1. Pendahuluan

1.1 Latar belakang

Dengan semakin berkembangnya teknologi informasi jumlah dokumen teks berupa digital semakin berkembang dengan pesat. Hal ini mengakibatkan munculnya suatu cabang ilmu baru dalam teknologi informasi yaitu pencarian informasi (*information retrieval*). Pencarian informasi berbasis *query (query based retrieval)* yang digunakan secara tradisional sangat berguna untuk pencarian terarah tetapi tidak begitu efisien. *Query based retrieval* berguna ketika kita mengetahui benar kejadian asli atau fakta yang dicari. Cara ini tidak begitu efektif ketika kita membutuhkan informasi yang spesifik/khusus dalam kategori yang besar dan juga tidak efisien untuk mendapatkan berita yang relevan dalam penelusuran suatu kejadian. Dengan menggunakan cara ini biasanya hanya dapat diketahui berita-berita tertentu saja, sedangkan berita-berita lama yang berkaitan dengan berita tersebut sulit untuk ditelusuri. Selain pemilihan *query* yang tidak tepat akan menyebabkan membanjirnya dokumen- dokumen yang tidak relevan.

Pengelompokan (*clustering*) dokumen merupakan sebuah cara yang dapat digunakan untuk mempermudah pencarian dokumen dalam database. *Clustering* merupakan proses pengelompokan data sehingga semua anggota dari bagian data memiliki kemiripan berdasarkan perhitungan jarak/kemiripan term-term yang sudah dilakukan pembobotan pada tahapan sebelumnya (*preprocessing*). *Clustering* merupakan salah satu metode *data mining* yang bersifat tanpa arahan (*unsupervised*). Ada dua jenis *clustering* yang [3]sering dipergunakan dalam proses pengelompokan data yaitu *hierarchical* (hirarki) data *clustering* dan *non-hierarchical* (non hirarki) *clustering*. *Single pass clustering* merupakan salah satu metode data *clustering* non hirarki (*partisioning*) yang berusaha mengelompokkan data yang ada ke dalam bentuk satu atau lebih *cluster*. Salah satu aplikasi dari *clustering* adalah *document clustering*.

Tahapan-tahapan *clustering* yang pertama adalah representasi dokumen. Tiap-tiap dokumen akan direpresentasikan kedalam matrik dokumen berupa bobot dari term pada masing-masing dokumen, kedua adalah penentuan tingkat kemiripan antar dokumenn, ketiga adalah penggunaan algoritma *clustering*, dan yang terakhir adalah tahap evaluasi. Evaluasi ini bertujuan untuk mengetahui kinerja dari algoritma yang digunakan untuk klasterisasi dokumen. Sebelum melakukan pengelompokan dokumen, diperlukan tahap *preprocessing* [2] terlebih dahulu. Pada tahap *preprocessing* yang dilakukan adalah *case folding*, ,*stopwords*, *stemming* dan *term weighting*. Metode *term weighting* yang digunakan adalah TF-IDF yang merupakan kombinasi antara TF (*Term Frequency*) dengan IDF (*Invers Document Frequency*).

Tugas Akhir ini menerapkan pengelompokan dokumen dengan menggunakan algoritma *Single Pass Clustering*. *Single pass clustering* merupakan [14] suatu tipe *clustering* yang berusaha melakukan pengelompokan data satu demi satu dan pembentukan kelompok dilakukan seiring dengan evaluasi setiap data yang dimasukkan ke dalam proses *cluster*.

1.2 Perumusan masalah

Berdasarkan latar belakang yang telah diuraikan, dapat diambil rumusan masalah sebagai berikut :

1. Bagaimana melakukan pengelompokan dokumen dengan menggunakan algoritma *Single Pass Clustering* ?
2. Bagaimana mengetahui dan menganalisis kualitas *cluster* yang dihasilkan dengan menggunakan metode *cohesion* dan *separation* dan menganalisis parameter-parameter dari algoritma *Single Pass Clustering* yang berpengaruh terhadap kualitas *cluster* ?

1.3 Batasan masalah

Adapun batasan-batasan masalah dalam pengerjaan Tugas Akhir ini adalah:

1. Tugas Akhir ini menggunakan dokumen berita berbahasa Indonesia yang telah mengalami *preprocessing* terlebih dahulu.
2. Tidak melakukan analisis terhadap data *preprocessing.*
3. Kualitas *cluster* yang dihasilkan tidak dibandingkan dengan algoritma lain.
4. Tidak menangani kesalahan pada penulisan kata dalam dokumen.

1.4 Tujuan

Berdasarkan pada masalah yang telah didentifikasikan di atas, maka tujuan Tugas Akhir ini adalah :

1. Menerapkan Algoritma *Single Pass Clustering* serta mengetahui performansinya sebagai salah satu metode data mining dalam mengelompokkan dokumen berita kejadian bahasa Indonesia.
2. Menganalisis kualitas *cluster* yang dihasilkan dengan melihat nilai *cohesion* dan *separation* serta mengetahui parameter yang berpengaruh untuk menghasilkan kualitas *cluster* yang baik.

1.5 Metodologi penyelesaian masalah

Metode yang digunakan untuk menyelesaikan Tugas Akhir ini adalah :

1. Studi literatur dan tinjauan pustaka tentang :
   a. Algoritma *Single Pass Clustering*.
   b. Tahapan-tahapan pengelompokan dokumen berita bahasa Indonesia.
   c. Tahapan *preprocessing* dokumen berita.
   d. *Cluster* analisis dengan metode *cohesion* dan *separation*.
2. Pengumpulan data
   Data yang digunakan sebagai sampel adalah dokumen berita bahasa Indonesia yang berhasil di download dari media online seperti www.kompas.com , www.republika.com, www.detik.com dan www.tempointeraktif.com Analisis dan perancangan sistem menggunakan algoritma *Single Pass Clustering*

3. Implementasi

a. Ekstraksi dokumen
   Proses ekstraksi ini bertujuan untuk menghasilkan *term-term* yang akan digunakan sebagai *prototype* bagi setiap dokumen yang berisi bobot term pada kumpulan dokumen. Ekstraksi dokumen dilakukan pada tahap *preprocessing.*

b. Penghitungan tingkat kemiripan
   Perbandingan tingkat *similarity* yang digunakan pada TA ini adalah menggunakan *standard cosine similarity.*[4]

c. Pengelompokan dokumen.
   Untuk melakukan pengelompokan dokumen algoritma yang digunakan adalah *Single Pass Clustering.* Dengan melakukan perbandingan tingkat *similarity* selanjutnya hasilnya akan dievaluasi untuk menentukan pasangan-pasangan dokumen yang dinyatakan mirip berdasarkan nilai *threshold* tertentu.

4. Testing dan analisis
   Untuk mengetahui kinerja algoritma *Single Pass Clustering* pada tahap uji coba,dilakukan pengukuran berdasarkan kualitas *cluster* menggunakan dua parameter, yakni *cohesion* dan *separation.* Selain itu juga akan dilihat dari sisi waktu, berapa lama waktu eksekusi untuk melakukan pengelompokan dokumen dengan algoritma *Single Pass Clustering.*

5. Pengambilan kesimpulan dan penyusunan laporan Tugas Akhir.