ANALISIS PENERAPAN ALGORITMA CURE PADA PENGELOMPOKKAN DOKUMEN

R. A. Isteri Yohana[1], Erwin Budi Setiawan[2], Erda Guslinar Perdana[3]

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

Abstrak

Seiring dengan perkembangan teknologi, semakin banyak informasi yang diberikan dalam bentuk dokumen teks digital. Dokumen teks memiliki informasi yang beraneka ragam, sehingga untuk memudahkan dalam pengambilan informasi agar sesuai dengan keinginan perlu adanya pengelompokan dokumen.Clustering adalah proses untuk mengelompokkan data ke dalam suatu klaster, sehingga objek pada suatu klaster memiliki kemiripan yang sangat besar dengan objek lain pada klaster yang sama, tetapi sangat tidak mirip dengan objek pada klaster yang lain. Clustering yang dilakukan terhadap dokumen-dokumen disebut document clustering.
Pada tugas akhir ini diimplementasikan suatu algoritma clustering yaitu algoritma cure. Algoritma cure merupakan algoritma yang bekerja dengan cara mengukur jarak antar dokumen dengan list poin perwakilan cluster yang sudah dipilih sebelumnya.
Pengujian yang dilakukan dengan menghitung jumlah cluster dan menghitung nilai cohesi dan separation cluster yang dihasilkan menggunakan algoritma cure . Berdasarkan nilai cohesi yang dihasilkan pengujian ini dalam membentuk cluster yang sesuai dengan kategori yang terdapat dari dataset diperoleh bahwa kualitas yang dihasilkan cukup bagus sekitar 0.0855. Sedangkan berdasarkan nilai separation yang dihasilkan pengujian ini yaitu 0.927039 meskipun membentuk cluster yang tidak sesuai dengan kategori dataset. Akan tetapi kualitas cluster yang dihasilkan cukup bagus, karena kualitas clustering baik jika semakin kecil nilai cohesi dan semakin besar nilai separation.

Kata Kunci : algoritma, cure, clustering, dokumen, cohesi, separation

Abstract

During the development of technology, there’s more information provided in the form of digital text documents. The document text has a lot of type of information, so to ease in retrieving information that match with the one we want, there’s need for grouping of document. Clustering is a process for classifying data into a cluster, so the objects in a cluster has a very large similarity with other objects in the same clusters, but has very litle similarity to the object on the other clusters. Clustering that performed on the documents referred as document clustering.
In this final task a clustering algorithm is implemented, that is Cure Algorithm. Cure algorithm is an algorithm that works by measuring the distance between documents with points representative list of the cluster that has been previously selected.
Testing is done by counting the number of cluster and calculate the value of cohesion and cluster separation that produce using Cure Algorithm. Based on the value of cohesion that is produced in this test in forming the right cluster with the category that is from dataset, it is obtained that the quality that have been produced is good enough around 0.0855. As based the value of separation that have been produced in this test is 0.927039 although forming clusters that do not fit with dataset category. But the cluster quality that have been produced is good enough, because the quality of clustering is good if the smaller cohesion value and the bigger separation value.

Keywords : cure algorithm, clustering, document, cohesion, separation.

1. Pendahuluan

1.1 Latar Belakang Masalah

Pada perkembangan teknologi sekarang ini begitu banyak dokumen teks dalam bentuk digital. Dengan banyaknya dokumen, informasi yang beredar juga semakin banyak sehingga untuk mempermudah mengambil suatu informasi sesuai yang dibutuhkan, perlu dilakukannya 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 satu kelas dan meminimumkan kesamaan antar kelas/*cluster*. *Clustering* juga dikenal sebagai *unsupervised learning* yang membagi data menjadi kelompok-kelompok atau *cluster*s berdasarkan suatu kemiripan atribut-atribut diantara data tersebut [3].

Clustering sendiri terbagi atas 3 yaitu *hierarchical* dan *partitional dan Clustering Large Data*. *Hierarchical clustering* adalah suatu struktur data berbentuk pohon yang disebut dendogram dimana data dikelompokkan secara bertingkat dari yang paling bawah dimana tiap instance data merupakan *cluster* sendiri, hingga tingkat yang paling atas. *Partitional clustering* yang mengelompokkan data ke dalam k *cluster* dimana k adalah banyak *cluster* dari input user. *Clustering Large Data* dibutuhkan untuk melakukan *cluster*ing pada data yang volumenya sangat besar sehingga tidak cukup ditampung dalam memori komputer pada suatu waktu [11].

Ada beberapa algoritma yang telah diimplementasikan untuk *clustering* misalnya algoritma Cobweb, algoritma K means dan lain-lain. Algoritma-algoritma tersebut masih memiliki beberapa kekurangan, sedangkan ada kendala yang sering terjadi dan harus dihadapi untuk menyelesaikan masalah pengelompokkan dokumen pada *clustering*. Kendala tersebut misalnya memiliki outlier atau jumlah dokumen yang besar. Pada tugas akhir ini, algoritma yang akan digunakan untuk membangun sistem pengelompokan dokumen adalah algoritma *clustering using representative* atau sering disebut dengan algoritma *cure*. Algoritma *cure* merupakan metode campuran *hierarchical* dan *partitional* yang berdasarkan poin perwakilan yang ditentukan sebelumnya[4]. Keuntungan dari algoritma ini dapat menangani data yang besar [4]. Data besar dapat berupa data yang berdimensi tinggi[6], dokumen memiliki banyak dimensi. Untuk itu, algoritma *cure* dipilih untuk melakukan pengelompokan dokumen karena dapat menangani jumlah data yang besar tanpa mengorbankan kualitas *clustering*. Untuk membuktikan kualitas *cluster* yang dihasilkan algoritma ini, maka pada tugas akhir ini akan dianalisa kualitas *cluster*.

.

1.2 Perumusan Masalah

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

1. Bagaimana menerapkan metode algoritma *CURE (clustering using representative)* untuk pengelompokkan dokumen bahasa indonesia?
2. Bagaimana menganalisis kualitas hasil *cluster* dengan mengukur nilai *cohesion* (kemiripan objek intra *cluster*) dan *separation* (ketidakmiripan objek antar *cluster*) serta pengaruh perubahan nilai *threshold* terhadap akurasi hasil *cluster* pada algoritma *CURE (clustering using representative)*?

1.3 Batasan Masalah

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

1. Data set yang digunakan berupa dokumen berita berbahasa Indonesia
2. Dokumen berita didapat dari artikel di www.kompas.com.
3. Data masukan sistem yang telah mengalami proses *text preprocessing* terlebih dahulu dan tidak dibahas lebih detail pada tugas akhir ini.

1.4 Tujuan

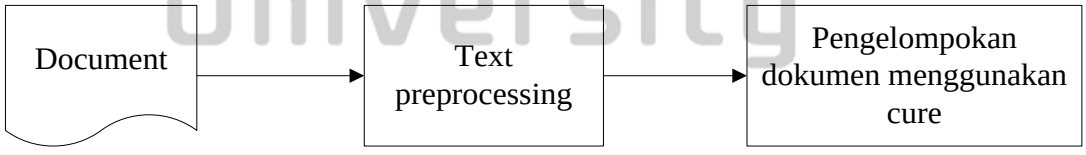
Adapun tujuan dalam pengerjaan Tugas Akhir ini adalah:

1. Menerapkan metode algoritma *CURE (clustering using representative)* untuk pengelompokkan dokumen bahasa Indonesia yaitu dengan mengimplementasikan algoritma *CURE* pada pengelompokkan dokumen.
2. Menganalisis kualitas hasil *cluster* dengan mengukur nilai *cohesion* (kemiripan objek intra *cluster*) dan *separation* (ketidakmiripan objek antar *cluster*) serta pengaruh perubahan parameter nilai P dan *threshold* terhadap akurasi hasil *cluster* pada algoritma *CURE (clustering using representative).*

1.5 Metodologi Penyelesaian Masalah

Metodologi yang digunakan untuk menyelesaikan permasalahan-permasalahan dalam Tugas Akhir ini terdiri dari 7 tahap, yaitu:

1. Studi pustaka/studi literature
   Tahapan ini adalah mengumpulkan bahan-bahan referensi Tugas Akhir yang meliputi algoritma *cure, data mining* , evaluasi kualitas *cluster* dan dokumen yang akan dijadikan dataset serta mempelajari dan memperdalam materi mengenai algoritma *cure(clustering using representative)*. Sumber referensi berupa buku, jurnal , e-book dan browse internet.
2. Tahap pengumpulan data dan *preprocessing* data
   Tahapan ini adalah melakukan pencarian dokumen berita berbahasa Indonesia. Setelah itu, dilakukan proses *text preprocessing* terhadap dokumen sehingga dokumen tersebut dapat digunakan sebagai data masukan untuk sistem yang akan dibangun.
3. Tahap perancangan sistem
   Tahapan ini melakukan proses perancangan sistem dengan membuat gambaran fungsi-fungsi yang akan membangun sistem pengelompokkan dokumen dengan algoritma *cure(clustering using representative)*. Berikut adalah rancangan system secara umum:



Gambar 1-1 Perancangan Sistem

4. Tahap implementasi pemrograman

Tahapan ini adalah melakukan proses pembangunan sistem berdasarkan rancangan yang telah dibuat sebelumnya menggunakan tools matlab.

5. Tahap pengujian sistem
   Tahapan ini melakukan pengujian pada sistem yang telah dibuat dengan menggunakan dataset yang telah disediakan dan mengevaluasi hasil *cluster* dengan melihat nilai *cohesion* dan *separation*.
6. Tahap analisis output sistem
   Tahapan ini melakukan analisis kualitas *cluster* yang dihasilkan oleh sistem.
7. Tahap Pembuatan Laporan.
   Tahapan ini melakukan 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 *cluster*ing dan algoritma *cure*.

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

Kesimpulan yang dapat diambil setelah melakukan implementasi, pengujian dan analisis pada tugas akhir ini adalah:

1. Nilai parameter P dan nilai threshold memiliki peranan terhadap hasil *cluster*ing menggunakan algoritma cure, yaitu:
    a. semakin besar nilai parameter P yang diinputkan menghasilkan jumlah *cluster* yang semakin sedikit, dan sebaliknya menghasilkan jumlah *cluster* yang semakin banyak. Sehingga dapat disimpulkan semakin besar nilai parameter P maka \jumlah *cluster* yang dihasilkan cenderung sedikit.
    b. semakin besar nilai threshold yang diinputkan menghasilkan jumlah *cluster* yang semakin banyak, dan sebaliknya menghasilkan jumlah *cluster* yang semakin sedikit. Sehingga dapat disimpulkan semakin besar nilai threshold, maka jumlah *cluster* yang dihasilkan cenderung banyak.
2. Berdasarkan hasil pengujian ini, yaitu membentuk *cluster* yang sesuai dengan kategori data yang diiputkan, diperoleh bahwa kualitas *cluster* yang dihasilkan kurang bagus.

5.2 Saran

Saran untuk pengembangan lebih lanjut mengenai penelitian yang dilakukan pada tugas akhir ini adalah melakukan normalisasi pada data inputan setelah tahapan preprocessing, hal ini bertujuan agar menghasilkan dataset yang lebih baik.

Daftar pustaka

[1]Away, Gunaidi A, 2006, *The Shortcut of Matlab Programming*, Penerbit Informatika,Bandung

[2]Ch.Milkha Harlian.2007.Text Mining.
http://lecturer.eepis-its.edu/~iwanarif/kuliah/dm/6*Text*%20Mining.pdf
diakses pada tanggal 24 Maret 2011

[3]Dunham, Margaret H.2003. *Data Mining Introductory and Advanced Topics.* New Jersey. Prentice Hall

[4]Guha, Sudipto, Ratogi, Rajeev. Shim, Kyuseok 1998, : An Efficient *Cluster*ing Algorithm for Large Databases. Proceeding of ACM SIGMOD Internasional Conference on Management of Data, pages 73-84, New York , 1998. ACM

[5]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
diakses 13 oktober 2010

[6]Heiki, David Hand; Smyth Mannila Padhraic. Principles of Data Mining. 2001

[7]Pressman, Roger S, 2002, *Rekayasa Sistem*,Penerbit Andi, Yogyakarta.

[8]__. Algoritma cure
http://en.wikipedia.org/wiki/CURE_data_*cluster*ing_algorithm
diakses 12 oktober 2010

[9]__. *Cluster*ing
http://rakaposhi.eas.asu.edu/cse494/notes/f02-*cluster*ing.ppt
diakses 13 oktober 2010

[10]__. *Cluster* analysis
http://en.wikipedia.org/wiki/*Cluster*_analysis/
diakses 13 oktober 2010

[11]__. Jain, A K, dkk.2000. *Data Clustering: A Review*, [online],
http://www.cs.rutgers.edu/~mlittman/courses/lightai03/jain99data.pdf
diakses 10 oktober 2010

[12]__. Pyle, D., 1999 *Persiapan. Data untuk Data Mining* Morgan Kaufmann *Publishers.*, Los Altos , CA.
http://en.wikipedia.org/wiki/Data_Pre-processing
diakses 13 oktober 2010

[13]__.Ramakrishnan, Raghu. 2001. *Computing Relevance Similarity: The Vector Space Model*, [online], (http://www.cs.wisc.edu/~cs784-1/handouts/ir2-vectorspace.ppt
Diakses 24 maret 2011

[14]__. Text Mining
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.74.3588&rep=rep1&type=pdf
diakses 20 oktober 2010.

[15]__.Tan, Steinbach dan Kumar. 2004. *Cluster Analysis: Basic Concepts and Algorithms*.
http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf
diakses 20 oktober 2010.