

BAB I

PENDAHULUAN

Bab satu menjelaskan latar belakang, rumusan masalah, tujuan, batasan masalah, metodologi penyusunan tugas akhir, dan sistematika pembahasan laporan yang dibuat.

1.1 Latar Belakang

Information Retrieval pada dasarnya merupakan proses untuk menentukan dokumen dalam koleksi yang harus ditemubalikkan untuk memenuhi keinginan pengguna akan informasi. Informasi yang diinginkan pengguna direpresentasikan dalam bentuk *query* dan mengandung satu atau lebih *term* yang akan digunakan dalam pencarian, selain dapat juga ditambahkan informasi lain seperti bobot tiap *term* tersebut. Oleh karena itu, keputusan menemubalikkan dokumen dibuat dengan membandingkan *term-term* dari *query* dengan *term* indeks yang terdapat dalam dokumen tersebut. Keputusan yang diambil dapat merupakan keputusan biner (*retrieve / reject*), atau melibatkan perkiraan *relevance degree* dokumen terhadap *query*.

Suatu sistem temu balik informasi dikatakan ideal jika sistem tersebut dapat menemukan seluruh dokumen yang relevan dan sistem hanya menemukan dokumen yang relevan saja. Akan tetapi, *term-term* yang terdapat di dokumen dan di *query* sering memiliki banyak varian morfologik, sehingga pasangan *term* seperti “*computing*” dan “*computation*” tidak akan dianggap ekuivalen oleh sistem tanpa suatu bentuk *Natural Language Processing* (NLP).

Pada beberapa kasus, varian morfologik dari *term-term* memiliki interpretasi semantic yang sama dan dapat dianggap ekuivalen oleh sistem. Karena alasan tersebut, teknik *stemming* atau *stemmers* dikembangkan dengan tujuan mereduksi *term* menjadi bentuk akarnya. Dalam penggunaannya pada sistem temu balik informasi, *stem* tidak harus menghasilkan kata-kata yang bermakna, sehingga kata “*computation*” dapat direduksi menjadi “*comput*”. Perkembangan teknik *stemming* dalam bidang *Information Retrieval* telah mengakibatkan munculnya banyak riset mengenai algoritma yang dianggap tepat sebagai implementasi teknik ini.

Stemming adalah proses pemotongan (pembuangan) *affix*, baik *prefix* maupun *suffix*, dari sebuah *term*. *Affix* sering mengandung informasi seperti bagian dari percakapan, *plurality*, dan atau *tenses* yang dapat menurunkan performansi sistem. Dalam konteks sistem temu balik informasi, *stemming* digunakan untuk mereduksi bentuk *term* untuk menghindari ketidakcocokan yang dapat mengurangi *recall*, di mana *term-term* yang berbeda namun memiliki makna dasar yang sama direduksi menjadi satu bentuk. Dalam beberapa bahasa, penerapan teknik *stemming* adalah sebuah keharusan untuk menjaga performansi sistem. Sebagai contoh, dalam bahasa Hebrew, teknik ini meningkatkan jumlah dokumen yang berhasil ditemukan sepuluh sampai lima puluh kali lipat. Akan tetapi dalam bahasa Inggris peningkatan performansi tidak sedrastis bahasa Hebrew, meskipun *stemming* tetap diperlukan untuk membentuk sebuah sistem temu balik yang reliabel.

Sebagai contoh sederhana, jika dicari suatu dokumen dengan judul “*How to Write*” dengan menggunakan *query* “*writing*”, dokumen yang dimaksud tidak akan pernah terdapat dalam hasil pencarian. Akan tetapi, jika *query* dipotong sehingga *writing* diubah menjadi *write*, maka pencarian akan berhasil. Hal ini tidak hanya berarti varian yang berbeda dari suatu *term* dapat direduksi sebagai satu bentuk representative, tetapi juga berdampak terhadap berkurangnya ukuran inverted file (jumlah *term* yang diperlukan untuk merepresentasikan suatu koleksi dokumen). Ukuran inverted file yang kecil dapat menghemat *storage space* dan mempercepat waktu pemrosesan.

Algoritma Porter dibuat oleh Martin Porter dan pertama kali dipublikasikan pada tahun 1980. Algoritma ini merupakan algoritma pembuangan *suffix* yang *context sensitive*, terdiri dari lima *step* linear untuk menghasilkan *stem* akhir. Algoritma Paice/Husk dibuat oleh Chris Paice dengan bantuan Gareth Husk, dipublikasikan pada tahun 1990. Paice *stemmer* merupakan *conflation based iterative stemmer* yang sangat kuat dan agresif. Sedangkan algoritma Lovins dibuat oleh Beth Lovins, dipublikasikan pada tahun 1968, dan melakukan penghapusan *endings* berdasarkan prinsip *longest-match*.

1.2 Rumusan Masalah

Masalah – masalah yang akan dikaji dalam Tugas Akhir ini adalah :

1. Bagaimana mengimplementasikan teknik *stemming* dalam suatu sistem temu balik informasi?
2. Bagaimana perbandingan algoritma-algoritma *stemming* yang diimplementasikan, dipandang dari pengaruhnya terhadap nilai *recall*, *precision*, *non-interpolated average precision*, nilai rata-rata *modified Hamming distance* antara *term* dan *stem* yang dihasilkan, rata-rata jumlah *term* dalam suatu *conflation class*, banyaknya *term* yang diubah oleh *stemmer*, dan faktor kompresi index pada suatu sistem temu balik informasi?

1.3 Tujuan

Tujuan utama Tugas Akhir ini adalah melakukan studi dan implementasi penggunaan teknik *stemming* untuk meningkatkan performansi suatu sistem temu balik informasi. Adapun tujuan lain yang ingin dicapai dalam pelaksanaan Tugas Akhir adalah sebagai berikut:

1. Memahami penggunaan teknik *stemming* pada suatu sistem temu balik informasi.
2. Melakukan implementasi dari algoritma-algoritma yang dipilih (Porter, Paice/Husk, dan Lovins).
3. Melakukan perbandingan algoritma-algoritma yang digunakan berdasarkan kriteria tertentu, seperti nilai *recall*, *precision*, *non-interpolated average precision*, nilai rata-rata *modified Hamming distance* antara *term* dan *stem* yang dihasilkan, rata-rata jumlah *term* dalam suatu *conflation class*, banyaknya *term* yang diubah oleh *stemmer*, dan faktor kompresi index.

1.4 Batasan Masalah

Batasan-batasan yang didefinisikan dalam pelaksanaan Tugas Akhir ini adalah:

1. Koleksi dokumen yang digunakan untuk Tugas Akhir ini merupakan berkas teks dengan *query* dan *relevance judgments* yang telah ditentukan sebelumnya.
2. Bahasa yang digunakan adalah bahasa Inggris.
3. Algoritma *stemming* yang diimplementasikan hanya meliputi algoritma Porter, algoritma Paice/Husk, dan algoritma Lovins.

1.5 Metodologi

Dalam penyusunan Tugas Akhir ini akan digunakan metodologi sebagai berikut:

1. Eksplorasi dan Studi Literatur

Tahap ini dilakukan dengan cara mempelajari literatur-literatur baik yang berupa buku (*textbook*), jurnal dan artikel ilmiah, maupun *website* yang berhubungan dengan *information retrieval systems* dan algoritma-algoritma *stemming*.

2. Analisis Penyelesaian Masalah

Melakukan analisis terhadap berbagai algoritma *stemming* dan teknik yang digunakan untuk membandingkan algoritma-algoritma tersebut.

3. Analisis dan Perancangan Perangkat Lunak

Melakukan analisis dan perancangan terhadap perangkat lunak sistem temu balik informasi yang akan dibangun, termasuk menentukan lingkungan pembuatan, bahasa pemrograman yang digunakan, arsitektur, fungsionalitas, dan antarmuka sistem.

4. Implementasi dan Pengujian Perangkat Lunak

Implementasi algoritma akan dilakukan berdasarkan hasil analisis dan perancangan algoritma pada tahap sebelumnya.

Pengujian algoritma akan dilakukan dengan menggunakan *input* berupa koleksi dokumen dan *relevance judgments*.

5. Evaluasi dan Analisis Hasil

Evaluasi dan analisis hasil dilakukan dengan membandingkan nilai *recall*, *precision*, faktor kompresi index, rata-rata jumlah *term* dalam suatu *conflation class*, banyaknya *term* yang diubah oleh *stemmer*, dan *modified Hamming distance* untuk tiap algoritma yang digunakan, serta membandingkan dengan sistem temu balik informasi yang tidak menggunakan teknik *stemming*.

1.6 Sistematika Pembahasan

Sistematika penulisan laporan tugas akhir ini adalah sebagai berikut:

1. **Bab I Pendahuluan**, berisi penjelasan mengenai latar belakang, rumusan masalah, tujuan, batasan masalah, metodologi, serta sistematika pembahasan yang digunakan untuk menyusun laporan tugas akhir.
2. **Bab II Landasan Teori**, berisi dasar teori yang digunakan dalam analisis, perancangan, dan implementasi tugas akhir.

3. **Bab III Analisis Penyelesaian Masalah**, berisi analisis terhadap penggunaan teknik *stemming* untuk meningkatkan performansi perangkat lunak sistem temu balik informasi.
4. **Bab IV Analisis dan Perancangan Perangkat Lunak**, berisi analisis dan perancangan perangkat lunak sistem temu balik informasi yang akan digunakan sebagai dasar tahap implementasi yang akan dilaksanakan berikutnya.
5. **Bab V Implementasi dan Pengujian Perangkat Lunak**, berisi implementasi perangkat lunak hasil perancangan beserta hasil pengujian perangkat lunak tersebut.
6. **Bab VI Penutup**, berisi kesimpulan dan saran yang didapatkan selama pelaksanaan Tugas Akhir.