IMPLEMENTASI METODE IMPROVED K-NEAREST NEIGHBOR DALAM PENGKLASIFIKASIAN ARTIKEL BAHASA INGGRIS

Tugas Akhir

Oleh

Fanny Pritami Widodo
22074208

Program Studi Teknik Informatika Fakultas Teknik

Universitas Kristen Duta Wacana

2010

IMPLEMENTASI METODE IMPROVED K-NEAREST NEIGHBOR DALAM PENGKLASIFIKASIAN ARTIKEL BAHASA INGGRIS

Tugas Akhir

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

Disusun Oleh :

Fanny Pritami Widodo
22074208

Program Studi Teknik Informatika Fakultas Teknik
Universitas Kristen Duta Wacana
2010

PERNYATAAN KEASLIAN TUGAS AKHIR

Saya menyatakan dengan sesungguhnya bahwa tugas akhir dengan judul :

Implementasi Metode Improved k-Nearest Neighbor Dalam Pengklasifikasian Artikel Bahasa Inggris

Yang saya kerjakan untuk melengkapi sebagian persyaratan menjadi Sarjana Komputer pada pendidikan sarjana Program Studi Teknik Informatika, Fakultas Teknik 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, 30 November 2010

Fanny Pritami Widodo

(_____)

22074208

INTISARI

IMPLEMENTASI METODE IMPROVED K-NEAREST NEIGHBOR DALAM PENGKLASIFIKASIAN ARTIKEL BAHASA INGGRIS

Semakin hari semakin banyak inovasi, perkembangan, dan temuan-temuan yang terkait dengan bidang Teknologi Informasi dan Komputer. Hal ini menyebabkan semakin banyaknya artikel-artikel ilmiah yang diterbitkan dalam proses mengembangkan dan menemukan hal-hal baru dalam bidang tersebut dan semakin sulitnya untuk mengorganisasi artikel-artikel tersebut secara efisien.

Untuk mempermudah pengorganisasian artikel tersebut, penulis membangun sebuah sistem pengelompokkan artikel yang dikenal dengan sistem klasifikasi. Penulis menggunakan metode klasifikasi yang mudah dan efektif yaitu *Improved k-Nearest Neighbor* yang dimodifikasi dari metode *k-Nearest Neighbor*. Modifikasi dilakukan dengan tujuan untuk meningkatkan performa dalam melakukan klasifikasi. Metode ini didasari oleh penelitian yang dilakukan oleh Muhammed Miah (2009). Dalam metode ini, tidak semua dokumen pelatihan akan dihitung kesamaannya dengan dokumen uji dan tidak menggunakan semua kata yang ada dalam dokumen pelatihan sehingga metode ini menjadi lebih efektif. Penulis menguji keakuratan yang dihasilkan oleh metode ini dan melakukan beberapa pengujian untuk mengetahui kondisi yang dapat meningkatkan keakuratan hasil dari sistem.

Sistem yang dibangun oleh dapat menghasilkan tingkat akurasi yang cukup tinggi dengan menggunakan nilai k sebesar 20, yaitu 83,33%. Modifikasi yang dilakukan tidak memberi manfaat yang signikan terhadap performa sistem.

HALAMAN PERSETUJUAN

Judul	:	Implementasi Metode Improved k-Nearest Neighbor Dalam Pengklasifikasian Artikel Bahasa Inggris
Nama	:	Fanny Pritami Widodo
NIM	:	22074208
Mata Kuliah	:	Tugas Akhir
Kode	:	TI2126
Semester	:	Gasal
Tahun Akademik	:	2010/2011

Telah diperiksa dan disetujui

Di Yogyakarta,

Pada Tanggal 30-11-2010

Dosen Pembimbing I	Dosen Pembimbing II
Lucia Dwi Krisnawati, S.S, M.A	Antonius Rachmat, S.Kom, M.Cs

HALAMAN PENGESAHAN

SKRIPSI

IMPLEMENTASI METODE K-NEAREST NEIGHBOR DALAM PENGKLASIFIKASIAN ARTIKEL BAHASA INGGRIS

Oleh : Fanny Pritami Widodo / 22074208

Dipertahankan di depan dewan Penguji Tugas Akhir/Skripsi
Program Studi Teknik Informatika Fakultas Teknik
Universitas Kristen Duta Wacana – Yogyakarta
Dan dinyatakan diterima untuk memenuhi salah satu
Syarat memperoleh gelar
Sarjana Komputer
Pada tanggal
22 Desember 2010

Yogyakarta, 22 Desember 2010
Mengesahkan,

Dewan Penguji :
1. Lucia Dwi Krisnawati, S.S, M.A.
2. Antonius Rachmat C,S.Kom., M.Cs.
3. Drs. R. Gunawan Santosa, M.Si.
4. Drs. Jong Jek Siang, M.Sc.

Dekan

(Ir. Henry Feriadi, Ms.Sc., Ph.D)

Ketua Program Studi

(Restyandito, S.Kom, M.SIS.)

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 Implementasi Metode Improved k-Nearest Neighbor Dalam Pengklasifikasian Artikel Bahasa Inggris.

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. Ibu Lucia Dwi Krisnawati, S.S, M.A. selaku pembimbing I yang telah memberukan bimbingannya dengan sabar dan baik kepada penulis, juga kepada
2. Bpk Antonius Rachmat, S.Kom, M.Cs. selaku dosen pembimbing II atas bimbingannya, petunjuk, dan masukan yang diberikan selama pengerjaan tugas akhir ini sejak awal hingga akhir, juga kepada
3. Deddy Wijaya Suliantoro, S.Kom. atas bimbingan dan masukannya yang diberikan saat akan mengajukan Tugas Akhir ini.
4. Keluarga tercinta yang member dukungan dan semangat.

5. Evarisma Wulandari, Danny Sebastian, Yohanes Septian, Shinta Marzelina, Lidya Agnes, dan Albert Michael yang selalu memberikan dukungan dan semangat.
6. Sahabat dan teman-teman yang telah memberikan masukan dan semangat.
7. 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,

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LISTING

Bab 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Semakin hari semakin banyak inovasi, perkembangan, dan temuan-temuan yang terkait dengan bidang Teknologi Informasi dan Komputer. Hal ini menyebabkan semakin banyaknya artikel-artikel ilmiah yang diterbitkan dalam proses mengembangkan dan menemukan hal-hal baru dalam bidang tersebut dan semakin sulitnya untuk mengorganisasi artikel-artikel tersebut secara efisien. Dengan mengelompokkan artikel-artikel tersebut maka akan memudahkan dalam mencari artikel yang diinginkan. Salah satu cara mendapatkan efisiensi dalam mengorganisasi artikel tersebut adalah dengan mengelompokkannya berdasarkan kategori tertentu. Pengelompokkan artikel ini dikenal dengan pengklasifikasian. Sejumlah metode klasifikasi telah banyak diterapkan untuk melakukan pengkategorian seperti *Naïve Bayes*, pohon keputusan, *k-Nearest Neighbor*, *Support Vector Machine*, dan sebagainya.

Salah satu metode klasifikasi yang mudah dan efektif adalah metode *k-Nearest Neighbor*. Dalam prosesnya, *k-Nearest Neighbor* memeriksa semua kata dalam dokumen pelatihan untuk menghitung kesamaannya dengan dokumen yang akan diklasifikasian (dokumen uji). Hal ini mengakibatkan lamanya waktu untuk melakukan klasifikasi artikel ilmiah jika jumlah dokumen pelatihan sangat banyak. Selain itu, jika korpus didominasi oleh satu atau beberapa kategori (label), maksudnya satu atau beberapa kategori memiliki jumlah dokumen yang sangat banyak, sedangkan kategori lain hanya memiliki sedikit dokumen, penggunaan

metode *k-Nearest Neighbor* yang menentukan kategori berdasarkan jumlah dokumen terbanyak akan menghasilkan kategori yang salah. Untuk itu, digunakan metode *Improved k-Nearest Neighbor* (k-NN_I) yang melakukan modifikasi pada metode *k-Nearest Neighbor* biasa sehingga dapat meningkatkan performa dalam melalukan klasifikasi. Metode ini didasari oleh penelitian yang dilakukan oleh Muhammed Miah (2009). Dalam metode ini, tidak semua dokumen pelatihan akan dihitung kesamaannya dengan dokumen uji dan tidak menggunakan semua kata yang ada dalam dokumen pelatihan sehingga metode ini menjadi lebih efektif. Dalam penelitian ini, penulis akan merancang sebuah sistem klasifikasi menggunakan metode *Improved k-Nearest Neighbor*.

1.2 Perumusan Masalah

Berdasarkan latar belakang masalah diatas, penulis akan merancang dan membangun sebuah sistem yang akan melakukan proses klasifikasi artikel ilmiah. Masalah yang akan diteliti adalah :

- Seberapa besar akurasi yang dihasilkan sistem dalam melakukan klasifikasi dengan menggunakan metode *Improved k-Nearest Neighbor*.
- Bagaimana pengaruh nilai k terhadap akurasi sistem.
- Bagaimana pengaruh kemunculan keyword dan judul terhadap akurasi sistem.
- Bagaimana pengaruh penghapusan kata yang memiliki TF dan DF tinggi terhadap akurasi sistem.

1.3 Batasan Masalah

Batasan dalam sistem ini adalah sebagai berikut:

- Sistem yang dibangun berupa aplikasi *web* yang diuji pada jaringan lokal.
- Data yang digunakan adalah artikel ilmiah dalam bentuk artikel berbahasa Inggris yang bersumber pada http://portal.acm.org .
- Artikel ilmiah diklasifikasikan dalam 5 kategori di bidang *IT*, yaitu *Software Engineering* (Rekayasa Perangkat Lunak), *Information Technology* (Teknologi Informasi), *Communication and Networking* (Komunikasi dan Jaringan), *Intelligent Technology* (Sistem Cerdas), dan *Computer Science* (Ilmu Komputer).
- Pembobotan yang digunakan adalah perhitungan *TF* dimana *TF* ini merupakan perbandingan dari jumlah banyaknya suatu kata terhadap jumlah semua kata yang terdapat dalam suatu dokumen.
- *Stoplist* yang digunakan dalam sistem klasifikasi ini terdiri dari 570 *stopwords* yang diperoleh dari http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop, 7 nama hari (dalam bahasa inggris), 12 nama bulan (dalam bahasa inggris), kata-kata yang sudah dipesan dalam MySQL sebanyak 131 kata yang diperoleh dari http://dev.mysql.com/doc/refirman/5.0/en/reserved-words.html.
- Pemilihan nilai *k* yang digunakan ditentukan oleh user sebelum melakukan proses klasifikasi.
- Tidak adanya tahap *stemming* dalam pre-prosessing dokumen.
- Bentuk masukan ke sistem berupa *single plained text* yang mengandung isi dari artikel ilmiah dan nama *file* berupa judul dari artikel tersebut, *keyword* artikel serta kategori dari artikel tersebut untuk dokumen pelatihan.
- Bentuk keluaran dari sistem berupa label kategori dari artikel yang dimasukkan untuk pengklasifikasian.

1.4 Hipotesis

- Implementasi metode *Improved k-Nearest Neighbor* dalam pengklasifikasian atikel ilmiah memberikan tingkat akurasi yang cukup tinggi ditinjau dari persentase jumlah dokumen yang diklasifikasian dengan benar.
- Nilai k yang semakin kecil menghasilan tingkat akurasi yang semakin tinggi.
- Dengan mempertimbangkan kemunculan *keyword* dan judul dari artikel, dapat menghasilkan tingkat akurasi yang lebih tinggi tanpa mempertimbangkan kemunculan *keyword* dan judul.
- Penghapusan kata yang mempunyai TF dan DF tinggi menghasilkan tingkat akurasi yang lebih tinggi.

1.5 Tujuan Penelitian

Tujuan dari penelitian ini adalah menghasilkan sistem klasifikasi artikel yang akurat menggunakan metode *k-Nearest Neighbor* untuk kategori di bidang Teknologi Informasi dan Komputer, khususnya untuk bidang *Software Engineering, Information Technology, Intelligent Technology, Communication and Networking,* dan *Computer Science*.

1.6 Metode Penelitian

Metode yang digunakan dalam penelitian ini adalah metode *Improved k-Nearest Neighbor*, sedangkan metode yang digunakan untuk mengumpulkan data adalah Studi Pustaka. Studi Pustaka yang dilakukan dengan mencari dan mempelajari buku-buku dan sumber dari internet yang berkaitan dengan klasifikasi dokumen.

1.7 Sistematika Penulisan

Penulisan laporan tugas akhir ini dibagi menjadi lima (5) bab, yaitu :

Bab 1 Pendahuluan, yang memberikan gambaran umum mengenai apa yang diteliti dalam penulisan tugas akhir ini. Pendahuluan memuat latar belakang masalah, perumusan masalah, batasan masalah, hipotesis, tujuan penelitian, metode penelitian, dan sistematika penulisan laporan.

Bab 2 Tinjauan Pustaka, yang terdiri dari tinjauan pustaka dan landasan teori. Tinjauan pustaka menguraikan berbagai teori mengenai klasifikasi dengan metode *k-Nearest Neighbor* yang didapatkan dari berbagai sumber pustaka yang digunakan dalam melakukan penelitian. Landasan teori berisi konsep dan prinsip utama yang digunakan untuk memecahkan masalah penelitian.

Bab 3 Analisis dan Perancangan Sistem, mencakup tahap perancangan sistem yang akan dibuat seperti kebutuhan *hardware* dan *software*, spesifikasi sistem, arsitektur sistem, diagram *use case*, algoritma yang digunakan dalam membuat sistem,kamus data, skema *database*, rancangan antarmuka, dan rancangan pengujian sistem.

Bab 4 Implementasi dan Analisis Sistem, membahas implementasi dan pengujian sistem yang dibuat berdasarkan bab 3, beserta hasil dari sistem yang dijalankan dan analisis dari sistem yang dibuat.

Bab 5 Kesimpulan dan Saran, berisi kesimpulan dari hasil penelitian yang dilakukan dan saran untuk memberikan hasil yang lebih baik lagi dalam penelitian yang sejenis.

Bab 5

KESIMPULAN DAN SARAN

5.1 Kesimpulan

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

1. Sistem klasifikasi dengan metode *Improved k-Nearest Neighbor* memberikan tingkat akurasi yang cukup tinggi dengan menggunakan nilai *k* sebesar 20, yaitu 83,33% sesuai pengujian yang telah dilakukan.
2. Semakin kecil nilai *k* yang digunakan dalam melakukan proses klasifikasi, maka semakin akurat hasil dari sistem. Pernyataan ini tidak berlaku untuk k=1 hingga k=4.
3. Dari hasil *confusion matrix* selama evaluasi sistem, diperoleh 7,69% *Computer and Networking* terjadi kebingungan ke *Information Technology*, 10,99% *Computer Science* terjadi kebingungan ke *Intelligent Technology*, 18,57% *Information Technology* terjadi kebingungan ke *Intelligent Technology*, 2,38% *Intelligent Technology* terjadi kebingungan ke *Computer Science*, dan 14,29% *Software Engineering* terjadi kebingungan ke *Computer Science*.
4. Memperbesar nilai perbandingan tf untuk metadata tidak memberi pengaruh yang berarti dalam proses klasifikasi.
5. Penghilangan token dengan nilai perbandingan tf dan df yang tinggi, setelah penghapusan *stopword*, dapat meningkatkan keakuratan sistem, namun tidak terlalu signikan.

6. Algoritma pengecekan nilai jumlah perbandingan tf dengan nilai *similarity* terkecil dalam *buffer* pada metode *Improved k-Nearest Neighbor* tidak pernah berfungsi.

5.2 Saran

Sistem yang dibuat dapat dikembangkan lebih lanjut untuk mencapai hasil yang optimal. Saran yang diajukan penulis untuk pengembangan dan perbaikan sistem adalah :

1. Diperlukan perbaikan struktur data untuk mempercepat proses *text-preprocessing* karena semakin banyaknya kata dalam dokumen maka semakin lama waktu yang dibutuhkan untuk memprosesnya, dengan cara penggunaan *indexing* pada *database* dan lebih banyak penggunaan *stored procedure* untuk melakukan perhitungan sehingga mengurangi penggunaan *memory* dan mempercepat waktu yang dibutuhkan.
2. Dapat ditambahkan kemampuan untuk mem-*parsing keyword* dari artikel secara otomatis.
3. Dapat ditambahkan kemampuan untuk pengelolaan kategori dan *stoplist*.

DAFTAR PUSTAKA

Feldman R., & Sanger J. (2007). *The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press.

Han E., Karypis G., Kumar V. (2001). *Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification*. Diakses 13 Februari 2010, dari http://www.springerlink.com/index/25gnd0jb6nklffhh.pdf.

Han J. & Kamber M. (2006). *Data Mining : Concepts and Technique 2nd Edition*. San Francisco: Morgon Kaufmann Publishers.

Intan R. & Defeng A. (2006). HARD: Subject-Based Search Engine Menggunakan TF-IDF dan Jaccard's Coefficient. *Jurnal Teknik Industri*. 8(1). 61-72.

Manning C. D., Raghavan P., & Schütze H. (2008). *An Introduction to Information Retrieval*. New York: Cambridge University Press.

Miah M. (2009). *Improved k-NN Algorithm for Text Classification*. Diakses 27 Februari 2010, dari http://dbxlab.uta.edu/dbxlab/Muhammed/DMIN2009.pdf.

Ridok A., Furqon M.T. (2009). *Pengelompokan Dokumen Berbahasa Indonesia Menggunakan Metode K-NN*. Diakses 13 Februari 2010, dari http://matematika.brawijaya.ac.id/web/cms/index2.php?option=com_docman&task=doc_view&gid=319&Itemid=99999999.