IMPLEMENTASI METODE K- NEAREST NEIGHBOR DALAM PENENTUAN AUTO-TAGGING PADA ARTIKEL BLOG

Skripsi

Oleh:

Setiadi Yulianto

22084416

PROGRAM STUDI TEKNIK INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
2016

IMPLEMENTASI METODE K- NEAREST NEIGHBOR DALAM PENENTUAN AUTO-TAGGING PADA ARTIKEL BLOG

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:
Setiadi Yulianto
22084416

PROGRAM STUDI TEKNIK INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
2016

PERNYATAAN KEASLIAN SKRIPSI

Saya menyatakan dengan sesungguhnya bahwa skripsi dengan judul:

IMPLEMENTASI METODE K-NEAREST NEIGHBOR PADA PENENTUAN AUTO-TAGGING PADA ARTIKEL BLOG

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, 9 Juni 2016

SETIADI YULIANTO
22084416

HALAMAN PERSETUJUAN

Judul Skripsi	:	IMPLEMENTASI METODE K-NEAREST NEIGHBOR PADA PENENTUAN AUTO-TAGGING PADA ARTIKEL BLOG
Nama Mahasiswa	:	SETIADI YULIANTO
N I M	:	22084416
Matakuliah	:	Skripsi (Tugas Akhir)
Kode	:	TIW276
Semester	:	Genap
Tahun Akademik	:	2015/2016

Telah diperiksa dan disetujui di
Yogyakarta,
Pada tanggal 9 Juni 2016

Dosen Pembimbing I	Dosen Pembimbing II
Antonius Rachmat C., S.Kom.,M.Cs.	Rosa Delima, S.Kom., M.Kom.

HALAMAN PENGESAHAN

IMPLEMENTASI METODE K-NEAREST NEIGHBOR PADA PENENTUAN AUTO-TAGGING PADA ARTIKEL BLOG

Oleh: SETIADI YULIANTO / 22084416

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 memperoleh gelar
Sarjana Komputer
pada tanggal 3 Juni 2016

Yogyakarta, 9 Juni 2016
Mengesahkan,

Dewan Penguji:

1. Antonius Rachmat C., S.Kom., M.Cs.
2. Rosa Delima, S.Kom., M.Kom.
3. Aditya Wikan Mahastama, S.Kom., M.Cs.
4. R. Gunawan Santosa, Drs. M.Si.

Dekan	Ketua Program Studi
(Budi Susanto, S.Kom., M.T.)	(Gloria Virginia, Ph.D.)

UCAPAN TERIMA KASIH

Puji syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa atas berkat, rahmat, dan karunianya sehingga penulis dapat menyelesaikan Tugas Akhir dengan judul "Implementasi Metode K- Nearest Neighbor Dalam Penentuan Auto-Tagging Pada Artikel Blog" dengan baik.

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

Dalam menyelesaikan penelitian 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. Bapak Antonius Rachmat C, S.Kom, M.Cs selaku dosen pembimbing I yang pertama yang selalu sabar dalam membimbing penulis dalam mengerjakan penelitian dan penyusunan laporan Tugas Akhir.
2. Ibu Rosa Delima, S.Kom., M.Kom. selaku dosen pembimbing II yang selalu sabar dan baik membimbing penulis dalam mengerjakan penelitian dan penyusunan laporan Tugas Akhir.
3. Rekan-rekan penulis yang dengan senang hati memberikan arahan, saran, dan, sharing dalam pengerjaan Tugas Akhir maupun penulisan laporan Tugas Akhir.
4. Teman-teman yang selalu dengan senantiasa mendampingi dan menemani dalam pengerjaan maupun penyusunan laporan Tugas Akhir.
5. Pihak lain yang tidak dapat penulis sebutkan satu per satu, sehingga Tugas Akhir ini dapat terselesaikan dengan baik.

Penulis menyadari bahwa penelitian 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 nanti penulis dapat memberikan karya yang lebih baik lagi.

Akhir kata penulis meminta maaf bila ada kesalahan dalam penyusunan laporan maupun sewaktu penulis melakukan penelitian Tugas Akhir. Semoga penelitian dan laporan Tugas Akhir ini dapat berguna bagi kita semua.

Yogyakarta, 12 Mei 2016

Penulis

INTISARI

Implementasi Metode K- Nearest Neighbor Dalam Penentuan Auto-Tagging Pada Artikel Blog

Saat ini ketersediaan sumber informasi sangatlah besar. Informasi yang disajikan sebagian besar berbentuk teks dan elektronik. Kondisi ini menyebabkan kesulitan bagi pengguna untuk mendapatkan informasi yang dibutuhkan. Oleh karena itu dilakukanlah pengelompokan informasi menurut kemiripannya. Hal ini dinamakan label (*tagging*). Label adalah kata kunci yang tugasnya adalah menunjukkan potongan-potongan informasi yang dilakukan untuk membantu mengklasifikasi suatu data yang ada. Meskipun tagging mempermudah pencarian, sebagian orang melakukan tagging secara manual.

Pada penelitian ini akan diteliti mengenai pembuatan *auto-tagging* oleh sistem dengan metode *K- Nearest Neighbor* dengan menghitung nilai *TF-IDF* pada awalnya. Setelah itu akan dilakukan penghitungan *Cosine Similarity* dan perhitungan kembali *TF-IDF* baru yang kemudian terbentuklah *auto-tagging*. Dalam penelitian ini juga akan dilakukan perbandingan seberapa besar kemiripan hasil dari tag baru dengan tag awal dari artikel tersebut.

Penelitian membuktikan bahwa nilai dari K tidak mempengaruhi hasil keakuratan dari sistem sedangkan semakin besar *feature selection*, semakin baik pula akurasi yang dihasilkan. Penggunaan *feature selection* yang optimal berada di 90%. Pada penelitian antara hasil keakuratan antara penggunaan bobot (TF-IDF) dan tidak diberi bobot didapatkan bahwa pemberian bobot memiliki akurasi yang lebih baik.

Kata Kunci : *Auto-Tagging, K-Nearest Neighbor, TF-IDF ,Cosine Similarity*

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN

INTISARI

Implementasi Metode K- Nearest Neighbor Dalam Penentuan Auto-Tagging Pada Artikel Blog

Saat ini ketersediaan sumber informasi sangatlah besar. Informasi yang disajikan sebagian besar berbentuk teks dan elektronik. Kondisi ini menyebabkan kesulitan bagi pengguna untuk mendapatkan informasi yang dibutuhkan. Oleh karena itu dilakukanlah pengelompokan informasi menurut kemiripannya. Hal ini dinamakan label (*tagging*). Label adalah kata kunci yang tugasnya adalah menunjukkan potongan-potongan informasi yang dilakukan untuk membantu mengklasifikasi suatu data yang ada. Meskipun tagging mempermudah pencarian, sebagian orang melakukan tagging secara manual.

Pada penelitian ini akan diteliti mengenai pembuatan *auto-tagging* oleh sistem dengan metode *K- Nearest Neighbor* dengan menghitung nilai *TF-IDF* pada awalnya. Setelah itu akan dilakukan penghitungan *Cosine Similarity* dan perhitungan kembali *TF-IDF* baru yang kemudian terbentuklah *auto-tagging*. Dalam penelitian ini juga akan dilakukan perbandingan seberapa besar kemiripan hasil dari tag baru dengan tag awal dari artikel tersebut.

Penelitian membuktikan bahwa nilai dari K tidak mempengaruhi hasil keakuratan dari sistem sedangkan semakin besar *feature selection*, semakin baik pula akurasi yang dihasilkan. Penggunaan *feature selection* yang optimal berada di 90%. Pada penelitian antara hasil keakuratan antara penggunaan bobot (TF-IDF) dan tidak diberi bobot didapatkan bahwa pemberian bobot memiliki akurasi yang lebih baik.

Kata Kunci : *Auto-Tagging, K-Nearest Neighbor, TF-IDF ,Cosine Similarity*

BAB I
PENDAHULUAN

1.1. Latar Belakang Masalah

Saat ini ketersediaan sumber informasi sangatlah besar. Informasi yang disajikan sebagian besar berbentuk teks dan elektronik. Kondisi ini menyebabkan kesulitan bagi pengguna untuk mendapatkan informasi yang dibutuhkan. Oleh karena itu dilakukanlah pengelompokan informasi menurut kemiripannya. Hal ini dinamakan label (tagging).

Label (tagging) adalah kata kunci yang tugasnya adalah menunjukkan potongan-potongan informasi yang dilakukan untuk membantu mengklasifikasi suatu data yang ada. Label merupakan jenis metadata yang membantu untuk menjelaskan suatu hal dan memungkinkan hal tersebut ditemukan ketika melakukan pencarian. Disinilah peran penting label dalam mempermudah pencarian suatu informasi tersebut.

Menghadapi masalah tersebut penulis melakukan penelitian mengenai auto-tagging yang berguna untuk membuat tag otomatis sehingga orang tidak perlu membuat secara manual. Terdapat beberapa metode yang bisa dilakukan dalam melakukan tag otomatis antara lain *K-Means , K-Median, K-Medoid, K- Nearest Neighbor, Naïve Bayes*. Dari beberapa metode tersebut metode *KNN* merupakan metode yang paling efektif dalam menangani data yang besar dan dapat menghasilkan data yang lebih akurat dibanding metode yang lain (Miah, 2009). Oleh karena itu penulis pada penelitian kali ini menggunakan metode *K-Nearest Neighbor (KNN)*.

Penelitian ini akan menghasilkan aplikasi (sistem) yang dapat membantu secara otomatis dalam menentukan tagging secara lebih tepat / meningkatkan keakuratan dalam pemilihan kata untuk tagging. Dengan sistem ini penulis juga berharap adanya kesesuaian antara isi artikel blog dengan tagging yang dihasilkan oleh sistem.

1.2. Rumusan Masalah

Perumusan Masalah dalam penelitian ini adalah :

1. Bagaimana mengimplementasikan algoritma *K-Nearest Neighbor* untuk tagging otomatis (auto-tagging) pada artikel blog?
2. Seberapa akurat sistem dalam menghasilkan auto-tagging yang tepat pada blog?

1.3. Batasan Masalah

Batasan masalah dalam penelitian ini antara lain :

- Artikel yang digunakan berbahasa Indonesia dan memiliki tag.
- Artikel yang digunakan hanya mencakup olah raga dan diambil dari blog ligaolahraga.com , indoberita.com.
- Artikel blog diambil sebanyak 100 buah dengan 70 sebagai dokumen pelatihan dan 30 sebagai dokumen yang akan dilatih.
- Tidak dapat menggunakan kata majemuk.
- Diasumsikan tag asli artikel benar.

1.4. Tujuan Penelitian

Tujuan penelitian ini adalah sebagai berikut :

1. Mengimplementasikan *K-Nearest Neighbor* agar dapat melakukan auto-tagging pada artikel blog
2. Meneliti seberapa akurat sistem dalam menghasilkan auto-tagging yang tepat pada artikel blog

1.5. Metode Penelitian

Pada penelitian ini penulis akan melakukan beberapa hal seperti yang akan dijelaskan berikut :

1. Studi Pustaka

   Mencari informasi dan mempelajari teori-teori yang berkaitan dengan topik yang dipilih dari berbagai sumber. Sumber-sumber yang dipakai adalah dari buku ataupun internet.

2. Pengumpulan data

   Mengumpulkan data berupa artikel atau data lainnya yang mendukung penelitian penulis. Artikel yang diambil berjumlah sekitar 100 blog.

3. Perancangan

   Penulis melakukan perancangan sistem, menyediakan antarmuka pengguna yang akan digunakan sebagai perantara komunikasi antar pengguna dengan sistem serta merancang database.

4. Implementasi

   Penulis mengimplementasikan teori-teori serta metode *K-Nearest Neighbor* yang digunakan ke dalam bentuk program dengan bahasa pemrograman PHP.

5. Evaluasi

   Pengujian dilakukan dengan membandingkan hasil auto-tagging antara metode *K-Nearest Neighbor* dengan tag asli dari artikel blog yang telah dipersiapkan. Dengan ini maka dapat dilihat seberapa tingkat persamaan yang ada dari hasil tag metode tersebut dengan tag asli.

1.6. Sistematika Penulisan

Penulisan skripsi ini dibagi menjadi lima bagian sebagai berikut:

BAB 1 : PENDAHULUAN

Menjelaskan beberapa pokok mengenai latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, dan metode/pendekatan yang dilakukan dalam penelitian ini

BAB 2 : TINJAUAN PUSTAKA

Pada bab ini akan diuraikan tentang teori-teori atau dasar-dasar pengetahuan yang berhubungan dengan metode atau algoritma yang digunakan pada pembuatan sistem.

BAB 3 : PERANCANGAN SISTEM

Akan dibahas tentang perancangan sistem yang akan dibangun. Terlebih dijelaskan tentang input-output sistem, serta menjelaskan rancangan cara kerja sistem.

BAB 4 : IMPLEMENTASI DAN ANALISIS SISTEM

Bab ini akan menjelaskan hasil implementasi serta hasil analisa dari sistem yang telah dirancang.

BAB 5 : KESIMPULAN DAN SARAN

Pada bab ini akan menjawab/menyimpulkan kegiatan yang telah dilakukan selama masa penelitian. Juga dituliskan tentang saran kedepan yang dapat membangun sistem menjadi lebih baik.

BAB 5
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Dari hasil pengamatan dan penelitian yang penulis lakukan, terlihat bahwa pengimplementasian *K-Nearest Neighbor* pada sistem untuk melakukan auto-tagging sudah cukup bagus. Dalam penelitian yang dilakukan dapat disimpulkan :

1. Pengujian menggunakan 30 artikel untuk pengujian *feature selection* sebesar 10% sampai 100% didapatkan kesimpulan bahwa semakin besarnya *feature selection* maka semakin tinggi tingkat akurasi yang dihasilkan dan nilai optimal untuk *feature selection* adalah 90%.
2. Berdasarkan pengujian, nilai variabel K tidak berpengaruh terhadap akurasi sistem.
3. Rata-rata akurasi sistem adalah 62,28%.
4. Berdasarkan pengujian pemberian bobot untuk perhitungan, TF-IDF berpengaruh terhadap akurasi sistem.

5.2 Saran

Pada penelitian ini, sistem yang dihasilkan belum mampu untuk memproses kata majemuk. Kata majemuk merupakan salah satu faktor penting untuk menyempurnakan hasil dari penelitian ini. Oleh karena itu diharapkan dalam pengembangan berikutnya kata majemuk dapat diterapkan di dalam sistem sehingga tagging yang dihasilkan menjadi semakin baik.

Daftar Pustaka

Aggarwal, C. C., dan Zai, C. (2012). Mining Text Data. London : Springer.

Blood, R. (2000). Weblogs: a history and perspective. Diakses tanggal 1 Juli 2014 dari http://www.rebeccablood.net/essays/weblog_history.html

Brook, C. H., dan Montanez, N. (2006). Improved Annotation of the Blogosphere via Autotagging and Hierarchical Clustering.

Evan (2010) . Buku TA K-Nearest Neighbor (KNN). Diakses tanggal 30 Juni 2014 dari http://kuliahinformatika.wordpress.com/2010/02/13/buku-ta-k-nearest-neighbor-knn

Golder, S. A., dan Huberman, B. A. (2005). Usage patterns of collaborative tagging systems. Journal of Information Science.

Han, J., dan Kamber, M. (2011). Data Mining Concepts and Techniques (3rd ed). San Francisco : Morgan Kaufman Publisers

Huang, A.(2012). Similarity Measure for Text Dokumen Clustering. Diakses tanggal 1 Juli 2014 dari http://www.milanmirkovic.com/wp-content/uploads/2012/10/pg049_Similarity_Measures_for_Text_Document_Clustering.pdf

Intan, R., dan Defeng, A. (2006). HARD : Subject-based Search Engine menggunakan TF-IDF dan Jaccard's Coefficient. Diakses 30 Juni 2014 dari http://fportfolio.petra.ac.id/user_files/92-008/Rolly-TI-Jurnal-June%202006(new).pdf

Kim, J., Jin, D., Kim, K.Y., dan Choe, H.(2009). Automatic In-Text Keyword Tagging based on Information Retrieval.

Kurniawan, B., Effendi, S., dan Sitompul, O.S.(2012). Klasifikasi Konten Berita Dengan Metode Text Mining. Diakses tanggal 30 Juni 2014 dari http://download.portalgaruda.org/article.php?article=58993&val=4123&title

Larose, D. T. (2005). Discovering Knowledge in Data: An Introduction to Data Mining. New Jersey : John Wiley & Sons, Inc.

Medelyan, O., Frank, E., dan Witten, I.H.(2009). Human Competitive tagging using automatic keyphase extraction.