PENERAPAN ALGORITMA *COMPLETE LINKAGE* UNTUK *CLUSTERING* DOKUMEN TEKS

SKRIPSI

Oleh:

LINA KUSUMA DEWI

22 08 4407

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS TEKNOLOGI INFORMASI

UNIVERSITAS KRISTEN DUTA WACANA

YOGYAKARTA

2012

PENERAPAN ALGORITMA *COMPLETE LINKAGE* UNTUK *CLUSTERING* DOKUMEN TEKS

SKRIPSI

Diajukan kepada Program Studi Teknik Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana
Sebagai Salah Satu Syarat dalam Memperoleh Gelar
Sarjana Komputer

Disusun oleh:
LINA KUSUMA DEWI
22 08 4407

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
YOGYAKARTA
2012

PERNYATAAN KEASLIAN TUGAS AKHIR

Saya menyatakan dengan sesungguhnya bahwa tugas akhir dengan judul:

PENERAPAN ALGORITMA *COMPLETE LINKAGE* UNTUK *CLUSTERING* DOKUMEN TEKS

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

Yogyakarta, 18 Oktober 2012

LINA KUSUMA DEWI

22 08 4407

HALAMAN PERSETUJUAN

Judul	:	Penerapan Algoritma Complete Linkage untuk Clustering Dokumen Teks
Nama	:	Lina Kusuma Dewi
NIM	:	22 08 4407
Mata Kuliah	:	Skripsi (Tugas Akhir)
Kode	:	TIW276
Semester	:	Ganjil
Tahun Akademik	:	2012/2013

Telah diperiksa dan disetujui
di Yogyakarta,
Pada Tanggal 1-Nov-2012

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

HALAMAN PENGESAHAN

SKRIPSI

PENERAPAN ALGORITMA *COMPLETE LINKAGE* UNTUK *CLUSTERING* DOKUMEN TEKS

Oleh: LINA KUSUMA DEWI / 22 08 4407

Dipertahankan di depan Dewan Penguji Skripsi
Program Studi Teknik Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana – Yogyakarta
Dan dinyatakan diterima untuk memenuhi salah satu syarat
untuk memperoleh gelar
Sarjana Komputer
pada tanggal
30 November 2012

Yogyakarta, 14/12/12

Mengesahkan,

Dewan Penguji:

1. Budi Susanto, S.Kom., M.T.
2. Antonius Rachmat C., S.Kom., M.Cs.
3. Aditya Wikan Mahastama, S.Kom
4. Rosa Delima, S.Kom, M.Kom.

Dekan

(Drs. Wimmie Handiwidjojo, MIT)

Ketua Program Studi

(Nugroho Agus Haryono, S.Si.,M.Si.)

UCAPAN TERIMA KASIH

Puji dan syukur penulis panjatkan kepada Tuhan kehadirat Tuhan Yang Maha Esa yang telah melimpahkan rahmat dan anugerah sehingga penulis dapat menyelesaikan Tugas Akhir dengan judul Penerapan Algoritma Complete Linkage untuk Clustering Dokumen Teks.

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 terima-kasih kepada :

1. Bpk. Budi Susanto, S.Kom., M.T. selaku dosen pembimbing I yang telah memberikan bimbingan dengan sabar dan baik kepada penulis, juga kepada
2. Bpk. Antonius Rachmat C., S.Kom, M.Cs selaku dosen pembimbing II atas bimbingan, petunjuk, dan saran yang diberikan selama pengerjaan Tugas Akhir ini sejak awal hingga akhir, juga kepada
3. Dosen-dosen Universitas Kristen Duta Wacana yang telah membantu memberikan pengarahan dan masukan kepada penulis.
4. Keluarga tercinta yang setia memberikan dukungan, doa, dan semangat.
5. Yanuar Budi Prasetyo atas saran, dukungan, dan semangat yang diberikan.
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 memohon maaf yang sebesar-besarnya. Dan semoga dapat berguna bagi kita semua.

Yogyakarta, 18 Oktober 2012

Penulis

INTISARI

Penerapan Algoritma *Complete Linkage* untuk *Clustering* Dokumen Teks

Perkembangan teknologi yang semakin cepat cenderung membawa dampak positif bagi kehidupan manusia. Manusia dapat memperoleh informasi dengan mudah melalui berbagai macam media. Semakin banyak informasi yang ada, semakin mudah informasi tersebut didapatkan, namun semakin sulit untuk mendapatkan informasi yang relevan. Proses seleksi atau pengelompokan informasi dibutuhkan untuk mendapatkan hasil yang relevan.

Melihat masalah di atas, penulis membangun sebuah aplikasi berbasis algoritma *Complete Linkage* untuk menemukan *cluster-cluster* pada dokumen teks. Proses untuk menemukan *cluster-cluster* tersebut menggunakan *Vector Space Model* untuk pembobotan dokumen, *Cosine Similarity* untuk menghitung nilai *similarity* antara dua dokumen, dan *thesaurus WordNet*. *WordNet* merupakan sebuah *database* yang menyimpan sinonim kata berbahasa Inggris. Dengan penggunaan *WordNet* diharapkan dapat meningkatkan kualitas *cluster* yang terbentuk.

Sistem yang dibangun telah mampu menemukan *cluster-cluster* dengan nilai purity sebesar 0,8308 untuk penggunaan *WordNet* dengan *cutting point* 0,02. Dengan uji coba tanpa menggunakan *WordNet* untuk *cutting point* yang sama, sistem menghasilkan *cluster-cluster* dengan nilai *purity* sebesar 0,9077.

Kata Kunci: *cluster*, *Complete Linkage*, *Cosine Similarity*, *cutting point*, dokumen teks, *purity*, *Vector Space Model*, *WordNet*

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR DAN GRAFIK

INTISARI

Penerapan Algoritma *Complete Linkage* untuk *Clustering* Dokumen Teks

Perkembangan teknologi yang semakin cepat cenderung membawa dampak positif bagi kehidupan manusia. Manusia dapat memperoleh informasi dengan mudah melalui berbagai macam media. Semakin banyak informasi yang ada, semakin mudah informasi tersebut didapatkan, namun semakin sulit untuk mendapatkan informasi yang relevan. Proses seleksi atau pengelompokan informasi dibutuhkan untuk mendapatkan hasil yang relevan.

Melihat masalah di atas, penulis membangun sebuah aplikasi berbasis algoritma *Complete Linkage* untuk menemukan *cluster-cluster* pada dokumen teks. Proses untuk menemukan *cluster-cluster* tersebut menggunakan *Vector Space Model* untuk pembobotan dokumen, *Cosine Similarity* untuk menghitung nilai *similarity* antara dua dokumen, dan *thesaurus WordNet*. *WordNet* merupakan sebuah *database* yang menyimpan sinonim kata berbahasa Inggris. Dengan penggunaan *WordNet* diharapkan dapat meningkatkan kualitas *cluster* yang terbentuk.

Sistem yang dibangun telah mampu menemukan *cluster-cluster* dengan nilai purity sebesar 0,8308 untuk penggunaan *WordNet* dengan *cutting point* 0,02. Dengan uji coba tanpa menggunakan *WordNet* untuk *cutting point* yang sama, sistem menghasilkan *cluster-cluster* dengan nilai *purity* sebesar 0,9077.

Kata Kunci: *cluster*, *Complete Linkage*, *Cosine Similarity*, *cutting point*, dokumen teks, *purity*, *Vector Space Model*, *WordNet*

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan teknologi yang semakin cepat cenderung membawa dampak positif bagi kehidupan manusia. Manusia dapat memperoleh informasi dengan mudah melalui berbagai macam media. Dengan kemudahan yang ditawarkan, proses untuk mendapatkan informasi atau data tidak menghabiskan banyak waktu. Namun, informasi dan data-data tersebut belum tentu memenuhi kriteria yang diinginkan. Semakin banyak informasi yang ada, semakin mudah informasi tersebut didapatkan, namun semakin sulit untuk mendapatkan informasi yang benar-benar dibutuhkan.

Untuk memperoleh informasi yang benar-benar sesuai, diperlukan proses seleksi terhadap semua informasi yang tersedia. Proses seleksi ini dapat dilakukan secara *manual* dengan membandingkan semua informasi tersebut. Akan tetapi, hal ini tidak disarankan karena proses seleksi tidak optimal dan bila informasi yang ada mencapai ribuan dokumen, waktu yang dibutuhkan untuk mengelompokkan dokumen akan sangat lama dan tidak menutup kemungkinan terjadi kesalahan dalam proses seleksi. Proses seleksi dapat dilakukan secara otomatis untuk mempermudah *user* dalam mencari dokumen yang berkualitas.

Proses seleksi atau *clustering* adalah suatu proses untuk menentukan kategori. Dokumen yang memiliki isi yang relatif sama dapat dikelompokkan dalam satu kelompok yang sama. Dokumen yang dimaksud dapat berupa berbagai macam dokumen, misalnya dokumen perpajakan, jurnal, dokumen keuangan, dsb. Dalam implementasinya di komputer, dokumen yang dimaksud adalah dokumen *plain text* yang artinya dokumen yang tidak memiliki format apapun. Penelitian ini memadukan metode *Vector Space Model* dengan algoritma *Complete Linkage*

guna memperoleh *cluster* dokumen yang memiliki kesamaan topik. Selain itu, penelitian ini menggunakan *thesaurus* dari *WordNet* guna meningkatkan kualitas *cluster* yang terbentuk.

1.2 Perumusan Masalah

Penelitian ini berfokus tentang seberapa akurat *clustering* dokumen teks dengan mengimplementasikan algoritma *Complete Linkage* dipadukan dengan kamus kata *WordNet*.

1.3 Batasan Masalah

Program *clustering* yang akan dibuat ini memiliki beberapa batasan masalah seperti berikut ini :

- Sistem dibuat dalam bentuk aplikasi berbasis *desktop* dengan dokumen teks berupa *file plain text* berekstensi .txt
- Korpus dokumen merupakan sekumpulan berita berbahasa Inggris yang diambil secara acak dari http://www.voanews.com
- Pembobotan kata diperoleh dengan menggunakan *tf-idf weighting* dan *similarity measure* diperoleh dengan *Cosine Similarity*
- Sistem menggunakan *thesaurus WordNet* yang diambil dari www.androidtech.com/downloads/wordnet20-from-prolog-all-3.zip untuk meningkatkan kualitas *cluster*

1.4 Hipotesis

Clustering dokumen teks dengan menggunakan *thesaurus* dari *WordNet* mampu memberikan hasil *cluster* dengan tingkat akurasi sebesar 75%.

1.5 Tujuan Penelitian

Penelitian ini memiliki tujuan untuk mengelompokkan dokumen teks dengan berbagai macam topik menjadi sekumpulan dokumen yang memiliki topik yang sama. Fokus penelitian ini yakni untuk meneliti tingkat keakuratan *cluster* yang terbentuk dengan menggunakan *thesaurus WordNet*.

1.6 Metode / Pendekatan

Metode yang digunakan dalam penelitian ini antara lain sebagai berikut :

a. Metode Pengumpulan Data

   Data-data yang digunakan dalam penelitian ini didapatkan penulis dengan cara melakukan studi pustaka melalui literatur-literatur yang mendukung penyelesaian penelitian ini. Selain itu, penulis juga mengumpulkan beberapa dokumen teks untuk dijadikan sebagai korpus dokumen. Data dari dokumen teks diambil dari sebuah situs berita berbahasa Inggris (www.voanews.com).

b. Metode Pengembangan Sistem

   Sistem akan dikembangkan dengan menggunakan algoritma *complete linkage*. Algoritma ini merupakan salah satu algoritma dalam *hierarchical clustering*.

c. Metode Evaluasi

   Metode evaluasi yang digunakan adalah dengan menghitung nilai *purity cluster* untuk beberapa *cutting point* tertentu .

1.7 Sistematika Penulisan

Untuk memudahkan dalam mendapatkan gambaran yang lengkap dan jelas mengenai penelitian yang akan dilakukan, penulis membagi laporan ini menjadi 5

(lima) bab yaitu Bab I Pendahuluan, Bab II Tinjauan Pustaka, Bab III Analisis dan Perancangan Sistem, Bab IV Implementasi dan Analisis Sistem, dan Bab V Kesimpulan dan Saran.

Dalam bab pertama menguraikan hal-hal seperti latar belakang masalah, perumusan masalah, batasan masalah, hipotesis, tujuan penelitian, metode/pendekatan yang digunakan serta sistematika penulisan laporan Tugas Akhir.

Bab kedua berisi tentang tinjauan pustaka serta landasan teori yang diperlukan untuk memecahkan masalah dan merumuskan hipotesis.

Bab ketiga berisikan tentang analisis teori yang digunakan dalam penelitian, uraian tentang variabel dan data yang akan dikumpulkan, flowchart dan arsitektur sistem, cara perancangan dan simulasi yang dilakukan .

Bab keempat berisi tentang hasil penelitian/implementasi serta pembahasan/analisis dari penelitian yang telah dilakukan dan dijelaskan secara terpadu.

Bab terakhir berisikan kesimpulan dari sistem yang telah dibuat dan saran yang akan berguna untuk pengembangan sistem dalam ma. Dengan adanya saran, diharapkan nantinya dapat membantu perusahaan untuk lebih meningkatkan kualitas sistem yang sudah ada sebelumnya.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan analisis dan implementasi sistem, maka diperoleh kesimpulan sebagai berikut:

- Dengan menggunakan algoritma *Complete Linkage* sistem dapat membentuk *cluster-cluster* dengan nilai *purity* tanpa penggunaan *WordNet* lebih tinggi daripada penggunaan *WordNet* meski selisihnya tidak mencapai 0,1.
- Dengan uji coba menggunakan *WordNet* didapatkan nilai *purity* sebesar 0,8308 untuk *cutting point* 0,02. Uji coba tanpa menggunakan *WordNet* menghasilkan nilai *purity* sebesar sebesar 0,9077 untuk *cutting point* yang sama.
- Pengujian dengan *WordNet* menghasilkan *term* yang lebih sedikit bila dibandingkan tanpa penggunaan *WordNet*, hal ini dikarenakan ada beberapa *term* yang dihapus dan digantikan oleh *term* baru dari *WordNet*. Pergantian *term* ini akan meningkatkan bobot *term* sehingga meningkatkan kemungkinan lolos *feature selection*.

5.2 Saran

Saran untuk pengembangan dan perbaikan sistem adalah:

- Perlunya perbaikan dalam menggunakan *feature selection*, misalnya dengan mengganti metode *feature selection* dengan metode yang lebih baik dari *Frequency-based Feature Selection*.
- Dapat ditambahkan dengan penggunaan *stemming* dan *lemmatization* untuk *pre-processing*.
- Dapat menambahkan tabel-tabel yang digunakan dalam *WordNet*.

DAFTAR PUSTAKA

Ackermann, M.R., Blomer, J., Kuntze, D., & Sohler, S. (2010). *Analysis of Agglomerative Clustering*. Diakses 25 Februari 2012 dari http://drops.dagstuhl.de/ opus/volltexte/2011/2994/pdf/30.pdf.

Agirre, E., & Lopez de Lacalle, O. *Clustering WordNet Word Senses*. (n.d.). Di akses 29 Februari 2012 dari http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.80.527&rep=rep1&type=pdf.

Frakes, W.B., & Baeza-Yates, R. (Eds). (1992). *Information Retrieval : Data Structure & Algorithms*. United States of America : Prentice-Hall Inc.

Grossman, D.A., & Frieder, O. (2004). *Information Retrieval Algorithms and Heuristics 2nd Ed*. Netherlands : Springer.

Han, J., & Kamber, M. (2006). *Data Mining Concepts & Techniques*. San Fransisco : Morgan Kaufmann.

Hartini, E. *Metode Clustering Hirarki*. (n.d.). Diakses 1 Februari 2012 dari http://www.batan.go.id/ppin/lokakarya/LKSTN_15/Entin.pdf

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

Moens, M., (2006). *Information Extraction: Algorithms and Prospects in a Retrieval Context*. Netherlands : Springer.

Passos, A., & Wainer, J. *WordNet-Based Metrics Do Not Seem to Help Document Clustering*. Di akses 29 Februari 2012 dari http://www.ic.unicamp.br/~tachard/docs/wncluster.pdf.

Sambamoorthi, N. (n.d.). *Hierarchical Cluster Analysis*. Di akses 4 September 2012 dari http://www.crmportals.com/hierarchical_cluster_analysis.pdf.

Soraya, Y. (2011). *Perbandingan Kinerja Metode Single Linkage, Metode Complete Linkage dan Metode K-Means dalam Analisis Cluster*. Di akses 8 Desember 2011 dari http://lib.unnes.ac.id/6803/1/8503.pdf.

Tan, P., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*. Boston : Addison-Wesley.