ANALISIS PENERAPAN ALGORITMA COMMITTEE CLUSTERING PADA PENGELOMPOKAN DOKUMEN

Siti Rahmawati[1], Imelda Atastina[2], Shaufiah[3]

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

Abstrak

Saat ini jumlah dokumen teks berbentuk digital semakin banyak dan beragam. Untuk memudahkan pengambilan informasi yang dibutuhkan dari dokumen teks tersebut, maka perlu dilakukan pengelompokkan dokumen sesuai dengan topiknya. Pengelompokkan ini dapat dilakukan dengan menggunakan teknik yang terdapat dalam data mining yaitu clustering. Clustering merupakan teknik yang bersifat unsupervised learning, yaitu data tidak diberi label sehingga algoritma clustering yang akan mengelompokkan dokumen berdasarkan nilai kesamaan. Untuk melakukan pengelompokkan dokumen tersebut, digunakan algoritma committee clustering yang bekerja dengan cara membangun pusat cluster dengan merata-ratakan nilai feature vector dari himpunan bagian dari anggota cluster yang disebut committee, yang akan bertugas untuk menentukan suatu dokumen untuk masuk ke dalam suatu cluster. Dengan memilih anggota committee secara hati-hati, feature dari pusat cluster akan cenderung mengarah kepada class target[8].

Dalam tugas akhir ini, jumlah cluster yang dibentuk oleh algoritma committee clustering disesuaikan dengan jumlah kategori dari data yang digunakan dan didapatkan nilai rata-rata silhouette coefficient sebesar 0.2296. Dengan demikian, kualitas cluster yang dihasilkan bersifat no structure.

Kata Kunci : algoritma, committee clustering, clustering, dokumen, feature,

Abstract

In this era, the number of digital text documents grows rapidly and in large number of variation. To facilitate the easiest way to retrieve information from those digital text documents, then classifying those documents are surely needed according to the topic. The classifying can be conducted by using data mining technique that called clustering. Clustering is an unsupervised learning technique which have meaning that the data is not being labeled, but then the clustering algorithm will classify documents based on their similarity. To make the document clustering, 'committee clustering' algorithm is used, which is working by building a cluster center with uniformly averaged value of feature vector of a subset of cluster members that called the committee. This committee will be responsible for determining whether a document is a part of a cluster or not. By selecting the committee members carefully, the feature of the center cluster will tend to lead to the target class [8].

In this final project, the number of cluster which formed by the committee clustering algorithm is adjusted with number of categories of data that used. That number of clusters produced the average value of silhouette coefficient 0.2296. Thus the quality of the resulting clusters are no structure.

Keywords : algorithm, committee clustering, clustering, document, feature,

1. Pendahuluan

1.1 Latar belakang

Saat ini jumlah dokumen teks berbentuk digital semakin banyak. Dokumen-dokumen tersebut menawarkan berbagai informasi yang beragam. Untuk memudahkan pengambilan informasi yang dibutuhkan dari dokumen teks tersebut, maka perlu dilakukan pengelompokkan dokumen sesuai dengan topiknya. Pengelompokkan ini dapat dilakukan dengan menggunakan teknik yang terdapat dalam *data mining* yaitu *clustering*. *Clustering* adalah proses mengelompokkan objek berdasarkan informasi yang diperoleh dari data yang menjelaskan hubungan antar objek dengan prinsip untuk memaksimalkan kesamaan antar anggota dalam satu kelompok dan meminimumkan kesamaan antar kelompok [16]. *Clustering* merupakan teknik yang bersifat *unsupervised learning*, yaitu data tidak diberi label sehingga data menemukan pola distribusinya sendiri. Dengan demikian, algoritma *clustering* akan bekerja untuk menggelompokkan dokumen berdasarkan kesamaan anggotanya [9].

Terdapat tantangan yang harus dihadapi dalam menyelesaikan masalah pengelompokkan dokumen, antara lain jumlah dokumen yang besar dengan dimensi data yang tinggi sehingga algoritma *clustering* yang digunakan harus mampu menghasilkan kompleksitas waktu dan ruang yang efisien. Selain itu, algoritma yang digunakan pun harus dapat menghasilkan *cluster* dengan kualitas yang baik. Oleh karena itu, dalam tugas akhir ini akan dibangun sistem pengelompokkan dokumen dengan menggunakan algoritma *committee clustering*. Algoritma ini akan membangun pusat *cluster* dengan merata-ratakan nilai *feature vector* dari himpunan bagian dari anggota *cluster* yang disebut *committee*. *Committee* ini, nantinya akan bertugas untuk menentukan suatu dokumen untuk masuk ke dalam suatu *cluster*. Dengan memilih anggota *committee* secara hati-hati, *feature* dari pusat *cluster* akan cenderung mengarah kepada *class* target [8]. Dengan demikian, algoritma ini akan mampu menghasilkan kualitas *cluster* yang baik.

1.2 Perumusan masalah

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

1. Bagaimana melakukan pengelompokkan dokumen dengan menggunakan algoritma *committee clustering*?.
2. Bagaimana mengetahui dan menganalisis kualitas *cluster* yang dihasilkan dengan menggunakan metode *silhouette coefficient* dan menganalisis parameter-parameter dari algoritma *committee 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 *proses text preprocessing* terlebih dahulu.
2. Kualitas *cluster* yang dihasilkan tidak dibandingkan dengan algoritma lain.

1.4 Tujuan

1. Melakukan pengelompokkan dokumen dengan menggunakan algoritma *committee clustering*.
2. Mengetahui dan menganalisis kualitas *cluster* yang dihasilkan dengan menggunakan metode *silhouette coefficient* dan menganalisis parameter-parameter dari algoritma *committee clustering* yang berpengaruh terhadap kualitas *cluster*.

1.5 Metodologi penyelesaian masalah

Metodologi yang digunakan untuk menyelesaikan permasalahan-permasalahan dalam Tugas Akhir ini terdiri dari 6 tahap, yaitu:

1. Studi pustaka
   Pada tahapan ini akan dilakukan pendalaman materi mengenai algoritma *committee clustering*.
2. Tahap pengumpulan data dan *preprocessing* data
   Pada tahapan ini dilakukan pencarian dokumen berita berbahasa Indonesia. Setelah itu, dilakukan proses *preprocessing* terhadap dokumen sehingga dokumen tersebut dapat digunakan sebagai masukan untuk sistem yang akan dibangun.
3. Tahap perancangan sistem
   Pada tahapan ini dilakukan proses perancangan sistem dengan membuat gambaran fungsi-fungsi yang akan membangun sistem pengelompokkan dokumen dengan algoritma *committee clustering*.
4. Tahap implementasi pemrograman
   Pada tahapan ini dilakukan proses pembangunan sistem berdasarkan rancangan yang telah dibuat sebelumnya dengan menggunakan *tools matlab*.
5. Tahap pengujian sistem
   Pada tahapan ini dilakukan proses pengujian sistem yang menggunakan sejumlah dokumen dengan mengganti-ganti nilai parameter *k* dan nilai *threshold*, dan dievaluasi dengan menggunakan metode *silhouette coefficient*.
6. Tahap analisis output sistem
   Pada tahapan ini dilakukan analisa kualitas *cluster* yang dihasilkan oleh sistem.
7. Tahap Pembuatan Laporan.
   Pada tahap ini dilakukan penyusunan laporan akhir dan pengumpulan dokumentasi berdasarkan analisa hasil pengujian tugas akhir ini.

1.6 Sistematika Penulisan

Tugas akhir ini disusun dengan sistematika sebagai berikut:

1. **Pendahuluan**
   Bab ini menguraikan tugas akhir ini secara umum, meliputi latar belakang, perumusan masalah, batasan masalah, tujuan dan metodologi penyelesaian masalah.

2. **Dasar Teori**
   Bab ini membahas mengenai uraian teori yang berhubungan dengan *clustering* dan algoritma *committee clustering*.

3. **Analisis Perancangan Dan Implementasi**
   Bab ini berisi analisis kebutuhan dari sistem yang kemudian dituangkan ke dalam suatu sistem pemodelan secara terstruktur. Dari tahap analisis kemudian dilanjutkan ke tahap perancangan dan implementasi.

4. **Analisis Hasil Pengujian**
   Bab ini membahas mengenai pengujian yang dilakukan terhadap sistem yang telah dibangun. Pengujian dilakukan dengan mengganti-ganti nilai parameter yang terdapat dalam sistem. Tahap pengujian dilanjutkan dengan tahap analisis hasil pengujian.

5. **Kesimpulan**
   Berisi kesimpulan dari penulisan Tugas Akhir ini dan saran-saran yang diperlukan untuk pengembangan lebih lanjut.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Setelah semua tahapan pengerjaan TA ini selesai dilakukan, diperoleh beberapa kesimpulan:

1. Sulit menentukan kombinasi yang tepat untuk nilai parameter *k*, *threshold* α dan β untuk membentuk jumlah *cluster* yang diinginkan sehingga pengujian yang dilakukan bersifat *trial and error*.
2. Cluster yang dihasilkan dipengaruhi oleh nilai parameter k, *threshold* α dan β.
    - Jika semakin besar nilai parameter *k* maka akan semakin besar pula kemungkinan calon kandidat *committee* yang dihasilkan merupakan *cluster* yang besar dan ketat.
    - Semakin besar nilai *threshold* α maka kualitas *cluster* yang dihasilkan akan semakin menurun karena *committee* yang dihasilkan akan semakin mirip
    - Semakin besar *nilai threshold* β maka kualitas *cluster* yang dihasilkan akan semakin baik, karena akan semakin ketat suatu *committee* dalam menyeleksi dokumen.
3. Berdasarkan nilai *silhouette* yang dihasilkan dari pengujian ini yaitu membentuk *cluster* sesuai dengan kategori yang terdapat dari data, diperoleh bahwa kualitas *cluster* yang dihasilkan kurang bagus.

5.2 Saran

Saran untuk pengembangan tugas akhir ini yaitu lebih ditekankan lagi terhadap proses *preprocessing* yang dilakukan terhadap dataset dengan melakukan proses seleksi *feature* (*unsupervised feature selection*). Hal ini dikarenakan, dataset hasil *preprocessing* yang akan digunakan juga mempengaruhi performansi sistem. Selain itu, dapat dicoba dengan menggunakan dokumen-dokumen berdasarkan kejadian.

Daftar Pustaka

[1] Barakbah, Ali Ridho. 2006. *Clustering*. Surabaya: EEPIS
http://lecturer.eepis-its.edu/~entin/Machine%20Learning/Minggu%206%20*Cluster*ing.pdf
Tanggal akses: 10 maret 2010.

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

[3] 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
Tanggal akses: 20 maret 2010.

[4] 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.

[5] Hotho, Andreas., Alexander Maedche dan Steffen Staab. ____. *Ontology-based Text Document Clustering*. Germany:Universitas of Karlsruhe
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.14.8083&rep=rep1&type=pdf
Tanggal akses: 25 juli 2010

[6] Jain, A.K., M.N. Murty dan P.J. Flynn. 1999. *Data Clustering: A Review.*
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.18.2720&rep=rep1&type=pdf
Tanggal Akses: 15 Maret 2010.

[7] Pantel, Patrick dan Dekang Lin. 2002. *Document Clustering with Committees*.
http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.87.3216
Tanggal akses: 9 maret 2010.

[8] Pantel, Patrick dan Dekang Lin. 2002. *Efficiently Clustering Documents with Committees*. Canada: University of Alberta.
http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.110.6785
Tanggal akses: 23 oktober 2009.

[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] Rousseeuw , Peter J. 1986. *Silhouette (Clustering).*
http://en.wikipedia.org/wiki/*Silhouette*_(*cluster*ing)
Tanggal akses: 26 Maret 2010.

[11] Tan, Steinbach dan Kumar. 2004. *Cluster Analysis: Basic Concepts and Algorithms.*
http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf
Tanggal akses: 26 Maret 2010.

[12] ___. 2002. *Clustering*
http://rakaposhi.eas.asu.edu/cse494/notes/f02-*cluster*ing.ppt
Tanggal akses: 26 maret 2010

[13] ___. *Cluster analysis.*
http://en.wikipedia.org/wiki/*Cluster*_analysis
Tanggal akses: 26 maret 2010

[14] ___. *Data mining*
http://en.wikipedia.org/wiki/Data_mining
Tanggal akses: 26 maret 2010

[15] ___. *Text Clustering.*
http://userweb.cs.utexas.edu/~mooney/ir-course/slides/Text*Cluster*ing.ppt
Tanggal akses: 26 maret 2010

[16] ___. *Text Preprocessing*
http://pinkynet.web.id/2009/09/08/text-prepocessing-text-mining/
Tanggal akses: 10 maret 2010