KATEGORISASI DOKUMEN WEB SECARA OTOMATIS BERDASARKAN FOLKSONOMY MENGGUNAKAN MULTINOMIAL NAIVE BAYES CLASSIFIER (AUTOMATIC FOLKSONOMY CATEGORIZATION OF WEB DOCUMENTS USING MULTINOMIAL NAIVE BAYES CLASSIFIER)

Hendy Irawan[1], -[2]

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

Abstrak

Folksonomy merupakan metode kategorisasi dokumen yang tidak hierarkis, menyamaratakan kedudukan setiap kategori, dan judul kategori ditentukan secara bebas oleh siapa saja yang memasukkan sebuah dokumen di dalam kategori-kategori tersebut. Pembuatan kategorisasi dilakukan secara otomatis pada saat dokumen dimasukkan, yaitu dengan cara mengetikkan daftar kategori yang kira-kira cocok untuk dokumen tersebut. Situs del.icio.us (http://del.icio.us) merupakan salah satu social bookmarking site terpopuler yang menggunakan folksonomy. Penggunaan folksonomy, meski sangat mudah, juga mempunyai beberapa kelemahan, yaitu penggunaan tag yang berbeda-beda untuk konsep yang sama, penggunaan tag yang sama untuk konsep yang berbeda-beda, tidak adanya pengendalian mutu, dan lain-lain. Di sini penulis mencoba memberikan solusi untuk sebagian masalah tersebut yaitu dengan cara menganalisa isi dari dokumen Web yang ditunjuk dan mengkategorisasikan link tersebut secara otomatis ke beberapa tag menggunakan multinomial naive Bayes classifier.

Bayes classifier bekerja berdasarkan sekumpulan bukti (evidence) dan kelas (class). Dengan melakukan pelatihan (training) terhadap sebagian data sampel, dapat ditentukan probabilitas kepastian (likelihood probability) dari sebuah bukti jika diberikan kelas tertentu. Bayes classifier juga menggunakan probabilitas sebelumnya (prior probability) dari sebuah kelas, yang perhitungannya dapat didasarkan dari sampel data tersebut. Dari analisa sampel data tersebut, jika diberikan sebuah dokumen baru yang terdiri dari sekumpulan bukti, probabilitas setiap kelas terhadap dokumen tersebut (posterior probability) dapat ditentukan.

Sistem ini diimplementasikan menggunakan PHP 5, Apache, dan MySQL. Kesimpulan yang didapatkan dari penelitian ini adalah metode Bayes dapat digunakan untuk melakukan kategorisasi dokumen secara otomatis maupun sebagai alat bantu untuk kategorisasi manual.

Kata Kunci : naive Bayes, text classification, folksonomy, indexing

Abstract

Folksonomy is a non-hierarchical document categorizing system, that treats every category in a flat manner, dan every category is entered freely by anyone who submitted a document in these categories. Categorization is done automatically at the time a document is submitted, by entering the list of categories that best fit the document. del.icio.us (http://del.icio.us) site is one of the most popular social bookmarking sites that uses folksonomy.

Usage of folksonomy, although very easy, also has its weaknesses, such as use of different tags for the same concept, use of the same tag for different concepts, no quality control, etc. We try to provide a solution for some of these problems by analyzing Web documents' contents and categorizing them automatically using multinomial naive Bayes algorithm.

Bayes classifier works by using a set of evidences and a set of classes. By training the system using sample data, we can determine the probability of an evidence given a particular class. Bayes classifier also uses prior probability of a class, which can be calculated from sample data. From these analysis, when given a new document which is formed by a set of evidences (words), the probabilities of each class given that document (posterior probabilities) can be determined.

This system is implemented using PHP 5, Apache, and MySQL. The conclusion from building this system is that the Bayes method can be used to automatically categorize documents and also as an assistive tool for manual categorization.

Keywords : naive Bayes, text classification, folksonomy, indexing

Bab I
Pendahuluan

1.1 Latar Belakang

Perkembangan World Wide Web dalam 10 tahun terakhir ini semakin pesat, dan sekarang keberadaan Internet sudah tidak bisa diabaikan lagi. Dengan banyaknya informasi yang tersedia di Internet, berbagai layanan telah dikembangkan untuk membantu pengguna Internet dalam menemukan informasi yang mereka inginkan.

Layanan yang paling populer adalah mesin pencari atau *search engine*, misalnya Google (http://www.google.com), Ask Jeeves (http://www.ask.com), dan MSN Search (http://www.msnsearch.com), yang dapat digunakan untuk mencari informasi berdasarkan kata-kata kunci tertentu. Meski sangat berguna, kadang-kadang kita ingin mencari informasi berdasarkan kategori tertentu. Untuk itu ada layanan direktori, misalnya Yahoo! (http://www.yahoo.com) dan dmoz.org Open Directory (http://www.dmoz.org), yang berisi daftar situs Web berdasarkan klasifikasi tertentu. Layanan direktori ini sebenarnya sangat berguna, namun memiliki banyak kelemahan, di antaranya adalah kurang terupdate, struktur kategori yang statis, dan kurangnya staf untuk merawat ribuan kategori dan subkategori yang terdapat dalam sebuah layanan direktori.

Kategori situs yang akhir-akhir ini populer adalah *social networking sites*, misalnya Friendster (http://www.friendster.com) dan GaulDong (http://www.gauldong.net). Pesatnya perkembangan *social sites* membuat banyak layanan lain bermunculan, salah satunya adalah del.icio.us (http://del.icio.us) yang merupakan *social bookmarking site* berdasarkan *folksonomy*. Setiap pengguna del.icio.us bebas mengirim link/bookmark ke sebuah situs yang dikategorikan ke satu atau lebih *tag* (istilah del.icio.us untuk kategori). *Folksonomy* merupakan teknik klasifikasi yang tidak hierarkis, melainkan semua kategori disamaratakan, dan kategori dibuat secara bebas berdasarkan masukan dari pengguna.

Penggunaan *folksonomy*, meski sangat mudah, juga mempunyai beberapa kelemahan, yaitu penggunaan *tag* yang berbeda-beda untuk konsep yang sama, penggunaan *tag* yang sama untuk konsep yang berbeda-beda, tidak adanya pengendalian mutu, dan lain-lain. Di sini penulis mencoba memberikan solusi untuk sebagian masalah tersebut yaitu dengan cara menganalisa isi dari dokumen Web yang ditunjuk dan mengkategorisasikan link tersebut secara otomatis ke beberapa *tag* menggunakan multinomial naive Bayes classifier untuk keperluan ini.

1.2 Perumusan Masalah

Permasalahan yang dijadikan objek penelitian tugas akhir ini adalah menitikberatkan pada analisa penggunaan multinomial naive Bayes classifier untuk kategorisasi dokumen Web secara otomatis. Dari penelitian ini diharapkan dapat diketahui bagaimana performansi algoritma tersebut dalam melakukan kategorisasi otomatis dokumen Web.

1.3 Tujuan

Tujuan atau hasil akhir yang ingin dicapai dari tugas akhir ini adalah:

1. Mengimplementasikan sebuah sistem kategorisasi dokumen Web dengan fasilitas minimal, bernama Gado-gado.
2. Menerapkan multinomial naive Bayes classifier untuk kategorisasi dokumen secara otomatis.
3. Melakukan analisa terhadap akurasi kategorisasi otomatis yang dilakukan dibandingkan dengan kategorisasi secara manual, serta analisa performansi dan efisiensi.

1.4 Batasan Masalah

Untuk menghindari meluasnya materi pembahasan tugas akhir ini, maka penulis membatasi permasalahan dalam tugas akhir ini hanya mencakup hal-hal berikut:

1. Aplikasi akan dibangun dengan menggunakan PHP 5.0.3 sebagai web scripting language, Apache 2.0 sebagai web server, MySQL sebagai database management system, dan Windows XP Professional sebagai operating system.
2. Aktivitas yang dilakukan dalam sistem ini meliputi login user, penambahan link, menampilkan daftar link dalam sebuah tag, dan melakukan uji coba.
3. Dokumen yang dapat diterima hanya dalam bahasa Inggris (variasi apa pun) atau Indonesia (tidak menggunakan dua bahasa dalam satu dokumen), dengan encoding berikut: ISO-8859-1, WIN-1252, atau UTF-8.
4. Ukuran keseluruhan database maksimal 50 MB.

1.5 Metodologi Penyelesaian Masalah

Berikut ini adalah metodologi penyelesaian masalah yang dipergunakan dalam tugas akhir ini:

1. Studi literatur

Bertujuan untuk mempelajari dasar teori dari literatur-literatur tentang :

- Folksonomy
- Indexing
- Inverted index
- Klasifikasi teks
- Teori probabilitas
- Bayes theorem
- Naive Bayes classifier
- Multinomial naive Bayes classifier

2. Pengumpulan data untuk inputan
3. Studi perancangan perangkat lunak

   Bertujuan untuk menentukan metodologi pengembangan perangkat lunak yang digunakan dengan menggunakan metode terstruktur dan melakukan perancangan sistem.

4. Pembuatan perangkat lunak

   Bertujuan untuk melakukan implementasi metode pada perangkat lunak sesuai dengan analisa perancangan yang telah dilakukan.

5. Pengujian perangkat lunak
6. Analisa terhadap hasil pengujian
7. Pengambilan kesimpulan dan penyusunan laporan

1.6 Sistematika Penulisan

BAB I PENDAHULUAN (INTRODUCTION)

Berisi latar belakang, perumusan masalah, tujuan pembahasan., metodologi penyelesaian masalah dan sistematika penulisan.

BAB II LANDASAN TEORI (FUNDAMENTAL THEORIES)

Penjelasan mengenai *folksonomy*, *indexing*, *inverted index*, *text classifier*, teori probabilitas, Bayes *theorem*, *naive* Bayes, *multinomial naive* Bayes.

BAB III ANALISIS DAN PERANCANGAN SISTEM (SYSTEM ANALYSIS AND DESIGN)

Membahas tentang perancangan awal sistem dengan metode Test-Driven Development (TDD) menggunakan bahasa pemodelan UML.

BAB IV IMPLEMENTASI DAN PENGUJIAN (IMPLEMENTATION AND TESTING)

Menyajikan hasil pengujian dan analisa terhadap kategorisasi dokumen Web secara otomatis berdasarkan folksonomy menggunakan multinomial naive Bayes classifier.

BAB V KESIMPULAN DAN SARAN (CONCLUSIONS AND SUGGESTIONS)

Berisi kesimpulan dan saran pengembangan.

Bab V
Kesimpulan dan Saran

5.1 Kesimpulan

Dari hasil pembangunan sistem ini serta dari hasil uji coba yang telah dilakukan, dapat ditarik beberapa kesimpulan sebagai berikut:

1. Metode multinomial naive Bayes yang dipakai oleh Gado-gado dapat digunakan sebagai metode klasifikasi atau kategorisasi teks dengan akurasi hampir 70% (sesuai dengan hasil pengujian pada Bab IV).
2. Algoritma multinomial naive Bayes yang dipakai oleh Gado-gado memiliki tingkat kegagalan sekitar 7%.
3. Akurasi algoritma tersebut berlaku untuk penggunaan 15 tag (kategori).
4. Kategorisasi otomatis yang dipakai bisa digunakan untuk mendukung atau sebagai sistem pemberi saran bagi kategorisasi manual. Sejumlah responden dibutuhkan untuk menilai apakah kategorisasi otomatis yang dihasilkan telah sesuai dengan yang diinginkan, karena hasil kategorisasi bersifat subjektif.
5. Pada operasi indexing, rata-rata lama waktu pemrosesan adalah 4,22 detik per dokumen.

5.2 Saran

Dari hasil pembangunan sistem ini serta dari hasil uji coba yang telah dilakukan, saran-saran yang dapat diberikan antara lain sebagai berikut:

1. Dalam menggunakan metode multinomial naive Bayes, perlu dilakukan atau diadakan teknik yang menambahkan fungsionalitas kategorisasi selain dengan multinomial naive Bayes apalagi metode ini gagal melakukan fungsinya.
2. Metode multinomial naive Bayes baik digunakan untuk mendukung kategorisasi teks secara manual sebagai fungsionalitas pendukung.
3. Waktu indexing yang lama membutuhkan konfigurasi sistem hardware yang cukup tinggi agar dapat meningkatkan kinerja atau dapat ditambahkan optimalisasi dari segi perangkat lunak.

Daftar Pustaka

1. Baeza-Yates and Ribeiro-Neto. Modern Information Retrieval. Addison-Wesley. ISBN 020139829X.
2. C.J. van Rijsbergen. *Information Retrieval*. Butterworths. 1979 (second edition).
3. Mathes, Adam (December 2004). Folksonomies - Cooperative Classification and Communication Through Shared Metadata. Computer Mediated Communication - LIS590CMC. Graduate School of Library and Information Science. University of Illinois Urbana-Champaign. [http://www.adammathes.com/]
4. McCallum, Andrew, et al. (1998). A Comparison of Event Models for Naive Bayes Text Classification. AAAI/ICML-98 Workshop on Learning for Text Categorization. Technical Report WS-98-05, 1998.
5. Peng, Fuchun, et al. (2004). Augmenting Naive Bayes Classifiers with Statistical Language Models. Kluwer Academic Publishers, Netherlands. Information Retrieval, 7, 317-345, 2004.
6. Rennie, Jason D. M., et al. (2003). Tackling The Poor Assumptions of Naive Bayes Text Classifiers. 2003. Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.
7. Schneider, Karl-Michael (2004). On Word Frequency Information and Negative Evidence in Naive Bayes Text Classification. EsTAL – España for Natural Language Processing, October 20-22, 2004, Alicante, Spain.
8. Sebastiani, Fabrizio (2002). Machine Learning in Automated Text Categorization. ACM Computing Surveys, Vol. 34, No. 1, March 2002, pp. 1–47.
9. Smith, Gene (Aug 3, 2004). “Atomiq: Folksonomy: social classification.” [http://atomiq.org/archives/2004/08/folksonomy_social_classification.html]
10. Witten, Moffat, and Bell. Managing Gigabytes. Morgan Kaufmann. ISBN 1558605703.