1. Pendahuluan

1.1 Latar Belakang Masalah

Pada perkembangan teknologi sekarang ini begitu banyak dokumen teks dalam bentuk digital. Dengan banyaknya dokumen, informasi yang beredar juga semakin banyak sehingga untuk mempermudah mengambil suatu informasi sesuai yang dibutuhkan, perlu dilakukannya pengelompokkan dokumen sesuai dengan topiknya. Pengelompokkan ini dapat dilakukan dengan menggunakan teknik yang terdapat dalam *data mining* yaitu *clustering. Clustering* adalah proses mengelompokkan objek berdasarkan informasi yang diperoleh dari data yang menjelaskan hubungan antar objek dengan prinsip untuk memaksimalkan kesamaan antar anggota satu kelas dan meminimumkan kesamaan antar kelas/*cluster*. *Clustering* juga dikenal sebagai *unsupervised learning* yang membagi data menjadi kelompok-kelompok atau *cluster*s berdasarkan suatu kemiripan atribut-atribut diantara data tersebut [3].

Clustering sendiri terbagi atas 3 yaitu *hierarchical* dan *partitional dan Clustering Large Data*. *Hierarchical clustering* adalah suatu struktur data berbentuk pohon yang disebut dendogram dimana data dikelompokkan secara bertingkat dari yang paling bawah dimana tiap instance data merupakan *cluster* sendiri, hingga tingkat yang paling atas. *Partitional clustering* yang mengelompokkan data ke dalam k *cluster* dimana k adalah banyak *cluster* dari input user. *Clustering Large Data* dibutuhkan untuk melakukan *cluster*ing pada data yang volumenya sangat besar sehingga tidak cukup ditampung dalam memori komputer pada suatu waktu [11].

Ada beberapa algoritma yang telah diimplementasikan untuk *clustering* misalnya algoritma Cobweb, algoritma K means dan lain-lain. Algoritma-algoritma tersebut masih memiliki beberapa kekurangan, sedangkan ada kendala yang sering terjadi dan harus dihadapi untuk menyelesaikan masalah pengelompokkan dokumen pada *clustering*. Kendala tersebut misalnya memiliki outlier atau jumlah dokumen yang besar. Pada tugas akhir ini, algoritma yang akan digunakan untuk membangun sistem pengelompokan dokumen adalah algoritma *clustering using representative* atau sering disebut dengan algoritma *cure*. Algoritma *cure* merupakan metode campuran *hierarchical* dan *partitional* yang berdasarkan poin perwakilan yang ditentukan sebelumnya[4]. Keuntungan dari algoritma ini dapat menangani data yang besar [4]. Data besar dapat berupa data yang berdimensi tinggi[6], dokumen memiliki banyak dimensi. Untuk itu, algoritma *cure* dipilih untuk melakukan pengelompokan dokumen karena dapat menangani jumlah data yang besar tanpa mengorbankan kualitas *clustering*. Untuk membuktikan kualitas *cluster* yang dihasilkan algoritma ini, maka pada tugas akhir ini akan dianalisa kualitas *cluster*.

.

1.2 Perumusan Masalah

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

1. Bagaimana menerapkan metode algoritma *CURE (clustering using representative)* untuk pengelompokkan dokumen bahasa indonesia?
2. Bagaimana menganalisis kualitas hasil *cluster* dengan mengukur nilai *cohesion* (kemiripan objek intra *cluster*) dan *separation* (ketidakmiripan objek antar *cluster*) serta pengaruh perubahan nilai *threshold* terhadap akurasi hasil *cluster* pada algoritma *CURE (clustering using representative)*?

1.3 Batasan Masalah

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

1. Data set yang digunakan berupa dokumen berita berbahasa Indonesia
2. Dokumen berita didapat dari artikel di www.kompas.com.
3. Data masukan sistem yang telah mengalami proses *text preprocessing* terlebih dahulu dan tidak dibahas lebih detail pada tugas akhir ini.

1.4 Tujuan

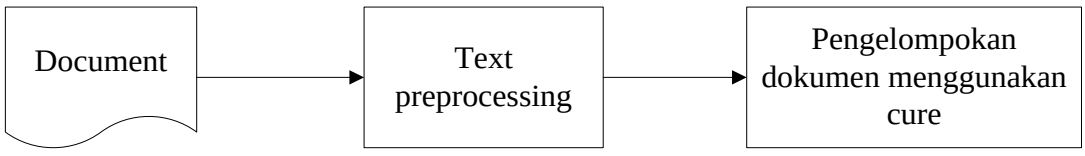
Adapun tujuan dalam pengerjaan Tugas Akhir ini adalah:

1. Menerapkan metode algoritma *CURE (clustering using representative)* untuk pengelompokkan dokumen bahasa Indonesia yaitu dengan mengimplementasikan algoritma *CURE* pada pengelompokkan dokumen.
2. Menganalisis kualitas hasil *cluster* dengan mengukur nilai *cohesion* (kemiripan objek intra *cluster*) dan *separation* (ketidakmiripan objek antar *cluster*) serta pengaruh perubahan parameter nilai P dan *threshold* terhadap akurasi hasil *cluster* pada algoritma *CURE (clustering using representative).*

1.5 Metodologi Penyelesaian Masalah

Metodologi yang digunakan untuk menyelesaikan permasalahan-permasalahan dalam Tugas Akhir ini terdiri dari 7 tahap, yaitu:

1. Studi pustaka/studi literature
   Tahapan ini adalah mengumpulkan bahan-bahan referensi Tugas Akhir yang meliputi algoritma *cure, data mining* , evaluasi kualitas *cluster* dan dokumen yang akan dijadikan dataset serta mempelajari dan memperdalam materi mengenai algoritma *cure(clustering using representative)*. Sumber referensi berupa buku, jurnal , e-book dan browse internet.
2. Tahap pengumpulan data dan *preprocessing* data
   Tahapan ini adalah melakukan pencarian dokumen berita berbahasa Indonesia. Setelah itu, dilakukan proses *text preprocessing* terhadap dokumen sehingga dokumen tersebut dapat digunakan sebagai data masukan untuk sistem yang akan dibangun.
3. Tahap perancangan sistem
   Tahapan ini melakukan proses perancangan sistem dengan membuat gambaran fungsi-fungsi yang akan membangun sistem pengelompokkan dokumen dengan algoritma *cure(clustering using representative)*. Berikut adalah rancangan system secara umum:



Gambar 1-1 Perancangan Sistem

4. Tahap implementasi pemrograman

Tahapan ini adalah melakukan proses pembangunan sistem berdasarkan rancangan yang telah dibuat sebelumnya menggunakan tools matlab.

5. Tahap pengujian sistem
   Tahapan ini melakukan pengujian pada sistem yang telah dibuat dengan menggunakan dataset yang telah disediakan dan mengevaluasi hasil *cluster* dengan melihat nilai *cohesion* dan *separation*.
6. Tahap analisis output sistem
   Tahapan ini melakukan analisis kualitas *cluster* yang dihasilkan oleh sistem.
7. Tahap Pembuatan Laporan.
   Tahapan ini melakukan penyusunan laporan akhir dan pengumpulan dokumentasi berdasarkan analisa hasil pengujian tugas akhir ini.

1.6 Sistematika Penulisan

Tugas akhir ini disusun dengan sistematika sebagai berikut:

1. Pendahuluan
   Bab ini menguraikan tugas akhir ini secara umum, meliputi latar belakang, perumusan masalah, batasan masalah, tujuan dan metodologi penyelesaian masalah.

2. Dasar Teori
   Bab ini membahas mengenai uraian teori yang berhubungan dengan *cluster*ing dan algoritma *cure*.

3. Analisis Perancangan Dan Implementasi
   Bab ini berisi analisis kebutuhan dari sistem yang kemudian dituangkan ke dalam suatu sistem pemodelan secara terstruktur. Dari tahap analisis kemudian dilanjutkan ke tahap perancangan dan implementasi.

4. Analisis Hasil Pengujian
   Bab ini membahas mengenai pengujian yang dilakukan terhadap sistem yang telah dibangun. Pengujian dilakukan dengan mengganti-ganti nilai parameter yang terdapat dalam sistem. Tahap pengujian dilanjutkan dengan tahap analisis hasil pengujian.

5. Kesimpulan
   Berisi kesimpulan dari penulisan Tugas Akhir ini dan saran-saran yang diperlukan untuk pengembangan lebih lanjut.