Abstrak

Mendapatkan hubungan semantik antara kata-kata dalam sebuah representasi dokumen merupakan masalah yang sulit. *Latent Semantic Indexing* (LSI) adalah salah satu algoritma reduksi dimensi yang paling dikenal. 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*.

Hasil pengujian dari tugas akhir ini menunjukkan bahwa *Hybrid Document Indexing* dengan menggunakan *preprocessing stemming* terbukti bisa menemukan dokumen yang relevan walau tidak mengandung *term* dari *query* yang diinputkan akan tetap terambil. Selain itu, akurasi dari hasil pencarian dengan menggunakan metode ini menghasilkan nilai precision, recall dan F-Measure yang di atas 0,50. Pada percobaan beberapa jumlah dataset, Semakin banyak jumlah dataset maka waktu proses indexing dan searching akan semakin lama. Peningkatan lama proses ini dikarenakan dengan semakin banyaknya jumlah dokumen maka akan semakin besar dimensi pada LSI ditambah pemrosesan tf-idf sehingga waktu proses menjadi lebih lama.

Kata kunci: *Information Retrieval, Latent Semantic Indexing, Hybrid Document Indexing*