Abstrak

Clustering (pengelompokan dokumen) merupakan salah satu teknik yang dapat digunakan untuk memudahkan *user* dalam menemukan dokumen *web* yang diinginkan dari sejumlah *retrieved documents* yang dihasilkan *search engine*. Teknik ini mengelompokkan dokumen berdasarkan kategori tertentu, sehingga penelusuran *user* terhadap *retrieved documents* akan lebih mengerucut.

Algoritma *Semantic Hierarchical Online Clustering* (SHOC) merupakan salah satu algoritma *clustering* yang mengelompokkan dokumen *web* hasil pencarian ke dalam *cluster* tertentu berdasarkan frase-frase kunci yang terdapat dalam dokumen tersebut. Tugas Akhir ini mengimplementasi dan menganalisis *clustering* hasil pencarian *search engine* dengan menggunakan algoritma SHOC.

Hasil pengujian menunjukkan bahwa algoritma SHOC mampu memisahkan *retrieved documents* yang relevan dan tidak dengan performansi yang dipengaruhi oleh kualitas hasil pencarian dan karakteristik dokumen. Algoritma SHOC akan optimal untuk mengelompokkan dokumen-dokumen yang saling berbagi frase kunci. Dan untuk pengaruh kualitas *search engine*, *precision* yang terlalu kecil akan menyebabkan banyaknya *cluster* "sampah" terbentuk, sedangkan *recall* yang terlalu kecil akan mengurangi ketepatan pembentukan cluster. Untuk menangani kualitas *search engine* yang kurang baik, nilai *threshold cluster quality* pada algoritma SHOC perlu diset sesuai dengan kualitas *search engine*, sehingga dokumen yang relevan bisa tetap dikelompokkan.

Kata kunci: *search engine*, *retrieved documents*, *clustering*, frase kunci, algoritma SHOC