1. Pendahuluan

1.1 Latar belakang

Ketika *user* melakukan pencarian dengan menggunakan suatu *term* pada *search engine*, sering kali *user* mendapatkan sangat banyak dokumen hasil pencarian (*retrieved documents*). Dari *retrieved documents* tersebut, tidak semua mengacu pada hasil yang diinginkan. Bahkan mungkin saja hasil yang diinginkan bukan terdapat pada dokumen hasil pencarian yang pertama, kedua, atau ketiga, melainkan pada dokumen di bagian akhir pencarian, sehingga mau tidak mau *user* dipaksa menelusuri satu per satu *retrieved documents* dari awal sampai akhir untuk mendapatkan dokumen yang diinginkan.

Salah satu solusi untuk masalah ini adalah dengan mengelompokkan *retrieved documents* ke dalam kelompok-kelompok tertentu. Pengelompokan dilakukan berdasarkan kemiripan topik yang dibahas dalam suatu dokumen, sehingga *user* hanya perlu mengakses kelompok dokumen yang memuat topik sesuai dengan topik yang diinginkan, bukannya memeriksa isi *retrieved documents* satu per satu. Teknik untuk mengelompokkan dokumen ini disebut *document clustering*. *Document clustering* juga bisa diterapkan dalam mengelompokkan hasil pencarian *search engine*, sehingga dapat menyelesaikan permasalahan di atas.

Banyak metode atau algoritma dapat diterapkan dalam *document clustering*, salah satunya adalah algoritma *Semantic Hierarchical Online Clustering* (SHOC). Algoritma ini merupakan algoritma yang berbasis pada frase kunci yang ada dalam dokumen. Awalnya, algoritma SHOC diterapkan untuk bahasa dengan jumlah alfabet yang banyak seperti bahasa Mandarin, namun dalam Tugas Akhir ini diterapkan algoritma SHOC pada bahasa Inggris, dengan kemudian menganalisis performansinya berdasarkan dua parameter kualitas *clustering* yakni F-Measure dan Dunn Index [4]. F-Measure menunjukkan ketepatan pembagian dokumen ke dalam *cluster*, sedangkan Dunn Index mengukur jarak dokumen intra dan inter *cluster*. Sebelumnya, belum pernah diterapkan analisis dengan kedua parameter tersebut pada algoritma SHOC.

1.2 Perumusan masalah

Mengacu pada latar belakang di atas, ada beberapa masalah yang akan diselesaikan dalam Tugas Akhir ini, diantaranya yaitu :

1. Bagaimana menerapkan *web search results clustering* dengan algoritma SHOC?
2. Sejauh mana performansi algortitma SHOC dalam membagi dokumen hasil pencarian ke dalam *cluster* yang tepat?
3. Bagaimana analisis terhadap kondisi yang mempengaruhi kualitas *web search results clustering* dengan algoritma SHOC.

1.3 Batasan masalah

Berikut adalah batasan masalah pada Tugas Akhir ini :

1. *Document collection* yang dipakai merupakan dokumen teks bahasa Inggris dalam bentuk file HTML (selain HTML 5).
2. *Document collection* merupakan dokumen teks bertopik yang biasa digunakan untuk pengujian *search engine* yang didapat dari ftp://ftp.cs.cornell.edu/pub/smart/med.
3. Pengujian dilakukan secara *offline*.

1.4 Tujuan

Berikut adalah tujuan pengerjaan Tugas Akhir ini :

1. Menerapkan algoritma SHOC untuk mengelompokkan hasil pencarian *search engine*.
2. Menganalisis ketepatan algoritma SHOC dalam mengelompokkan dokumen ke dalam *cluster* tertentu. Parameter analisis yang digunakan adalah F-Measure (berdasarkan *precision* dan *recall clustering*), Dunn Index (berdasarkan *distance measure* dokumen yang dikelompokkan).
3. Menganalisis kondisi yang mempengaruhi performansi *web search results clustering* dengan algoritma SHOC dan bagaimana kondisi optimalnya.

1.5 Metodologi penyelesaian masalah

Berikut adalah metodologi yang digunakan dalam penyelesaian masalah dalam Tugas Akhir ini :

1. Studi literatur.
   Pencarian materi-materi dan referensi yang berkaitan dengan permasalahan yang dibahas, seperti materi tentang *information retrieval*, *document clustering*, algoritma SHOC, dan lain-lain.
2. Analisis dan perancangan kebutuhan sistem.
   Merupakan tahap perancangan sistem yang akan dibuat, yakni sebuah *search engine* yang menerapkan *document clustering* dengan algoritma SHOC, serta membuat analisis terhadap kebutuhan sistem yang akan dibuat.
3. Implementasi sistem.
   Mengimplementasikan sistem ke dalam program untuk kemudian diuji dan dianalisis. Proses implementasi dilakukan dengan tahapan :
   a. mengumpulkan *document collection*,
   b. membangun sebuah *search engine*, dan
   c. menerapkan algoritma SHOC pada *post-processing search engine* tersebut untuk mengelompokkan hasil pencarian.
4. Pengujian sistem.
   Menguji sistem yang telah diimplementasikan dan menganalisis hasil performansi dari *search engine* yang menerapkan *document clustering* dengan algoritma SHOC. Performansi dilihat dari waktu *clustering* dan ketepatan pengelompokan dokumen ke dalam *cluster*s. Pengujian dilakukan dengan memasukkan kata kunci tertentu pada *search engine*,

kemudian mengukur waktu *clustering* (waktu pengelompokkan dokumen, tidak termasuk waktu pencarian), dan melihat hasil akhir pengelompokkan *cluster* untuk kemudian dicocokkan dengan pengelompokkan yang diharapkan (*clustering judgement*). Pengujian dilakukan berkali-kali dengan kata kunci yang berbeda-beda.

5. Analisis hasil pengujian dan pengambilan keputusan.
   Analisis terhadap hasil pengujian sistem dan perumusan kesimpulan terhadap hasil analisis. Aspek yang dianalisis adalah waktu *clustering* untuk melihat efisiensi algoritma SHOC, serta ketepatan *clustering* untuk melihat efektifitas dari algoritma SHOC. Ketepatan *clustering* dilihat dari parameter F-Measure dan Dunn Index.
6. Penyusunan laporan Tugas Akhir.
   Penyusunan laporan semua tahap yang telah dilakukan mulai dari tahap studi literatur sampai perumusan kesimpulan.