ANALISIS DAN IMPLEMENTASI ALGORITMA JELITA ASIAN DAN COVER COEFFICIENT-BASED CLUSTERING METHOD (C3M) PADA PENGELOMPOKAN DOKUMEN TEKS BERBAHASA INDONESIA

Moh. Eko Arifianto[1], Suyanto[2], -[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

---

Abstrak

Pengelompokan dokumen merupakan hal yang biasa ditemukan dalam kehidupan sehari-hari dan dianggap penting sehingga dimanfaatkan diberbagai bidang salah satunya teknologi. Dengan adanya pengelompokan ini, maka dokumen-dokumen yang dianggap relevan maka akan dijadikan dalam satu kelompok

Cover Coefficient Clustering Method (C3M) adalah salah satu algoritma pengelompokan atau clustering dokumen yang menggunakan model probalistik, kesamaan term ,dan seed dokumen sebagai poin penting dalam menentukan inisialisasi awal dari pembentukan clusternya, dimana dokumen yang dikelompokan ini telah diproses terlebih dahulu agar bisa diolah menggunakan algoritma C3M. Pemprosesan ini disebut juga preprocessing atau indexing dokumen. Dalam Indexing , ada beberapa tahap yang biasa dilakukan, salah satunya stemming. Pada Tugas Akhir ini, Algoritma Jelita Asian digunakan pada tahap stemming kata dengan dilanjutkan penggunaan Algoritma C3M dalam pengclusteran dokumen abstrak ITTELKOM.

Pengujian yang dilakukan dalam Tugas Akhir ini yaitu untuk mengclusterkan dokumen dengan Algoritma C3M , menganalisa konsep Cover Coefficient C3M dan tahapan Algoritma C3M serta menganalisa hasil kualitas cluster yang dihasilkan oleh perpaduan Jelita Asian dengan C3M menggunakan nilai Silhouette Coefficient. Dan didapatkan bahwa kualitas cluster yang dihasilkan termasuk dalam kualitas yang lemah berdasarkan nilai rata-rata silhouette yang diperoleh.

Kata Kunci : C3M, Jelita Asian, CBR , SilhouetteCoefficient , Clustering, Indexing

---

Abstract

Clustering documents are a commonly found in our life and considered an important things so it is used in various fields, like technology. Basic of the idea on clustering dokumen, that is grouping the document which have relevan with other document. So, in the end of step, each document will be join with other relevan documents.

Cover Coefficient Clustering Method (C3M) is one of Clustering Algorithm which used probabilistic model, term similarities and document seed as important poins in decide the first of initialization from forming the cluster where the documents before this have been processed so they can be processed by using C3M Algorithm. It is called data preprocessing or indexing document. In indexing, one of the step is stemming. In this Final Project, Jelita Asian Algorithm used on stemming step and C3M Algorithm used on clustering document.

Tests performed in this Final Project are clustering document by C3M Algorithm, analyze the Cover Coefficient C3M and the steps of C3M and also analyze the result of cluster quality by combine of Asian Jelitausing Silhouette Coefficient. And found that the quality of the resulting cluster is included in weak by the average silhouette values

Keywords : C3M, Jelita Asian, CBR , SilhouetteCoefficient , Clustering, Indexing

---

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 berbagi macam karakteristik, salah satunya yaitu *Partitional Clustering* yang mempunyai konsep menjadikan setiap dokumennya berada tepat pada satu cluster saja sehingga *overlapping* atau beririsan antar *cluster*nya 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*. Algoritm *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 *Prepocessing* data yang harus dilakukan lebih dahulu dengan tujuan data yang akan dikelompokan sudah dalam bentuk atau format yang sesuai untuk diolah dengan Algoritma *C3M*. *Prepocessing* 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 diatas:

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 abtrak TA ITTELKOM kedalam cluster-cluster terbentuk menggunakan algoritma C3M yang sebelumnya dokumen tersebut telah di*indexing* dengan Jelita Asian. Serta menghitung dan mengalisa konsep *Cover Coefficient* dari Algoritma C3M terhadap jumlah cluster yang terbentuk dan pemilihan dokumen sebagi pusat cluster sebagai acuan bagi dokumen lainnya
4. Mengukur dan menganalia 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. **Pedahuluan**

   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

5. Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan pengujian yang telah dilakukan pada Tugas Akhir ini, dapat diambil kesimpulan bahwa :

a) Penggunaan *Stemming* pada tahap *indexing* menjadikan antar dokumen mempunyai kesamaan *term* lebih tinggi jika mengacu pada nilai ketidakrelevanan *(decoupling)* tiap dokumen.

b) Dokumen yang tidak mempunyai keterkaitan sama sekali dengan dokumen lainnya (*Ragbag)* dapat diminimalisir dengan penggunaan kumpulan term unik dari tiap dokumen

c) Jumlah *cluster* akan cenderung semakin banyak seiring dengan banyaknya dokumen yang mempunyai tingkat kesamaan sedikit terhadap dokumen lainnya. Dan tingkat kesamaan ini berpengaruh terhadap tinggi rendahnya nilai *Seed Power* masing-masing dokumen

d) Kualitas *Cluster* yang didapatkan dengan menggunakan acuan nilai *Silhouette Coefficient* pada Tugas Akhir ini golongkan kedalam kategori lemah dan tidak terstruktur. Hal ini dikarenakan adanya dokumen yang masuk kedalam suatu cluster yang kurang tepat.

5.2 Saran

Saran dari penulis untuk keperluan penelitian lebih lanjut :

a) Pengembangan terhadap algoritma C3M agar dapat menangani kasus overlapping yang menjadikan setiap dokumen berpeluang masuk kedalam beberapa cluster yang terbentuk

b) Pembaharuan persamaan dalam menentukan seed dokumen atau dokumen pusat

c) Penggunaan konsep *incremental* clustering sehingga jikaada penambahan dokumen baru , maka hanya proses clustering dilakukan hanya kepada dokumen baru tersebut

Daftar Pustaka

[1] Adipathy, Asriko.2010.*Analisis dan implementasi perbandingan stemming dengan menggunakan algoritma jelita Asian dan arifin&seriono pada information retieval*.Institut Teknologi Telkom.Bandung

[2] Asian, Jelita. 2007. *Effective Techniques for Indonesian Text Retrieval*. School of Computer Science and Information Technology. Australia.

[3] Asian,Jelita., E,Hugh., Tahaghoghi,S. 2005. *Stemming Indonesian*. School of Computer Science and Information Technology. Australia.

[4] Can, Fazli. 1991. *Experiments on Incremental Clustering*. Miami University. Ohio.

[5] Can, Fazli. 1993. *Incremental Clustering for Dynamic Information Processing*. Miami University. Ohio.

[6] Can, Fazli., Altingovde, Ismail Sengor. 2003. *Efficiency and Effectiveness of Query Processing in Cluster Based-Retrieval*. Bilkent University. Turkey

[7] Can, Fazli., Ozkarahan Esen, A. 1985. *Concepts Of The Cover Coefficient- Based Clustering Methodology*. Tempe. Arizona

[8] Can, Fazli., Ozkarahan Esen, A. 1990. *Concepts And Effectiveness Of The Cover Coefficient Based Clustering Methodology For Text Databases*. Miami University. Ohio

[9] Http://boynurah.wordpress.com/2009/07/16/stemming-bahasa-indonesia/ Diakses Februari 2012

[10] Http://imeldas.blog.ittelkom.ac.id/blog/files/2010/03/Dami2_EksplorasiData2b.pdf Diakses Mei 2012

[11] Http://live-hadi.blogspot.com/2009/04/information-retrieval.html Diakses Februari 2012

[12] Http://math.stackexchange.com/questions/102924/cosine-similarity-distance-and-triangle-equation Diakses Mei 2012

[13] Http://yusufxyz.files.wordpress.com/2012/04/text-operation1.ppt Diakses Mei 2012

[14] Myke,Harista.2010.Analisis Penerapan Algoritma Cover Coefficient-Based Incremental Clustering Methodology (C2ICM) Dalam Pengelompokkan Dokumen Teks.ITTELKOM.Bandung

[15] Roesita,Ranny.2009.*Analisis dan implementasi Cluster Based Retrieval Menggunakan Metode C3M pada Dokumen Teks Berbahasa Inggris*. ITTelkom.Bandung

[16] Vural, A. 2002. *Online New Event Detection and Clustering Using the Concepts of the Cover Coefficient-Based Clustering methodology*. The Institute of Engineering and Science of Bilken University. Turkey