IMPLEMENTASI DAN ANALISIS ALGORITMA SINGLE PASS CLUSTERING UNTUK PENGELOMPOKAN DOKUMEN BERITA KEJADIAN BAHASA INDONESIA
IMPLEMENTATION AND ANALYSIS OF SINGLE PASS CLUSTERING ALGORITHM FOR INDONESIAN NEWS EVENT DOCUMENT CLUSTERING

David Widyarto[1], Imelda Atastina[2], Angelina Prima Kurniati[3]

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

Abstrak

Dengan semakin berkembanganya teknologi informasi, jumlah dokumen teks berupa digital semakin berkembang dengan pesat. Untuk melakukan pencarian informasi tentang kejadian-kejadian baru yang berkaitan dengan topik tertentu dalam kumpulan dokumen akan lebih sulit. Clustering merupakan salah satu metode data mining yang bersifat unsupervised learning untuk mengelompokkan dokumen berdasarkan kemiripannya. Untuk melakukan pngelompokan tersebut, digunakan salah satu algoritma clustering yaitu single pass clustering. Single pass clustering merupakan suatu tipe clustering yang berusaha melakukan pengelompokan data satu demi satu dan pembentukan kelompok dilakukan seiring dengan pengevaluasian setiap data yang dimasukkan ke dalam proses cluster. Jumlah cluster yang dihasilkan sangat tergantung pada nilai threshold. Kemiripan antar dokumen yang digunakan adalah standard cosine similarity.

Dalam Tugas Akhir ini kualitas cluster yang dihasilkan diukur dengan dua parameter yaitu cohesion dan separation. Nilai threshold 0.03 menghasilkan nilai rata-rata cohesion sebesar 0.13863266 dan rata-rata separation sebesar 0.009379814.

Kata Kunci : single pass clustering, unsupervised learning,cosine similarity, threshold, cohesion, separation.

Abstract

The growing up of information technology, couse of the number of digital text documents also grow up rapidly. To retrieve of information about current event related to specific topics in a documents collection will be more difficult. Clustering is one of data mining method which is unsupervised learning to classify documents based on similarity. One of clustering method is single pass clustering. Single pass clustering is a type of clustering algorithm that try to create group of data one by one and the formation of the group performed in line with the evaluation of any data entered into the cluster process. The number of clusters generated is depended on the threshold value. The similarity between the documents uses the standard cosine similarity.

In this final project, the quality of the clusters result is measured by two parameters cohesion and separation. Value of 0.03 and 0.029, can produce the average of cohesion 0.13863266 and the average of separation 0.009379814.

Keywords : single pass clustering, unsupervised learning, cosine similarity, threshold, cohesion, separation.

1. Pendahuluan

1.1 Latar belakang

Dengan semakin berkembangnya teknologi informasi jumlah dokumen teks berupa digital semakin berkembang dengan pesat. Hal ini mengakibatkan munculnya suatu cabang ilmu baru dalam teknologi informasi yaitu pencarian informasi (*information retrieval*). Pencarian informasi berbasis *query (query based retrieval)* yang digunakan secara tradisional sangat berguna untuk pencarian terarah tetapi tidak begitu efisien. *Query based retrieval* berguna ketika kita mengetahui benar kejadian asli atau fakta yang dicari. Cara ini tidak begitu efektif ketika kita membutuhkan informasi yang spesifik/khusus dalam kategori yang besar dan juga tidak efisien untuk mendapatkan berita yang relevan dalam penelusuran suatu kejadian. Dengan menggunakan cara ini biasanya hanya dapat diketahui berita-berita tertentu saja, sedangkan berita-berita lama yang berkaitan dengan berita tersebut sulit untuk ditelusuri. Selain pemilihan *query* yang tidak tepat akan menyebabkan membanjirnya dokumen- dokumen yang tidak relevan.

Pengelompokan (*clustering*) dokumen merupakan sebuah cara yang dapat digunakan untuk mempermudah pencarian dokumen dalam database. *Clustering* merupakan proses pengelompokan data sehingga semua anggota dari bagian data memiliki kemiripan berdasarkan perhitungan jarak/kemiripan term-term yang sudah dilakukan pembobotan pada tahapan sebelumnya (*preprocessing*). *Clustering* merupakan salah satu metode *data mining* yang bersifat tanpa arahan (*unsupervised*). Ada dua jenis *clustering* yang [3]sering dipergunakan dalam proses pengelompokan data yaitu *hierarchical* (hirarki) data *clustering* dan *non-hierarchical* (non hirarki) *clustering*. *Single pass clustering* merupakan salah satu metode data *clustering* non hirarki (*partisioning*) yang berusaha mengelompokkan data yang ada ke dalam bentuk satu atau lebih *cluster*. Salah satu aplikasi dari *clustering* adalah *document clustering*.

Tahapan-tahapan *clustering* yang pertama adalah representasi dokumen. Tiap-tiap dokumen akan direpresentasikan kedalam matrik dokumen berupa bobot dari term pada masing-masing dokumen, kedua adalah penentuan tingkat kemiripan antar dokumenn, ketiga adalah penggunaan algoritma *clustering*, dan yang terakhir adalah tahap evaluasi. Evaluasi ini bertujuan untuk mengetahui kinerja dari algoritma yang digunakan untuk klasterisasi dokumen. Sebelum melakukan pengelompokan dokumen, diperlukan tahap *preprocessing* [2] terlebih dahulu. Pada tahap *preprocessing* yang dilakukan adalah *case folding*, ,*stopwords*, *stemming* dan *term weighting*. Metode *term weighting* yang digunakan adalah TF-IDF yang merupakan kombinasi antara TF (*Term Frequency*) dengan IDF (*Invers Document Frequency*).

Tugas Akhir ini menerapkan pengelompokan dokumen dengan menggunakan algoritma *Single Pass Clustering*. *Single pass clustering* merupakan [14] suatu tipe *clustering* yang berusaha melakukan pengelompokan data satu demi satu dan pembentukan kelompok dilakukan seiring dengan evaluasi setiap data yang dimasukkan ke dalam proses *cluster*.

1.2 Perumusan masalah

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

1. Bagaimana melakukan pengelompokan dokumen dengan menggunakan algoritma *Single Pass Clustering* ?
2. Bagaimana mengetahui dan menganalisis kualitas *cluster* yang dihasilkan dengan menggunakan metode *cohesion* dan *separation* dan menganalisis parameter-parameter dari algoritma *Single Pass Clustering* yang berpengaruh terhadap kualitas *cluster* ?

1.3 Batasan masalah

Adapun batasan-batasan masalah dalam pengerjaan Tugas Akhir ini adalah:

1. Tugas Akhir ini menggunakan dokumen berita berbahasa Indonesia yang telah mengalami *preprocessing* terlebih dahulu.
2. Tidak melakukan analisis terhadap data *preprocessing*.
3. Kualitas *cluster* yang dihasilkan tidak dibandingkan dengan algoritma lain.
4. Tidak menangani kesalahan pada penulisan kata dalam dokumen.

1.4 Tujuan

Berdasarkan pada masalah yang telah didentifikasikan di atas, maka tujuan Tugas Akhir ini adalah :

1. Menerapkan Algoritma *Single Pass Clustering* serta mengetahui performansinya sebagai salah satu metode data mining dalam mengelompokkan dokumen berita kejadian bahasa Indonesia.
2. Menganalisis kualitas *cluster* yang dihasilkan dengan melihat nilai *cohesion* dan *separation* serta mengetahui parameter yang berpengaruh untuk menghasilkan kualitas *cluster* yang baik.

1.5 Metodologi penyelesaian masalah

Metode yang digunakan untuk menyelesaikan Tugas Akhir ini adalah :

1. Studi literatur dan tinjauan pustaka tentang :
   a. Algoritma *Single Pass Clustering*.
   b. Tahapan-tahapan pengelompokan dokumen berita bahasa Indonesia.
   c. Tahapan *preprocessing* dokumen berita.
   d. *Cluster* analisis dengan metode *cohesion* dan *separation*.
2. Pengumpulan data
   Data yang digunakan sebagai sampel adalah dokumen berita bahasa Indonesia yang berhasil di download dari media online seperti www.kompas.com , www.republika.com, www.detik.com dan www.tempointeraktif.com Analisis dan perancangan sistem menggunakan algoritma *Single Pass Clustering*

3. Implementasi

a. Ekstraksi dokumen
Proses ekstraksi ini bertujuan untuk menghasilkan *term-term* yang akan digunakan sebagai *prototype* bagi setiap dokumen yang berisi bobot term pada kumpulan dokumen. Ekstraksi dokumen dilakukan pada tahap *preprocessing.*

b. Penghitungan tingkat kemiripan
Perbandingan tingkat *similarity* yang digunakan pada TA ini adalah menggunakan *standard cosine similarity.*[4]

c. Pengelompokan dokumen.
Untuk melakukan pengelompokan dokumen algoritma yang digunakan adalah *Single Pass Clustering.* Dengan melakukan perbandingan tingkat *similarity* selanjutnya hasilnya akan dievaluasi untuk menentukan pasangan-pasangan dokumen yang dinyatakan mirip berdasarkan nilai *threshold* tertentu.

4. Testing dan analisis
Untuk mengetahui kinerja algoritma *Single Pass Clustering* pada tahap uji coba,dilakukan pengukuran berdasarkan kualitas *cluster* menggunakan dua parameter, yakni *cohesion* dan *separation.* Selain itu juga akan dilihat dari sisi waktu, berapa lama waktu eksekusi untuk melakukan pengelompokan dokumen dengan algoritma *Single Pass Clustering.*

5. Pengambilan kesimpulan dan penyusunan laporan Tugas Akhir.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan percobaan dan analisis yang dilakukan, maka dapat diambil beberapa kesimpulan sebagai berikut:

1. Jumlah *cluster* yang dihasilkan berpengaruh terhadap waktu eksekusi program. Semakin banyak jumlah *cluster* yang dihasilkan maka semakin sedikit waktu yang dibutuhkan.
2. Jumlah dan kualitas *cluster* dipengaruhi oleh nilai parameter *threshold*. Semakin besar nilai *threshold* maka jumlah *cluster* yang dihasilkan juga semakin besar. Pemilihan nilai *threshold* yang tepat akan menghasilkan kualitas *cluster* yang bagus.
3. Pada nilai *threshold* yang sama, urutan eksekusi dokumen berpengaruh terhadap jumlah dan kualitas *cluster* yang dihasilkan. Hal ini disebabkan oleh bobot masing-masing dokumen yang berbeda. Perbedaan bobot dokumen berpengaruh terhadap nilai *update cluster centroid* baru.

5.2 Saran

Sebagai saran untuk perkembangan Tugas Akhir selanjutnya, perlu dilakukan pertimbangan analisis sebagai berikut:

1. Pemilihan metode evaluasi *cluster* lain untuk pengujian kualitas *cluster* terhadap jumlah *cluster* yang berbeda-beda.
2. Sistem yang dibuat tidak hanya untuk melakukan pengelompokan dokumen, tetapi perlu diuji juga dengan *search engine* yang mampu me-*retrieve* dokumen dari hasil kluterisasi.

Daftar Pustaka

[1] Arifin, Agusta Zainal dan Ari Novan Setiono, *Klasifikasi Dokumen Berita Kejadian Berbahasa Indnesia dengan Slagoritma Single Pass Clutering*
http://www.its.ac.id/personal/files/pub/667-agusza-SITIAKlasifikasiEvent.pdf
diakses pada tanggal 28 Nov 2009

[2] Ch, Milkha Harlian. 2007. *Text Mining*.
http://lecturer.eepis-its.edu/~iwanarif/kuliah/dm/6Text%20Mining.pdf
Tanggal akses: 23 oktober 2009.

[3] Chu, Hio Hong Steven 2009, *Cluster Analysis: Basic Concepts,* Nyanyang technological university,
https://svn.mosuma.net/r4000/doc/course/ci6227/public/lectures/lecture07*cluster*.pdf
diakses pada tanggal 29 september 2010

[4] Frakes, Willam B dan Ricardo Baeza-Yates 1992, *Information Retrieval data Structur and Algorithms.*
http://www.google.com/books?id=IHpggU5LZDsC&printsec=frontcover&hl=id&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false
diakses pada tanggal 20 agustus 2010

[5] Garcia, E., *An information retrieval tutorial on cosine similarity measures, dot products and term weight calculations*, 2006, http://www.miislita.com/information-retrieval-tutorial/cosine-*similarity*-tutorial.html
Diakses pada tanggal 21 Maret 2010

[6] Han, Jiawei dan Micheline Kamber. 2006. *Data preprocessing*. Urbana-Champaign: Departement of Computer Science University of Illinois
http://www.cs.uiuc.edu/homes/hanj/cs412/slides/02.ppt
Tanggal akses: 20 Maret 2010.

[7] Klampanos A., Jose J. M., and van Rijsbergen 2006. *Single Pass Clustering for Peer-to-Peer Information Retrieval: The Effect of Document Ordering.*

[8] Mobasher, Bamshad, *Single Pass Clustering Technique.* 2002, School of CTI, DePaul University
http://maya.cs.depaul.edu/classes/ds575/single-pass.html
diakses pada tanggal 15 juni 2010

[9] Pantel, Patrick Andre. 2003. *Clustering by Committee*. Canada: University of Alberta
http://www.patrickpantel.com/download/papers/2003/cbc.pdf
Tanggal akses: 27 oktober 2009.

[10] Papka, Ron dan James Alan. 1998, *On-Line New Event Detection using Single Pass Clustering.* Center for Intellegent Information Retrieal Departement of Computer Science University of Massachusetts
http://ciir.cs.umass.edu/pubfiles/ir-123.pdf
diakses pada tanggal 3 Mei 2010

[11] Tan, Steinbach dan Kumar. 2004. *Cluster Analysis: Basic Concepts and Algorithms*.
http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf
diakses pada tanggal Mei 2010.

[12] Tan, Pang-ning, Michael Steinbach, dan Vipin Kumar. Introduction to Data
mining. Pearson education, Inc.2006.

[13] ___. *Text Clustering*.
http://userweb.cs.utexas.edu/~mooney/ircourse/slides/Text*Cluster*ing.ppt
Tanggal akses: 24maret 2010

[14] ___Single pass *cluster*ing
http://yudiagusta.wordpress.com/2010/07/19/single-pass-*cluster*ing/
diakses pada tanggal 25 April 2010

[15] ____. *Cluster*ing
http://compdiag.molgen.mpg.de/ngfn/docs/2007/sep/*Cluster*ing.pdf
diakes pada tanggal 14 Maret 2010