer and Ignite Seattle.

Books and Writing

Fox's book, Marketing in the Age of Google,[10] was published in May 2010 by Wiley, provides a blueprint for incorporating search strategy into organizations of all levels. She writes regularly for a number of offline and online publications, such as O'Reillly Radar[11] and Search Engine Land.[12] She also regularly writes about holistic marketing strategies that integrate searcher behavior at Nine By Blue[13] and search-friendly best practices for developers at Jane andRobot.[14]

References

[1] DannySullivan,8GooglerAlternativesToSuperstarMattCutts(http://searchengineland.com/061201-084842.php),SearchEngineLand, Dec. 1, 2006

[2] "Google Webmaster Central" (http://www.google.com/webmasters/). www.google.com. . Retrieved 2008-08-31.

[3] Kessler,Michelle(October5,2007)."Fridaysgofromcasualtoe-mail-free-USATODAY.com"(http://www.usatoday.com/money/ workplace/2007-10-04-no-email_N.htm). usatoday.com. . Retrieved 2007-11-15.

[4] Vanessa Fox Speaker Bio (http://www.pubcon.com/bios/vanessa_fox.htm), PubCon, accessed April 8, 2007

[5] Profile, Vanessa Fox (http://www.searchenginestrategies.com/sew/ny07/vfox.html), Search Engine Strategies, accessed April 8, 2007

[6] Blog (http://www.seomoz.org/blog/vanessa-fox-the-person-to-whom-webmasters-owe-the-greatest-debt-of-gratitude), SEOmoz.org, accessed June 14, 2007

[7] Seattle's Sisters of Search (http://www.seattle24x7.com/people/searchsisters.htm), Seattle24x7, accessed April 13, 2007

[8] A Conversation with Google's Webmaster Central (http://www.seattle24x7.com/up/googlecentral.htm), Seattle24x7, accessed April 13, 2007

[9] UW MSIM Board (http://ischool.uw.edu/msim_board.aspx)

[10] Marketing in the Age of Google (http://marketingintheageofgoogle.com), accessed June 25, 2010

[11] O'Reilly Radar profile (http://radar.oreilly.com/vanessa/), accessed June 25, 2010

[12] Search Engine Land profile (http://searchengineland.com/author/vanessa-fox/), accessed June 25, 2010

[13] Nine By Blue (http://ninebyblue.com/blog), accessed June 25, 2010

[14] Jane and Robot (http://janeandrobot.com), accessed June 25, 2010

External links

- Vanessa Fox personal blog (http://www.ninebyblue.com/blog/)
- JaneandRobot.com (http://www.janeandrobot.com)
- Author page (http://searchengineland.com/author/vanessa-fox) at Search Engine Land
- Google Webmaster Central blog (http://googlewebmastercentral.blogspot.com/)
- Vanessa Fox Clarifies Sitemaps, PageRank (http://videos.webpronews.com/2006/12/06/ vanessa-fox-clarifies-the-role-of-google-sitemaps/) (video interview), Web Pro News, April 12, 2007.

Search engines

A **web search engine** is designed to search for information on the World Wide Web and FTP servers. The search results are generally presented in a list of results often referred to as SERPS, or "search engine results pages". The information may consist of web pages, images, information and other types of files. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain real-time information by running an algorithm on a web crawler.

History

Timeline (full list)		
Year	**Engine**	**Current status**
1993	W3Catalog	Inactive
	AIiweb	Inactive
1994	WebCrawIer	Active, Aggregator
	Go.com	Active, Yahoo Search
	Lycos	Active
1995	AItaVista	Inactive(URLredirectedtoYahoo!)
	Daum	Active
	Magellan	Inactive
	Excite	Active
	SAPO	Active
	Yahoo!	Active, Launched as a directory
1996	Dogpile	Active, Aggregator
	Inktomi	Acquired by Yahoo!
	HotBot	Active (lycos.com)
	Ask Jeeves	Active(ask.com,Jeeveswentaway)
1997	Northern Light	Inactive
	Yandex	Active
1998	GoogIe	Active
	MSN Search	Active as Bing
1999	AlltheWeb	Inactive(URLredirectedtoYahoo!)
	GenieKnows	Active, rebranded Yellowee.com
	Naver	Active
	Teoma	Active
	Vivisimo	Inactive
2000	Baidu	Active
	Exalead	Acquired by Dassault Systèmes
2002	Inktomi	Acquired by Yahoo!
2003	Info.com	Active

2004	Yahoo! Search	Active, Launched own web search (see Yahoo! Directory, 1995)
	A9.com	Inactive
	Sogou	Active
2005	AOL Search	Active
	Ask.com	Active
	GoodSearch	Active
	SearchMe	Closed
2006	wikiseek	Inactive
	Quaero	Active
	Ask.com	Active
	Live Search	Active as Bing, Launched as rebranded MSN Search
	ChaCha	Active
	Guruji.com	Active
2007	wikiseek	Inactive
	Sproose	Inactive
	Wikia Search	Inactive
	Blackle.com	Active
2008	Powerset	Inactive (redirects to Bing)
	Picollator	Inactive
	Viewzi	Inactive
	Boogami	Inactive
	LeapFish	Inactive
	Forestle	Inactive (redirects to Ecosia)
	VADLO	Active
	Duck Duck Go	Active, Aggregator
2009	Bing	Active, Launched as rebranded Live Search
	Yebol	Active
	Megafore	Active
	Mugurdy	Inactive due to a lack of funding
	Goby	Active
2010	Black Google Mobile	Active
	Blekko	Active
	Cuil	Inactive
	Yandex	Active, Launched global (English) search
	Yummly	Active
2011	Interred	Active
2012	Volunia	Active , only *Power User*

During the early development of the web, there was a list of webservers edited by Tim Berners-Lee and hosted on the CERN webserver. One historical snapshot from 1992 remains.[1] As more webservers went online the central list could not keep up. On the NCSA site new servers were announced under the title "What's New!"[2]

The very first tool used for searching on the Internet was Archie.[3] The name stands for "archive" without the "v". It was created in 1990 by Alan Emtage, Bill Heelan and J. Peter Deutsch, computer science students at McGill University in Montreal. The program downloaded the directory listings of all the files located on public anonymous FTP (File Transfer Protocol) sites, creating a searchable database of file names; however, Archie did not index the contents of these sites since the amount of data was so limited it could be readily searched manually.

The rise of Gopher (created in 1991 by Mark McCahill at the University of Minnesota) led to two new search programs, Veronica and Jughead. Like Archie, they searched the file names and titles stored in Gopher index systems. Veronica (*V*ery *E*asy *R*odent-*O*riented *N*et-wide *I*ndex to *C*omputerized *A*rchives) provided a keyword search of most Gopher menu titles in the entire Gopher listings. Jughead (*J*onzy's *U*niversal *G*opher *H*ierarchy *E*xcavation *A*nd *D*isplay) was a tool for obtaining menu information from specific Gopher servers. While the name of the search engine "Archie" was not a reference to the Archie comic book series, "Veronica" and "Jughead" are characters in the series, thus referencing their predecessor.

In the summer of 1993, no search engine existed yet for the web, though numerous specialized catalogues were maintained by hand. Oscar Nierstrasz at the University of Geneva wrote a series of Perl scripts that would periodically mirror these pages and rewrite them into a standard format which formed the basis for W3Catalog, the web's first primitive search engine, released on September 2, 1993.[4]

In June 1993, Matthew Gray, then at MIT, produced what was probably the first web robot, the Perl-based World Wide Web Wanderer, and used it to generate an index called 'Wandex'. The purpose of the Wanderer was to measure the size of the World Wide Web, which it did until late 1995. The web's second search engine Aliweb appeared in November 1993. Aliweb did not use a web robot, but instead depended on being notified by website administrators of the existence at each site of an index file in a particular format.

JumpStation (released in December 1993[5]) used a web robot to find web pages and to build its index, and used a web form as the interface to its query program. It was thus the first WWW resource-discovery tool to combine the three essential features of a web search engine (crawling, indexing, and searching) as described below. Because of the limited resources available on the platform on which it ran, its indexing and hence searching were limited to the titles and headings found in the web pages the crawler encountered.

One of the first "full text" crawler-based search engines was WebCrawler, which came out in 1994. Unlike its predecessors, it let users search for any word in any webpage, which has become the standard for all major search engines since. It was also the first one to be widely known by the public. Also in 1994, Lycos (which started at Carnegie Mellon University) was launched and became a major commercial endeavor.

Soon after, many search engines appeared and vied for popularity. These included Magellan, Excite, Infoseek, Inktomi, Northern Light, and AltaVista. Yahoo! was among the most popular ways for people to find web pages of interest, but its search function operated on its web directory, rather than full-text copies of web pages. Information seekers could also browse the directory instead of doing a keyword-based search.

In 1996, Netscape was looking to give a single search engine an exclusive deal to be the featured search engine on Netscape's web browser. There was so much interest that instead a deal was struck with Netscape by five of the major search engines, where for $5 million per year each search engine would be in rotation on the Netscape search engine page. The five engines were Yahoo!, Magellan, Lycos, Infoseek, and Excite.[6][7]

Search engines were also known as some of the brightest stars in the Internet investing frenzy that occurred in the late 1990s.[8] Several companies entered the market spectacularly, receiving record gains during their initial public offerings. Some have taken down their public search engine, and are marketing enterprise-only editions, such as Northern Light. Many search engine companies were caught up in the dot-com bubble, a speculation-driven market

boom that peaked in 1999 and ended in 2001.

Around 2000, Google's search engine rose to prominence. The company achieved better results for many searches with an innovation called PageRank. This iterative algorithm ranks web pages based on the number and PageRank of other web sites and pages that link there, on the premise that good or desirable pages are linked to more than others. Google also maintained a minimalist interface to its search engine. In contrast, many of its competitors embedded a search engine in a web portal.

By 2000, Yahoo! was providing search services based on Inktomi's search engine. Yahoo! acquired Inktomi in 2002, and Overture (which owned AlltheWeb and AltaVista) in 2003. Yahoo! switched to Google's search engine until 2004, when it launched its own search engine based on the combined technologies of its acquisitions.

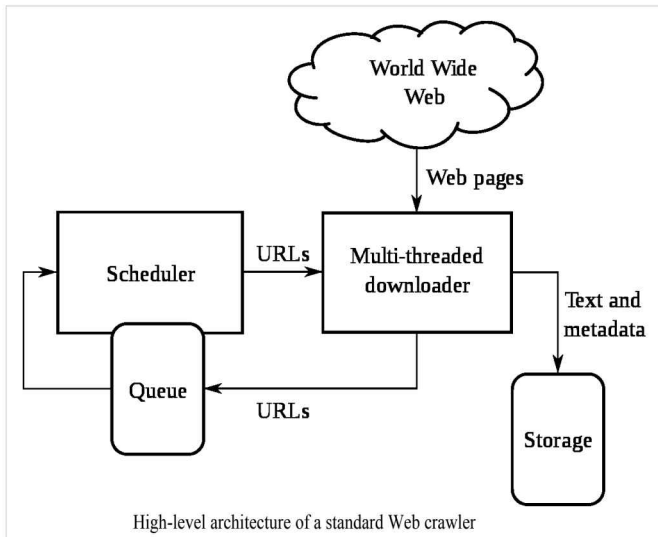
Microsoft first launched MSN Search in the fall of 1998 using search results from Inktomi. In early 1999 the site began to display listings from Looksmart blended with results from Inktomi except for a short time in 1999 when results from AltaVista were used instead. In 2004, Microsoft began a transition to its own search technology, powered by its own web crawler (called msnbot).

Microsoft's rebranded search engine, Bing, was launched on June 1, 2009. On July 29, 2009, Yahoo! and Microsoft finalized a deal in which Yahoo! Search would be powered by Microsoft Bing technology.

How web search engines work

A search engine operates in the following order:

1. Web crawling
2. Indexing
3. Searching



High-level architecture of a standard Web crawler

Web search engines work by storing information about many web pages, which they retrieve from the HTML itself. These pages are retrieved by a Web crawler (sometimes also known as aspider) — an automated Web browser which follows every link on the site. Exclusions can be made by the use of robots.txt. The contents of each page are then analyzed to determine how it should be indexed (for example, words are extracted from the titles, headings, or special fields called meta tags). Data about web pages are stored in an index database for use in later queries. A query can be a single word. The purpose of an index is to allow information to be found as quickly as possible. Some search engines, such as Google, store all or part of the source page (referred to as a cache) as well as information about the web pages, whereas others, such as AltaVista, store every word of every page they find. This cached page always holds the actual search text since it is the one that was actually indexed, so it can be very useful when the content of the current page has been updated and the search terms are no longer in it. This problem might be considered to be a mild form of linkrot, and Google's handling of it increases usability by satisfying user expectations that the search terms will be on the returned webpage. This satisfies the principle of least astonishment since the user normally expects the search terms to be on the returned pages. Increased search relevance makes these cached pages very useful, even beyond the fact that they may contain data that may no longer be available elsewhere.

When a user enters a query into a search engine (typically by using keywords), the engine examines its index and provides a listing of best-matching web pages according to its criteria, usually with a short summary containing the document's title and sometimes parts of the text. The index is built from the information stored with the data and the

method by which the information is indexed. Unfortunately, there are currently no known public search engines that allow documents to be searched by date. Most search engines support the use of the boolean operators AND, OR and NOT to further specify the search query. Boolean operators are for literal searches that allow the user to refine and extend the terms of the search. The engine looks for the words or phrases exactly as entered. Some search engines provide an advanced feature called proximity search which allows users to define the distance between keywords. There is also concept-based searching where the research involves using statistical analysis on pages containing the words or phrases you search for. As well, natural language queries allow the user to type a question in the same form one would ask it to a human. A site like this would be ask.com.

The usefulness of a search engine depends on the relevance of the **result set** it gives back. While there may be millions of web pages that include a particular word or phrase, some pages may be more relevant, popular, or authoritative than others. Most search engines employ methods to rank the results to provide the "best" results first. How a search engine decides which pages are the best matches, and what order the results should be shown in, varies widely from one engine to another. The methods also change over time as Internet usage changes and new techniques evolve. There are two main types of search engine that have evolved: one is a system of predefined and hierarchically ordered keywords that humans have programmed extensively. The other is a system that generates an "inverted index" by analyzing texts it locates. This second form relies much more heavily on the computer itself to do the bulk of the work.

Most Web search engines are commercial ventures supported by advertising revenue and, as a result, some employ the practice of allowing advertisers to pay money to have their listings ranked higher in search results. Those search engines which do not accept money for their search engine results make money by running search related ads alongside the regular search engine results. The search engines make money every time someone clicks on one of these ads.

Market share

Search engine	Market share in May 2011		Market share in December 2010[9]	
Google	82.80%		84.65%	
Yahoo!	6.42%		6.69%	
Baidu	4.89%		3.39%	
Bing	3.91%		3.29%	
Ask	0.52%		0.56%	
AOL	0.36%		0.42%	

Google's worldwide market share peaked at 86.3% in April 2010.[10] Yahoo!, Bing and other search engines are more popular in the US than in Europe.

According to Hitwise, market share in the U.S. for October 2011 was Google 65.38%, Bing-powered (Bing and Yahoo!) 28.62%, and the remaining 66 search engines 6%. However, an Experian Hit wise report released in August 2011 gave the "success rate" of searches sampled in July. Over 80 percent of Yahoo! and Bing searches resulted in the users visiting a web site, while Google's rate was just under 68 percent.[11] [12]

In the People's Republic of China, Baidu held a 61.6% market share for web search in July 2009.[13]

Search engine bias

Although search engines are programmed to rank websites based on their popularity and relevancy, empirical studies indicate various political, economic, and social biases in the information they provide.[14][15] These biases could be a direct result of economic and commercial processes (e.g., companies that advertise with a search engine can become also more popular in its organic search results), and political processes (e.g., the removal of search results in order to comply with local laws).[16] Google Bombing is one example of an attempt to manipulate search results for political, social or commercial reasons.

References

[1] World-Wide Web Servers (http://www.w3.org/History/19921103-hypertext/hypertext/DataSources/WWW/Servers.html)

[2] What's New! February 1994 (http://home.mcom.com/home/whatsnew/whats_new_0294.html)

[3] "Internet History - Search Engines" (from Search Engine Watch), Universiteit Leiden, Netherlands, September 2001, web: LeidenU-Archie (http://www.internethistory.leidenuniv.nl/index.php3?c=7).

[4] Oscar Nierstrasz (2 September 1993). "Searchable Catalog of WWW Resources (experimental)" (http://groups.google.com/group/comp.infosystems.www/browse_thread/thread/2176526a36dc8bd3/2718fd17812937ac?hl=en&lnk=gst&q=Oscar+ Nierstrasz#2718fd17812937ac). .

[5] Archive of NCSA what's new in December 1993 page(http://web.archive.org/web/20010620073530/http://archive.ncsa.uiuc.edu/SDG/Software/Mosaic/Docs/old-whats-new/whats-new-1293.html)

[6] "Yahoo! And Netscape Ink International Distribution Deal" (http://files.shareholder.com/downloads/YHOO/701084386x0x27155/ 9a3b5ed8-9e84-4cba-a1e5-77a3dc606566/YHOO_News_1997_7_8_General.pdf).

[7] *Browser Deals Push Netscape Stock Up 7.8%* (http://articles.latimes.com/1996-04-01/business/fi-53780_1_netscape-home). Los Angeles Times. 1 April1996.

[8] Gandal, Neil (2001). "The dynamics of competition in the internet search engine market". *International Journal of Industrial Organization* **19** (7): 1103–1117. doi:10.1016/S0167-7187(01)00065-0.

[9] Net Marketshare - World (http://marketshare.hitslink.com/search-engine-market-share.aspx?qprid=4)

[10] Net Market share - Google (http://marketshare.hitslink.com/report.aspx?qprid=5&qpcustom=Google - Global&qptimeframe=M& qpsp=120&qpnp=25)

[11] "Google Remains Ahead of Bing, But Relevance Drops" (http://news.yahoo.com/ google-remains-ahead-bing-relevance-drops-210457139.html). August 12, 2011..

[12] Experian Hitwise reports Bing-powered share of searches at 29 percent in October 2011 (http://www.hitwise.com/us/about-us/ press-center/press-releases/bing-powered-share-of-searches-at-29-percent), Experian Hitwise, November 16, 2011

[13] Search Engine Market Share July 2009 | Rise to the Top Blog (http://risetothetop.techwyse.com/internet-marketing/ search-engine-market-share-july-2009/)

[14] Segev, Elad (2010). Google and the Digital Divide: The Biases of Online Knowledge, Oxford: Chandos Publishing.

[15] Vaughan, L. & Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes, Information Processing & Management, 40(4), 693-707.

[16] Berkman Center for Internet & Society (2002), "Replacement of Google with Alternative Search Systems in China: Documentation and Screen Shots" (http://cyber.law.harvard.edu/filtering/china/google-replacements/), Harvard Law School.

- GBMW: Reports of 30-day punishment, re: Carmaker BMW had its German website bmw.de delisted from Google, such as: Slashdot-BMW (http://slashdot.org/article.pl?sid=06/02/05/235218) (05-Feb-2006).
- INSIZ: Maximum size of webpages indexed by MSN/Google/Yahoo! ("100-kb limit"): Max Page-size (http://www.sitepoint.com/article/indexing-limits-where-bots-stop) (28-Apr-2006).

Further reading

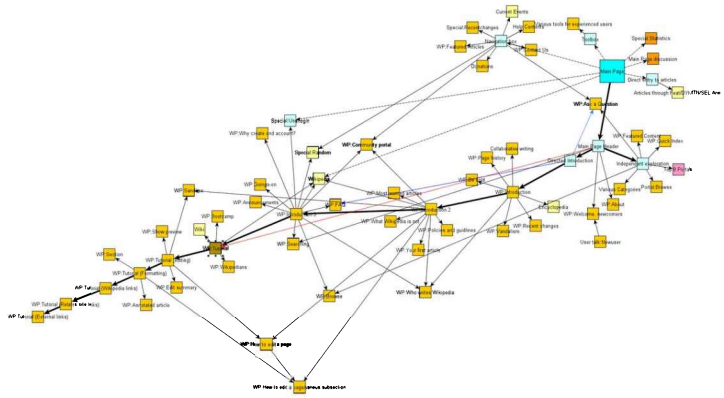
- For a more detailed history of early search engines, see Search Engine Birthdays (http://searchenginewatch.com/showPage.html?page=3071951)(fromSearchEngineWatch),ChrisSherman,September2003.
- Steve Lawrence; C. Lee Giles (1999). "Accessibility of information on the web". *Nature* **400** (6740): 107–9. doi:10.1038/21987. PMID 10428673.
- Bing Liu (2007), *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data* (http://www.cs.uic.edu/~liub/WebMiningBook.html). Springer, ISBN 3540378812
- Bar-Ilan, J. (2004). The use of Web search engines in information science research. ARIST, 38, 231-288.
- Levene, Mark (2005). *An Introduction to Search Engines and Web Navigation*. Pearson.
- Hock, Randolph (2007). *The Extreme Searcher's Handbook*. ISBN 978-0-910965-76-7
- Javed Mostafa (February 2005). "Seeking Better Web Searches" (http://www.sciam.com/article.cfm?articleID=0006304A-37F4-11E8-B7F483414B7F0000). *Scientific American Magazine*.
- Ross, Nancy; Wolfram, Dietmar (2000). "End user searching on the Internet: An analysis of term pair topics submitted to the Excite search engine". *Journal of the American Society for Information Science* **51** (10): 949–958. doi:10.1002/1097-4571(2000)51:10<949::AID-ASI70>3.0.CO;2-5.
- Xie, M. et al (1998). "Quality dimensions of Internet search engines". *Journal of Information Science* **24** (5): 365–372. doi:10.1177/016555159802400509.
- *Information Retrieval: Implementing and Evaluating Search Engines* (http://www.ir.uwaterloo.ca/book/). MIT Press. 2010.

External links

- Search Engines (http://www.dmoz.org/Computers/Internet/Searching/Search_Engines//) at the Open Directory Project

Site map

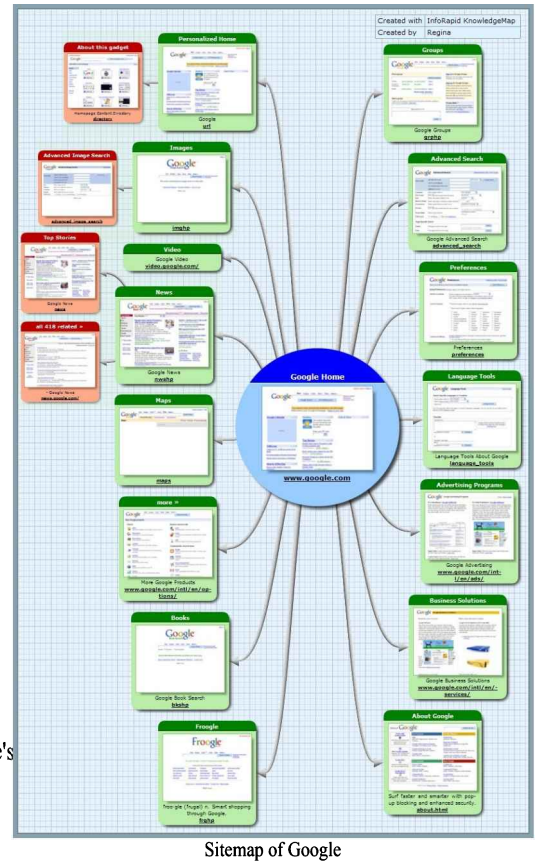
A **site map** (or **sitemap**) is a list of pages of a web site accessible to crawlers or users. It can be either a document in any form used as a planning tool for web design, or a web page that lists the pages on a web site, typically organized in hierarchical fashion. This helps visitors and search engine bots find pages on the site.



A site map of what links from the English Wikipedia's Main Page.

While some developers argue that **site index** is a more appropriately used term to relay page function, web visitors are used to seeing each term and generally associate both as one and the same. However, a site index is often used to mean an A-Z index that provides access to particular content, while a site map provides a general top-down view of the overall site contents.

XML is a document structure and encoding standard used, amongst many other things, as the standard for webcrawlers to find and parse sitemaps. There is an example of an XML sitemap below (missing link to site). The instructions to the sitemap are given to the crawler bot by a Robots Text file, an example of this is also given below. Site maps can improve search engine optimization of a site by making sure that all the pages can be found. This is especially important if a site uses a dynamic access to content such as Adobe Flash or JavaScript menus that do not include HTML links.



Sitemap of Google

They also act as a navigation aid [1] by providing an overview of a site's

Benefits of XML sitemaps to search-optimize Flash sites

Below is an example of a validated XML sitemap for a simple three page web site. Sitemaps are a useful tool for making sites built in Flash and other non-html languages searchable. Note that because the website's navigation is built with Flash (Adobe), the initial homepage of a site developed in this way would probably be found by an automated search program (ref: bot). However, the subsequent pages are unlikely to be found without an XML sitemap.

XML sitemap example:

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>http://www.example.com/?id=who</loc>
    <lastmod>2009-09-22</lastmod>
    <changefreq>monthly</changefreq>
    <priority>0.8</priority>
  </url>
  <url>
    <loc>http://www.example.com/?id=what</loc>
    <lastmod>2009-09-22</lastmod>
    <changefreq>monthly</changefreq>
    <priority>0.5</priority>
  </url>
  <url>
    <loc>http://www.example.com/?id=how</loc>
    <lastmod>2009-09-22</lastmod>
    <changefreq>monthly</changefreq>
    <priority>0.5</priority>
  </url>
</urlset>
```

XML Sitemaps

Google introduced Google Sitemaps so web developers can publish lists of links from across their sites. The basic premise is that some sites have a large number of dynamic pages that are only available through the use of forms and user entries. The Sitemap files contains URLs to these pages so that web crawlers can find them[2]. Bing, Google, Yahoo and Ask now jointly support the Sitemaps protocol.

Since Bing, Yahoo, Ask, and Google use the same protocol, having a Sitemap lets the four biggest search engines have the updated page information. Sitemaps do not guarantee all links will be crawled, and being crawled does not guarantee indexing. However, a Sitemap is still the best insurance for getting a search engine to learn about your entire site.[3]

XML Sitemaps have replaced the older method of "submitting to search engines" by filling out a form on the search engine's submission page. Now web developers submit a Sitemap directly, or wait for search engines to find it.

XML (Extensible Markup Language) is much more precise than HTML coding. Errors are not tolerated, and so syntax must be exact. It is advised to use an XML syntax validator such as the free one found at: http://validator. w3.org

There are automated XML site map generators available (both as software and web applications) for more complex sites.

More information defining the field operations and other Sitemap options are defined at http://www.sitemaps.org (Sitemaps.org: Google, Inc., Yahoo, Inc., and Microsoft Corporation)

See also Robots.txt, which can be used to identify sitemaps on the server.

References

[1] Site Map Usability (http://www.useit.com/alertbox/sitemaps.html) Jakob Nielsen's Alertbox, August 12, 2008
[2] "WordPress Plugin: Google XML Sitemaps" (http://linksku.com/10/wordpress-plugins). Linksku. .
[3] Joint announcement (http://www.google.com/press/pressrel/sitemapsorg.html) from Google, Yahoo, Bing supporting Sitemaps

External links

- CommonOfficialWebsite(http://www.sitemaps.org/)-JointlymaintainedwebsitebyGoogle,Yahoo,MSN for an XML sitemap format.
- /Sitemapgenerators(http://www.dmoz.org/Computers/Internet/Searching/Search_Engines/Sitemaps)atthe Open Directory Project
- Tools and tutorial (http://www.scriptol.com/seo/simple-map.html) Helping to build a cross-systemssitemap generator.

Sitemaps

The **Sitemaps** protocol allows a webmaster to inform search engines about URLs on a website that are available for crawling. A Sitemap is an XML file that lists the URLs for a site. It allows webmasters to include additional information about each URL: when it was last updated, how often it changes, and how important it is in relation to other URLs in the site. This allows search engines to crawl the site more intelligently. SitemapsareaURLinclusion protocol and complement robots.txt, a URL exclusion protocol.

Sitemaps are particularly beneficial on websites where:

- some areas of the website are not available through the browsable interface, or
- webmasters use rich Ajax, Silverlight, or Flash content that is not normally processed by search engines.

The webmaster can generate a Sitemap containing all accessible URLs on the site and submit it to search engines. Since Google, Bing, Yahoo, and Ask use the same protocol now, having a Sitemap would let the biggest search engines have the updated pages information.

Sitemaps supplement and do not replace the existing crawl-based mechanisms that search engines already use to discover URLs. Using this protocol does not guarantee that web pages will be included in search indexes, nor does it influence the way that pages are ranked in search results.

History

Google first introduced Sitemaps 0.84 [1] in June 2005 so web developers could publish lists of links from across their sites. Google, MSN and Yahoo announced joint support for the Sitemaps protocol [2] in November 2006. The schema version was changed to "Sitemap 0.90", but no other changes were made.

In April 2007, Ask.com and IBM announced support [3] for Sitemaps. Also, Google, Yahoo, MS announced auto-discovery for sitemaps through robots.txt. In May 2007, the state governments of Arizona, California, Utah and Virginia [4] announced they would use Sitemaps on their web sites.

The Sitemaps protocol is based on ideas[5] from "Crawler-friendly Web Servers".[6]

File format

The Sitemap Protocol format consists of XML tags. The file itself must be UTF-8 encoded. Sitemaps can also be just a plain text list of URLs. They can also be compressed in .gz format.

A sample Sitemap that contains just one URL and uses all optional tags is shown below.

```
<?xml version="1.0" encoding="utf-8"?>

<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
    <url>
        <loc>http://example.com/</loc>
        <lastmod>2006-11-18</lastmod>
        <changefreq>daily</changefreq>
        <priority>0.8</priority>
    </url>
</urlset>
```

Element definitions

The definitions for the elements are shown below[7]:

Element	Required?	Description
<urlset>	Yes	The document-level element for the Sitemap. The rest of the document after the '<?xml version>' element must be contained in this.
<url>	Yes	Parent element for each entry. The remaining elements are children of this.
<loc>	Yes	Provides the full URL of the page, including the protocol (e.g. http, https) and a trailing slash, if required by the site's hosting server. This value must be less than 2,048 characters.
<lastmod>	No	The date that the file was last modified, in ISO 8601 format. This can display the full date and time or, if desired, may simply be the date in the format YYYY-MM-DD.
<changefreq>	No	How frequently the page may change: • always • hourly • daily • weekly • monthly • yearly • never 'Always' is used to denote documents that change each time that they are accessed. 'Never' is used to denote archived URLs (i.e. files that will not be changed again). This is used only as a guide for crawlers, and is not used to determine how frequently pages are indexed.
<priority>	No	The priority of that URL relative to other URLs on the site. This allows webmasters to suggest to crawlers which pages are considered more important. The valid range is from 0.0 to 1.0, with 1.0 being the most important. The default value is 0.5. Rating all pages on a site with a high priority does not affect search listings, as it is only used to suggest to the crawlers how important pages in the site are to one another.

Support for the elements that are not required can vary from one search engine to another.[7]

Sitemap index

The Sitemap XML protocol is also extended to provide a way of listing multiple Sitemaps in a 'Sitemap index' file. The maximum Sitemap size of 10 MB or 50,000 URLs means this is necessary for large sites. As the Sitemap needs to be in the same directory as the URLs listed, Sitemap indexes are also useful for websites with multiple subdomains, allowing the Sitemaps of each subdomain to be indexed using the Sitemap index file and robots.txt.

Other formats

Text file

The Sitemaps protocol allows the Sitemap to be a simple list of URLs in a text file. The file specifications of XML Sitemaps apply to text Sitemaps as well; the file must be UTF-8 encoded, and cannot be more than 10 MB large or contain more than 50,000 URLs, but can be compressed as a gzip file.[7]

Syndication feed

A syndication feed is a permitted method of submitting URLs to crawlers; this is advised mainly for sites that already have syndication feeds. One stated drawback is this method might only provide crawlers with more recently created URLs, but other URLs can still be discovered during normal crawling.[7]

Search engine submission

If Sitemaps are submitted directly to a search engine (pinged), it will return status information and any processing errors. The details involved with submission will vary with the different search engines. The location of the sitemap can also be included in the `robots.txt` file by adding the following line to `robots.txt`:

```
Sitemap: <sitemap_location>
```

The <sitemap_location> should be the complete URL to the sitemap, such as: *http://www.example.org/sitemap.xml* (however, see the discussion). This directive is independent of the user-agent line, so it doesn't matter where it is placed in the file. If the website has several sitemaps, this URL can simply point to the main sitemap index file.

The following table lists the sitemap submission URLs for several major search engines:

Search engine	Submission URL	Help page	
Google	**http://www.google.com/webmasters/tools/ping?sitemap=**	Submitting a Sitemap (http:// www.google.com/support/webmasters/bin/answer.py?hl=en&answer=34575	)

Yahoo!	—	Site Explorer has moved to Bing Webmaster Tools (http://developer.yahoo.com/search/siteexplorer/V1/ping.html)	
Ask.com	**http://submissions.ask.com/ping?sitemap=**	Q: Does Ask.com support sitemaps? (http://about.ask.com/en/docs/about/webmasters.shtml#22)	
Bing (Live Search)	**http://www.bing.com/webmaster/ping.aspx?siteMap=**	Bing Webmaster Tools (http://www.bing.com/webmaster)	
Yandex	—	Sitemaps files (http://help.yandex.com/webmaster/?id=1115259)	} Sitemap URLs submitted using the sitemap submission URLs need to be URL-encoded, replacing : with `%3A`, / with `%2F`, etc.[7] **Sitemap limits** Sitemap files have a limit of 50,000 URLs and 10 megabytes per sitemap. Sitemaps can be compressed using gzip, reducing bandwidth consumption. Multiple sitemap files are supported, with a Sitemap index file serving as an entry point. Sitemap index files may not list more than 50,000 Sitemaps and must be no larger than 10MiB (10,485,760 bytes) and can be compressed. You can have more than one Sitemap index file.[7] As with all XML files, any data values (including URLs) must use entity escape codes for the characters ampersand (&), single quote ('), double quote ("), less than (<), and greater than (>). **References**

External links

- Official page (http://www.sitemaps.org) (set up by Google, Yahoo & MSN)
- Google, Yahoo, MSN joint announcement in Nov '06 (http://www.google.com/press/pressrel/ sitemapsorg.html)
- Official Blog (http://sitemaps.blogspot.com/)
- Google Sitemaps newsgroup (archived) (http://groups.google.com/group/google-sitemaps)
- Google Sitemaps newsgroup (http://groups.google.com/group/Google_Webmaster_Help-Sitemap)
- Sitemap Gen (http://goog-sitemapgen.sourceforge. net/) Python script to generate Sitemaps by Google
- sitemap_gen.py (http://www.bashkirtsev.com/2009/05/14/sitemap/) Python script to generate Sitemaps by Google with MemoryError fixed
- Search::Sitemap (http://search.cpan.org/~jasonk/ Search-Sitemap/) Perl Library for manipulating Sitemaps
- PHP Sitemap Class (http://www.php-ease.com/classes/sitemap.html) A PHP Class for creating sitemaps
- XmlSitemapGenerator.org (http://www.XmlSitemapGenerator.org) A free online tool for creating sitemaps

Methods of website linking

This article pertains to methods of hyperlinking to/of different websites, often used in regard to search engine optimization (SEO). Many techniques and special terminology about linking are described below.

Reciprocal link

A **reciprocal link** is a mutual link between two objects, commonly between two websites to ensure mutual traffic. For example, Alice and Bob have websites. If Bob's website links to Alice's website, and Alice's website links to Bob's website, the websites are reciprocally linked. Website owners often submit their sites to reciprocal link exchange directories in order to achieve higher rankings in the search engines. Reciprocal linking between websites is an important part of the search engine optimization process because Google uses link popularity algorithms (defined as the number of links that lead to a particular page and the anchor text of the link) to rank websites for relevancy.

Resource linking

Resource links are a category of links, which can be either one-way or two-way, usually referenced as "Resources" or "Information" in navbars, but sometimes, especially in the early, less compartmentalized years of the Web, simply called "links". Basically, they are hyperlinks to a website or a specific webpage containing content believed to be beneficial, useful and relevant to visitors of the site establishing the link.

In recent years, resource links have grown in importance because most major search engines have made it plain that—in Google's words--"quantity, quality, and relevance of links count towards your rating."[1]

The engines' insistence on resource links being relevant and beneficial developed because many artificial link building methods were employed solely to "spam" search-engines, i.e. to "fool" the engines' algorithms into awarding the sites employing these unethical devices undeservedly high page ranks and/or return positions.

Despite cautioning site developers (again quoting from Google) to avoid "'free-for-all' links, link popularity schemes, or submitting your site to thousands of search engines (because) these are typically useless exercises that don't affect your ranking in the results of the major search engines[2] -- at least, not in a way you would likely consider to be positive,"[3] most major engines have deployed technology designed to "red flag" and potentially penalize sites employing such practices.

Forum signature linking

Forum signature linking is a technique used to build backlinks to a website. This is the process of using forum communities that allow outbound hyperlinks in a member's signature. This can be a fast method to build up inbound links to a website; it can also produce some targeted traffic if the website is relevant to the forum topic. It should be stated that forums using the nofollow attribute will have no actual Search Engine Optimization value.

Blog comments

Leaving a comment on a blog can result in a relevant do-follow link to the individual's website. Most of the time, however, leaving a comment on a blog turns into a no-follow link, which is almost useless in the eyes of search engines, such as Google and Yahoo! Search. On the other hand, most blog comments get clicked on by the readers of the blog if the comment is well-thought-out and pertains to the discussion of the other commenters and the post on the blog.

Directory link building

Website directories are lists of links to websites, which are sorted into categories. Website owners can submit their site to many of these directories. Some directories accept payment for listing in their directory, while others are free.[4]

References

[1] "Link schemes" (http://www.google.com/support/webmasters/bin/answer.py?answer=66356) Google webmaster central

[2] "Study on differences in link value" (http://wiep.net/link-value-factors/)

[3] "Search Engine Optimization (SEO)" (http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=35291) Google webmaster central

[4] Wendy Boswell. "What is a Web Directory" (http://websearch.about.com/od/enginesanddirectories/a/subdirectory.htm). About.com. . Retrieved 2011-04-27.

External links

- All Wikipedia Links Are Now NOFOLLOW (http://www.searchenginejournal.com/all-wikipedia-links-are-now-nofollow/4288/) *Search Engine Journal*, January 21, 2007.

Deep linking

On the World Wide Web, **deep linking** is making a hyperlink that points to a specific page or image on a website, instead of that website's main or home page. Such links are called *deep links*.

Example

This link: http://en.wikipedia.org/wiki/Deep_linking is an example of a deep link. The URL contains all the information needed to point to a particular item, in this case the English Wikipedia article on deep linking, instead of the Wikipedia home page at http://www.wikipedia.org/ [1].

Deep linking and HTTP

The technology behind the World Wide Web, the Hypertext Transfer Protocol (HTTP), does not actually make any distinction between "deep" links and any other links—all links are functionally equal. This is intentional; one of the design purposes of the Web is to allow authors to link to any published document on another site. The possibility of so-called "deep" linking is therefore built into the Web technology of HTTP and URLs by default—while a site can attempt to restrict deep links, to do so requires extra effort. According to the World Wide Web Consortium Technical Architecture Group, "any attempt to forbid the practice of deep linking is based on a misunderstanding of the technology, and threatens to undermine the functioning of the Web as a whole". [2]

Usage

Some commercial websites object to other sites making deep links into their content either because it bypasses advertising on their main pages, passes off their content as that of the linker or, like *The Wall Street Journal*, they charge users for permanently valid links.

Sometimes, deep linking has led to legal action such as in the 1997 case of Ticketmaster versus Microsoft, where Microsoft deep-linked to Ticketmaster's site from its Sidewalk service. This case was settled when Microsoft and Ticketmaster arranged a licensing agreement.

Ticketmaster later filed a similar case against Tickets.com, and the judge in this case ruled that such linking was legal as long as it was clear to whom the linked pages belonged.[3] The court also concluded that URLs themselves were not copyrightable, writing: "A URL is simply an address, open to the public, like the street address of a building, which, if known, can enable the user to reach the building. There is nothing sufficiently original to make the URL a copyrightable item, especially the way it is used. There appear to be no cases holding the URLs to be subject to copyright. On principle, they should not be."

Deep linking and web technologies

Websites which are built on web technologies such as Adobe Flash and AJAX often do not support deep linking. This can result in usability problems for people visiting such websites. For example, visitors to these websites may be unable to save bookmarks to individual pages or states of the site, web browser forward and back buttons may not work as expected, and use of the browser's refresh button may return the user to the initial page.

However, this is not a fundamental limitation of these technologies. Well-known techniques, and libraries such as SWFAddress[4] and History Keeper, now exist that website creators using Flash or AJAX can use to provide deep linking to pages within their sites.[5][6][7][8]

Court rulings

Probably the earliest legal case arising out of deep-linking was the 1996 Scottish case of *The Shetland Times vs The Shetland News* where the *Times* accused the *News* of appropriating stories on the *Times*' website as its own.[9]

In the beginning of 2006 in a case between the search engine Bixee.com and job site Naukri.com, the Delhi High Court in India prohibited Bixee.com from deeplinking to Naukri.com.[10]

In December 2006, a Texas court ruled that linking by a motocross website to videos on a Texas-based motocross video production website did not constitute fair use. The court subsequently issued an injunction.[11] This case, SFX Motor Sports Inc., v. Davis, was not published in official reports, but is available at 2006 WL 3616983.

In a February 2006 ruling, the Danish Maritime and Commercial Court (Copenhagen) found systematic crawling, indexing and deep-linking by portal site ofir.dk of real estate site Home.dk not to conflict with Danish law or the database directive of the European Union. The Court even stated that search engines are desirable for the functioning of the Internet of today. And that one, when publishing information on the Internet, must assume—and accept—that search engines deep link to individual pages of one's website.[12]

Opt out

Web site owners wishing to prevent search engines from deep linking are able to use the existing Robots Exclusion Standard (`/robots.txt` file) to specify their desire or otherwise for their content to be indexed. Some feel that content owners who fail to provide a `/robots.txt` file are implying that they do not object to deep linking either by search engines or others who might link to their content. Others believe that content owners may be unaware of the Robots Exclusion Standard or may not use robots.txt for other reasons. Deep linking is also practiced outside the search engine context, so some participating in this debate question the relevance of the Robots Exclusion Standard to controversies about Deep Linking. The Robots Exclusion Standard does not programmatically enforce its directives so it does not prevent search engines and others who do not follow polite conventions from deep linking.

References

[1] http://www.wikipedia.org/
[2] Bray, Tim (Sept. 11, 2003). ""DeepLinking" in the World Wide Web" (http://www.w3.org/2001/tag/doc/deeplinking.html). W3.. Retrieved May 30, 2007.
[3] Finley, Michelle (Mar. 30, 2000). "Attention Editors: Deep Link Away". Wired News.
[4] SWFAddress Deep Linking Dynamic Paging Tutorial (http://squibl.com/blog/2010/10/14/swfaddress-and-deep-linking/), SQUIBL
[5] "Deep-linking to frames in Flash websites" (http://www.adobe.com/devnet/flash/articles/deep_linking.html). .
[6] "Deep Linking for Flash and Ajax" (http://www.asual.com/swfaddress/). .
[7] "History Keeper – Deep Linking in Flash & JavaScript" (http://www.unfocus.com/projects/historykeeper/). .
[8] "Deep Linking for AJAX" (http://blog.onthewings.net/2009/04/08/deep-linking-for-ajax/). .
[9] For a more extended discussion, see generally the Wikipedia article Copyright aspects of hyperlinking and framing.
[10] "High Court Critical On Deeplinking" (http://www.efytimes.com/efytimes/fullnews.asp?edid=9018&magid=). EFYtimes.com. Dec. 29, 2005. . Retrieved May 30, 2007.

[11] Declan McCullagh. "Judge: Can't link to Webcast if copyright owner objects" (http://news.com.com/2100-1030_3-6145744.html). News.com. . Retrieved May 30, 2007.

[12] "Udskrift af SØ- & Handelsrettens Dombog" (http://www.bvhd.dk/uploads/tx_mocarticles/S_-_og_Handelsrettens_afg_relse_i_Ofir-sagen.pdf) (in Danish). bvhd.dk. Feb. 24, 2006. . Retrieved May 30, 2007.

External links

- Linking Law (http://www.netlitigation.com/netlitigation/linking.htm) - Netlitigation's summary and case law archive.
- American Library Association (http://www.ala.org/ala/aboutala/offices/oif/ifissues/deeplinking.cfm) - list of (mostly deep) links to articles about deep linking
- Discussion of the *Shetland Times vs Shetland News* case, 1996 (http://www.ariadne.ac.uk/issue6/copyright/)
- Report on the Indian Court Ruling (http://yro.slashdot.org/yro/06/01/09/1146224.shtml?tid=95&tid=17)
- Report on Danish Court Ruling (http://newmediatrends.fdim.dk/2006/02/ danish-court-approves-of-deep-linking.html)
- Cory Doctorow on fan-made radio podcasts: "What deep linking means." (http://www.boingboing.net/2006/06/22/cory_on_fanmade_radi.html) from BoingBoing
- Deep Linking is Good Linking (http://www.useit.com/alertbox/20020303.html) - Usability implications of deep links

Backlink

Backlinks, also known as **incoming links**, **inbound links**, **inlinks**, and **inward links**, are incoming links to a website or web page. In basic link terminology, a **backlink** is any link received by a web node (web page, directory, website, or top level domain) from another web node.[1]

Inbound links were originally important (prior to the emergence of search engines) as a primary means of web navigation; today, their significance lies in search engine optimization (SEO). The number of backlinks is one indication of the popularity or importance of that website or page (for example, this is used by Google to determine the PageRank of a webpage). Outside of SEO, the backlinks of a webpage may be of significant personal, cultural or semantic interest: they indicate who is paying attention to that page.

Search engine rankings

Search engines often use the number of backlinks that a website has as one of the most important factors for determining that website's search engine ranking, popularity and importance. Google's description of their PageRank system, for instance, notes that *Google interprets a link from page A to page B as a vote, by page A, for page B.*[2] Knowledge of this form of search engine rankings has fueled a portion of the SEO industry commonly termed linkspam, where a company attempts to place as many inbound links as possible to their site regardless of the context of the originating site.

Websites often employ various search engine optimization techniques to increase the number of backlinks pointing to their website. Some methods are free for use by everyone whereas some methods like linkbaiting requires quite a bit of planning and marketing to work. Some websites stumble upon "linkbaiting" naturally; the sites that are the first with a tidbit of 'breaking news' about a celebrity are good examples of that. When "linkbait" happens, many websites will link to the 'baiting' website because there is information there that is of extreme interest to a large number of people.

There are several factors that determine the value of a backlink. Backlinks from authoritative sites on a given topic are highly valuable.[3] If both sites have content geared toward the keyword topic, the backlink is considered relevant

and believed to have strong influence on the search engine rankings of the webpage granted the backlink. A backlink represents a favorable 'editorial vote' for the receiving webpage from another granting webpage. Another important factor is the anchor text of the backlink. Anchor text is the descriptive labeling of the hyperlink as it appears on a webpage. Search engine bots (i.e., spiders, crawlers, etc.) examine the anchor text to evaluate how relevant it is to the content on a webpage. Anchor text and webpage content congruency are highly weighted in search engine results page (SERP) rankings of a webpage with respect to any given keyword query by a search engine user.

Increasingly, inbound links are being weighed against link popularity and originating context. This transition is reducing the notion of *one link, one vote* in SEO, a trend proponents hope will help curb link spam as a whole.

Technical

When HTML (HyperTextMarkupLanguage) was designed, there was no explicit mechanism in the design to keep track of backlinks in software, as this carried additional logistical and network overhead.

Most Content management systems include features to track backlinks, provided the external site linking in sends notification to the target site. Most wiki systems include the capability of determining what pages link internally to any given page, but do not track external links to any given page.

Most commercial search engines provide a mechanism to determine the number of backlinks they have recorded to a particular web page. For example, Google can be searched using Google:link:http://www.wikipedia.org|link:wikipedia.org to find the number of pages on the Web pointing to http:// wikipedia.org/.Google only shows a small fraction of the number of links pointing to a site. It credits many more backlinks than it shows for each website.

Other mechanisms have been developed to track backlinks between disparate webpages controlled by organizations that aren't associated with each other. The most notable example of this is TrackBacks between blogs.

References

[1] Lennart Björneborn and Peter Ingwersen (2004). "Toward a Basic Framework for Webometrics" (http://www3.interscience.wiley.com/ cgi-bin/abstract/109594194/ABSTRACT). *Journal of the American Society for Information Science and Technology* **55** (14): 1216–1227. doi:10.1002/asi.20077. .

[2] Google's overview of PageRank (http://www.google.com/intl/en/technology/)

[3] "Does Backlink Quality Matter?" (http://www.adgooroo.com/backlink_quality.php). *Adgooroo*. 2010-04-21. . Retrieved 2010-04-21.

[[de:Rückverweis]

URL redirection

URL redirection, also called **URL forwarding**, is a World Wide Web technique for making a web page available under more than one URL address. When a web browser attempts to open a URL that has been redirected, a page with a different URL is opened. For example, www.example.com [1] is redirected to www.iana.org/domains/example/ [2]. Similarly, **Domain redirection** or **domain forwarding** is when all pages in a URL domain are redirected to a different domain, as when wikipedia.com [3] and wikipedia.net [4] are automatically redirected to wikipedia.org [5]. URL redirection can be used for URL shortening, to prevent broken links when web pages are moved, to allow multiple domain names belonging to the same owner to refer to a single web site, to guide navigation into and out of a website, for privacy protection, and for less innocuous purposes such as phishing attacks using URLs that are similar to a targeted web site.

Purposes

There are several reasons to use URL redirection:

Similar domain names

A user might mis-type a URL—for example, "example.com" and "exmaple.com". Organizations often register these "mis-spelled" domains and re-direct them to the "correct" location: example.com. The addresses example.com and example.net could both redirect to a single domain, or web page, such as example.org. This technique is often used to "reserve" other top-level domains (TLD) with the same name, or make it easier for a true ".edu" or ".net" to redirect to a more recognizable ".com" domain.

Moving a site to a new domain

A web page may be redirected for several reasons:

- a web site might need to change its domain name;
- an author might move his or her pages to a new domain;
- two web sites might merge.

With URL redirects, incoming links to an outdated URL can be sent to the correct location. These links might be from other sites that have not realized that there is a change or from bookmarks/favorites that users have saved in their browsers.

The same applies to search engines. They often have the older/outdated domain names and links in their database and will send search users to these old URLs. By using a "moved permanently" redirect to the new URL, visitors will still end up at the correct page. Also, in the next search engine pass, the search engine should detect and use the newer URL.

Logging outgoing links

The access logs of most web servers keep detailed information about where visitors came from and how they browsed the hosted site. They do not, however, log which links visitors left by. This is because the visitor's browser has no need to communicate with the original server when the visitor clicks on an outgoing link.

This information can be captured in several ways. One way involves URL redirection. Instead of sending the visitor straight to the other site, links on the site can direct to a URL on the original website's domain that automatically redirects to the real target. This technique bears the downside of the delay caused by the additional request to the original website's server. As this added request will leave a trace in the server log, revealing exactly which link was followed, it can also be a privacy issue.[1]

The same technique is also used by some corporate websites to implement a statement that the subsequent content is at another site, and therefore not necessarily affiliated with the corporation. In such scenarios, displaying the warning causes an additional delay.

Short aliases for long URLs

Web applications often include lengthy descriptive attributes in their URLs which represent data hierarchies, command structures, transaction paths and session information. This practice results in a URL that is aesthetically unpleasant and difficult to remember, and which may not fit within the size limitations of microblogging sites. URL shortening services provide a solution to this problem by redirecting a user to a longer URL from a shorter one..

Meaningful, persistent aliases for long or changing URLs

Sometimes the URL of a page changes even though the content stays the same. Therefore URL redirection can help users who have bookmarks. This is routinely done on Wikipedia whenever a page is renamed.

Manipulating search engines

Some years ago, redirect techniques were used to fool search engines. For example, one page could show popular search terms to search engines but redirect the visitors to a different target page. There are also cases where redirects have been used to "steal" the page rank of one popular page and use it for a different page, usually involving the 302 HTTP status code of "moved temporarily."[2][3]

Search engine providers noticed the problem and took appropriate actions . Usually, sites that employ such techniques to manipulate search engines are punished automatically by reducing their ranking or by excluding them from the search index.

As a result, today, such manipulations usually result in less rather than more site exposure.

Satire and criticism

In the same way that a Google bomb can be used for satire and political criticism, a domain name that conveys one meaning can be redirected to any other web page, sometimes with malicious intent. The website `shadyurl.com` offers a satirical service that will create an apparently "suspicious and frightening" redirection URL for even benign webpages. For example, an input of `en.wikipedia.org` generates `5z8.info/hookers_e4u5_inject_worm`.

Manipulating visitors

URL redirection is sometimes used as a part of phishing attacks that confuse visitors about which web site they are visiting . Because modern browsers always show the real URL in the address bar, the threat is lessened. However, redirects can also take you to sites that will otherwise attempt to attack in other ways. For example, a redirect might take a user to a site that would attempt to trick them into downloading antivirus software and ironically installing a trojan of some sort instead.

Removing `referer` information

When a link is clicked, the browser sends along in the HTTP request a field called referer which indicates the source of the link. This field is populated with the URL of the current web page, and will end up in the logs of the server serving the external link. Since sensitive pages may have sensitive URLs (for example, `http://company.com/plans-for-the-next-release-of-our-product`), it is not desirable for the `referer` URL to leave the organization. A redirection page that performs referrer hiding could be embedded in all external URLs, transforming for example `http://externalsite.com/page` into

`http://redirect.company.com/http://externalsite.com/page`. This technique also eliminates other potentially sensitive information from the referer URL, such as the session ID, and can reduce the chance of phishing by indicating to the end user that they passed a clear gateway to another site.

Techniques

There are several techniques to implement a redirect. In many cases, Refresh meta tag is the simplest one. However, there exist several strong opinions discouraging this method.[4]

Manual redirect

The simplest technique is to ask the visitor to follow a link to the new page, usually using an HTML anchor as such:

```
Please follow <a href="http://www.example.com/">this link</a>.
```

This method is often used as a fall-back for automatic methods — if the visitor's browser does not support the automatic redirect method, the visitor can still reach the target document by following the link.

HTTP status codes 3xx

In the HTTP protocol used by the World Wide Web, a **redirect** is a response with a status code beginning with *3* that induces a browser to go to another location, with annotation describing the reason, which allows for the correct subsequent action (such as changing links in the case of code 301, a permanent change of address)

The HTTP standard [10] defines several status codes [11] for redirection:

- 300 multiple choices (e.g. offer different languages)
- 301 moved permanently
- 302 found (originally temporary redirect, but now commonly used to specify redirection for unspecified reason)
- 303 see other (e.g. for results of cgi-scripts)
- 307 temporary redirect

All of these status codes require that the URL of the redirect target be given in the Location: header of the HTTP response. The 300 multiple choices will usually list all choices in the body of the message and show the default choice in the Location:header.

Within the 3xx range, there are also some status codes that are quite different from the above redirects (they are not discussed here with their details):

- 304 not modified
- 305 use proxy

This is a sample of an HTTP response that uses the 301 "moved permanently" redirect:

```
HTTP/1.1 301 Moved Permanently

Location: http://www.example.org/

Content-Type: text/html

Content-Length: 174

<html>

<head>
```

```
<title>Moved</title>

</head>

<body>

<h1>Moved</h1>

<p>This page has moved to <a href="http://www.example.org/">http://www.example.org/</a>.</p>

</body>

</html>
```

Using server-side scripting for redirection

Often, web authors don't have sufficient permissions to produce these status codes: The HTTP header is generated by the web server program and not read from the file for that URL. Even for CGI scripts, the web server usually generates the status code automatically and allows custom headers to be added by the script. To produce HTTP status codes with cgi-scripts, one needs to enable non-parsed-headers.

Sometimes, it is sufficient to print the "Location: 'url'" header line from a normal CGI script. Many web servers choose one of the 3xx status codes for such replies.

Frameworks for server-side content generation typically require that HTTP headers be generated before response data. As a result, the web programmer who is using such a scripting language to redirect the user's browser to another page must ensure that the redirect is the first or only part of the response. In the ASP scripting language, this can also be accomplished using the methods `response.buffer=true` and `response.redirect "http://www.example.com/"`. Using PHP, one can use the header function as follows:

```
header('HTTP/1.1 301 Moved Permanently');

header('Location: http://www.example.com/');

exit();
```

According to the HTTP protocol, the Location header must contain an absolute URI.[5] When redirecting from one page to another within the same site, it is a common mistake to use a relative URI. As a result most browsers tolerate relative URIs in the Location header, but some browsers display a warning to the end user.

There are other methods that can be used for performing redirects, but they do not offer the flexibility that mod_rewrite offers. These alternative rules use functions within mod_alias:

```
Redirect permanent /oldpage.html http://www.example.com/newpage.html

Redirect 301 /oldpage.html http://www.example.com/newpage.html
```

To redirect a requests for any non-canonical domain name using .htaccess or within a <Directory> section in an Apache config file:

```
RewriteEngine on

RewriteCond %{HTTP_HOST}
^([^.:]+\.)*oldsite\.example\.com\.?(:[0-9]*)?$ [NC]

RewriteRule ^(.*)$ http://newsite.example.net/$1 [R=301,L]
```

Use of .htaccess for this purpose usually does not require administrative permissions. However, .htaccess can be disabled by your host, and so may not work (or continue to work) if they do so.

In addition, some server configurations may require the addition of the line:

```
Options +FollowSymLinks
```

ahead of the "RewriteEngine on" directive, in order to enable the mod_rewrite module.

When you have access to the main Apache config files (such as httpd.conf), it is best to avoid the use of .htaccess files.

If the code is placed into an Apache config file and not within any <Directory> container, then the RewriteRule pattern must be changed to include a leading slash:

```
RewriteEngine on

RewriteCond %{HTTP_HOST} ^([^.:]+\.)*oldwebsite\.com\.?(:[0-9]*)?$ [NC]

RewriteRule ^/(.*)$ http://www.preferredwebsite.net/$1 [R=301,L]
```

Refresh Meta tag and HTTP refresh header

Netscape introduced a feature to refresh the displayed page after a certain amount of time. This method is often called meta refresh. It is possible to specify the URL of the new page, thus replacing one page after some time by another page:

- HTML <meta> tag[13]
- An exploration of dynamic documents [14]
- Meta refresh

A timeout of 0 seconds means an immediate redirect. Meta Refresh with a timeout of 0 seconds is accepted as a 301 permanent redirect by Google, allowing to transfer PageRank from static html files.[6]

This is an example of a simple HTML document that uses this technique:

```
<html>

<head>

<meta http-equiv="Refresh" content="0; url=http://www.example.com/" />

</head>

<body>

<p>Please follow <a href="http://www.example.com/">this link</a>.</p>
```

```
</body>

</html>
```

- This technique is usable by all web authors because the meta tag is contained inside the document itself.
- The meta tag must be placed in the "head" section of the HTML file.
- The number "0" in this example may be replaced by another number to achieve a delay of that many seconds.
- This is a proprietary extension to HTML introduced by Netscape but supported by most web browsers. The manual link in the "body" section is for users whose browsers do not support this feature.

This is an example of achieving the same effect by issuing an HTTP refresh header:

```
HTTP/1.1 200 ok

Refresh: 0; url=http://www.example.com/

Content-type: text/html

Content-length: 78

Please follow <a href="http://www.example.com/">this link</a>!
```

This response is easier to generate by CGI programs because one does not need to change the default status code. Here is a simple CGI program that effects this redirect:

```
#!/usr/bin/perl

print "Refresh: 0; url=http://www.example.com/\r\n";

print "Content-type: text/html\r\n";

print "\r\n";

print "Please follow <a href=\"http://www.example.com/\">this link</a>!"
```

Note: Usually, the HTTP server adds the status line and the Content-length header automatically.

This method is considered by the W3C to be a poor method of redirection, since it does not communicate any information about either the original or new resource, to the browser (or search engine). The W3C's Web Content Accessibility Guidelines (7.4) [16] discourage the creation of auto-refreshing pages, since most web browsers do not allow the user to disable or control the refresh rate. Some articles that they have written on the issue include W3C Web Content Accessibility Guidelines (1.0): Ensure user control of time-sensitive content changes [17] and Use standard redirects: don't break the back button! [18]

This example works best for a refresh, or in simple terms - a redirect for webpages, as follows, however, for a refresh under 4 seconds, your webpage will not be given priority listing on search engines. For some users, this is preferred not to be listed. Inline, you will find the time as in seconds:

CONTENT="2

this number can be adjusted to suit your needs. Place in

your head:

```
<HTML>

<HEAD>

   <META HTTP-EQUIV="refresh" CONTENT="2;URL=http://www.example.com/example.html">

</HEAD>
```

JavaScript redirects

JavaScript offers several ways to display a different page in the current browser window. Quite frequently, they are used for a redirect. However, there are several reasons to prefer HTTP header or the refresh meta tag (whenever it is possible) over JavaScript redirects:

- Security considerations
- Some browsers don't support JavaScript
- many web crawlers don't execute JavaScript.

Frame redirects

A slightly different effect can be achieved by creating a single HTML frame that contains the target page:

```
<frameset rows="100%">

  <frame src="http://www.example.com/">

</frameset>

<noframes>

  <body>Please follow <a href="http://www.example.com/">link</a>!</body>

</noframes>
```

One main difference to the above redirect methods is that for a frame redirect, the browser displays the URL of the frame document and not the URL of the target page in the URL bar.

This technique is commonly called *cloaking*. This may be used so that the reader sees a more memorable URL or, with fraudulent intentions, to conceal a phishing site as part of website spoofing.[7]

Redirect loops

It is quite possible that one redirect leads to another redirect. For example, the URL http://www.wikipedia.com/ wiki/URL_redirection (note the differences in the domain name) is first redirected to http://www.wikipedia.org/ wiki/URL_redirection and again redirected to the correct URL: http://en.wikipedia.org/wiki/URL_redirection. This is appropriate: the first redirection corrects the wrong domain name, the second redirection selects the correct language section, and finally, the browser displays the correct page.

Sometimes, however, a mistake can cause the redirection to point back to the first page, leading to an infinite loop of redirects. Browsers usually break that loop after a few steps and display an error message instead.

The HTTP standard [11] states:

> A client SHOULD detect infinite redirection loops, since such loops generate network traffic for each redirection.
>
> Previous versions of this specification recommended a maximum of five redirections; some clients may exist that implement such a fixed limitation.

Services

There exist services that can perform URL redirection on demand, with no need for technical work or access to the webserver your site is hosted on.

URL redirection services

A **redirect service** is an information management system, which provides an internet link that redirects users to the desired content. The typical benefit to the user is the use of a memorable domain name, and a reduction in the length of the URL or web address. A redirecting link can also be used as a permanent address for content that frequently changes hosts, similarly to the Domain Name System.

Hyperlinks involving URL redirection services are frequently used in spam messages directed at blogs and wikis. Thus, one way to reduce spam is to reject all edits and comments containing hyperlinks to known URL redirection services; however, this will also remove legitimate edits and comments and may not be an effective method to reduce spam.

Recently, URL redirection services have taken to using AJAX as an efficient, user friendly method for creating shortened URLs.

A major drawback of some URL redirection services is the use of delay pages, or frame based advertising, to generate revenue.

History

The first redirect services took advantage of top-level domains (TLD) such as ".to" (Tonga), ".at" (Austria) and ".is" (Iceland). Their goal was to make memorable URLs. The first mainstream redirect service was V3.com that boasted 4 million users at its peak in 2000. V3.com success was attributed to having a wide variety of short memorable domains including "r.im", "go.to", "i.am", "come.to" and "start.at". V3.com was acquired by FortuneCity.com, a large free web hosting company, in early 1999. In 2001 emerged .tk (Tokelau) as a TLD used for memorable names.[8] As the sales price of top level domains started falling from $70.00 per year to less than $10.00, the demand for memorable redirection services eroded.

With the launch of TinyURL in 2002 a new kind of redirecting service was born, namely URL shortening. Their goal was to make long URLs short, to be able to post them on internet forums. Since 2006, with the 140 character limit on the extremely popular Twitter service, these short URL services have seen a resurgence.

Referrer Masking

Redirection services can hide the referrer by placing an intermediate page between the page the link is on and its destination. Although these are conceptually similar to other URL redirection services, they serve a different purpose, and they rarely attempt to shorten or obfuscate the destination URL (as their only intended side-effect is to hide referrer information and provide a clear gateway between other websites.)

This type of redirection is often used to prevent potentially-malicious links from gaining information using the referrer, for example a session ID in the query string. Many large community websites use link redirection on external links to lessen the chance of an exploit that could be used to steal account information, as well as make it clear when a user is leaving a service, to lessen the chance of effective phishing .

Here is a simplistic example of such a service, written in PHP.

```
<?php

$url = htmlspecialchars($_GET['url']);

header( 'Refresh: 0; url=http://'.$url );

?>

<!-- Fallback using meta refresh. -->

<html>

 <head>

  <title>Redirecting...</title>

  <meta http-equiv="refresh" content="0;url=http://<?php echo $url; ?>">

 </head>

 <body>

 Attempting to redirect to <a href="http://<?php echo $url; ?>">http://<?php echo $url; ?></a>.

 </body>

</html>
```

References

[1] "Google revives redirect snoopery" (http://blog.anta.net/2009/01/29/509/). *blog.anta.net*. 2009-01-29. ISSN 1797-1993. . Retrieved 2009-01-30.

[2] Google's serious hijack problem (http://www.pandia.com/sw-2004/40-hijack.html)

[3] Stop 302 Redirects and Scrapers from Hijacking Web Page PR - Page Rank (http://www.loriswebs.com/hijacking_web_pages.html)

[4] Core Techniques for Web Content Accessibility Guidelines 1.0 section 7 (http://www.w3.org/TR/WCAG10-CORE-TECHS/ #auto-page-refresh), w3.org, published 2000-11-6, fetched 2009-12-15.

[5] R. Fielding, et al., Request for Comments: 2616, Hypertext Transfer Protocol — HTTP/1.1, published 1999-07, §14.30 "Location" (http:// www.w3.org/Protocols/rfc2616/rfc2616-sec14.html#sec14.30), fetched 2008-10-07

[6] Google and Yahoo accept undelayed meta refreshs as 301 redirects, 3 September 2007, http://sebastians-pamphlets.com/ google-and-yahoo-treat-undelayed-meta-refresh-as-301-redirect/

[7] Anti-Phishing Technology" (http://www.sfbay-infragard.org/Documents/phishing-sfectf-report.pdf), Aaron Emigh, Radix Labs, 19 January 2005

[8] "Net gains for tiny Pacific nation" (http://news.bbc.co.uk/2/hi/technology/6991719.stm). *BBC News*. 2007-09-14. . Retrieved 2010-05-27.

External links

- Mapping URLs to Filesystem Locations (http://httpd.apache.org/docs/1.3/urlmapping.html)
- Paper on redirection spam (UC Davis) (http://www.cs.ucdavis.edu/~hchen/paper/www07.pdf)
- Servlet redirect example (http://www.jsptube.com/examples/response-sendredirect-servlet.html)
- Servlet forward example (http://www.jsptube.com/examples/servlet-forward.html)
- Security vulnerabilities in URL Redirectors (http://projects.webappsec.org/URL-Redirector-Abuse) The Web Application Security Consortium Threat Classification
- 301 Redirects for moved pages using .htaccess (http://www.dancatts.com/articles/ htaccess-301-redirects-for-moved-pages.php)
- 301-redirect.info, site summarizing redirection methods in Apache, PHP, ASP, JPs or ColdFusion programming (http://www.301-redirect.info/)
- Redirecting your visitors to your preferred domain (http://911-need-code-help.blogspot.com/2011/03/ redirecting-visitors-to-preferred.html) using 301 permanent redirects — rationale and mod_rewrite/PHP/ASP.NET implementation