SISTEM PENCARIAN INFORMASI MENGGUNAKAN HYBRID MODEL: VECTOR SPACE MODEL DAN JACCARD

Skripsi

oleh
SYAMSUL ARIES NUR F
22074298

PROGRAM STUDI TEKNIK INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
2012

SISTEM PENCARIAN INFORMASI MENGGUNAKAN HYBRID MODEL: VECTOR SPACE MODEL DAN JACCARD

Skripsi

Diajukan kepada Program Studi Teknik Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana
Sebagai Salah Satu Syarat dalam Memperoleh Gelar
Sarjana Komputer

Disusun oleh

SYAMSUL ARIES NUR F
22074298

PROGRAM STUDI TEKNIK INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
2012

PERNYATAAN KEASLIAN SKRIPSI

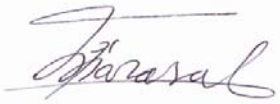
Saya menyatakan dengan sesungguhnya bahwa skripsi dengan judul:

SISTEM PENCARIAN INFORMASI MENGGUNAKAN HYBRID MODEL: VECTOR SPACE MODEL DAN JACCARD

yang saya kerjakan untuk melengkapi sebagian persyaratan menjadi Sarjana Komputer pada pendidikan Sarjana Program Studi Teknik Informatika Fakultas Teknologi Informasi Universitas Kristen Duta Wacana, bukan merupakan tiruan atau duplikasi dari skripsi kesarjanaan di lingkungan Universitas Kristen Duta Wacana maupun di Perguruan Tinggi atau instansi manapun, kecuali bagian yang sumber informasinya dicantumkan sebagaimana mestinya.

Jika dikemudian hari didapati bahwa hasil skripsi ini adalah hasil plagiasi atau tiruan dari skripsi lain, saya bersedia dikenai sanksi yakni pencabutan gelar kesarjanaan saya.

Yogyakarta, 28 November 2012



SYAMSUL ARIES NUR F
22074298

HALAMAN PERSETUJUAN

Judul Skripsi	:	SISTEM PENCARIAN INFORMASI MENGGUNAKAN HYBRID MODEL: VECTOR SPACE MODEL DAN JACCARD
Nama Mahasiswa	:	SYAMSUL ARIES NUR F
N I M	:	22074298
Matakuliah	:	Skripsi (Tugas Akhir)
Kode	:	TIW276
Semester	:	Gasal
Tahun Akademik	:	2012/2013

Telah diperiksa dan disetujui di
Yogyakarta,
Pada tanggal 28 November 2012

Dosen Pembimbing I	Dosen Pembimbing II
Antonius Rachmat C., SKom.,M.Cs	Joko Purwadi, M.Kom

HALAMAN PENGESAHAN

SISTEM PENCARIAN INFORMASI MENGGUNAKAN HYBRID MODEL: VECTOR SPACE MODEL DAN JACCARD

Oleh: SYAMSUL ARIES NUR F / 22074298

Dipertahankan di depan Dewan Penguji Skripsi
Program Studi Teknik Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana - Yogyakarta
Dan dinyatakan diterima untuk memenuhi salah satu syarat memperoleh gelar
Sarjana Komputer
pada tanggal 26 November 2012

Yogyakarta, 29 November 2012
Mengesahkan,

Dewan Penguji:

1. Antonius Rachmat C., SKom.,M.Cs
2. Joko Purwadi, M.Kom
3. Aditya Wikan Mahastama, S.Kom
4.

Dekan	Ketua Program Studi
(Drs. Wimmie Handiwidjojo, MIT.)	(Nugroho Agus Haryono, M.Si)

UCAPAN TERIMAKASIH

Puji dan syukur saya haturkan kepada Tuhan Yang Maha Esa karena berkat bimbingan dan tuntunan-Nya saya bisa menyelesaikan tugas akhir ini. Saya juga mengucapkan terimakasih kepada Ibu, saudara, dosen pembimbing, dan sahabat-sahabat saya karena selama ini terus memberikan dukungan kepada saya.

INTISARI
SISTEM PENCARIAN INFORMASI MENGGUNAKAN HYBRID MODEL, VECTOR SPACE MODEL DAN JACCARD

Pemanfaatan teknologi komputer dan internet dalam menyimpan dan meneruskan informasi terus berkembang. Sehingga terdapat banyak sekali informasi yang disimpan di dalamnya. Apabila dilakukan suatu pencarian informasi dengan cara manual maka akan membutuhkan waktu yang sangat lama. Selain itu manusia juga memerlukan informasi yang sesuai dengan kebutuhannya. Untuk itu diperlukan suatu metode yang bisa digunakan untuk mendapatkan kembali informasi yang sering disebut juga dengan sistem pencarian informasi.

Pada tugas akhir ini, penulis membuat sistem pencarian informasi dengan menggabungkan *vector space model* dan *jaccard similarity measure* (*hybrid model*). Tahap pertama adalah memberikan bobot vektor pada dokumen dan *query*. Selanjutnya menghitung nilai relevansi antara dokumen dan *query* dengan menggunakan *jaccard similarity measure*. Semakin tinggi nilai relevansi yang didapat maka dokumen tersebut semakin relevan.

Hasil dari pencarian informasi yang dilakukan dengan *hybrid model* kemudian dievaluasi untuk mengetahui seberapa bagus performa yang dimiliki dengan menggunakan metode pembanding, yaitu: *inner product* dan *cosine similarity measure*. Hasil evaluasi menujukkan bahwa *hybrid model* menunjukkan precision yang masih dibawah *inner product* dan *cosine similarity measure* dalam menangani dokumen berita.

Kata kunci: *information retrieval, vector space model, jaccard similarity measure, inner product, cosine similarity measure*

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

INTISARI
SISTEM PENCARIAN INFORMASI MENGGUNAKAN HYBRID MODEL, VECTOR SPACE MODEL DAN JACCARD

Pemanfaatan teknologi komputer dan internet dalam menyimpan dan meneruskan informasi terus berkembang. Sehingga terdapat banyak sekali informasi yang disimpan di dalamnya. Apabila dilakukan suatu pencarian informasi dengan cara manual maka akan membutuhkan waktu yang sangat lama. Selain itu manusia juga memerlukan informasi yang sesuai dengan kebutuhannya. Untuk itu diperlukan suatu metode yang bisa digunakan untuk mendapatkan kembali informasi yang sering disebut juga dengan sistem pencarian informasi.

Pada tugas akhir ini, penulis membuat sistem pencarian informasi dengan menggabungkan *vector space model* dan *jaccard similarity measure* (*hybrid model*). Tahap pertama adalah memberikan bobot vektor pada dokumen dan *query*. Selanjutnya menghitung nilai relevansi antara dokumen dan *query* dengan menggunakan *jaccard similarity measure*. Semakin tinggi nilai relevansi yang didapat maka dokumen tersebut semakin relevan.

Hasil dari pencarian informasi yang dilakukan dengan *hybrid model* kemudian dievaluasi untuk mengetahui seberapa bagus performa yang dimiliki dengan menggunakan metode pembanding, yaitu: *inner product* dan *cosine similarity measure*. Hasil evaluasi menujukkan bahwa *hybrid model* menunjukkan precision yang masih dibawah *inner product* dan *cosine similarity measure* dalam menangani dokumen berita.

Kata kunci: *information retrieval, vector space model, jaccard similarity measure, inner product, cosine similarity measure*

BAB 1

PENDAHULUAN

1.1. Latar Belakang Masalah

Manusia selalu membutuhkan informasi di dalam kehidupan sehari-hari. Pada awalnya manusia mendapatkan informasi melalui buku, surat kabar, peninggalan bersejarah, dan sebagainya. Seiring dengan perkembangan teknologi, teknik yang digunakan untuk mendapatkan kembali informasi menjadi semakin baik. Bayangkan jika informasi dicari dengan membuka ratusan berkas atau buku secara manual. Hal itu bisa memakan banyak waktu. Selain itu ketepatan informasi yang didapatkan juga sangat penting. Seiring dengan perkembangan teknologi komputer dan internet metode untuk mendapatkan informasi juga dikembangkan menjadi semakin baik. Sehingga manusia bisa mendapatkan informasi yang tepat. Kebutuhan akan informasi yang bagus memicu pengembangan berbagai metode untuk mendapatkan kembali informasi atau disebut dengan sistem pencarian informasi (*information retrieval*).

Sistem pencarian informasi mengambil informasi yang berasal dari data teks. Proses pencarian dimulai ketika pengguna memasukkan kata kunci (*query*) ke dalam sistem. *Query* adalah data teks yang berisi informasi yang diinginkan pengguna. Kemudian dilakukan proses pencarian untuk mendapatkan hasil yang sesuai dengan *query* yang diberikan. Dengan sistem pencarian informasi maka proses mendapatkan informasi bisa menjadi lebih mudah dan cepat.

Saat ini ada banyak sekali metode yang dipakai dalam sistem pencarian informasi. Metode-metode tersebut terus dikembangkan dan diperbaharui untuk mendapatkan hasil pencarian yang lebih baik. Maka dari itu penulis akan

membuat sistem pencarian dokumen teks dengan menggabungkan *vector space model* dan *jaccard similarity measure* atau bisa juga disebut dengan *hybrid model*. Kemudian untuk mengukur tingkat akurasi, *precision* dari hasil pencarian dari *jaccard similarity measure* akan dibandingkan dengan hasil pencarian dari *cosine similarity measure*.

1.2. Perumusan Masalah

Masalah yang akan dibahas di dalam penelitian ini adalah sebagai berikut :

a. Berapa waktu yang dibutuhkan dalam proses pencarian yang dilakukan dengan hybrid model, *inner product* dan *cosine similarity measure*?
b. Bagaimana performa hasil pencarian dengan menggunakan *hybrid model* dibandingkan dengan *inner product* dan *cosine similarity measure*?

1.3. Batasan Masalah

a. Tidak ada proses *stemming* yang dilakukan terhadap dokumen teks dalam perhitungan nilai relevansi dokumen tersebut terhadap suatu *query*.
b. Dokumen yang digunakan adalah teks berita dan berbahasa indonesia.
c. Sistem pencarian informasi dibuat berbasiskan web dan dijalanan secara offline.
d. Kata kunci pencarian menggunakan hubungan antar token atau (*or*).
e. Jumlah maksimum kolom per tabel pada *database* mysql adalah 4096 kolom.

1.4. Hipotesis

Dugaan hasil akhir dari penelitian yang akan dilakukan adalah sebagai berikut :

a. Sistem IR yang dibuat dengan metode *hybrid model* ini mampu menampilkan dokumen-dokumen teks yang isinya relevan dengan *query* yang dimasukkan.

b. Sistem pencarian informasi yang menggunakan *jaccard similarity measure* bisa memberikan hasil pencarian yang lebih baik jika dibandingkan dengan yang menggunakan *cosine similarity measure.*

1.5. Tujuan Penelitian

a. Tujuan Penulisan

- Tujuan bagi mahasiswa :

  Mahasiswa mampu menerapkan ilmu yang didapat selama menempuh kuliah di Universitas Kristen Duta Wacana Yogyakarta dalam sebuah karya ilmiah.

- Tujuan bagi universitas :

  Universitas dapat mengkaji sejauh mana kemampuan mahasiswa dalam mengimplementasikan ilmu yang telah diberikan selama kuliah.

b. Tujuan Penelitian

- Tujuan Utama :

  Menerapkan dan meneliti keakuratan hybrid model, *vector space model* dan *jaccard similarity measure* untuk merangking dokumen berdasarkan kata kunci (*query*) tertentu.

- Sub-Tujuan :

  a. Mempermudah proses pencarian dokumen.

1.6. Metode/Pendekatan

Berikut ini adalah tiga jenis metodologi beserta tahapan yang akan digunakan oleh peneliti dalam melaksanakan penelitian ini :

a. Metode Pengumpulan Data

Untuk mengumpulkan data penulis mengambil artikel dari *website* detik.com. Artikel yang digunakan terdiri dari enam kategori yaitu : keuangan, internet dan teknologi, kesehatan, berita umum, olahraga, dan sosok.

b. Metode Pengembangan Sistem

Metode yang digunakan untuk mengembangkan sistem yang dipakai sebagai instrument penelitian adalah *hybrid model, vector space model* dan *jaccard similarity measure*. Metode ini digunakan dalam proses pencarian informasi pada sistem yang akan dibuat.

c. Metode Evaluasi

Evaluasi hasil pencarian dilakukan dengan menggunakan metode *eleven-point interpolated average precision*. Bahan yang digunakan untuk evaluasi adalah *query* dan dokumen yang relevan dengan *query* tersebut. Keduanya telah ditentukan sebelumnya. Hasil dari perhitungan nilai *eleven-point interpolated average precision* itulah yang akan menentukan tingkat akurasi dari sistem yang dibuat.

1.7. Sistematika Penulisan

Laporan Tugas Akhir ini terdiri dari lima bab sebagai berikut :

Bab 1 : Pendahuluan. Berisi latar belakang masalah, perumusan masalah, batasan masalah, hipotesis, tujuan penelitian, metode/pendekatan dan sistematika penulisan.

Bab 2 : Landasan Teori. Bagian ini berisi dasar teori yang diperlukan dalam membuat tugas akhir ini yaitu: teori sistem pencarian informasi, *vector space model, jaccard similarity measure, cosine similarity measure* beserta metode evaluasi yang digunakan untuk mengukur kemampuan dari sistem pencarian informasi pada tugas akhir ini.

Bab 3 : Gambaran Sistem. Bagian ini berisi tentang analisis dan perancangan sistem yang meliputi arsitektur sistem, *database* sistem, spesifikasi sistem, perancangan proses, dan perancangan antar muka.

Bab 4 : Implementasi dan Analisis Sistem. Bab ini akan memuat hasil riset/implementasi serta pembahasan/analisis dari riset yang dilakukan.

Bab 5 : Kesimpulan dan Saran. Bagian ini berisi kesimpulan yang didapatkan setelah melakukan analisis pada sistem dan saran untuk pengembangan program pada masa yang akan datang.

BAB 5

KESIMPULAN DAN SARAN

5.1. Kesimpulan

Dari penelitian yang dilakukan oleh penulis, didapatkan hasil pengujian melalui metode evaluasi yang digunakan. Melalui analisis hasil pengujian, didapatkan kesimpulan sebagai berikut :

1. *Jaccard similarity measure* memiliki *precision* yang paling jelek dengan rata-rata yang lebih rendah 0.16 poin dari *inner product* dan lebih rendah 0.18 poin dari *cosine similarity measure* ketika menangani dokumen berita.
2. Cosine similarity measure memiliki *precision* terbaik dengan rata-rata lebih tinggi 0.18 poin dari *jaccard similarity measure* dan lebih tinggi 0.02 poin dari *inner product*.
3. *Jaccard similarity measure*, *inner product*, dan *cosine similarity measure* memiliki waktu pencarian yang relatif sama, yaitu antara 2 sampai 8 detik.

5.2. Saran

Berikut ini adalah saran yang diperlukan untuk melakukan pengembangan proses pencarian sehingga memberikan hasil yang lebih baik:

1. Perlu dilakukan penelitian untuk mencari isi korpus data atau jenis dokumen teks yang paling sesuai dengan metode *jaccard similarity measure*.
2. Korpus data yang digunakan perlu dikembangkan sehingga tidak terbatas pada dokumen teks (.txt).

DAFTAR PUSTAKA

Baeza-Yates, R., & Riberio-Neto, B. (1999). *Modern Information Retrieval*. Harlow: Addisson-Wesley.

Cha, S. (2007). *Comprehension Survey on Distance/Similarity Measures between Probability Density Functions.* International Journal of Mathematical Models and Methods in Applied Science. Vol 1, 300-307.

Chaer, A. (1988). *Tata Bahasa Praktis bahasa Indonesia*. Jakarta: Bhratara Karya Aksara.

Grossman, D.A., & Frieder, F. (2004). *Information Retrieval Algorithm and Heuristics*. Dordrecht: Springer.

Jones, W.P., & Furnas G.W. (1987). *Pictures of Relevance: A Geometric Analysis of Similarity Measure*. Journal of The American Society for Information Science. New York: Wiley & Sons Inc. Vol 38, 420-442.

Kowalski, G. J., & Maybury, M. T. (2002). *Information Storage and Retrieval System*. New York: Kluwer Academic Publisher.

Leydesdorff, L. (2008). *On the Normalization of Author Co-Citation Data: Salton's Cosine versus the Jaccard Index*. Journal of the American Society of Information Science & technology. New York: Wiley & Sons Inc. Vol 59, 77-85.

Manning, C.D., Raghavan, P., & Schűtze, H. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Moens, M. (2006). *Information Extraction: Algorithms and Prospects in a Retrieval Context*. Dordrecht: Springer.

Santhisree, K., & Damodaram, A. (2011). *SSM-DBSCAN and SSM-OPTIC: Incorporating a new similarity measure for Density based Clustering of Web Usage data*. International Journal on Computer Science and Engineering. India: Engg Journals Publications. Vol 3, 3071-3083.

Tan, P., Steinbach, M., & Kumar, V. (2006). *Introduction to Data mining*. Boston: Addison-Wesley.

Vikas, O., & Arora, P. (2010). *Ranking Strategy Using Hybrid Model*. International Journal of Computer Application. Volume 5, 10-15.

Witten, I.H., & Frank E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. San Fransisco: Morgan Kaufmann Publishers.

Zang, J. (2008). *Visualization for Information Retrieval*. Berlin: Springer.