

1. Pendahuluan

1.1 Latar Belakang

Pengelompokan dokumen atau yang biasa dikenal dengan klasterisasi merupakan teknik penting yang telah lama digunakan dalam bidang data mining dan karena perannya yang begitu besar, maka teknik klasterisasi ini juga digunakan dalam *Information Retrieval*. Klasterisasi mempunyai pengertian bahwa setiap dokumen mempunyai karakteristik masing-masing dalam hal persamaan dan perbedaannya, sehingga akan lebih baik jika dokumen-dokumen yang relevan dikelompokkan kedalam *cluster* yang sama.

Klasterisasi berpengaruh terhadap keluaran sistem terhadap dokumen-dokumen yang tersedia. Dengan diterapkannya klasterisasi dokumen ini, harapan pastinya yaitu supaya prosentase user dalam mendapatkan dokumen yang sesuai dan relevan dengan pencariannya semakin lebih besar karena dokumen yang dicari setidaknya berada dalam kelompok yang hampir sama dengan dokumen lainnya yang mempunyai kesamaan isi.

Klasterisasi terbagi menjadi berbagai macam karakteristik, salah satunya yaitu *Partitional Clustering* yang mempunyai konsep menjadikan setiap dokumennya berada tepat pada satu cluster saja sehingga *overlapping* atau beririsan antar *clusternya* tidak terjadi. Penerapan konsep *Partitional Clustering* ini digunakan oleh Algoritma *C3M (Cover Coefficient Clustering Method)* yang menggunakan model probalistik, kesamaan term (*term similarities*), dan *seed document* sebagai inisialisasi awal dari setiap *cluster*. Algoritma *C3M* termasuk bagian dari *Cluster Based Retrieval (CBR)* yang mempunyai karakter utama yaitu membutuhkan *query* inputan dari user dalam pencarian dokumen yang tersedia. Sehingga *query* inputan tersebut mempunyai keterbatasan dalam penulisannya seperti halnya *search engine* yang biasa digunakan dalam keseharian.

Dalam *Cluster Based Retrieval*, Klasterisasi merupakan tahap lanjut dari *Preprocessing* data yang harus dilakukan lebih dahulu dengan tujuan data yang akan dikelompokkan sudah dalam bentuk atau format yang sesuai untuk diolah dengan Algoritma *C3M*. *Preprocessing* dalam Tugas Akhir ini biasa disebut dengan *indexing* dokumen yang terdiri dari beberapa tahap ; *break into tokens*, *stoplist* atau *stopword*, dan *stemming*.

Stemming merupakan salah satu bagian dari *indexing* yang bertujuan untuk menjadikan setiap kata atau term yang ada dalam setiap data atau dokumen kedalam kata dasarnya dengan menghilangkan semua imbuhan yang melekat pada masing-masing kata tersebut. Imbuhan terdiri dari awalan (*prefix*), sisipan (*infix*), akhiran (*suffix*) dan kombinasi antara awalan dan akhiran (*confix*).

Dalam *Stemming* untuk Bahasa Indonesia, ada beberapa macam metode yang telah diciptakan dan dikembangkan, salah satunya yaitu Algoritma Jelita Asian yang merupakan pengembangan dari Algoritma Nazief Adriani.

Pada tugas akhir ini, penelitian yang dilakukan yaitu mengenai penggunaan algoritma C3M (*Cover Coefficient Clustering Method*) pada pengelompokan dokumen berupa abstrak Tugas Akhir mahasiswa ITTELKOM berbahasa Indonesia yang dikombinasikan algoritma Jelita Asian pada tahap *stemming* data dengan harapan mendapatkan hasil yang sesuai dengan kebutuhan user, yaitu cluster yang terbentuk berisikan dokumen-dokumen yang relevan satu sama lainnya.

1.2 Rumusan Masalah

Berikut merupakan rumusan masalah yang mengacu pada latar belakang di atas:

1. Bagaimana pengaruh indexing terhadap hasil dengan tahap *stemming* menggunakan Algoritma Stemmer Jelita Asian terhadap hasil clustering menggunakan Algoritma C3M
2. Bagaimana hasil penghilangan konsep Information Retrieval berupa inputan query dari user jika diterapkan dalam Tugas Akhir ini pada Tahap Clustering dokumen?
3. Bagaimana proses clustering dengan menggunakan algoritma C3M dan pengaruhnya terhadap jumlah cluster yang terbentuk dan pemilihan dokumen pusat cluster sebagai acuan bagi dokumen lainnya?
4. Bagaimana kualitas *cluster* yang dihasilkan oleh algoritma C3M dengan kombinasi Algoritma Stemmer Jelita Asian jika dihitung dengan menggunakan *silhouette coefficient*?

1.3 Tujuan

Tujuan dari pembuatan Tugas Akhir ini adalah

1. Menganalisa pengaruh indexing dengan tahap *stemming* menggunakan Algoritma Stemmer Jelita Asian terhadap hasil clustering menggunakan Algoritma C3M
2. Menganalisa pengaruh dari penghilangan query inputan user khususnya pada tahap clustering dokumen dengan konsep Cluster Based Retrieval yang seharusnya menggunakan query inputan user.
3. Mengelompokkan dokumen abstrak TA ITTELKOM kedalam cluster-cluster terbentuk menggunakan algoritma C3M yang sebelumnya dokumen tersebut telah di *indexing* dengan Jelita Asian. Serta menghitung dan menganalisa konsep *Cover Coefficient* dari Algoritma C3M terhadap jumlah cluster yang terbentuk dan pemilihan dokumen sebagai pusat cluster sebagai acuan bagi dokumen lainnya
4. Mengukur dan menganalisa kualitas Cluster yang dihasilkan Algoritma C3M dengan menggunakan *Silhouette Coefficient*.

1.4 Batasan Masalah

Batasan masalah pada tugas akhir ini, antara lain :

1. Data yang digunakan sebanyak 325 Dokumen Tugas Akhir Mahasiswa ITTELKOM berupa .txt yang diambil secara acak dari masing-masing fakultas.
2. Dokumen harus berisi minimal satu kata.
3. Istilah term dalam Tugas Akhir ini yaitu 1 kata
4. Fokus algoritma yang digunakan adalah Algoritma Jelita Asian pada tahap stemming dan Algoritma C3M pada tahap clustering

1.5 Metode Penyelesaian Masalah

Metode penyelesaian masalah yang digunakan dalam Tugas Akhir ini, antara lain :

1. Studi Literatur

Mengumpulkan informasi dan referensi dari buku maupun artikel dan paper-paper yang ada serta memahaminya sehingga dapat digunakan dasar teori dalam penyusunan Tugas Akhir yang berkaitan dengan algoritma Jelita Asian dan C3M, algoritma yang akan digunakan dalam pengelompokan dokumen abstrak TA ITTELKOM

2. Analisis dan Perancangan Sistem

Tahap ini Merupakan tahap perancangan sistem yang dibuat, yakni sebuah perangkat lunak yang akan menerapkan algoritma Jelita Asian dan C3M.

3. Implementasi Sistem

Melakukan coding menggunakan tools yang sesuai untuk membangun sistem sesuai dengan rancangan pada tahap sebelumnya.

4. Pengujian Sistem

Pada tahap ini, dilakukan pengujian terhadap sistem yang telah dibangun. Hal yang diujikan ialah seperti yang telah dipaparkan pada tahap perancangan.

5. Analisis hasil pengujian

Dari tahap pengujian sistem yang dilakukan sebelumnya, dilakukan analisis terhadap pengaruh digunakannya Algoritma Jelita Asian terhadap hasil clustering yang didapatkan dengan menggunakan C3M hasilnya lebih baik dan juga pengaruh konsep Cover Coefficient dari Algoritma C3M terhadap jumlah cluster yang terbentuk dan pemilihan dokumen pusat sebagai acuan bagi dokumen lainnya.

6. Penyusunan Laporan Tugas Akhir

Pada tahap ini, dilakukan penyusunan laporan akhir dan pengumpulan dokumentasi yang diperlukan, format laporan mengikuti kaidah penulisan yang benar dan yang sesuai dengan ketentuan-ketentuan yang telah ditetapkan oleh institusi.

1.6 Sistematika Penulisan

Sistematika penulisan dalam penyusunan tugas akhir ini terbagi menjadi beberapa BAB , diantaranya BAB :

1. Pendahuluan

Bab ini berisikan latar belakang masalah, rumusan masalah, tujuan ,batasan masalah, metode penyelesaian masalah, dan sistematika penulisan yang secara umum, subbab tersebut merupakan gambaran awal dan juga fokus penulis terhadap penelitian ini

2. Landasan Teori

Bab ini berisikan Penjelasan mengenai semua teori yang penulis gunakan dalam pembuatan sistem atau pun dokumentasi yang berhubungan dengan tugas akhir ini.

3. Analisis dan Perancangan

Bab ini berisikan perancangan sistem yang akan digunakan dalam tahap implementasi, yaitu dengan menggunakan algoritma Jelita Asian dan C3M

4. Implementasi dan Pengujian

Bab ini berisikan implementasi sistem sesuai rancangan sebelumnya yang kemudian dijelaskan secara rinci dalam analisa pengujian sistem

5. Kesimpulan dan Saran

Bab ini berisikan kesimpulan yang didapat penulis dalam penelitian ini serta saran untuk mengembangkan penelitian ini lebih lanjut