PERBANDINGAN PONTE-CROFT DAN JELINEK-MERCER SMOOTHING DALAM QUERY LIKELIHOOD MODEL

TUGAS AKHIR

Disusun oleh :

Martin Eric Gunawan

22074206

Program Studi Teknik Informatika Fakultas Teknologi Informasi

Universitas Kristen Duta Wacana

Tahun 2011

PERBANDINGAN PONTE-CROFT DAN JELINEK-MERCER SMOOTHING DALAM QUERY LIKELIHOOD MODEL

TUGAS AKHIR

Diajukan kepada Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana
Sebagai salah satu syarat dalam memperoleh gelar
Sarjana Komputer

Disusun oleh :
Martin Eric Gunawan
22074206

Program Studi Teknik Informatika
Universitas Kristen Duta Wacana
Tahun 2011

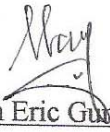
PERNYATAAN KEASLIAN TUGAS AKHIR

Saya menyatakan dengan sesungguhnya bahwa tugas akhir dengan judul :

Penambahan Jelinek-Mercer Smoothing Dalam Query Likelihood Model Untuk Peningkatan Akurasi Sistem Temu Kembali Informasi

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 plagiasi atau tiruandari skripsi lain, saya bersedia dikenai sanksi yakni pencabutan gelar kesarjanaan saya.

Yogyakarta, Juli 2011



(Martin Eric Gunawan)

22074206

HALAMAN PENGESAHAN

SKRIPSI

PERBANDINGAN PONTE-CROFT DAN JELINEK-MERCER SMOOTHING DALAM QUERY LIKELIHOOD MODEL

Oleh : Martin Eric Gunawan / 22 07 4206

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
26 Juli 2011

Yogyakarta, 11/8/2011

Mengesahkan,

Dewan Penguji:

1. Lucia Dwi Krisnawati, S.S., M.A.
2. Budi Susanto, S. Kom., M.T.
3. Antonius R, S. Kom, M. Cs
4. Yuan Lukito, S.Kom

Dekan

(Drs. Wimmie Handiwidjojo, MIT)

Ketua Program Studi

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

HALAMAN PERSETUJUAN

Judul	: Penambahan Jelinek-Mercer Smoothing dalam Query Likelihood Model untuk Peningkatan Akurasi Sistem Temu Kembali Informasi
Nama	: Martin Eric Gunawan
NIM	: 22074206
Mata Kuliah	: Tugas Akhir
Kode	: TI 2126
Semester	: Genap
Tahun Akademik	: 2010/2011

Telah diperiksa dan disetujui

Di Yogyakarta,

Pada tanggal 18 Juli 2011

Dosen Pembimbing 1	Dosen Pembimbing 2
Lucia D. Krisnawati	Bambang Soesanto,

UCAPAN TERIMA KASIH

Puji dan Syukur saya haturkan kepada Tuhan yang selalu membimbing saya dalam mengerjakan Tugas Akhir ini. Saya juga tidak lupa mengucapkan terima kasih kepada Papah, Mamah, dan Oh Henry yang sudah mendukung saya dalam doa sehingga saya dapat menyelesaikan Tugas Akhir ini. Saya juga tidak lupa mengucapkan terima kasih kepada Rudy yang telah menjadi sahabat dan teman untuk diskusi dalam mengerjakan Tugas Akhir ini. Terima kasih juga saya sampaikan kepada Ko Deddy yang telah membimbing dan memberikan masukan untuk pembangunan sistem yang saya buat.

Tidak Lupa saya juga berterima kasih kepada Andhri yang telah membantu saya dalam memberikan informasi dan trik-trik untuk optimalisasi *database* dan *code*. Saya tidak lupa mengucapan terima kasih kepada Aka yang telah menemani dan menyemangati saya dalam mengerjakan Tugas Akhir ini. Saya juga mengucapkan banyak terima kasih kepada Ibu Lucia Dwi Krisnawati dan Bapak Budi Susanto yang telah menjadi dosen pembimbing saya. Beliau-beliau inilah yang telah membantu saya dengan cara member masukan-masukan, kritik maupun saran terhadap Tugas Akhir saya.

Saya juga ingin mengucapkan terima kasih kepada keluarga kedua saya yaitu keluarga PSPP (Pusat Studi dan Pengembangan Perdamaian) dan Duta Voice. Semua rekan-rekan yang ada di 2 unit tersebut sangat mendukung saya dalam pengerjaan Tugas Akhir ini. Saya juga mendapat pengalaman yang sangat berharga dari setiap kegiatan yang saya ikuti selama bergabung dalam kedua unit tersebut. Satu lagi yang perlu saya ucapkan terima kasih yaitu para sahabat-sahabatku KC (Kebo Ceria) dan Hexa. Terima kasih semuanya, semoga kita semua sukses ^_^

Yogyakarta, Juli 2011

Penulis

INTISARI

Informasi merupakan hal yang sangat penting bagi semua orang. Informasi tersebut sebagian besar digunakan untuk menunjang pekerjaannya ataupun untuk menambah pengetahuan umum seseorang. Jumlah informasi yang sangat besar dan banyak merupakan suatu latar belakang untuk membuat suatu mesin pencarian informasi. Dewasa ini, mesin pencarian informasi tersebut makin lama makin canggih dan menghasilkan informasi yang sangat relevan bagi penggunanya. Mesin pencari informasi yang biasa dikenal dengan sistem temu kembali informasi (*Information Retrieval*) dapat menggunakan metode-metode tertentu. Tidak semua metode dapat menjawab kebutuhan dari pengguna.

Pada penelitian ini penulis membandingkan 2 (dua) metode *smoothing* pada *Query Likelihood Model* yang biasa digunakan untuk membangun *Information Retrieval* (IR) antara lain *Ponte Croft experience* dan *Jelinek Mercer*. Kedua metode tersebut memiliki kegunaan memberi bobot pada setiap token di tiap dokumen, sehingga sistem dapat menyajikan dokumen-dokumen yang kira-kira sesuai dengan kebutuhan pengguna.

Dari penelitian yang dibuat penulis ini terlihat bahwa performa dari kedua metode tersebut hampir sama. Kedua metode tersebut memiliki perbedaan hanya diposisi peringkat dokumen yang dianggap sistem paling relevan, sementara untuk jumlah dokumen yang ditemukan kembali sama.

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Pencarian informasi yang sulit menjadi pokok masalah pada saat *user* mencari suatu informasi dari suatu dokumen yang sangat besar dan banyak. Sistem pencarian informasi terus berkembang seiring berjalannya waktu, dengan penerapan suatu metode tertentu membuat sistem pencarian informasi tersebut menjadi otomatis dan mempermudah *user* untuk mencari informasi.

Information Retrieval (*IR*) merupakan suatu sistem pencari otomatis yang dapat menyelesaikan permasalahan pencarian yang lambat tersebut. Sistem ini biasanya berbasis *web* yang memiliki korpus yang sangat besar. Penerapan langkah – langkah *IR* antara lain pemotongan kata (tokenisasi), penghapusan *stopword*, pemberian bobot pada setiap kata hingga proses evaluasi dengan memperhitungkan *precision* dan *recall.*

Pada penelitian ini, penulis akan menggunakan metode *Query Likelihood Model* yang merupakan salah satu metode *IR* yang dapat memecahkan masalah pencarian informasi yang lambat tersebut. Pada penelitian ini pula, penulis akan menambahkan metode *Jelinek Mercer Smoothing* agar mendapatkan hasil yang lebih optimal.

Penulis berharap bahwa dengan penambahan *Jelinek – Mercer Smoothing* pada *Query Likelihood Model, user* mendapatkan hasil yang lebih cepat dan lebih relevan.

1.2 Rumusan Masalah

1. Bagaimana perbandingan kinerja metode smoothing *Ponte-Croft experience* dan *Jelinek-Mercer Smoothing Method*?

1.3 Batasan Masalah

Dalam tugas akhir ini, penulis membatasi penelitian dengan beberapa batasan yang didefinisikan di bawah ini:

1. Dokumen teks yang menjadi korpus data adalah *file plain text* yang berekstensi .txt.
2. Jumlah maksimal *file* .txt yang diteliti adalah 110 *file*.
3. Kamus data yang akan dipakai akan diambil dari *http://internasional.kompas.com/* dengan topik
4. *Query* yang diberikan tidak memperhitungkan operator boolean.
5. Tidak dilakukan proses *stemming* untuk setiap *token*.
6. Dalam pencarian, sistem bersifat contextual
7. Panjang *query* dibatasi maksimal 5 kata.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini yaitu dapat membandingkan performa dari metode *Ponte-Croft experience* dan metode *smoothing Jelinek-Mercer*.

1.5 Hipotesis

Dengan penelitian ini, penulis menduga dengan penggunaan metode *Query Ponte-Croft experience* nilai presisinya akan lebih baik daripada menggunakan metode *smoothing Jelinek-Mercer*.

1.6 Metode / Pendekatan

Tahap pertama yang dilakukan penulis yaitu mengumpulkan kamus data dengan cara menyalin dari situs – situs *web* dan kemudian akan disimpan kedalam *plain text* dengan format .txt. Proses kedua yang dilakukan yaitu menerapkan metode *Query Likelihood Model* pada program yang akan dibuat kemudian kamus data dimasukkan dan dilakukan proses evaluasi dengan memperhitungkan *precision* dan *recall*.

Proses ketiga yang dilakukan setelah metode *Query Likelihood Model* diterapkan adalah penulis akan menambahkan metode *Jelinek – Mercer Smoothing* pada program yang diteliti. Penulis akan menggunakan kamus data yang sama ketika melakukan penelitian dan evaluasi terhadap metode *Query Likelihood Model*. Hasil evaluasi dari *Query Likelihood Model* dan dengan penambahan *Smoothing* akan dibandingkan dan akan diperoleh hasilnya.

1.7 Sistematika Penulisan

Laporan yang akan dibuat oleh penulis terdiri dari 5 bab. Bab 1 PENDAHULUAN berisi tentang gambaran umum dari penelitian yang akan dilakukan oleh penulis. Bab 2 LANDASAN TEORI yang berisi tentang dasar teori dan hasil penelitian yang ditulis dalam jurnal – jurnal mengenai topik penelitian yang sama yang diambil oleh penulis. Bab 3 RANCANGAN SISTEM yang berisi tentang cara kerja sistem yang akan dibuat oleh penulis. Bab 4 IMPLEMENTASI DAN ANALISIS SISTEM yang berisi tentang hasil penelitian yang dilakukan dan yang telah diterapkan kedalam *program*. Bab 5 KESIMPULAN DAN SARAN yang berisi tentang jawaban dari hipotesis yang telah dibuat pada bab 1 dan saran – saran yang diberikan oleh penulis agar penerapan metode ini dapat menjadi lebih baik.

BAB 5
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Setelah melakukan penelitian mengenai penerapan metode *Ponte-Croft experience* dan metode *smoothing Jelinek-Mercer* dengan korpus dari http://internasional.kompas.com , kedua metode tersebut memiliki hasil temu kembali yang hampir sama. Hasil temu kembali dari kedua metode tersebut sangat dipengaruhi oleh probabilitas suatu token yang muncul. Jika token *query* jarang muncul dalam korpus, maka nilai probabilitasnya akan menjadi semakin kecil juga.

Pembobotan token untuk metode *Ponte-Croft* experience sangat dipengaruhi oleh probabilitas munculnya suatu token. Sedangkan metode *Jelinek-Mercer* dipengaruhi oleh perubahan nilai konstanta λ. Dari rentang nilai konstanta λ yang diuji ($0{,}1 \leq \lambda \leq 1$) maka dapat disimpulkan bahwa nilai konstanta λ semakin besar maka hasilnya akan mendekati sama dengan perhitungan bobot yang menggunakan *Ponte-Croft experience*. Hasil temu kembali informasi untuk kedua metode tersebut tidak terpengaruh dengan adanya *query* yang berupa kata-kata berulang.

Kedua metode ini lebih cocok diterapkan pada sistem yang jarang melakukan *update* data sebab waktu yang diperlukan untuk update data cukup lama. Waktu pengindeksan akan terus meningkat sesuai dengan dokumen yang dimasukkan ke dalam korpus sistem.

5.2 Saran

Untuk pengembangan sistem selanjutnya yang lebih baik, penulis memiliki beberapa saran diantaranya adalah pada metode *Jelinek Mercer* dapat diubah untuk

menguji nilai konstanta. Berdasarkan jurnal yang ditulis oleh Zhou dan Lafferty, semakin kecil nilai λ maka semakin cocok untuk *query* yang lebih panjang, tetapi jika nilai λ semakin besar maka semakin cocok untuk *query* yang pendek.

Penerapan kedua metode ini lebih baik tidak menggunakan basis data untuk media penyimpanan, sebab jika disimpan dalam basis data maka waktu yang diperlukan untuk melakukan *indexing* dan perhitungan akan sangat lama dan kurang efisien. Proses *indexing* untuk kedua metode tersebut menjadi sangat lama sebab setiap dokumen memiliki semua token yang tidak berada pada dokumen itu sendiri. Saran dari penulis adalah penyimpanan informasi token pada saat proses *indexing* dan perhitungan bobot token lebih baik disimpan pada suatu *file* teks sehingga ketika *user* akan melakukan *indexing* hanya memanggil *file* tersebut.

Penulis juga memiliki saran untuk masukkan kata kunci dapat menggunakan *operator boolean* sehingga dapat menggabungkan beberapa *query* sekaligus. Kemudian untuk panjang *query* dapat dibuat lebih dari 5 kata sehingga hasil temu kembalinya akan lebih mencerminkan *information need* dari *user*.

DAFTAR PUSTAKA

Grossman, David. A.,& Frieder, O.(2004). *Information Retrieval, Algorithms and Heuristic*. 2nd Edition. Springer.

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

Zhou, Croft.(2005). *Document Quality Model for Web Ad Hoc Retrieval.* University of Massachusetts, Amherst.

Zhai, Lafferty. *The Dual Role of Smoothing in the Language Modeling Approach.* Carnegie Mellon University.

Ponte, Croft. (1998). *A Language Model Approach to Information Retrieval.*

University of Massachusetts, Amherst.