TESIS

INTEGRASI PEMBOBOTAN TF IDF PADA METODE K-MEANS UNTUK CLUSTERING DOKUMEN TEKS

DEDDY WIJAYA SULIANTORO

No. Mhs. : 105301466/PS/MTF

PROGRAM STUDI MAGISTER TEKNIK INFORMATIKA

PROGRAM PASCA SARJANA

UNIVERSITAS ATMA JAYA YOGYAKARTA

2012

UNIVERSITAS ATMA JAYA YOGYAKARTA

PROGRAM PASCASARJANA

PROGRAM STUDI MAGISTER TEKNIK INFORMATIKA

PENGESAHAN TESIS

Nama : DEDDY WIJAYA SULIANTORO

Nomor Mahasiswa : 105301466/PS/MTF

Konsentrasi : Enterprise Information System

Judul Tesis : Integrasi Pembobotan TF IDF pada Metode k-Means untuk clustering Dokumen Teks

Nama Pembimbing	Tanggal	Tanda Tangan
Irya Wisnubhadra	16/1/2012	
Ernawati	17/1/2012	

UNIVERSITAS ATMA JAYA YOGYAKARTA

PROGRAM PASCASARJANA

PROGRAM STUDI MAGISTER TEKNIK INFORMATIKA

PENGESAHAN TESIS

Nama : DEDDY WIJAYA SULIANTORO

Nomor Mahasiswa : 105301466/PS/MTF

Konsentrasi : Enterprise Information System

Judul Tesis : Integrasi Pembobotan TF IDF pada Metode *k-Means* untuk *clustering* Dokumen Teks

Nama Penguji	Tanggal	Tanda Tangan
Inya Wisnubhadra	2/2/2012	
Dra. Ernawati, MT	30/1/2012	
Paulus Mudjihartono, ST, MT	30/1/2012	

Ketua Program Studi

.....

HALAMAN PERNYATAAN

Dengan ini saya menyatakan bahwa tesis ini tidak meniru atau menduplikasi karya yang sebelumnya pernah diajukan untuk memperoleh gelar kesarjanaan di suatu perguruan tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan di dalam daftar pustaka.

Yogyakarta, 25 Desember 2011

Deddy Wijaya Suliantoro

INTISARI

Pada era teknologi informasi seperti saat ini, dokumen teks, berita, dan jurnal-jurnal sudah cenderung tersimpan dalam format digital karena mudah dan cepat dalam penyimpanan. Terlalu banyaknya dokumen teks yang tersimpan dalam komputer membuat pencarian informasi menjadi sulit. Kesulitan dalam pencarian informasi yang sesuai dengan kebutuhan seringkali menjadi penghambat proses pembelajaran.

Pengelompokan data (*clustering*) menjadi salah satu solusi untuk mengorganisasi dokumen-dokumen teks digital yang berjumlah besar. Metode *k-Means clustering*, dipadukan dengan pembobotan TF-IDF yang diimplementasikan ke dalam sebuah aplikasi pengelompokan dokumen dapat melakukan proses *clustering* dokumen teks secara otomatis sehingga menghemat waktu untuk mengelompokkan dokumen secara manual.

Hasil proses *clustering* dari aplikasi pengelompokan dokumen ini menunjukkan tingkat presisi yang tinggi dalam mengelompokkan dokumen-dokumen berdasarkan isi dokumen tersebut sehingga diyakini mampu membantu proses pencarian informasi dari data yang terlalu banyak.

ABSTRAK

In today's information-technology era, documents containing texts, news, and journals are commonly recorded in digital format since it is easy and fast in keeping the data. With too many documents saved in a computer, searching information turns to be more difficult/complex. Such difficulty in finding information needed often becomes an obstacle in learning process.

Clustering becomes one of the solutions to organize digital documents in large numbers of data. The k-Means clustering method, combined with TF-IDF weighting which is then implemented in a document clustering application, can automatically process the clustering, which then save more time compared to manual clustering.

The result of this document clustering application shows a high precision in classifying documents based on their content. It then helps with the process of information finding from a great amount of data.

KATA HANTAR

Puji syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa yang telah memberikan rahmat dan hidayah, sehingga penulis dapat menyelesaikan tesis yang berjudul Integrasi Pembobotan TF IDF pada metode k-Means untuk clustering dokumen teks. Tesis ini penulis susun sebagai salah satu syarat untuk mencapai derajat sarjana S2 di Program Studi Magister Teknik Informatika Universitas Atma Jaya Yogyakarta.

Dalam kesempatan ini penulis ingin mengucapkan terima kasih yang sebesar-besarnya kepada berbagai pihak yang telah membantu dalam penyelesaian tesis ini, yaitu:

1. Pak Irya, selaku dosen pembimbing I yang selalu memberikan waktu dan bimbingannya kepada penulis dalam menyelesaikan tesis ini.
2. Ibu Dra. Ernawati, M.T. selaku dosen pembimbing II dan Ketua Program Studi Magister Teknik Informatika Universitas Atma Jaya Yogyakarta.
3. Teman-teman Pascasarjana MTF angkatan September 2010 yang selalu memberi bantuan materi dan semangat bagi penulis.
4. Karyawan Tata Usaha Program Pascasarjana yang selalu membantu menyediakan informasi dan membantu proses administrasi.
5. Kedua orang tua penulis yang selalu memberikan semangat dan kekuatan dalam mengerjakan tesis.
6. Semua pihak yang secara langsung maupun tidak langsung membantu proses penyelesaian tesis ini

7. Terakhir dan terhebat, kepada Tuhan Yesus Kristus yang tanpa-Nya, penulis tidaklah mampu menyelesaikan tesis ini.

Penulis sadari bahwa tesis ini masih memiliki banyak kekurangan. Untuk itu, saran dan kritik yang membangun dari pembaca sangat penulis nantikan demi perbaikan dan pengembangan di masa mendatang.

Yogyakarta, Desember 2011

Penulis

DAFTAR ISI

HALAMAN PENGESAHAN DOSEN PEMBIMBING ..... **ii**
HALAMAN PENGESAHAN TIM PENGUJI ..... **iii**
HALAMAN PERNYATAAN ..... **iv**
INTISARI ..... **v**
ABSTRAK ..... **vi**
KATA HANTAR ..... **vii**
DAFTAR TABEL ..... **xi**
DAFTAR GAMBAR ..... **xiii**
BAB 1 PENDAHULUAN **1**
A. Latar Belakang 1
B. Perumusan Masalah 3
C. Batasan Masalah 3
D. Keaslian Penelitian 4
E. Manfaat yang Diharapkan 4
F. Tujuan Penelitian 4
G. Sistematika Penulisan 5
BAB 2 TINJAUAN PUSTAKA **7**
A. Tinjauan Pustaka 7
B. Landasan Teori 9
1. Information Retrieval 9
2. Clustering 10
3. Pembobotan TF-IDF 11
4. k-Means clustering 12
5. Precision dan recall 20
C. Hipotesis 22
BAB 3 METODOLOGI PENELITIAN **23**
A. Bahan atau Materi Penelitian 23
B. Alat Penelitian 23

C. Langkah-Langkah Penelitian 23
D. Analisis, Perancangan, Implementasi, dan Pengujian Perangkat Lunak ... 24
1. Analisis Kebutuhan 24
2. Perancangan Perangkat Lunak 26
3. Implementasi Perangkat Lunak 39
4. Implementasi Antarmuka Perangkat Lunak 41
5. Pengujian Fungsionalitas Perangkat Lunak 45
BAB 4 HASIL PENELITIAN DAN PEMBAHASAN 47
A. Hasil Penelitian 47
B. Pembahasan 66
BAB 5 KESIMPULAN DAN SARAN 69
A. Kesimpulan 69
B. Saran 69

DAFTAR TABEL

Tabel 1. Penelitian terdahulu mengenai *k-Means* dan TF-IDF 8
Tabel 2. Perhitungan bobot tiap token dalam *lexicon* 15
Tabel 3. perhitungan jarak dokumen dengan *centroid* 17
Tabel 4. Pembagian dokumen ke dalam *cluster* 18
Tabel 5. perhitungan *centroid* C1 yang baru 18
Tabel 6. Perhitungan *centroid* C2 yang baru 18
Tabel 7. Jarak *centroid* lama dan *centroid* baru 19
Tabel 8. Implementasi Perangkat Lunak 39
Tabel 8 (Lanjutan). Implementasi Perangkat Lunak 40
Tabel 9. Pengujian Fungsionalitas Perangkat Lunak 45
Tabel 10. Hasil Pengujian Pertama 49
Tabel 11. Hasil Pengujian Kedua 49
Tabel 12. Hasil Pengujian Ketiga 50
Tabel 13. Pengujian *clustering* Berita Olahraga 52
Tabel 14. Pengelompokan korpus secara manual ke dalam 2 *cluster* 54
Tabel 15. Pengelompokan korpus secara manual ke dalam 3 *cluster* 54
Tabel 16. Hasil Percobaan Pertama 55
Tabel 17. Hasil Perbandingan *clustering* Manual dan Percobaan Pertama 55
Tabel 18. Hasil Percobaan Kedua 56
Tabel 19. Hasil Perbandingan *clustering* Manual dan Percobaan Pertama 56
Tabel 20. Hasil Percobaan Ketiga 57
Tabel 21. Hasil Percobaan Keempat 57
Tabel 22. Hasil Percobaan Pertama Berita Olahraga 58
Tabel 23. Hasil Percobaan Kedua Berita Olahraga 59
Tabel 24. Hasil Percobaan Ketiga Berita Olahraga 60
Tabel 25. Hasil *clustering* Berita Umum 62
Tabel 26. Hasil *clustering* Korpus Besar 64

Tabel 26. Hasil *clustering* Kedua Korpus Besar 65
Tabel 27. Rata-rata Precision dan *Recall* 68

DAFTAR GAMBAR

Gambar 1. Ilustrasi proses *k-Means clustering*. 13
Gambar 2. Precision dan *Recall* 21
Gambar 3. Use Case Diagram 24
Gambar 4. Rancangan Arsitektur 26
Gambar 5. ERD Aplikasi *clustering* Dokumen Teks *k-Means* 27
Gambar 7. Alur input Dokumen 30
Gambar 8. Alur pembangunan indeks dokumen bagian A 32
Gambar 9. Alur pembangunan indeks dokumen bagian B 33
Gambar 10. Alur Penghitungan TF, DF, dan IDF 35
Gambar 11. Rancangan *form* Main 37
Gambar 12. Rancangan *Form* InsertDoc 38
Gambar 13. Rancangan *Form Cluster*s 38
Gambar 14. *Form* Main 41
Gambar 15. *Form InsertDoc* 42
Gambar 16. *Form Cluster*s 43
Gambar 17. *Form* ShowDoc 44
Gambar 18. Hasil *clustering* dari aplikasi *clustering* dokumen (1) 47
Gambar 19. Hasil *clustering* dari aplikasi *clustering* dokumen (2) 48
Gambar 20. Hasil pengujian pertama 6 dokumen percobaan 49
Gambar 21. Hasil pengujian kedua 6 dokumen percobaan 50
Gambar 22. Hasil pengujian ketiga terhadap berita olahraga 51
Gambar 23. Hasil pengujian terhadap rumus normalisasi TF-IDF 53
Gambar 24. Hasil *clustering* Pertama berita olahraga 59
Gambar 25. Hasil *clustering* Ketiga berita olahraga threshold 50% 61
Gambar 26. Hasil *clustering* Ketiga berita olahraga threshold 25% 61
Gambar 27. Hasil *clustering* Berita Umum threshold 50% 62
Gambar 28. Hasil *clustering* Berita Umum threshold 75% 63
Gambar 29. Hasil *clustering* Korpus Besar 64

Gambar 30. Hasil *clustering* Kedua Korpus Besar threshold 50%..... 65
Gambar 31. Hasil *clustering* Kedua Korpus Besar threshold 25%..... 66