PENGELOMPOKAN TEKS MENGGUNAKAN ALGORITMA CANOPY CLUSTERING

Agha Dwi Nugraha[1], Angelina Prima Kurniati[2], Intan Nurma Yulita[3]

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

Abstrak

Peningkatan jumlah dokumen dalam format teks yang cukup signifikan belakangan ini membuat proses pengelompokan dokumen (document clustering) menjadi penting. Pengelompokan dokumen bertujuan membagi dokumen kedalam beberapa kelompok (cluster) sehingga dokumen-dokumen yang mempunya tingkat kesamaan tinggi termasuk dalam cluster yang sama dan yang mempunyai mempunya kesamaan rendah termasuk dalam cluster yang berbeda. Untuk melakukan pengelompokan tersebut, digunakan salah satu algoritma clustering yaitu Canopy Clustering. Canopy Clustering merupakan pengembangan dari Kmeans clustering. Algoritma ini dapat mengatasi permasalah yang terdapat pada K-means dalam masalah akurasi dan waktu proses untuk set data yang besar. Clustering dari nilai parameter T. Parameter ini berfungsi sebagai ukuran cluster pada pembentukan Canopy. Untuk mengukur similarity antar dokumen sebelum proses clustering digunakan Euclidean distance.

Pada tugas akhir ini cluster yang dihasilkan diukur akurasinya menggunakan precision, recall, dan F1-measure . Berdasarkan percobaan yang dilakukan bahwa Canopy Clustering dengan menggunakan K-means lebih tinggi tingkat akurasinya dan lebih sedikit waktu prosesnya dibandingkan dengan Algoritma K-means murni.

Kata Kunci : Canopy Clustering, K-means , Clustering

Abstract

An increasing number of documents in text format significantly lately makes the process of grouping documents (document clustering) becomes important. Grouping the document aims to divide the document into several groups (clusters) so that the documents possessed a high degree of similarity are included in the same cluster and possessed similarities that have low included indifferent clusters. To perform such clustering,clustering algorithms used one of the CanopyClustering. Canopy Clustering is a development of the Kmeans clustering. This algorithm can overcome the problems found on the Kmeans in amatter of accuracy and processing time for large data sets. Clustering of the value of the parameter T.This parameter serves as the cluster size on the formation of Canopy. To measure the similarity between the documents before the clustering process used Euclidean distance.

In this final cluster resulting accuracy is measured using precision,recall, and F1-measure. Based on experiments conducted that Canopy Clustering using K-means higher level of accuracy andless time to process compared to the K-means algorithm .

Keywords : Canopy Clustering, K-means, Clustering

1. Pendahuluan

1.1. **Latar Belakang**

Peningkatan jumlah dokumen dalam format teks yang cukup signifikan belakangan ini membuat proses pengelompokan dokumen (*document clustering*) menjadi penting. Pengelompokan dokumen bertujuan membagi dokumen kedalam beberapa kelompok (*cluster*) sehingga dokumen-dokumen yang mempunya tingkat kesamaan tinggi termasuk dalam *cluster* yang sama dan yang mempunyai mempunya kesamaan rendah termasuk dalam cluster yang berbeda.

Pada umumnya metode *clustering* teks menggunaan Algoritma *K-Means* [2]. *K-means* merupakan suatu metode memilih secara acak *k* buah data sebagai . Kemudian menempatkan data dalam *cluster* yang terdekat dihitung dari titik tengan *cluster* (*centroid*). Centroid baru ditentukan akan ditentukan bila semua data telah ditempatkan dalam *cluster* terdekat. Proses penentuan *centroid* dan penentuan data dalam *cluster* diulangi sampe nilai centroid konvergen. Namun kelemahan K-means adalah pada saat jumlah data set banyak. Pada jumlah data set yang banyak membutuhkan banyak iterasi hingga mencapai nilai yang konvergen sehingga membutuhkan waktu yang lama. Untuk mengatasi kelemahan tersebut telah ditemukan metode *Canopy Clustering[1]*. Maka dalam tugas akhir ini menggunakan metode *Canopy Clustering* dengan *K-means* untuk teks clustering.

Metode *Canopy Clustering* ini sering digunakan untuk membagi data set yang besar memnajdi beberapa kelompok (canopy) sebelum dilakukan teknil *clustering* yang lebih ketat, seperti *K-means*. Penghitungan jarak terdekat dihitung pada data yang ada dalam tiap *canopy* menggunakan *Euclidean Distence[3]*. Setiap dokumen dapat berada dibeberapa *l* dan menimal 1 buah *canopy. Clustering* selanjutnya digunakan K-means untuk setiap *canopy*- nya.

1.2. **Perumusan Masalah**

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

1. Bagaimana melakukan *pengelompokan* dokumen dengan menggunakan algoritma *Canopy Clustering* dengan *Kmeans*?
2. Bagaimana waktu proses teks *clustering* dengan algoritma *K-means* setelah menggunakan metode Canopy Clustering ?
3. Bagaimana performasi Metode *Canopy K-means* pada system teks *clustering* berdasarkan *confusion matriks* dengan parameter *F-measure* ?

1.3. **Batasan Masalah**

Adapun yang menjadi batasan-batasan masalah dalam pengerjaan tugas akhir ini adalah:

1. Tugas akhir ini menggunakan dokumen berita berbahasa Indonesia yang telah mengalami *text preprocessing* terlebih dahulu.

2. Tidak melakukan analisa terhadap *text preprocessing*.
3. Dokumen didapat dari artikel berita Berbahasa Indonesia.

1.4. **Tujuan**

Berdasarkan rumusan masalah yang telah diuraikan di atas, maka tujuan dari tugas akhir ini adalah:

1. Mengimplementasikan algoritma *Canopy Clustering* dalam mengelompokkan dokumen.
2. Mengetahui dan menganalisis waktu proses peng-*cluster* menggunakan Algoritma *K-means* setelah menggunakan metode *Canopy clutering*.
3. Menganalisa hasil implementasi metode *Canopy K-means* dalam teks *clustering* berparameter *F-measure* dilihat dari *precison* dan *recall*.

1.5. **Metodologi Penyelesaian**

Metodologi yang digunakan untuk menyelesaikan permasalahan dalam Tugas Akhir ini terdiri dari :

1. Studi literatur
Pencarian referensi dan sumber-sumber yang berhubungan dengan teks *clustering*.

2. Tahap pengumpulan data dan *preprocessing* data
Pada tahap ini akan dibangun model berkaitan dengan requirement-requirement yang dibutuhkan pada saat implementasi, mulai dari analisis kebutuhan, desain database, desain aplikasi/*interface*.

3. Tahap perancangan sistem
Pada tahap ini akan dibangun model berkaitan dengan requirement-requirement yang dibutuhkan pada saat implementasi, mulai dari analisis kebutuhan, desain database, desain aplikasi/*interface*.

4. Tahap Implementasi
Pada tahap ini, akan dilakukan implementasi sistem yang mampu mengklasifikasikan dokumen teks secara otomatis dengan menggunakan metode *K-means dan Canopy Clustering*.

5. Tahap Pengujian Sistem
Pada tahap ini akan dilakukan pengujian program yang telah diimplementasikan dengan melakukan memasukkan dokumen teks lalu melihat hasil klasifikasi.

6. Tahap Analisis Hasil Pengujian
Setelah sistem telah sempurna maka akan dianalisis hasil dari klasifikasi dan kestabilan dari sistem.

7. Tahap Pembuatan Laporan

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

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Dari hasil pengujian dan analisa yang dikerjakan dalam tugas akhir ini dapat diperoleh kesimpulan sebagai berikut:

1. Penentuan nilai parameter T pada *Canopy* merupakan bagian terpenting yang merupakan inputan awal dari proses *clustering* dalam algoritma *K-means*.
2. Nilai parameter T pada *Canopy K-means* berpengaruh terhadap nilai akurasi dan waktu, pada saat nilai parameter T kecil, akan menghasilkan waktu yang besar dan akurasi yang buruk dibandingkan dengan *K-means* . Untuk nilai parameter T yang besar dihasilkan akurasi yang lebih baik dan waktu yang lebih cepat dibandingkan dengan *K-means.* Akan teteapi apabila nilai T terlalu besar maka akan dihasilkan waktu dan akurasi yang tidak berbeda jauh dengan *K-means.* Maka dibutuhkan nilai parameter T yang tepat untuk medapat kan akurasi dan waktu proses yang lebih baik.
3. Nilai k pada *Canopy K-means* tidak terlihat perbedaan waktu yang signifikan dengan *K-means*.
4. Secara keseluruhan *Canopy K-means* mempunyai keunggulan dalam waktu proses.

5.2 Saran

Saran-saran untuk pengembangan tahap selanjutnya antara lain:

1. *Dataset* hasil *preprocessing* sangat mempengaruhi performansi sistem. Oleh karena itu sangat disarankan agar *dataset* yang digunakan dilakukan proses *feature selection* pada saat *preprocessing* agar *dataset* yang terbentuk lebih baik.
2. Mencari perhitungan parameter T yang tepat. Hasil yang dihasilkan pun akan lebih baik.

Daftar Pustaka

[1] McCallum, Nigam and Ungar: "Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching" http://www.kamalnigam.com/papers/canopy-kdd00.pdf. , diakses tanggal 14 Oktober 2010.

[2] Weisstein, Eric W. "K-Means Clustering Algorithm." From MathWorld--A Wolfram http://mathworld.wolfram.com/K-MeansClusteringAlgorithm.html, diakses tanggal 14 Oktober 2010

[3] Gower,JC.1981. Euclidean Distance. http://www.convexoptimization.com/TOOLS/Gower2.pdf, diakses pada tanggal 7 Januari 2011

[4] Ch, Milkha Harlian. 2007. Text Mining. http://lecturer.eepis-its.edu/~iwanarif/kuliah/dm/6Text%20Mining.pdf, diakses pada tanggal 23 Februari 2011..

[5] Gerberry, David J.2007. Clustering Algorithms [downloaded] http://www.math.purdue.edu/~gerberry/Research/Clustering.pdf, diakses pada tanggal 15 Oktober 2010

[6] Han, Jiawei dan Micheline Kamber. 2006. Cluster Analysis. Urbana-Champaign: Departement of Computer Science University of Illinois http://www.cs.uiuc.edu/homes/hanj/cs412/slides/07.ppt, diaskses 13 Februari 2011

[7] Hotho, Andreas., Alexander Maedche dan Steffen Staab. ____. Ontologybased
Text Document Clustering. Germany:Universitas of Karlsruhe http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.14.8083&rep=rep1&type=pdf, diaskses pada tanggal 20 Februari 2011

[8] Evaluation Method in Text Categorization. http://datamin.ubbcluj.ro/wiki/index.php/Evaluation_methods_in_text_categorization, diakses tanggal 14 Februari 2010.

[9] Marti Hearst: What Is Text Mining http://www.people.ischool.berkeley.edu/~hearst/text-mining.html, diakses tanggal diakses tanggal 14 Oktober 2010.

[10] Beil, Florian, Martin Ester, Xiaowei xu. 2002. Frequent Term-Based Text Clustering. ACM.

[11] K-Means Demo http://www.cs.cmu.edu/~zhuxj/courseproject/kmeansdemo/Kmeans.html, diakses tanggal 14 Oktober 2010.

[12] Manning C. D. and Schutze H., 1999. Foundations of Statistical Natural Language Processing [M]. Cambridge:MIT Press.