ANALISIS DAN IMPLEMENTASI HIERARCHICAL AGGLOMERATIVE CLUSTERING PADA DOKUMEN BERITA BERBAHASA INDONESIA

Donny Iswan Situngkir[1], Ema Rachmawati[2], Warih Maharani[3]

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

Abstrak

Berita saat ini tidak hanya disebarkan melalui media elektronik dan media cetak, namun juga disebarkan melalui dunia internet. Sehingga jumlah berita yang tersedia sangatlah besar dan butuh waktu lama untuk mengelompokkannya secara manual. Clustering dapat digunakan sebagai solusi dari permasalahan tersebut. Salah satu metode yang dapat digunakan untuk clustering dokumen berita ini adalah metode Hierarchical Agglomerative Clustering (HAC). Pada tugas akhir ini metode HAC yang digunakan ialah Single Link, Complete Link, dan Average Link.

Metode HAC mempunyai kemampuan menggabungkan data dengan membuat hirarki, dimana data yang memiliki kemiripan akan ditempatkan di hirarki yang berdekatan dan yang tidak memiliki kemiripan ditempatkan pada hirarki yang berjauhan. Setiap dokumen akan dianggap sebagai sebuah cluster, kemudian digabungkan sesuai dengan metode HAC yang digunakan, berulang hingga jumlah cluster sesuai dengan yang diinginkan.

Hirarki yang terbentuk akan dihitung kualitasnya menggunakan cophenetic correlation coefficient, sementara kualitas cluster yang terbentuk akan dihitung menggunakan silhouette coefficient dan purity.

Kata Kunci : Clustering, HAC, cophenetic correlation coefficient, silhouette coefficient,

Abstract

News today is not only disseminated through electronic media and print media, but also disseminated through the internet. So the amount of news available is enormous and it took a long time to group them manually. Clustering can be used as a solution to these problems. One method that can be used for word document clustering is a method Hierarchical Agglomerartive Clustering (HAC). In this final project HAC method used is Single link, Complete link, and Average link.

HAC method has the ability to combine data to create hierarchies, which have similar data will be placed in an adjacent hierarchy and that it bears no resemblance to the hierarchy that placed far apart. Each document will be considered as a cluster, then combined according to the HAC method used, repeated until the number of clusters as desired.

Hierarchy that is formed will be calculated using the cophenetic correlation coefficient of quality, while the quality of the formed clusters will be calculated using silhouette coefficient and purity.

Keywords : clustering, cophenetic correlation coefficient, silhouette coefficient, purity

1. Pendahuluan

1.1 Latar Belakang Masalah

Berita merupakan salah satu bentuk penyebaran informasi. Di dalam era yang berkembang saat ini, berita tidak hanya disebarkan melalui media elektronik dan media cetak seperti televisi dan koran. Akan tetapi, saat ini berita juga disebarkan melalui dunia internet. Sehingga, jumlah berita yang tersedia sangatlah besar dan menyulitkan bagi pembaca berita untuk mendapatkan secara tepat berita yang diinginkannya. Hal inilah yang kemudian mendorong kebutuhan untuk menemukan dan mengelola berita tersebut agar lebih terstruktur.

Clustering dapat menjadi alternatif untuk mengelompokkan berita. Secara umum *clustering* ada dua jenis, yaitu *hierarchical clustering* dan *partitioned clustering*. Pada tugas akhir ini digunakan *hierarchical clustering.* Ide dari *hierarchical clustering* adalah adanya pengelompokan data dengan membuat hirarki, dimana data yang memiliki kemiripan akan ditempatkan pada hirarki yang berdekatan dan yang tidak memiliki kemiripan pada hirarki yang berjauhan[4]. *Hierarchical clustering* mempunyai kelebihan dari *partitioned clustering.* Pertama, kefleksibelannya terhadap tingkat *granularity*, kedua *hierarchical clustering* mudah mengadaptasi berbagai ukuran kemiripan dan jarak sehingga dapat diimplementasikan ke berbagai tipe atribut[5], dan yang terakhir ialah melalui metode ini kita dapat melihat dokumen yang paling dekat dengan sebuah dokumen.

Hierarchical agglomerative clustering merupakan suatu pengelompokan hirarki yang bersifat *bottom up* dimana keberadaan setiap titik data dalam *cluster* ditentukan oleh *proximity* antar *cluster* tersebut. Metode *hierarchical agglomerative clustering* yang digunakan pada tugas akhir ini adalah *single linkage* (jarak terkecil), *complete linkage* (jarak terjauh), dan *average linkage* (jarak rata-rata). Metode ini berawal dari objek-objek individual yang paling mirip dikelompokkan dan kelompok-kelompok awal ini digabungkan sesuai dengan kemiripannya, berulang hingga jumlah *cluster* sesuai dengan yang diinginkan.

Dengan metode ini, dokumen direpresentasikan ke dalam bentuk hirarki *cluster* yang selanjutnya akan dikelompokkan ke dalam kelompok-kelompok yang berbeda. Selain itu juga akan dihitung *silhouette coefficient* untuk mengukur seberapa baik sebuah *hierarchical clustering* memenuhi kesesuaian data. Setelah itu dilakukan analisa dari hasil yang diperoleh melalui pengelompokan yang menggunakan metode *single linkage, complete linkage,* dan *average linkage* untuk mengetahui hirarki yang terbaik.

1.2 Perumusan Masalah

Berdasarkan pada latar belakang masalah diatas, maka permasalahan yang diangkat dalam tugas akhir ini ialah Jumlah berita yang tersedia sekarang sangatlah banyak, dan memerlukan waktu yang sangat lama apabila ingin mengelompokkannya secara manual sehingga diperlukan sebuah sistem yang dapat mengelompokkan berita tersebut.

Batasan masalah pada tugas akhir ini adalah:

- Dataset yang digunakan merupakan artikel berita berbahasa Indonesia dari portal berita *online* dengan jumlah 100 artikel
- Aplikasi bekerja secara *offline*
- *Stemming* yang digunakan adalah stemming bahasa Indonesia sehingga tidak mengatasi kata asing yang terdapat dalam dokumen
- Metode pembobotan yang digunakan ialah TF•IDF

1.3 Tujuan

Secara umum tujuan dari yang ingin dicapai dalam tugas akhir ini adalah :

1. Mengetahui serta menerapkan metode *Hierarchical Agglomerative Clustering* yaitu metode *single linkage, complete linkage,* dan *average linkage* dalam proses *clustering* dari dokumen berita berbahasa Indonesia.
2. Melakukan pengujian dan analisis berdasarkan *silhouette coefficient, purity,* dan *cophenetic correlation coefficient* dari implentasi metode *single linkage, complete linkage,* dan *average linkage* pada *clustering* dokumen.

1.4 Hipotesis

Agglomerative Hierarchical Clustering dapat digunakan untuk mengklaster dokumen-dokumen ke dalam klasternya. Metode *average linkage* memiliki kualitas *cluster* lebih baik dari *single linkage* maupun *complete linkage* karena menghitung jarak rata-rata antar *cluster* sehingga mengurangi pengaruh *outlier*.

1.5 Metodologi Penyelesaian Masalah

Metode yang digunakan dalam menyelesaikan tugas akhir ini adalah menggunakan metode studi pustaka atau studi literatur dan analisis dengan langkah kerja sebagai berikut :

1. Mencari dan mempelajari referensi bahan-bahan yang berhubungan dengan tugas akhir ini seperti *Text Mining, Hierarchical Agglomerative Clustering, single linkage,*

complete linkage, *average linkage*, dan bahan lain yang berhubungan dengan tugas akhir ini

2. Melakukan pencarian data yang akan dikelompokkan
3. Merancang aplikasi untuk melakukan pengelompokan data dan mengimplementasikannya ke dalam perangkat lunak
4. Melakukan pengujian sistem dengan data yang diperoleh
5. Melakukan analisis dari hasil pengujian
6. Membuat kesimpulan dari hasil implementasi dan analisis
7. Menyusun laporan tugas akhir

1.6 Sistematika Penulisan

Tugas Akhir ini disusun berdasarkan sistematika sebagai berikut :

BAB I : Pendahuluan

Pada bab ini berisi latar belakang masalah, perumusan masalah yang akan dibahas, batasan masalah, tujuan yang akan dicapai, metodologi penyelesaian, serta sistematika penulisan.

BAB II : Dasar Teori

Pada bab ini berisi dasar teori yang digunakan dalam membangun sistem untuk Tugas Akhir ini.

BAB III : Analisis dan Perancangan Sistem

Pada bab ini berisi analisis sistem yang meliputi gambaran umum dan analisa kebutuhan sistem, serta perancangan sistem

BAB IV : Implementasi dan Pengujian

Pada bab ini diuraikan mengenai hasil yang didapatkan dari *clustering* dokumen menggunakan metode *single linkage*, *complete linkage*, dan *average linkage*. Pengukuran performansi diukur menggunakan *silhouette coefficient*, *purity*, dan *cophenetic correlation coefficient*.

BAB V : Penutup

Bab ini akan berisi kesimpulan dan saran dari hasil pengujian yang dilakukan serta diberikan saran-saran untuk pengembangan lebih lanjut perangkat lunak ini.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Dari hasil pengujian yang telah dilakukan dapat diambil kesimpulan :

1. Tinggi rendahnya kualitas *cluster* dapat dipengaruhi oleh parameter masukan baik dari metode pengelompokan, jumlah *cluster*, maupun jumlah dokumen yang dikelompokan.
2. Penambahan jumlah dokumen dan Semakin kecil jumlah cluster memperbesar kemungkinan nilai purity lebih kecil dari 1 yang berarti ada dokumen yang berbeda kategori berada dalam cluster yang sama.
3. Berdasarkan hasil *clustering* dari keseluruhan skenario pengujian, metode *hierarchical agglomerative clustering* dengan pendekatan metode *average linkage clustering* menghasilkan performansi yang lebih baik dibandingkan metode *single linkage clustering* dan *complete linkage clustering* dimana hasil dari *clustering*-nya menghasilkan *cluster* yang berisi dokumen-dokumen yang berasal dari kategori yang sama.

5.2 Saran

1. Dilakukan perbandingan dengan menggunakan metode *clustering* secara partisi untuk membandingkan performansi dari masing-masing metode *clustering*.

DAFTAR PUSTAKA

[1] Al-Zoubi, M. B., & al Rawi, M. (2008). *An Efficient Approach for Computing Silhouette Coefficients.* Computer Science .

[2] Aranganayagi, S., & Thangavel, K. (2007). *Clustering Categorical Data using Silhouette Coefficients as a Relocating Measure.*

[3] Arumugavelu, S. (2007). *SIMD Algorithms for single link and complete link pattern clustering.*

[4] Berkhin, P. (2002). *A Survey of Clustering Data Mining Techniques.*

[5] Dubes, R., & Jain, A. (1988). *Algorithms for Clustering Data.*

[6] Farris, J. S. *On The Cophenetic Correlation Coefficient.* New York: Departement of Biological Sciences.

[7] Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook.* New York: Cambridge University Press.

[8] Fung, B. C., Wang, K., Ester, M., & Fraser, S. *Hierarchical Document Clustering.*

[9] Hasniawati, h. (2007). *Image Clustering berdasarkan warna untuk identifikasi buah dengan metode valley tracing.*

[10] He, Q. (1999). *A Review of Clustering Algorithms as Applied in IR.*

[11] Jain, A., Murty, M., & Flynn, P. (1999). *Data Clustering; A Review.*

[12] Karhendana, A. Analisis Cluster dan Representasi dokumen. *Dalam Pemanfaatan Document Clustering pada Agregator Berita.*

[13] Li, Y., Luo, C., & Chung, S. M. (2008). *Text Clustering with Feature Selection by Using Statistical Data.*

[14] Manning, C. D., Raghavan, P., & Schutze, H. (2009). *An Introduction to Information Retrieval.* Cambridge, England: Cambridge University Press.