PENGELOMPOKAN DOKUMEN TEKS DENGAN MENGGUNAKAN METODE FUZZY C-MEANS

Tugas Akhir

Oleh

Yessika Naftali Budiono

22074217

Program Studi Teknik Informatika Fakultas Teknologi Informasi

Universitas Kristen Duta Wacana

Tahun 2011

PENGELOMPOKAN DOKUMEN TEKS DENGAN MENGGUNAKAN METODE FUZZY C-MEANS

Tugas Akhir

Diajukan kepada Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana
Sebagai salah satu syarat dalam memperoleh gelar
Sarjana Komputer

Disusun oleh:

Yessika Naftali Budiono
22074217

Program Studi Teknik Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana
Tahun 2011

PERNYATAAN KEASLIAN TUGAS AKHIR

Saya menyatakan dengan sesungguhnya bahwa tugas akhir dengan judul:

PENGELOMPOKAN DOKUMEN TEKS DENGAN MENGGUNAKAN METODE FUZZY C-MEANS

Yang saya kerjakan untuk melengkapi sebagian persyaratan menjadi Sarjana Komputer pada pendidikan sarjana Program Studi Teknik Informatika, Fakultas Teknologi Informasi Universitas Kristen Duta Wacana, bukan merupakan tiruan atau duplikasi dari skripsi kesarjanaan di lingkungan Universitas Kristen Duta Wacana maupun di Perguruan Tinggi atau instansi manapun, kecuali bagian yang sumber informasinya dicantumkan sebagaimana mestinya.

Jika dikemudian hari didapati bahwa hasil skripsi ini adalah hasil plagiasi atau tiruan dari skripsi lain, saya bersedia dikenai sanksi yakni pencabutan gelar kesarjanaan saya.

Yogyakarta, 15 Juli 2011

(Yessika Naftali Budiono)
22074217

INTISARI

Pada zaman perkembangan teknologi saat ini, semakin banyak pula data-data penting yang tersimpan dalam bentuk digital pada suatu sistem penyimpanan tertentu. Sebagian besar metode *clustering* ditujukan kepada data terstruktur, yaitu data yang berupa angka dan mempunyai nilai yang jelas. Sedangkan dokumen berupa teks termasuk data tidak terstruktur yang sulit dikelompokkan langsung secara manual berdasarkan kemiripan isi dokumen. Maka dari itu, penulis ingin melakukan penelitian dalam pengelompokan dokumen.

Pada kasus ini, penulis melakukan sebuah penelitian pengelompokan dokumen teks dengan menggunakan metode *Fuzzy C-Means*. Penelitian dilakukan terhadap dokumen sederhana dan dokumen kompleks. Dokumen sederhana dibuat oleh penulis secara manual yang berisi sebuah kalimat singkat dengan kata-kata yang sederhana. Sedangkan dokumen kompleks diambil dari situs berita dengan cara *copy-paste* isi berita kemudian disimpan dalam bentuk *.txt sebanyak 100 dokumen. Kemudian penulis memasukkan dokumen tersebut satu persatu ke dalam sistem diikuti dengan proses tokenisasi. Setelah tokenisasi, sistem melakukan pembobotan *tf-idf*, *feature selection*, kemudian menciptakan dokumen vektor dan menghitung magnitude tiap dokumen. Dari nilai magnitude tersebut baru dilakukan pengelompokan dengan menggunakan metode *Fuzzy C-Means*.

Setelah didapat hasil pengelompokan dokumen, ternyata metode *Fuzzy C-Means* cukup bagus dalam melakukan pengelompokan dokumen sederhana dengan nilai *purity* sebesar 0,83. Namun pada dokumen yang kompleks, metode tersebut kurang mampu untuk melakukan pengelompokan dokumen dengan baik. Hal tersebut disimpulkan dari nilai *purity* yang rendah yaitu pada kisaran 0,3 sampai 0,4.

**Kata kunci : ***Clustering* dokumen teks, pengelompokan dokumen teks, *Fuzzy C-Means.*

FORMULIR PERBAIKAN (REVISI) TUGAS AKHIR

Dengan ini kami menyatakan bahwa mahasiswa yang melakukan Tugas Akhir dibawah ini:

Nama Mahasiswa : YESSIKA NAFTALI BUDIONO
NIM : 22074217
Judul Tugas Akhir :
PENGELOMPOKAN DOKUMEN TEKS DENGAN MENGGUNAKAN METODE FUZZY C-MEANS

Tgl. Pendadaran : 27 Juli 2011
Tgl. Revisi : 10-8-2011

Telah melakukan perbaikan tugas akhir dengan lengkap. Demikian pernyataan kami agar dapat dipergunakan sebagaimana mestinya.

Yogyakarta, 10-8-2011

Dosen Pembimbing Tugas Akhir I	Dosen Pembimbing Tugas Akhir II
(Antonius Rachmat C, S.Kom, M.Cs)	(Rudi Susanto, Ma, M.J.)

HALAMAN PENGESAHAN

SKRIPSI

PENGELOMPOKAN DOKUMEN TEKS DENGAN MENGGUNAKAN METODE FUZZY C-MEANS

Oleh: Yessika Naftali Budiono / 22074217

Dipertahankan di depan dewan Penguji Tugas Akhir/Skripsi
Program Studi Teknik Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana – Yogyakarta
Dan dinyatakan diterima untuk memenuhi salah satu
Syarat memperoleh gelar
Sarjana Komputer
Pada tanggal
27/7/2011

Yogyakarta, 11/8/2011
Mengesahkan,

Dewan Penguji:

1. Antonius Rachmat C, S.Kom., M.Cs
2. Budi Susanto, S.Kom., M.T
3. Drs. R. Gunawan Santoso, M.Si.
4. Yuan Lukito, S.Kom.

Dekan

Drs. Wimmie Handiwidjojo, MIT.

Ketua Program Studi

Nugroho Agus H., S.Si., M.Si

HALAMAN PERSETUJUAN

Judul	: Pengelompokan Dokumen Teks dengan Menggunakan Metode Fuzzy C-Means
Nama	: Yessika Naftali Budiono
NIM	: 22074217
Mata Kuliah	: Tugas Akhir
Kode	: TI2126
Semester	: Genap
Tahun Akademik	: 2010/2011

Telah diperiksa dan disetujui
Di Yogyakarta,
Pada Tanggal 15-7-2011

Dosen Pembimbing I	Dosen Pembimbing II
Antonius Rachmat, S.Kom, M.Cs	**Budi Susanto, S.Kom, M.T**

UCAPAN TERIMA KASIH

Puji dan syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa yang telah melimpahkan rahmat dan anugerah, sehingga penulis dapat menyelesaikan Tugas Akhir dengan judul Pengelompokan Dokumen Teks dengan Menggunakan Metode Fuzzy C-Means dengan baik dan tepat waktu.

Penulisan laporan ini merupakan kelengkapan dan pemenuhan dari salah satu syarat dalam memperoleh gelar Sarjana Komputer. Selain itu bertujuan melatih mahasiswa untuk dapat menghasilkan suatu karya yang dapat dipertanggungjawabkan secara ilmiah, sehingga dapat bermanfaat bagi penggunanya.

Dalam menyelesaikan pembuatan program dan laporan Tugas Akhir ini, penulis telah banyak menerima bimbingan, saran dan masukan dari berbagai pihak, baik secara langsung maupun secara tidak langsung. Untuk itu dengan segala kerendahan hati, pada kesempatan ini penulis menyampaikan ucapan terimakasih kepada :

1. Bpk Antonius Rachmat C, S.Kom, M.Cs selaku dosen pembimbing I yang telah memberikan bimbingannya, petunjuk dan masukan yang diberikan selama proses pengerjaan Tugas Akhir ini.
2. Bpk Budi Susanto, S.Kom., M.T. selaku dosen pembimbing II yang telah memberikan bimbingannya dengan sabar, jelas dan baik kepada penulis.
3. Keluarga tercinta yang selalu memberi dukungan dan semangat.
4. Ryan sebagai orang terdekat yang selalu memberikan dukungan dan semangat.
5. Bpk Andronicus Riyono, Willi, Andhri, Lusi, Veni, Veve, Amel, Eric, Rudy, dan teman-teman yang telah memberikan selalu memberikan semangat kepada penulis.
6. Pihak lain yang tidak dapat penulis sebutkan satu per satu, sehingga Tugas Akhir ini dapat terselesaikan dengan baik.

Penulis menyadari bahwa program dan laporan Tugas Akhir ini masih jauh dari sempurna. Oleh karena itu, penulis sangat mengharapkan kritik dan saran yang membangun dari pembaca sekalian. Sehingga suatu saat penulis dapat memberikan karya yang lebih baik lagi.

Akhir kata penulis ingin meminta maaf bila ada kesalahan baik dalam penyusunan laporan maupun yang pernah penulis lakukan sewaktu membuat program Tugas Akhir. Sekali lagi penulis mohon maaf yang sebesar-besarnya. Dan semoga dapat berguna bagi kita semua.

Yogyakarta, 15 Juli 2011

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1
PENDAHULUAN

1.1. Latar Belakang Masalah

Pada zaman perkembangan teknologi saat ini, semakin banyak pula data-data penting yang tersimpan dalam bentuk digital pada suatu sistem penyimpanan tertentu. Data yang tersimpan mempunyai dua sifat, yaitu data terstruktur dan data tidak terstruktur. Dokumen berupa teks termasuk data tidak terstruktur yang sulit dikelompokkan langsung secara manual berdasarkan kemiripan isi dokumen. Maka dari itu, pengelompokan dokumen sangat diperlukan agar lebih mudah dalam mengenali dokumen.

Pengelompokan dokumen tersebut didasarkan pada seberapa mirip isi dokumen satu dengan dokumen yang lainnya. Agar dapat dikelompokkan secara otomatis, tingkat kemiripan tersebut harus dihitung. Tingkat kemiripan biasanya dihitung berdasaran kata-kata yang terdapat dalam dokumen satu dengan dokumen yang lain.

Metode dalam pengelompokan data secara komputer sudah banyak tersedia, contohnya pengelompokan data dengan menggunakan metode *C-Means Clustering, K-Means Clustering, Single Link Clustering*, dan sebagainya. Dalam penelitian ini, penulis memfokuskan pada kasus pengelompokan dokumen teks dengan menggunakan metode *Fuzzy C-Means*.

Untuk mengetahui seberapa baik kualitas dari suatu metode pengelompokan data, maka dibutuhkan adanya tahap evaluasi. Ada berbagai macam rumus penghitungan untuk evaluasi pengelompokan data. Pada kasus ini, penulis menggunakan *Purity* untuk menganalisa hasil dari pengelompokan dokumen teks dengan *Fuzzy C-Means*. *Purity* merupakan tingkat kemurnian dari suatu *cluster*. Tingkat kemurnian yang dimaksud dihitung dari seberapa banyak anggota dari kelas yang sama, yang terkumpul pada suatu *cluster*.

1.2. Perumusan Masalah

Seberapa besar tingkat *purity* yang dihasilkan oleh sistem pengelompokan dokumen dengan menggunakan *Fuzzy C-Means* untuk dokumen berbahasa Indonesia?

1.3. Batasan Masalah

- Penelitian dilakukan hanya dibatasi pada pengelompokan dokumen teks dalam bahasa Indonesia saja karena penggunakan *stopword* dapat lebih mudah dipahami,
- Tidak dilakukan *stemming* karena *stemming* dalam bahasa Indonesia masih kurang konsisten.
- Dokumen *sample* dibatasi pada dokumen teks (*.txt*) yang tidak berfotmat apapun pada dokumen berita seputar teknologi.
- Studi kasus dilakukan dengan mengambil contoh data dari website berita http://tekno.kompas.com, http://tekno.liputan6.com, http://www.tempointeraktif.com/teknologi/, dan http://techno.okezone.com dengan cara *copy-paste* isi berita kemudian disimpan dalam format *.txt.
- Pengambilan dokumen sample untuk penelitian sebanyak 100 dokumen.
- Dokumen akan dikelompokan berdasarkan mobile, hardware, software, website, dan antariksa.

1.4. Tujuan Penelitian

Tujuan penelitian ini adalah untuk mengetahui lebih dalam tentang *Fuzzy C-Means* dan meneliti apakah metode tersebut dapat mengelompokan dokumen yang berupa teks dengan baik.

1.5. Metode Penelitian

Tahap-tahap yang dilakukan penulis dalam melakukan penelitian ini antara lain:

- Membuat rencana perancangan sistem
- Mengumpulkan dokumen-dokumen *sample* sebanyak 100 dokumen berupa dokumen teks.
- Proses pengujian sistem pengelompokan dokumen dengan menggunakan metode *Fuzzy C-Means*.
- Melakukan tahap pengujian dan menghitung tingkat *purity*-nya.
- Membuat laporan dari hasil pengujian sistem.

1.6. Sistematika Penulisan

Penulis membagi Laporan Tugas Akhir ini menjadi 5 bab. Bab pertama merupakan Bab Pendahuluan. Bab ini berisi tentang penjelasan umum mengenai penelitian yang akan dilakukan oleh penulis serta apa yang akan dibuat oleh penulis pada Tugas Akhir ini. Bab ini terdiri dari 6 sub bab, yaitu: Latar Belakang Masalah, Perumusan Masalah, Batasan Masalah, Tujuan Penelitian, Metode Penelitian, dan Sistematika Penulisan.

Bab kedua terdiri dari 2 sub bab yaitu: Tinjauan Pustaka dan Landasan Teori. Tinjauan Pustaka membahas tentang penelitian yang dilakukan oleh pihak lain tentang pengelompokan data. Landasan Teori membahas tentang konsep yang digunakan dalam proses penelitian, seperti penjelasan tentang bagaimana dokumen teks dapat dikelompokkan berdasarkan nilai yang dimiliki tiap dokumen, dan bagaimana konsep pengelompokan dokumen yang dilakukan dengan *Fuzzy C-Means*.

Kemudian bab ketiga, terdiri dari Analisis dan Perancangan Sistem. Dalam bab ini, penulis akan memaparkan tentang spesifikasi sistem yang dirancang sampai pada tahap perancangan alur kerja sistem pengelompokan dokumen dengan *Fuzzy C-Means*.

Bab keempat terdiri dari Implementasi dan Analisi Sistem yang merupakan implementasi dari perancangan sistem pada bab 3. Sedangkan bab kelima berisi kesimpulan dan saran penulis selama pembuatan sistem pengelompokan dokumen dengan menggunakan *Fuzzy C-Means*.

BAB 5
KESIMPULAN DAN SARAN

5.1. Kesimpulan

Berdasarkan pengujian dan analisis yang dilakukan oleh penulis, maka dapat diambil kesimpulan yaitu penerapan metode *Fuzzy C-Means* ternyata kurang baik untuk pengelompokan data berupa dokumen teks yang kompleks. Hal ini dibuktikan dari nilai *purity* yang dihasilkan. Dalam percobaan sederhana pun, sistem tidak dapat menghasilkan nilai *purity* yang sempurna.

Waktu yang digunakan untuk pengelompokan dokumen teks dalam setiap percobaan berbeda-beda berdasarkan dari besarnya nilai *feature selection* yang telah ditentukan, karena semakin besar *feature selection* maka semakin banyak pula token yang dihitung.

Sedangkan pengaruh *feature selection* terhadap hasil pengelompokkan dokumen yaitu semakin besar nilai *feature selection* maka semakin besar pula nilai *purity* yang dihasilkan, tetapi kenaikkan nilai *purity*-nya tidak terlalu banyak.

5.2. Saran

Untuk pengembangan sistem yang lebih baik dimasa yang akan datang, maka penulis memberikan saran sebaiknya melakukan penelitian terhadap metode yang lain untuk pengelompokan data berupa dokumen teks.

DAFTAR PUSTAKA

Kang, B.Y., Kim, D.W., Li Qing. (2005). Spatial Homogeneity-Based Fuzzy c-Means Algorithm for Image Segmentation. *Lecture Notes in Artificial Intelligence* (vol. 3613, pp. 462-469). China.

Khoiruddin, A.A. (2007). *Menentukan Nilai Akhir Kuliah dengan Fuzzy C-Means*. Seminar Nasional Sistem dan Informatika 2007; Bali, 16 November 2007. Universitas Islam Indonesia.

Kusumadewi Sri. (2007). *Klasifikasi Kandungan Nutrisi Bahan Pangan Menggunakan Fuzzy C-Means*. Seminar Nasional Aplikasi Teknologi Informasi 2007; Yogyakarta, 16 Juni 2007. Universitas Islam Indonesia.

Manning, C.D., Raghavan, P., Schutze, H. (2009). *Introduction to Information Retrieval*. England: Cambridge University Press.

Miyamoto, S., et. al. (2008). Algorithm for fuzzy clustering: methods in c-means clustering with application.

Veges, K.E., Pope Nigel. (2006). *Business Application and Computational Intelligence*. America: Ideal Group Publishing.

Weiss, S.M., Indurkhya, N., Zhang, T., Damerau, F.J. (2005). *Text Mining: Predictive Methods for Analyzing Unstructured Information*. New York: Springer.