IMPLEMENTASI ALGORITMA OKAPI BM25 DAN K-MEANS UNTUK MENCARI RELEVANSI ARTIKEL PADA BEBERAPA SITUS BERITA

Tugas Akhir

Oleh :

Danny Sebastian

22074222

Program Studi Teknik Informatika Fakultas Teknologi Informasi

Universitas Kristen Duta Wacana Yogyakarta

Tahun 2011

IMPLEMENTASI ALGORITMA OKAPI BM25 DAN K-MEANS UNTUK MENCARI RELEVANSI ARTIKEL PADA BEBERAPA SITUS BERITA

Tugas Akhir

Diajukan kepada Fakultas Teknik Teknologi Informasi

Universitas Kristen Duta Wacana Sebagai salah satu syarat

dalam memperoleh gelar Sarjana Komputer

Oleh :

Danny Sebastian

22074222

Program Studi Teknik Informatika Fakultas Teknologi Informasi

Universitas Kristen Duta Wacana Yogyakarta

Tahun 2011

PERNYATAAN KEASLIAN TUGAS AKHIR

Saya menyatakan dengan sesungguhnya bahwa tugas akhir dengan judul :

IMPLEMENTASI ALGORITMA OKAPI BM25 DAN K-MEANS UNTUK MENCARI RELEVANSI ARTIKEL PADA BEBERAPA SITUS BERITA

Yang saya kerjakan untuk melengkapi sebagian persyaratan menjadi Sarjana komputer pada pendidikan sarjana Program Studi Teknik Informatika, Fakultas Teknik 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 skripsi ini adalah hasil plagiasi atau tiruan dari skripsi lain, saya bersedia dikenai sanksi yakni pencabutan gelar kesarjanaan saya.

Yogyakarta, 10 Juni 2011

(Danny Sebastian)

22074222

HALAMAN PERSETUJUAN

Judul	: IMPLEMENTASI ALGORITMA OKAPI BM25 DAN K-MEANS UNTUK MENCARI RELEVANSI ARTIKEL PADA BEBERAPA SITUS BERITA
Nama	: Danny Sebastian
NIM	: 22074222
Mata Kuliah	: Tugas Akhir
Kode	: TI2126
Semester	: Genap
Tahun akademik	: 2010/2011

Telah diperiksa dan disetjui
Di Yogyakarta, 10 Juni 2011
Pada tanggal

Dosen Pembimbing I	Dosen Pembimbing II
Antonius Rachmat, S.Kom, M.Cs	Willy Sudiarto Raharjo, S.Kom, M.Cs

HALAMAN PENGESAHAN

SKRIPSI

IMPLEMENTASI ALGORITMA OKAPI BM25 DAN K-MEANS UNTUK MENCARI RELEVANSI ARTIKEL PADA BEBERAPA SITUS BERITA

Oleh Danny Sebastian / 22074222

Dipertahankan di depan dewan Penguji Tugas Akhir
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
10 Juni 2011

Yogyakarta, 20/6/2011
Mengesahkan,

Dewan Penguji

1. Willy Sudiarto R, S.Kom, M.Cs.
2. Nugroho Agus Haryono, S.Si, M.Si.
3. Drs. R. Gunawan Santosa, M.Si.

Dekan

(Drs. Wimmie Handiwidjojo, MIT.)

Ketua Program Studi

(Nugroho Agus Haryono, S.Si,MSi.)

UCAPAN TERIMA KASIH

Terima kasih. Yang pertama dan yang terutama kepada Tuhan Yesus Kristus atas segala hikmat, berkat dan kasih karunia-Nya dalam hidup saya, segala kemuliaan hanya bagi Engkau.

Kepada Papa, Mama, dan Cicik atas segala doa, kasih sayang dan kesabaran yang selalu melimpah bagi saya. Terima kasih atas pengalaman hidup yang Papa dan Mama ajarkan bagi saya. Tuhan selalu berkati Papa, Mama, dan Cicik.

Terima kasih juga kepada dosen pembimbing I, Bapak Antonius Rachmat C. Terima kasih atas segala kesabaran selama membimbing pembuatan Tugas Akhir dan pengetahuan yang telah Bapak berikan. Tuhan selalu Berkati Pak Anton sekeluarga.

Kepada dosen pembimbing II, Bapak Willy Sudiarto R. Terima kasih atas segala kesabaran dan pengetahuan yang diberikan. Tuhan selalu Berkati Pak Willy Sekeluarga.

Kepada teman-teman yang telah memberikan penghiburan dikala susah. Tertawa itu menular, dan hati yang gembira adalah obat yang manjur. Tuhan berkati kalian semua.

Kepada pihak-pihak yang tidak dapat saya sebutkan satu persatu yang telah membantu baik secara langsung ataupun tidak langsung. Tuhan berkati kalian semua.

Akhir kata saya, selaku penulis ingin meminta maaf bila ada kesalahan baik dalam penyusunan laporan maupun yang pernah saya lakukan sewaktu membuat program Tugas Akhir.

Yogyakarta, 30 Mei 2011

Penulis

INTISARI

IMPLEMENTASI ALGORITMA OKAPI BM25 DAN K-MEANS UNTUK MENCARI RELEVANSI ARTIKEL PADA BEBERAPA SITUS BERITA

Media informasi semakin berkembang pesat. Banyak media elektronik yang dibangun untuk meningkatkan penyebaran informasi. Seiring dengan perkembangan teknologi, tuntutan masyarakat akan kebutuhan media informasi yang semakin mudah diakses pun semakin meningkat. Media informasi yang mulai dipilih sebagai alternative penyampaian informasi adalah *website*. Seiring dengan berkembangnya *website*, semakin banyak pula *website* bertemakan berita yang bermunculan. Dengan semakin banyaknya webste berita, *website-website* berita harus bersaing untuk meningkatkan jumlah pengunjung dengan cara meningkatkan jumlah artikel berita yang mereka miliki. Oleh karena itu, dibutuhkan cara pencarian artikel guna memudahkan pengguna mencari artikel.

Untuk mempermudah hal tersebut penulis membangun sebuah sistem yang dapat mengumpulkan artikel dari beberapa situs berita. Penulis menggunkan 2 algoritma yang efektif untuk memudahkan pencarian artikel, yaitu algoritma Okapi BM25 dan K-Means. Kombinasi kedua algoritma tersebut bertujuan untuk mendapatkan hasil pencarian yang lebih efektif. Dalam penelitian ini, penulis melakukan pembandingan antara pencarian menggunakan Okapi BM25 tanpa menggunakan clustering dan pencarian menggunakan Okapi BM25 dengan menggunakan clustering.

Hasil yang didapat dari penelitian ini adalah kombinasi kedua algoritma tidak memberikan manfaat yang signifikan terhadap performa system ketika dibandingkan dengan pencarian tanpa metode clustering. Pada penelitian ini pemrosesan tanpa clustering memiliki nilai rata-rata precision dan recall sebesar 72.73% dan 97.38%. sedangkan pemrosesan dengan menggunakan clustering memiliki rata-rata precision dan recall sebesar 72.76% dan 80.10%.

DAFTAR ISI

LAMPIRAN

DAFTAR TABEL

DAFTAR GAMBAR

BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Sekarang ini, media informasi semakin berkembang pesat. Banyak media elektronik yang dibangun untuk meningkatkan penyebaran informasi. Seiring dengan perkembangan teknologi, tuntutan masyarakat akan kebutuhan media informasi yang semakin mudah diakses pun semakin meningkat. Media informasi yang mulai dipilih sebagai alternative penyampaian informasi adalah *website*. Seiring dengan berkembangnya *website*, semakin banyak pula *website* bertemakan berita yang bermunculan. Dengan semakin banyaknya webste berita, *website-website* berita harus bersaing untuk meningkatkan jumlah pengunjung dengna cara meningkatkan jumlah artikel. Oleh karena itu, dibutuhkan cara pencarian artikel guna memudahkan pengguna mencari artikel. Cara pencarian itu dikembangkan dengan metode *information retrieval*.

Information retrieval adalah salah satu metode untuk menampilkan sebuah dokumen yang sesuai dengan permintaan user. *Information retrieval* memiliki banyak metode, beberapa diantaranya adalah Okapi BM11, Okapi BM15, Vector Space Model, Okapi BM25, dan lain sebagainya. Sedangkan clustering merupakan pengelompokan dokumen sehingga dokumen dapat lebih mudah untuk dicari. Beberapa metode clustering adalah K-means, C-Means, Fuzzy C-Means, dan lain sebagainya.

Penulis melihat terlalu banyak artikel berita sehingga membuat masyarakat yang mencari informasi kesulitan mendapatkan artikel yang sesuai. Sistem yang dibuat adalah sebuah *website* dengan layanan yang mengumpulkan artikel berita dari beberapa sumber *website* berita secara online, dan menampilkan artikel berita berdasarkan kata kunci yang dimasukkan oleh pengguna. Maka di angkat kasus:

“Implementasi algoritma Okapi BM25 dan K-Means untuk mencari relevansi artikel pada beberapa *website* berita”.

1.2 Perumusan Masalah

Dengan demikian, masalah yang akan diteliti dirumuskan sebagai berikut:

- Bagaimana implementasi algoritma Okapi BM25 untuk mencari artikel berita yang memiliki relevansi dengan kueri yang dimasukkan pengguna.
- Bagaimana implementasi algoritma K-Means untuk mengelompokkan artikel berita kedalam cluster.
- Bagaimana keakuratan pencarian berita berdasarkan *recall* dan *precision* dari algoritma yang digunakan.

1.3 Batasan Masalah

Pada permasalahan ini, batasan masalah yang digunakan dalam pembangunan aplikasi adalah :

- Sistem yang dibangun berupa aplikasi web yang akan diuji pada jaringan local.
- Artikel berita yang diambil berasal dari RSS dan di dapat dari situs berita http://www.bbc.co.uk/indonesia , www.kompas.com, dan www.vivanews.com.
- Aplikasi yang di bangun akan melakukan pemrosesan untuk berita dengan waktu terbit 3 hari sebelumnya.
- Artikel berita yang diproses adalah artikel berita dalam bahasa Indonesia.
- Metode pembobotan yang digunakan adalah metode TF-IDF.
- Masukan sistem adalah sebuah kueri yang memiliki batas maksimum 5 kata.
- Penelitian ini tidak memperhatikan proses caching atau teknik-teknik lain

yang dapat mempercepat proses *information retrieval*.

- Sistem yang dibangun tidak melihat konteks dan makna pada query.

1.4 Hipotesis

Hipotesis sementara penelitian ini adalah:

- Sistem IR yang dibangun dengan menggunakan metode Okapi BM25 dapat menampilkan dokumen yang relevan dengan kueri yang dimasukkan oleh pengguna.
- Sistem IR yang dibangun dengan menggunakan kombinasi metode Okapi BM25 dan K-Means dapat menampilkan dokumen yang relevan dengan kueri yang dimasukkan oleh pengguna.

1.5 Tujuan Penelitian

- Menentukan relevansi artikel berita dengan kueri dengan menggunakan algoritma Okapi BM25.
- Menentukan cluster untuk artikel berita dengan algoritma K-Means.
- Memudahkan masyarakat yang ingin mencari berita yang relevan dengan kueri dari beberapa *website* berita.

1.6 Metode / Pendekatan

Pada program aplikasi ini, beberapa metode / pendekatan digunakan untuk membantu penyelesaian masalah, yaitu :

- Pencarian *website* di internet yang menyajikan informasi berita.
- Pencarian jurnal di internet tentang Okapi BM25 dan K-Means
- Perancangan Program

  Berdasarkan data yang diperoleh, akan dibuat suatu rincian sistem yang

meliputi :

- Penentuan urutan proses-proses yang terjadi dalam sistem.
- Perancangan antar muka dan gambaran kerja aplikasi yang akan dibangun.

- Pengujian Program

  Pengujian program aplikasi dilakukan dengan dua cara :

  - Melihat *precision* dan *recall* untuk proses pencarian kueri dari pengguna tanpa menggunakan proses clustering.
  - Melihat *precision* dan *recall* untuk proses pencarian kueri dari pengguna dengan menggunakan proses clustering.

1.7 Sistematika Penulisan

Sistematika penulisan laporan tugas akhir ini akan dibagi menjadi beberapa bab, yaitu :

Bab 1 Pendahuluan, berfungsi untuk memberikan gambaran umum tentang penelitian yang dilakukan oleh penulis. Pada bagian pendahuluan ini berisi tentang latar belakang masalah, perumusan masalah, batasan masalah, hipotesis, tujuan penelitian, metodologi penelitian, dan sistematika penulisan.

Bab 2 Tinjauan Pustaka, yang berisi 2 bagian utama, yaitu tinjauan pustaka dan landasan teori. Bagian tinjauan pustaka akan menguraikan teori yang didapatkan dari berbagai sumber pustaka yang penulis gunakan dalam penelitian ini. Untuk landasan teori, akan memuat penjelasan tentang konsep dan prinsip utama yang digunakan dalam pemecahan masalah.

Bab 3 Perancangan Sistem, bab ini berisi mengenai kebutuhan hardware dan software minimum yang digunakan penulis dan dibutuhkan pengguna, spesifikasi sistem yang akan dibuat, arsitektur sistem, diagram use case, algoritma dan flowchart, kamus data, diagram skema, rancangan antarmuka sistem dan rancangan pengujian

terhadap sistem.

Bab 4 Implementasi dan Analisis Sistem, berisi pembahasan implementasi dan pengujian sistem yang ada pada bab 3, beserta hasil print-screen dan hasil analisis dari sistem yang dibuat.

Bab 5 Kesimpulan dan Saran, merupakan bagian kesimpulan dari hasil penelitian yang dilakukan, dan juga berisi saran untuk riset.

BAB 5
KESIMPULAN & SARAN

5.1 Kesimpulan

Berdasarkan penelitian yang telah dilakukan, penulis mengambil beberapa kesimpulan, antara lain:

1. Pada kasus yang penulis angkat, Metode Okapi BM25 mampu menampilkan dokumen yang relevan terhadap sebuah kueri pada peringkat atas.
2. Pada kasus yang penulis angkat, Metode K-Means dapat menghasilkan cluster yang berisikan dokumen yang mirip satu sama lain ketika memiliki nilai centroid awal yang sesuai.
3. Metode Okapi BM25 menganggap kueri sebagai token yang tidak memiliki kaitan satu sama lain, sehingga urutan kueri tidak mempengaruhi perhitungan nilai relevansi/similarity.
4. Pada beberapa kasus, pengelompokan korpus data dengan menggunakan clustering dapat meningkatkan kinerja Okapi BM25. Pada korpus yang memiliki kelompok dokumen, dimana terdapat banyak kata yang sama pada dokumen yang terdapat pada kelompok dokumen yang berbeda.
5. Pada metode K-Means, dimensi vector posisi yang besar dapat mengurangi keakuratan clustering.
6. Pada penelitin yang dilakukan terhadap korpus data dengan tanggal 3-5 April 2011, 8-10 April 2011, 15-17 April 2011, 3-5 Mei 2011, 6-8 Mei 2011, dan 9-11 Mei 2011, Pemrosesan tanpa clustering memiliki nilai precision dan recall sebesar 72.73% dan 97.38%. sedangkan pemrosesan dengan menggunakan clustering memiliki precision dan recall sebesar 72.76% dan 80.10%.

5.2 Saran

Hal yang dapat diterapkan untuk pengembangan system adalah:

1. Penggunaan operator Boolean untuk pencarian dokumen.

2. Pada clustering dengan menggunakan K-Means, pemilihan centroid awal dan jumlah cluster sangat berpengaruh. Untuk meningkatkan hasil clustering, jumlah cluster yang dipilih sebaiknya merupakan jumlah kelompok dokumen yang mirip pada korpus data. Dan centroid awal sebaiknya diambil dari masing-masing kelompok dokumen tersebut.

DAFTAR PUSTAKA

Candra Dewi, Rukmana. (2010). ***Pencarian Dokumen Teks dengan Metode Okapi BM25.*** Universitas Kristen Duta Wacana, Yogyakarta.

Clough, Paul. (2001). ***Filtering Meter Corpus Documents Using a Probabilistic IR System(Okapi)***. Regent Court, University of Steffield, 2011 Portobello Street, Sheffield S1 4DP.

Grossman, David A, Dan Ophinr Frieder. (2004). ***Information Retrieval, algorithm and Heuristic***. 2nd Edition. Springer.

Kathuria, Ashish. Jansen , Bernard J. and Hafernik ,Carolyn. Spink, Amanda. (2010).***Classifying the User Intent of Web Queries Using K-Means Clustering***.

Mandala, Rila (2006). ***Evaluasi Efektifitas metode machine-learning pada Search-engine.*** URL: *http://journal.uii.ac.id* hal.13

Maning, Christoper D. (2008). ***Introduction to Information Retrieval.*** Cambridge University Press, New York.

Ramos, Juan. ***Using TF-IDF to Determine Word Relevance in Document Queries***. Department of Computer Science, Rutgers University, 23515 BPO Way, Piscataway, NJ, 08855