PENGELOMPOKKAN OTOMATIS DOKUMEN BAHASA INDONESIA MENGGUNAKAN METODE HILL CLIMBING

Nigel Steven Souisa[1], Ema Rachmawati[2], Warih Maharani[3]

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

Abstrak

Dokumen dengan kategori yang sama dalam jumlah yang besar sulit untuk dibedakan kesamaannya antara dokumen yang satu dengan dokumen yang lainnya. Salah satu cara yang dapat digunakan untuk mengatasi masalah ini adalah dengan document clustering. Untuk jumlah cluster-nya, user bahkan tidak mengetahui berapa jumlah yang tepat untuk melakukan clustering dokumen-dokumen tersebut. Untuk itu diperlukan metode clustering yang dapat menghasilkan jumlah cluster secara otomatis. Satu dari banyak metode yang dapat digunakan untuk menghasilkan jumlah cluster secara otomatis adalah Hill Climbing.

Hill Climbing akan melakukan identifikasi terhadap pergerakan varian dari tiap tahap pembentukan cluster dan menganalisis polanya agar dapat menemukan nilai global optimum sehingga jumlah cluster akan terbentuk secara otomatis. Sedangkan untuk metode clustering yang digunakan adalah salah satu metode dalam Hierarchical Agglomerative, yaitu Centroid Linkage Hierarchical Method (CLHM). Setiap dokumen akan dianggap sebagai sebuah cluster, kemudian digabungkan dengan algoritma CLHM yang berulang hingga jumlah cluster sesuai dengan yang diinginkan. Hasil dari cluster yang terbentuk akan dihitung kualitasnya dengan parameter evaluasi dan nilai purity.

Kata Kunci : Clustering, CLHM, Hill Climbing, varians, purity

Abstract

Documents of the same category in large numbers is difficult to distinguish the similarity between documents with the other documents. One way that can be used to overcome this problem is to document clustering. For the number of clusters, users do not even know how much the right to perform clustering documents. So, It required a clustering method which can produce the number of clusters automatically. One of the many methods that can be used is Hill Climbing.

Keywords : Clustering, CLHM, Hill Climbing, variance, purity

1. Pendahuluan

1.1 Latar Belakang Masalah

Dunia pendidikan tidak terlepas dari tugas dan ujian. Dengan banyaknya tugas yang diberikan, terkadang membuat pengajar sulit untuk memeriksa apakah tugas yang dikerjakan oleh pelajar merupakan hasil sendiri atau meniru hasil tugas pelajar yang lain. Banyak cara yang digunakan pelajar untuk meniru hasil tugas pelajar yang lain, seperti melakukan perubahan susunan kalimat, mengganti suatu kata dengan sinonimnya, merubah kalimat aktif menjadi pasif, atau hanya menambah sedikit kata-kata. Hal inilah yang mempersulit pengajar dalam mengetahui apakah pelajar tersebut mengerjakan tugasnya secara mandiri atau tidak. *Clustering* dapat menjadi solusi dari masalah tersebut, dimana *clustering* dapat digunakan untuk menganalisis data dengan cara mengelompokkan objek ke dalam kelompok-kelompok berdasarkan suatu kemiripan tertentu sehingga semua anggota dari setiap partisi mempunyai kemiripan[5].

Sebuah *cluster* adalah sekumpulan objek yang digabung bersama karena kemiripan atau kedekatannya. Teknik *Clustering* sangat berguna karena akan mentranslasi ukuran persamaan yang intuitif menjadi ukuran yang kuantitatif. Dalam tugas akhir ini, metode *clustering* yang digunakan adalah CLHM (*Centroid Linkage Hierarchical Method*). Metode ini akan mengelompokkan dokumen berdasarkan nilai *centroid* terdekat. Untuk jumlah *cluster*-nya *user* tidak perlu menentukan berapa jumlah yang tepat, karena dengan metode *Hill Climbing*, jumlah cluster akan dipilih sehingga dihasilkan jumlah *cluster* secara otomatis.

Metode *Hill Climbing* ini berfungsi untuk melakukan identifikasi terhadap pergerakan varians yang dihasilkan dari tiap tahap pembentukan *cluster*. *Hill Climbing* akan mencari pada tahap mana terdapat *global optimum*, dengan cara menganalisis pola varians tersebut. Pengelompokkan dengan CLHM dan proses *Hill Climbing Automatic Clustering* sangat memudahkan *user* karena menghasilkan *cluster* secara otomatis dan tepat.

1.2 Perumusan Masalah

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

Batasan masalah pada tugas akhir ini adalah:

a. Dataset yang digunakan adalah dokumen abstrak tugas akhir dengan jumlah 50 dokumen.

b. Aplikasi bekerja secara *offline*

c. Stemming yang digunakan adalah stemming bahasa Indonesia sehingga tidak mengatasi kata asing yang terdapat dalam dokumen

d. Dokumen yang digunakan hanya dalam bentuk format .txt.

1.3 Tujuan

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

1. Mengetahui pengaruh *Hill Climbing* dalam *clustering* dokumen bahasa Indonesia.
2. Melakukan analisis hasil *clustering* metode *Hill Climbing*.

1.4 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, centroid linkage, Hill Climbing,* 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

1.5 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 analisis kebutuhan sistem, serta perancangan sistem

BAB IV : Implementasi dan Pengujian

Pada bab ini akan diuraikan mengenai hasil yang didapatkan dari *clustering* dokumen otomatis menggunakan metode *Hill Climbing* dan akan dilakukan analisis parameter evaluasi hasil *clustering* dan nilai *purity*-nya.

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 dan analisis yang telah dilakukan, maka dapat diambil kesimpulan :

1. *Clustering* dengan menggunakan algoritma *Centroid Linkage Hierarchical Method* dengan analisis pola varians *Hill Climbing* dapat digunakan untuk mengelompokkan dokumen bahasa Indonesia secara otomatis dimana *cluster* yang dihasilkan berasal dari kategori yang sama.
2. Penambahan jumlah dokumen akan memperbesar kemungkinan nilai *purity* = 1 yang berarti semua dokumen berada dalam *cluster* yang sesuai. Hal ini berpotensi untuk implementasi program dalam skala yang lebih luas.
3. Perubahan nilai alpha = 2, 3, atau 4 tidak mempengaruhi hasil *clustering* karena semakin besar nilai alpha hanya merubah nilai minimum, sementara nilai optimum selalu terletak pada tahap yang sama.

5.2 Saran

Dalam tugas akhir ini terdapat beberapa kelebihan dan kekurangan yang membutuhkan saran-saran untuk semakin mengembangkan tugas akhir ini sehingga bisa menjadi lebih sempurna. Adapun saran-saran yang diberikan adalah sebagai berikut :

1. Dilakukan perbandingan dengan menggunakan metode *cluster*ing secara *partitioning* untuk membandingkan performansi dan kualitas terdahap proses *clustering.*
2. Dilakukan perbandingan dengan menggunakan metode *stemming* yang berbeda, seperti *Potter stemming,* Arifin dan Setiono, dan sebagainya.

DAFTAR PUSTAKA

[1] Agusetia, Usmaida (2007), *Web Mining* Untuk Pencarian Berdasarkan Kata Kunci Dengan Teknik *Clustering*, Tugas Akhir Jurusan Teknologi Informasi Politeknik Elektronika Negeri Surabaya, Surabaya.

[2] A.R. Barakbah, *Clustering*, In. Workshop Data Mining 2006, Jurusan Teknologi Informasi Politeknik Elektronika Negeri Surabaya,ITS.

[3] A.R. Barakbah, K. Arai, *Identifying moving variance to make automatic clustering for normal data set*, In. Proc. IECI Japan Workshop 2004 (IJW 2004), Musashi Institute of Technology, Tokyo.

[4] Benjamin C. M. Fung, Ke Wang, and Martin Ester, Simon Fraser. *Hierarchical Document Clustering.* Canada

[5] Eldira, H. Web Mining untuk pencarian Dokumen Bahasa Inggris menggunakan *Hill Climbing Automatic Clustering*.

[6] Hasniawati Helmy (2007), *Image Clustering* Berdasarkan Warna Untuk Identifikasi Buah Dengan Metode *Valley Tracing*, Tugas Akhir Jurusan Teknologi Informasi Politeknik Elektronika Negeri Surabaya, Surabaya.

[7] Jain, A.K. Murty M.N and P.J Flynn. *Data Clustering: AReview.* The Ohio State University.

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

[9] Martiana, Entin. Mesin Pencari Dokumen dengan Peng*cluster*an Secara Otomatis.

[10] Wijanarto (2010). *Hill Climbing and Least-Cost*.