ANALISIS ALGORITMA MIXED HEURISTIC UNTUK PENCOCOKAN STRING PADA WEB SEARCH ENGINE

Ressy Dwitias Sari[1], Yanuar Firdaus A.w.[2], M. Dickson Ichdayantomsee[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Mesin pencari dirancang untuk membantu seseorang menemukan file dan dokumen dari dokumen yang sangat banyak dan disimpan dalam komputer, misalnya sebuah server. Algoritma pencocokan string merupakan komponen dasar dalam melakukan pencarian data di mesin pencari. Masalah utama dalam pencarian string adalah untuk mencari sebuah string yang terdiri dari beberapa karakter (yang biasa disebut pattern) dalam sejumlah besar text [8]. Sehingga pencarian yang dilakukan membutuhkan waktu yang banyak. Dibutuhkan algoritma yang dapat bekerja secara cepat dan mengurutkan dokumen sesuai tingkat kecocokan. Salah satu algoritma yang cocok adalah Mixed Heuristic. Algoritma ini melakukan pencarian pattern atau query tidak hanya terhadap 1 kata, tetapi dapat berupa kalimat yang lebih dari 1 kata. Pola pencocokan string melakukan penggeseran pattern sebanyak 1 kata, jika kecocokan belum ditemukan. Selain itu, algoritma ini juga melakukan perangkingan dokumen yang relevan. Dalam tugas akhir ini, dipaparkan analisis tingkat akurasi menggunakan precision dan recall dari hasil yang diberikan oleh mesin pencari dengan menggunakan algoritma Mixed Heuristic untuk pencocokan string. Dan analisis hasil perangkingan yang diberikan oleh mesin pencari.

Kata Kunci : mesin pencari, pencocokan string, mixed heuristic

Abstract

Search engines are designed to help people search files and documents from many other documents and then stored in a computer, such as a server. String matching algorithm is the basic component for data searching in search engines. The main problem in the search string is finding a string that consists of several characters (called pattern) in a large number of texts [8]. As a result it takes a lot of time to get the search done. It requires an algorithm that can work quickly and can sort the documents according to the level of compatibility. One of the algorithms that match is the Mixed Heuristic. These algorithms perform a search pattern or query not just against a word, but can be a sentence of more than one word. The matching string system makes the pattern shift as much as one word, if the match is not found. In addition, these algorithms also perform ranking of relevant documents. In this thesis, it shows the analysis of the level of accuracy using precision and recall of the results given by search engines by using the Mixed Heuristic algorithms for string matching, and the analysis of documents from the results given by search engines.

Keywords : search engine, string matching, mixed heuristic

1. PENDAHULUAN

1.1 Latar Belakang

Setiap orang ingin mendapatkan informasi yang akurat dan cepat. Pemanfaatan internet kerap sekali dilakukan oleh masyarakat masa kini untuk mendapatkan informasi yang diinginkan. Faktanya pengguna internet yang membaca berita *online* sebesar 47 persen dan yang memanfaatkan mesin pencari sebesar 56 persen [18]. Walaupun untuk mendapatkan data bisa dengan berbagai cara, teks tetap menjadi bentuk utama dalam bertukar informasi [9].

Untuk membaca berita *online* sering sekali *user* memanfaatkan mesin pencari agar cepat mendapatkan berita yang diinginkan. Tetapi kerap sekali berita yang didapatkan tidak sesuai dengan yang diinginkan. Untuk itu, penulis mencoba membuat mesin pencari dengan menggunakan dataset media masa elektronik. Mesin pencari ini akan melakukan pencocokan string dengan *keyword* yang di-*input* oleh *user*. Agar hasil yang diberikan mesin pencari sesuai dengan yang diinginkan *user*.

Algoritma pencocokan string merupakan komponen dasar dalam melakukan pencarian data. Banyak metode atau algoritma yang melakukan pencarian dengan mencocokkan string. Salah satunya algoritma dengan teknik *Heuristic* seperti *Brute Force,* merupakan algoritma yang sering digunakan untuk melakukan pencocokan string. Algoritma ini bekerja dengan cara mencocokkan *pattern* ke awal kalimat, jika belum mendapatkan kecocokan, geser *pattern* satu karakter ke kanan, ini dilakukan sampai kecocokan ditemukan.Masalah utama dalam pencarian string adalah untuk mencari sebuah string yang terdiri dari beberapa karakter (yang biasa disebut *pattern*) dalam sejumlah besar text [8]. Sehingga pencarian yang dilakukan membutuhkan waktu yang banyak.

Untuk itu dibutuhkan mesin pencari yang dapat mencari dokumen secara cepat dan diurutkan sesuai tingkat kecocokan. Dengan mempertimbangkan masalah diatas maka pada Tugas Akhir ini penulis akan mengimplementasikan dan menganalisis bagaimana menggunakan algoritma *Mixed Heuristic* untuk pencocokan string pada *Search Engine* yang berbasis *Information Retrieval.*

Algoritma *Mixed Heuristic* ini merupakan algoritma yang mengadopsi dari algoritma *Heuristic* yang sudah dikenal, yaitu *Brute Force.* Algoritma *Mixed Heuristic* melakukan pencarian *pattern* atau *query* tidak hanya terhadap 1 kata, tetapi dapat berupa kalimat yang lebih dari 1 kata. Pola pencocokan string melakukan penggeseran *pattern* sebanyak 1 kata, jika kecocokan belum ditemukan.

Algoritma ini mampu untuk memberikan informasi yang lebih cocok dengan keinginan *user*[9]. *Search Engine* dapat digunakan dengan mudah karena pengaruh dari sifat algoritma *Mixed Heuristic, user* mendapatkan solusi yang diinginkan dengan cepat dan akurat[9]. Algoritma ini bekerja dengan mendapatkan solusi sebesar 100% (sama persis), jika tidak mendapatkan solusi sebesar 100% maka algoritma ini akan memberikan solusi dengan menggunakan perkiraan string yang cocok dan menghitung probabilitas keberhasilannya [9].

Dalam Tugas Akhir ini penulis menggunakan parameter *precision* dan *recall* sebagai alat ukur performansi dari algoritma *Mixed Heuristic*. Dan

memperhatikan tingkat keakuratan dari perangkingan dokumen. Diharapkan dengan menggunakan algoritma *Mixed Heuristic* pada *Search Engine* hasil dari pencocokan string lebih akurat, dimana informasi yang dibutuhkan *user* sesuai dengan yang diinginkan dan perangkingan dokumen sesuai dengan tingkat kecocokannya. Sehingga, *user* tidak perlu mengecek satu persatu dokumen yang diberikan, karena yang diinginkannya sudah berada di peringkat teratas.

1.2 Perumusan Masalah

Berdasarkan latar belakang permasalahan di atas, maka penulis merumuskan permasalahan yang akan diangkat dalam Tugas Akhir ini, yaitu :

- Bagaimana mengimplementasikan algoritma *mixed heuristic* untuk pencocokan string pada *web search engine* berbasis *Information Retrieval*?
- Bagaimana tingkat akurasi pencarian yang dihasilkan oleh algoritma *mixed heuristic* pada *web search engine* dilihat dari parameter *precision* dan *recall* berdasarkan jumlah dataset?
- Bagaimana tingkat relevansi dokumen yang dihasilkan oleh algoritma *mixed heuristic* pada *web search engine* berdasarkan perangkingan dokumen?

1.3 Tujuan

Berdasarkan perumusan masalah di atas, maka tujuan dari Tugas Akhir ini, yaitu :

- Mengimplementasikan algoritma *mixed heuristic* untuk pencocokan string pada *web search engine* berbasis *Information Retrieval*.
- Menganalisis tingkat akurasi yang dihasilkan algoritma *mixed heuristic* pada *web search engine* dilihat dari *precision* dan *recall* berdasarkan jumlah dataset dan jumlah term pada *query*.
- Menganalisis perangkingan dokumen yang dihasilkan oleh algoritma *mixed heuristic* pada *web search engine.*

1.4 Batasan Masalah

Adapun batasan masalah pada tugas akhir ini, yaitu:

- Dalam implementasinya, *text preprocessing* dan *indexing* menjadi bagian dalam Tugas Akhir ini, tetapi tidak menjadi pokok permasalahan Tugas Akhir.
- Dokumen yang digunakan untuk input *preprocessing* adalah file berekstensi *.txt.
- Dokumen merupakan berita Bahasa Indonesia yang diunduh dari www.kompas.com.
- Dokumen merupakan dokumen *free text (unstructured text).*
- Pengujian dilakukan secara *offline.*

1.5 Metodologi Penyelesaian Masalah

Metodologi yang digunakan dalam menyelesaikan Tugas Akhir ini, yaitu:

1. Studi Literatur
   Tahap ini akan melakukan pencarian referensi-referensi dan materi yang ada di internet serta memahami dan mempelajarinya sehingga dapat digunakan untuk menyelesaikan permasalahan dalam tugas akhir ini. Pencarian referensi berkaitan dengan pembangunan aplikasi search engine berbasis Web, *Information Retrieval*, Algortima Perankingan *retrieved dokumen* dari sebuah *query*, dan lebih mendalam mengenai algoritma *Mixed Heuristic*.
2. Pengumpulan Data
   Tahap ini akan melakukan pengumpulan data berupa dokumen berita Bahasa Indonesia Bahasa Indonesia yang diunduh dari www.kompas.com. Dokumen ini akan diolah pada tahap *pre-processing of information retrieval system*. Penulis juga membuat sekumpulan dokumen uji untuk membantu dalam skenario pengujian.
3. Analisis Kebutuhan Sistem
   Tahap ini akan melakukan analisis sistem dan menentukan kebutuhan dari sistem yang akan dibangun seperti kebutuhan fungsional sistem, spesifikasi perangkat lunak dan perangkat keras yang digunakan, serta pemodelan sistem yang akan dibangun.
4. Perancangan Sistem
   Tahap ini akan melakukan perancangan kebutuhan perangkat lunak dalam membangun aplikasi *search engine* berbasis Information Retrieval dan menerapkan Algoritma *Mixed Heuristic* dalam pencarian dan perankingan dokumen.
5. Implementasi dan Pengujian
   Tahap ini akan melakukan implementasi hasil perancangan. Pada Tugas Akhir ini juga tingkat relevansi perangkingan hasil *query*.
6. Analisis
   Tahap ini akan melakukan analisa hasil pengujian dan pengukuran performansi berdasar data yang diuji serta mengambil kesimpulan dari hasil yang telah dianalisa. Pengujian metode dilakukan dengan menggunakan sistem yang sebelumnya telah diimplementasikan pada tahap implementasi.
7. Pembuatan Laporan
   Tahap ini akan melakukan dokumentasi tahap-tahap kegiatan dan hasil yang di dapat ke dalam laporan Tugas Akhir. Dijelaskan pula mengenai langkah-langkah secara detail dalam menganalisis kebutuhan dari awal, perancangan sistem, implementasi, pengujian, serta analisisnya.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Setelah melakukan pengujian dan analisis, maka dapat diambil kesimpulan berikut ini:

1. Algoritma *Mixed Heuritic* dapat diterapkan untuk *search engine*, ditandai dengan nilai *recall* yang tinggi, tetapi masih memiliki kelemahan pada cara kerja.
2. Dengan menggunakan jumlah dataset besar, maka nilai *precision* dan *recall* relatif tinggi. Untuk jumlah term yang panjang pada *query*, dengan jumlah terpanjang lebih dari 2 term, maka nilai *precision* dan *recall* relatif rendah. Termasuk *query* yang menggunakan *stopword*.
3. Perangkingan dengan menggunakan algoritma *Mixed Heuristic* memberikan hasil yang baik, dengan nilai rata-rata 12% untuk hasil yang salah dari 20 teratas dokumen ter-*retrieve*.

5.2 Saran

Setelah menyelesaikan proses pengerjaan tugas akhir ini, penulis memiliki beberapa saran sebagai berikut:

1. Dalam penerapan algoritma *Mixed Heuristic*, lebih baik dikombinasikan dengan algoritma lain yang menangani proses *retrieve* dokumen.
2. Dapat dilakukan dengan mencoba dataset yang lebih besar.

Referensi

[1] Allauzen C, Raffinot M, "*Simple optimal string matching algorithm*" Journal of Algorithms 36(1): 102-116, 2000.

[2] Asian,Jelita dkk. ”*Stemming Indonesia*” [CRPITV3Asian.pdf]. RMIT University Melbourne.Australia

[3] Binstock and Rex 1995. “*Practical Algorithms for Programmers*”. Addison Wesley

[4] Baeza-Yates, Ricardo dan Berthier-Neto.1999, “*Modern Information Retrieval*”, New York: Addison Wesley, ACM Press

[5] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, “*Introduction to Information Retrieval”*, Cambridge University Press. 2008

[6] Craswell Nicholas Eric, “*Methods for Distributed Information Retrieval*”, The Australian National University, 2000

[7] Dictionary of Algorithm and Data Structure. National Institute of Standards and Technology

[8] Hartoyo Eko Gunocipto, Vembrina Yus Gias, Meilana Anggia Ferdina, “Analisis Algoritma Pencarian String”, Departemen Teknik Informatika, Institut Teknologi Bandung, Bandung, Indonesia, 2010

[9] Hassan Atif Agha, “*Mixed Heuristic Algorithm for Intelligent String Matching for Information Retrieval*”, UK, 2005

[10] Juariyah Oktaf, “Analisis dan Implementasi *Distributed Indexing* Dengan *Database* Tunggal”, Fakultas Informatika, Institut Teknologi Telkom, Bandung, 2010

[11] Kumar Narendra, Bibhu Vimal, Islam Mohammad, Bhardwaj Shashank, “*Approximate String Matching Algorithm*”, International Journal on Computer Science and Engineering, India, 2010

[12] Marlow, Kit, “*Information Retrieval Methods”,* 2003

[13] Martoglia Riccardo, “*Information Retrieval Techniques for Pattern Matching*”, Anno Accademico, 2004

[14] Snowbound. *Text Extraction - Extracting Text from Documents Like PDF, MS Word, Excel*. http://www.snowbound.com/solutions/text_extraction.html diakses tanggal 16 Oktober 2010.

[15] Suyanto, “*Artificial Intelligence*”, Penerbit Informatika, Bandung, Indonesia, 2007

[16] Waiyamai, Kitsana. *Introduction to Text Mining.* Dept of Computer Engineering, Faculty of Engineering, Kasetsart University, Bangkok, Thailand

[17] Wikipedia. Mesin Pencari, http://id.wikipedia.org/wiki/Mesin_pencari/ di akses tanggal 15 Oktober 2010

[18] Wordpress. 10 Fakta Pengguna Internet, http://adhimaswijaya.wordpress.com/2010/06/25/10-fakta-pengguna-internet/ di akses pada tanggal 15 Oktober 2010

[19] Zahid M. A. Hai, Mittal Ankush, Joshi R.C, “ A Heuristic Algorithm For Optimal Agreement Supertrees” http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.101.8663&rep=rep1&type=pdf di akses pada tanggal 15 Oktober 2010