PERBANDINGAN PERFORMANSI ALGORITMA K-MEANS DENGAN ALGORITMA SUFFIX TREE CLUSTERING DALAM PENGELOMPOKAN BERITA BERBAHASA INDONESIA

Deddy Rahman Prayer Sitio[1], Arie Ardiyanti Suryani[2], Moch Arif Bijaksana[3]

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

Abstrak

Jumlah dokumen sangat banyak dan harus ditangani memerlukan sistem pengorganisasian secara otomatis. Salah satu teknik dalam pengorganisasian dokumen adalah clustering. Pada tugas akhir dibandingkan performansi dari dua algoritma clustering yaitu K-means dan Suffix Tree Clustering dalam mengelompokkan berita dalam bahasa Indonesia. Pengukuran performansi yang dimaksud adalah recall, precision, F-measure dan variance. Dengan parameter tersebut akan terlihat algoritma terbaik dalam mengelompokkan berita dalam bahasa Indonesia. Algoritma K-means memiliki kerapatan cluster yang lebih baik dibandingkan dengan Suffix Tree Clustering, sedangkan Suffix Tree Clustering memiliki nilai F-measure yang lebih tinggi dibandingkan Kmeans.

Kata Kunci : clustering, K-means, Suffix Tree Clustering recall, precision, F-measure dan variance

Abstract

The number of documents is very much needed and should be handled automatically organizing system. Clustering is one of technique to organizing document. At the end of the task compared to the performance of two clustering algorithms namely K-means and Suffix Tree Clustering to cluster Indonesian news. Performance measurement in question is the recall, precision, F-measure and the variance. With these parameters would look best algorithm in classifying the news in Indonesian. Algorithm K- means cluster density better than the Suffix Tree Clustering, while the Suffix Tree Clustering has F-measure values are higher than K- means.

Keywords : clustering, K-means, Suffix Tree Clustering recall, precision, F-measure and variance

1. Pendahuluan

1.1 Latar belakang

Data Mining merupakan salah satu bidang yang berkembang pesat karena besarnya kebutuhan akan nilai tambah dari database skala besar sebagai tuntutan dari pertumbuhan teknologi informasi. *Data Mining* digunakan untuk menemukan informasi atau pola yang bermanfaat atau bernilai ekonomis yang tinggi melalui metode-metode dalam *Data Mining itu*. *Data mining* secara otomatis mengekstraksi informasi dari sekumpulan data yang berukuran besar. Karena itulah, Data Mining merupakan bagian dari *Knowledge Discovery in Database* atau disingkat sebagai KDD.

Dalam *data mining* diperkenalkan salah satu teknik pengelompokan yaitu *clustering*. Idenya adalah dengan mengelompokkan beberapa dokumen yang belum berlabel ke dalam kelompok-kelompok atau *clusters* berdasarkan kemiripan antar berita. Teknik ini bertujuan menemukan distribusi data serta mengelompokkannya berdasarkan kriteria homogenitas atau sejenis sehingga data dapat di *cluster* berdasarkan ditemukannya kesamaan antar data tersebut serta dapat menunjukkan perbedaan data antara satu *cluster* dengan data yang ada di cluster yang lainnya. Klaster yang baik adalah klaster yang memiliki persamaan (*similarity*) intraklaster yang tinggi dan perbedaan (*dissimilarity)* antarklaster yang tinggi.

Sekarang ini jumlah berita yang beredar yang bersifat *unsupervised* sangatlah tidak sedikit. Tenaga manusia kurang efisien untuk menanganinya. Untuk itu diperlukan perangkat yang dapat membantu dalam pengolahan berita secata otomatis, khususnya dalam pengelompokan berita yang belum berlabel (*unsupervised*). Perangkat dibuat dengan mengaplikasikan berbagai bidang ilmu dari mulai matematika, statistika, kecerdasan buatan *data mining*, *information retrieval*, dsb. Secara teoritis dapat membantu dalam menyelesaikan masalah pengelompokan data.

Perkembangan teknik pengelompokan dokumen teks telah menarik perhatian para penelitian untuk mengembangkan dan menemukan suatu teknik atau algoritma yang dapat memperbaiki teknik-teknik *clustering* yang sudah ada. *Suffix Tree Clustering* (STC) adalah algoritma pertama yang menggunakan frasa (multi-word terms) sehingga prosesnya lebih sederhana dibandingkan dengan algoritma yang lain. STC adalah algoritma incremental, kompleksitas waktu perhitungannya linear O(*n*) dan memenuhi *document clustering*. STC tidak memperlakukan dokumen sebagai kumpulan dari kata tetapi lebih dari sebuah string dan didasarkan pada *suffix tree* dalam mengenali kumpulan dokumen serta menggunakan informasi tersebut untuk membangun *cluster*. Oleh karena menggunakan frasa, maka STC mampu mengurangi dimensionalitas himpunan dokumen, hal ini akan mempengaruhi kecepatan dan efisiensi dari STC. Dengan mengembangkan algoritma STC untuk *document clustering*, akan diperoleh manfaat proses *clustering* memberikan

performasi yang baik, dengan adanya beberapa keunggulan yang dimiliki seperti yang telah disebutkan diatas. Algoritma lain yang sudah banyak diaplikasikan dalam Data Mining adalah *K-means*. *K-means* mengelompokkan data berdasarkan kedekatan *data point* dengan *centroid.* Jumlah *centroid* yang ditentukan adalah sebanyak jumlah *cluster* yang ingin dihasilkan. Proses pada tugas akhir ini adalah membandingkan performansi dari algoritma *K-means* dengan algoritma *Suffix Tree Clustering* dalam pengelompokan berita dalam bahasa Indonesia. *K-means* adalah algoritma yang *clustering* yang sederhana dan mudah untuk diimplementasikan. Dengan pembuatan sistem untuk mengelompokkan berita berbahasa Indonesia yang menggunakan algoritma tersebut diatas kita dapat mengetahui algoritma terbaik dalam pengelompokan berita. Dan menghasilkan berita yang telah terkelompokkan dengan *noise* yang sangat minimum lebih dapat digunakan untuk keperluan selanjutnya.

1.2 Perumusan masalah

Clustering dalam *data mining* digunakan untuk mengelompokkan data berupa berita yang bersifat *unsupervised.* Data-data yang ada belum memiliki kelasnya masing-masing. Berita akan dikelompokkan sesuai dengan kemiripan-kemiripan yang ada antara berita yang satu dengan berita yang lain sehingga membentuk kelompok-kelompok berita.
Permasalahannya adalah:

- Bagaimana mengelompokan berita dengan tepat dan memiliki akurasi yang tinggi.
- Membandingkan performansi dari kedua algoritma yang dipakai dalam pengelompokan berita dalam bahasa Indonesia.

Permasalahan dalam tugas akhir ini memiliki batasan sebagai berikut :

1. Tidak membahas sistem *data mining* secara keseluruhan, hanya salah satu
   fungsionalitas data mining yaitu *clustering*.
2. Berita yang digunakan adalah berita berbahasa Indonesia.
3. Berita yang digunakan tidak diambil langsung dari web, tetapi dari database.
4. Tiap berita hanya akan terkelompokkan ke dalam satu kelompok berita.
5. Tidak menerapkan proses stemming dalam dalam proses *pre-processing*
   data.

1.3 Tujuan

Tujuan dari pembuatan tugas akhir ini adalah sebagai berikut :

1. Merancang dan membangun perangkat lunak untuk mengelompokkan berita berbahasa Indonesia dengan algoritma *Suffix Tree Clustering* dan algoritma *K-means*.
2. Menganalisis dan mengevaluasi performansi system dalam hal ini yaitu akurasi *cluster* yang dihasilkan sistem dalam melakukan *clustering*.
3. Membandingkan performansi (tingkat akurasi) dari kedua algoritma dalam pengelompokan berita berbahasa Indonesia.

1.4 Metodologi penyelesaian masalah

Metodologi yang akan digunakan untuk menyelesaikan tugas akhir ini adalah:

1. Studi Literatur
   Pada tahap ini akan dilakukan pendalaman materi, identifikasi masalah, dan metodologi yang akan digunakan dalam pemecahan masalah dengan mencari informasi dan referensi dari berbagai sumber seperti artikel, informasi dari buku maupun internet.
2. Perancangan
   Melakukan analisa penerapan metode yang digunakan dan perancangan akan sistem yang akan diimplementasikan.
3. Implementasi
   Mengimplementasikan desain perangkat lunak yang telah dinyatakan fixed kedalam bahasa pemograman untuk menghasilkan suatu program yang dapat menganalisis berdasarkan perumusan masalah yang telah diuraikan diatas.
4. Pengujian
   Melakukan pengujian dari sistem yang telah dibangun pada tahap implementasi, serta melakukan perbaikan terhadap *bug* dan *error* yang ditemukan pada perangkat lunak yang dibangun.
5. Analisa dan pembuatan laporan
   Membuat analisis dari hasil implementasi yang telah dibuat sesuai dengan parameter yang telah ditentukan sebelumnya dan membuat laporan hasil analisa.

5. Kesimpulan dan saran

5.1 Kesimpulan

a) Algoritma k-means memiliki nilai variance yang baik dalam mengelompokkan berita berbahasa Indonesia. Dengan kata lain hasil pengelompokan oleh algoritma K-means memiliki kedekatan antar dokumen dalam satu kelompok dan kedekatan antar kelompok yang lebih baik (lebih rapat) dibanding hasil algoritma Suffix Tree Clustering. Namun dilain sisi dari segi performansi (dalam hal ini *f-measure*) Suffix tree clustering lebih baik dibandingkan algoritma K-means.
b) Untuk mendapatkan hasil klaster yang baik, maka k = N/2 , k = 5 , k = 7 dimana k : jumlah klaster (inputan user) dan N : jumlah dokumen yang akan dikelompokkan.

5.2 Saran

a) Diperlukannya proses pembobotan yang lebih baik agar memperoleh hasil klasterisasi yang lebih baik lagi.
b) Dalam pengujian algoritma k-means dan suffix tree clustering diperlukan waktu yang sangat lama maka menurut saya sangat diperlukan suatu teknik yang dapat mengurangi waktu komputasi dikarenakan pada dataset yang berjumlah banyak.

Daftar Pustaka

[1] B. Pierre, F. Paolo, S. Padhraic. "*Modeling the Internet and the Web* ", Chapter 4: Text Analysis (Page77-274). School of Information and Computer Science, University of California, Irvine, USA.

[2] Dawid Weiss . "*A CLUSTERING INTERFACE FOR WEB SEARCH RESULTS IN POLISH AND ENGLISH*", Poznań University of Technology, Poland. June 2001

[3] D. Christopher D, R. Prabhakar and S. Hinrich. "*An Introduction to Information Retrieval*", Cambridge University Press, Cambridge, England, 2006

[4] Fabrizio Sebastiani, "*Machine Learning in Automated Text Categorization*". Italy. March 2002

[5] Fung, B. C. M., Wang, K., Ester, M., 2005. *Hierarchical Document Clustering*. Encyclopedia of Data Warehouse and MiningVolume 1, Idea Group Reference, USA.

[6] http://people.revoledu.com/kardi/tutorial/kMean/index.html Tanggal akses: 20 Agustus 2009

[7] Hung Chim and Xiaotie Deng, "*A New Suffix Tree Similarity Measure for Document Clustering*". City University of Hong Kong. 2007

[8] J. Han and M. Khamber. "*Data Mining: Concepts and Techniques*", Morgan Kaufmann Publishers, San Francisco, USA, 2001

[9] Prayitno. "*Clustering*" from ilmukomputer.com

[10] Pressman Roger. S. "*Software Engineering a Practitioner Approach*", McGraw- Hill Inc, Sixth Edition, 2005.

[11] Oren Zamir(1998). *Web Document Clustering: A Feasibility Demonstration*. University of Washington. Seattle.

[12] Wei Ning, "Textmining and Organization in Large Corpus". Kongens Lyngby 2005.