BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Manusia tidak pernah luput dan lepas dari namanya permasalahan pencarian, dimanapun dankapanpun selalu ada model pencarian yang secara mutlak diperlukan.Dalam komputer, pencarianjuga menjadi hal yang sangat penting di dalamnya,dimana seluruh aktifitas IT (*Information Technology*) selalu berkaitan denganhal tersebut.

Saat ini teknologi informasi di Indonesia sudah berkembang sangat pesat.Masyarakat sudah banyak yang menggunakan teknologi informasi untuk berbagai keperluan, keuntungan dari penggunaan teknologi informasi diantaranya adalah peningkatan efesiensi dan efektifitas yang signifikan dalam menyampaikan berbagai informasi.

Dengan dasar peningkatan efisiensi dan efektifitas inilah,sistem temu kembali informasi merupakan salah satu teknologi yang sangat dibutuhkan dalam pencarian informasi yang cepat dan akurat. Penelitian mengenai sistem temu kembali telah banyak dilakukan baik pada media teks, suara, citra, maupun video.

Belakangan ini terdapat beragam sistem informasi berbasis teks, yaituinformasi yang disimpan dalam dokumen-dokumen berupa file text. Karenabanyaknya dokumen yang dapat disimpan, maka pengguna sistem informasimengalami kesulitan untuk mendapatkan informasi yang diinginkan dan penggunatidak dapat melihat dokumen satu demi satu untuk mendapatkan informasi yangtepat. Sehingga diperlukan suatu cara agar pengguna dapat mengakses informasisecara cepat dan tepat.

Pencarian terhadapseluruh isi dokumen yang tersimpan bukanlahsolusi yang tepat, mengingat pertumbuhanukuran data yang tersimpan umumnya.Temukembali informasi (*information retrieval*) bertujuan untuk membantupengguna dalam menemukan informasi yangrelevan dengan kebutuhan mereka dalamwaktu singkat. Akan tetapi banyak teknik-

teknikersebut yang tergantung pada bahasayang digunakan dalam dokumen. Untukmengembangkan teknik-teknik temu kembaliinformasi bagi dokumen teks berbahasaIndonesia, dibutuhkan perangkat pengujianuntuk Bahasa Indonesia. Salah satunyaadalah suatu koleksi dokumen dalam BahasaIndonesia sebagai pendekatan seragam dalamevaluasi sistem temu kembali informasi.

Model sistem temu kembali informasi menentukan detail sistem temu kembali informasi yaitu meliputi representasi dokumen maupun *query*, fungsi pencarian (*retrieval function*) dan notasi kesesuaian (*relevance notation*) dokumen terhadap *query*.

Model yang terdapat dalam Temu Kembali Informasi (*Information Retrieval*)terbagi dalam 3 model besar, yaitu:

1. *Boolean Model*

   Merupakan model IR sederhana yang merepresentasikan dokumen sebagai himpunan kata atau frase berdasarkan atas teori himpunan dan aljabar boolean.
2. *Vector Space Model*

   Merupakan model IR yang merepresentasikan dokumen dan *query* dalam bentuk vektor dimensional.
3. *Probabilistik Model*

   Merupakan model IR yang menggunakan *framework* probabilistik.

Salah satu model sistem temu kembali informasi yang paling awal digunakan adalah model *boolean*. Model *boolean*merepresentasikan dokumen sebagai suatu himpunan kata-kunci (*set of keywords*). Sedangkan *query* direpresentasikan sebagai ekspresi *boolean*. *Query* dalam ekspresi *boolean*merupakan kumpulan kata kunci yang saling dihubungkan melalui operator *boolean* seperti *AND, OR* dan *NOT* serta menggunakan tanda kurung untuk menentukan *scope* operator. Hasil pencarian dokumen dari model *boolean*adalah himpunan dokumen yang relevan (Mandala. 2002).

Model ruang vektor dan model probabilistik adalah model yang menggunakan pembobotan kata dan perangkingan dokumen. Hasil *retrieval* yang didapat dari model-model ini adalah dokumen terangking yang dianggap paling relevan terhadap *query*.

Dalam model ruang vektor, dokumen dan *query* direpresentasikan sebagai vektor dalam ruang vektor yang disusun dalam indeks *term*, kemudian dimodelkan dengan persamaan geometri. Sedangkan model probabilistik membuat asumsi-asumsi distribusi *term* dalam dokumen relevan dan tidak relevan dalam orde estimasi kemungkinan relevansi suatu dokumen terhadap suatu *query*.

Salah satu model sistem temu kembali informasi yang digunakan pada skripsi ini, yang paling sederhananamun produkif adalah model ruang vektor. Vektor model ini mempresentasikanterm yang terdapat pada dokumen dan query. Elemen vektor tersebut adalah bobot*term* yang menjadi penilaian dan perankingan dokumen. Dalam model ruang vektor ini hal yang perlu diperhatikan adalah pembobotan *term* (*term weighting*) (Arifin, 2002).

Metode pembobotan yang umumnya diunggulkan dalam beberapa penelitianmenggunakan model ruang vektor yaitu *Term Frequency Inverse DocumentFrequency* TF-IDF (Arifin, 2002). Dalam perhitungan bobot *term*, sekalipun *termfrequency* banyak digunakan, namun hal itu hanya mendukung proporsi jumlahdokumen yang dapat ditemukan kembali oleh proses pencarian pada sistem temukembali informasi. Sedangkan proporsi jumlah dokumen yang ditemukan dandianggap relevan untuk kebutuhan pengguna akan lebih meningkat bila vektorbobot tersebut menggunakan term yang jarang muncul pada koleksi dokumen.*Term* demikian diharapkan mampu mengelompokkan sejumlah dokumen yangmemuatnya, sehingga berbeda dengan seluruh anggota koleksi dokumen lain yangtidak memilikinya. Kriteria ini dapat diakomodasi dengan menghitung inversfrekuensi dokumen. Dengan digabungkannya kedua metode ini yaitu frekuensi kemunculan *term* (*Term Frequency*)dan invers frekuensi (*Inverse Document Frequency*) yang mengandung kata tersebut,diharapkan mampu meningkatkan proporsi jumlah dokumen yang dapatditemukan kembali dan yang dianggap relevan secara sekaligus. Sehingga kriteria*term* yang paling tepat adalah *term* yang sering muncul dalam dokumen secaraindividu, namun jarang dijumpai pada dokumen lainnya.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah, diperoleh perumusan masalah yaitu :

1. Bagaimana mengimplementasikan Information Retrieval dengan menggunakan metode metode model ruang vektor (Vector Space Model) pada perangkat lunak berupa mesin pencari?
2. Bagaimana mesin pencari dapat menghasilkan dokumen relevan dan teranking berdasarkan kata kunci yang dimasukan oleh pengguna pada pengelompokkan dokumen teks berbahasa Indonesia?

1.3 Batasan Masalah

Untuk menghindari melebarnya pembahasan yang ada, maka dibuatlah beberapa batasan sebagai berikut:

1. Aplikasi ini mencari dokumen pendek berupa abstrak skripsi berbahasa Indonesia yang berformatteks (.txt).
2. Query inputan yang digunakan adalah berupa kata, bukan simbol atau notasi tertentu.
3. Sistem yang akan dibangun dalam skripsi ini hanya akan melakukan proses *retrieval* terhadap dokumen yang disimpan dalam dataset lokal pada suatu direktori di *web server*.

1.4 Tujuan Penelitian

Tujuan yangakandicapai dari penulisan skripsi ini adalah:

1. Membangun mesin pencari *Information Retrieval Syatem*untuk pencarian dokumen-dokumen teks berbahsa Indonesia yang berupa abstrak skripsi mahasiswa Ilmu Komputer Universitas Pendidikan Indonesia.
2. Menghasilkan dokumen yang relevan berdasarkan *query* atau *keyword* yang dimasukan oleh pengguna, serta telah teranking sesuai tingkat pembobotannya.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat yaitu :

1. Aplikasi ini dapat menjadi solusi untuk memperoleh informasi yang sesuai dengan pengguna.
2. Memberikan informasi pada khalayak mengenai model ruang vector pada system temu balik informasi (*information retrieval*).

1.6 Sistematika Penulisan

Sistematika penulisan dalam pembahasan penelitian ini adalah sebagai berikut :

BAB I PENDAHULUAN

Bab ini membahas masalah yang meliputi latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian,sistematika penelitian .

BAB II TINJAUAN PUSTAKA

Bab ini memuat landasan teori mengenai pengertian sistem temu kembali informasi (*information retrieval*) secara umum, penjelasan tentang metode model ruang vector pada *information retrieval*, serta penjelasan mengenai pembobotan TF-IDF.

BAB III METODOLOGI PENELITIAN

Bab ini merupakan penjabaran dari desain penelitian, metode pengembangan perangkat lunak yang digunakan yaitu pendekatan terstruktur dengan model pengembangannya sekuensial linier serta alat dan bahan yang digunakan dalam penelitian.

BAB IV HASIL PENELITIAN DAN PEMBAHASAN

Di dalam bab ini, dibahas mengenai hasil penelitian sesuai dengan rumusan masalah berupa deskripsi sistem yang dibangun, penerapan metode *vector space model*di dalam sistem, analisa hasil uji metode *vector space model* dengan penerapan algoritma pembobotan TF-IDF.

BAB V` KESIMPULAN DAN SARAN

Bab ini berisikan kesimpulan dari hasil penelitian yang merupakan jawaban dari rumusan masalah beserta saran untuk pengembangan penelitian selanjutnya.