1. Pendahuluan

1.1 Latar Belakang

Ketersediaan informasi dalam jumlah besar di satu sisi merupakan berkah yang patut disyukuri pada zaman ini, karena dengannya kehidupan manusia akan semakin mudah. Namun, di sisi lain, melimpahnya informasi ini melahirkan permasalahan baru bagi manusia. Yaitu, perihal bagaimana mendapatkan informasi yang dibutuhkan secara tepat dan cepat diantara kumpulan berbagai informasi. Kalau hanya terdapat beberapa buah informasi, tentu hal ini dapat dengan mudah dicari dan ditentukan hanya dengan memeriksanya satu persatu. Namun bagaimana jika terdapat ribuan bahkan ratusan ribu informasi heterogen, sedangkan informasi yang kita butuhkan spesifik. Untuk itu dikembangkanlah cara untuk mengatasi permasalahan ini, yaitu Information Retrieval. Information Retrieval merupakan sebuah cabang Ilmu Komputer yang mempelajari bagaimana menemukan sebuah informasi yang sesuai dengan kebutuhan/relevan dari kumpulan besar koleksi informasi, biasanya dalam komputer.

Di dalam Information Retrieval sendiri terdapat berbagai macam metode atau model, mulai dari yang klasik sampai dengan yang modern. Yang paling sederhana contohnya dengan menggunakan metode boolean/exact matching, yaitu dengan mencocokan query dalam sebuah kumpulan informasi pada suatu proses pencarian (searching). Hasil dari pencarian ini terkadang tingkat relevansinya kurang, yang pada Information Retrieval diukur dengan nilai precision dan recall. Ketidak-akuratan ini disebabkan oleh tersedianya banyak cara bagi pemakai dalam menggunakan berbagai kata sebagai query (kata sinonim) dan karena terkadang suatu kata itu sendiri memiliki banyak makna (polisemi) . Sebuah terobosan hadir untuk mengatasi hal ini, salah satunya metode Latent Semantic Indexing (LSI). Metode ini akan melakukan pemetaan ulang terhadap term-document matrix yang dihasilkan dengan menggunakan algoritma SVD (Singular Value Decomposition) sehingga menghasilkan term-document matrix yang lebih sedikit dengan menghilangkan term yang tidak signifikan (noise term). LSI ini akan melakukan optimasi dengan mengikutsertrakan term yang memiliki kedekatan nilai semantik dengan query pada proses perangkingan.

Pada LSI, dokumen diindeks dengan menggunakan konsep *latent semantic*. LSI menunjukkan peningkatan kerja yang besar di atas representasi *tf-idf* pada koleksi dokumen kecil tetapi sering tidak berkinerja baik pada koleksi dokumen heterogen yang besar. LSI memetakan semua kata ke dalam dimensi matrik. Semakin besar jumlah dokumen semakin besar dimensi matrik yang terbentuk. Selain itu, informasi numerik dan singkatan dokumen yang mungkin indikator yang sangat baik dari topic tidak lagi didapatkan setelah menggunakan LSI. Hal ini disebabkan pada LSI, semua term yang meliputi kosakata noun maupun selain noun diproses dengan cara yang sama.

Pada tugas akhir ini akan dianalisa kinerja sebuah sistem *information retrieval* dengan menggunakan *Hybrid Document Indexing*. Pendekatan ini digunakan dalam pengindeksan dokumen untuk mengatasi masalah pada LSI. *Hybrid Document Indexing* tetap menggunakan konsep *latent semantic* dan juga mencoba untuk menjaga spesifik dari koleksi dokumen. *Hybrid Document Indexing* menggunakan kombinasi LSI untuk pembobotan kata yang mengandung noun dan selain noun pada dokumen akan dilakukan pembobotan *tf-idf*.

1.2 Perumusan Masalah

Dari latar belakang di atas maka dapat dirumuskan beberapa permasalahan pokok dalam tugas akhir ini, antara lain :

a. Bagaimana menjaga informasi numerik dan singkatan setelah dilakukan pengurangan dimensi.
b. Bagaimana cara membedakan kosakata *noun* dan kosakata selain *noun*.
c. Bagaimana kinerja sistem *information retrieval*, dilihat dari parameter *precision*, *recall* dan *F-Measure.*

1.3 Tujuan Pembahasan

Adapun tujuan yang ingin dicapai melalui tugas akhir ini antara lain :

a. Mengimplementasikan *Hybrid Document Indexing* pada *information retrieval* pada *dataset collection* yang digunakan berasal dari Cornell University yang berbahasa inggris
b. Menganalisi kinerja indexing setelah dilakukan pembagian kosakata menjadi *noun* dan selain *noun*.
c. Menganalisis kinerja dari *Hybrid Document Indexing* dengan parameter *precision*, *recall* dan *F-Measure.*

1.4 Batasan Masalah

Dalam Tugas Akhir ini, objek penelitian dibatasi dengan ruang lingkup sebagai berikut:

a. *Dataset Collection* yang digunakan berasal dari Cornell University, di alamat ftp.cs.cornell.edu/pub/smart berbahasa Inggris.
b. Melakukan penyaringan *stoplist* pada *preprocessing*. Sehingga kata-kata umum yang menyangkut tata bahasa, seperti kata konjungsi dan kata depan tidak dimasukan ke dalam indeks.
c. Dilakukan *stemming* pada *preprocessing* sehingga setiap kata dan turunannya dianggap sebagai term yang sama.
d. Pada saat proses pembedaan kosakata noun maupun selain noun, semua *term* yang tidak ada pada kamus data akan dianggap sebagai noun.

1.5 Metodologi Penyelesaian Masalah

Metodologi yang akan digunakan untuk menyelesaikan tugas akhir ini adalah :

1. Studi Literatur
   Mengumpulkan bahan-bahan referensi yang menunjang proses penelitian, yaitu yang berhubungan dengan *Information Retrieval, Latent Semantic Indexing, Hybrid Document Indexing, Spectral Methods* dan artikel-artikel terkait.
2. Analisis
   Pada tahap ini dilakukan analisis terhadap representasi *tf-idf* setelah dilakukan penggabungan dengan *spectral methods*. Pada tahapan ini juga akan dilakukan analisis kebutuhan perangkat lunak, serta perancangan dan desain perangkat lunak.
3. Implementasi
   Pada tahap ini dilakukan implementasi berdasarkan hasil rancangan dan melakukan evaluasi kinerja metode yang digunakan sebagai kajian selanjutnya.

4. Pengujian
   Menjelaskan tentang pengujian terhadap perangkat lunak untuk menguji fungsionalitas perangkat lunak. Pengujian perangkat lunak dilakukan terhadap data nyata berupa dokumen-dokumen tekstual.
5. Analisis hasil pengujian
   Melakukan analisis terhadap data dan hasil pengujian. Analisis dilakukan dengan melakukan uji coba terhadap koleksi dokumen untuk mengetahui kemampuan *Hybrid Document Indexing* dalam menemukan dokumen yang relevan walau tidak mengandung *term* dari *query* yang diinputkan akan tetap terambil serta untuk mengetahui. Parameter yang digunakan untuk mengukur akurasi yaitu waktu, *precision, recall* dan *F-Measure.* Dari hasil tahap ini, ditarik kesimpulan dan diusulkan saran untuk pengembangan lebih lanjut.

1.6 Sistematika Penulisan

BAB I **Pendahuluan**
Bab ini memaparkan latar belakang dilakukannya penelitian ini, perumusan masalah yang akan dibahas, hipotesa awal, pembatasan masalah, tujuan yang ingin dicapai melalui penelitian ini, metode penyelesaian masalah dan sistematika pembahasan.

BAB II **Dasar Teori**

Bab ini memuat berbagai dasar teori yang mendukung dan mendasari penulisan tugas akhir ini, yaitu mengenai konsep dari *Information Retrieval,* seperti *Indexing, Latent Semantic Indexing, Hybrid Document Indexing,* representasi *tf-idf* dan *Spectral Methods.*

BAB III **Analisis dan Perancangan Sistem**
Menganalisis kebutuhan sistem dan memuat pemilihan metode perancangan, yaitu dengan menggunakan teknik berorientasi objek, sehingga digunakan UML (*Unified Modeling Language*) sebagai bahasa permodelannya.

BAB IV **Pengujian dan Analisis**
Memuat spesifikasi perangkat keras dan lunak yang diperlukan agar sistem dapat berjalan, melakukan pengujian terhadap sistem dalam berbagai kondisi, dan analisis terhadap seluruh hasil pengujian.

BAB V **Kesimpulan dan Saran**

Berisi kesimpulan dari hasil penelitian tugas akhir ini serta saran-saran untuk pengembangan lebih lanjut.