PENERAPAN ALGORITMA TF-IDF UNTUK PENCARIAN KARYA ILMIAH

Abdul Azis Maarif
Jurusan Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
Jl. Nakula I No. 5-11 Semarang, Jl. Imam Bonjol No. 207 Semarang, 50131-Indonesia
E-mail: azismaarif@gmail.com

ABSTRAK

Sorting a scientific paper can be done easily by humans, but sorting of documents is done automatically by the computer will bring its own problems. Similarly, by measuring the level of similarity of a document with other documents, humans can easily measure whether a document has the level of similarity/similaritas with other documents. Keyword that is used in the extraction process document in the process of sorting the categories document. In order for the results of the measurement of level of similaritas documents with keywords to get optimal results then used algorithms for text mining algorithm used in the process whereby the TF-IDF (Term Frequency – Inversed Document Frequency) of the model IR (information retrieval) as a measure of the level of similaritas between documents with keywords obtained from the extraction of the text in the document. The purpose of this study was to apply the algorithm TF-IDF that can be used to find the scientific papers as the measuring level similaritas between documents with keywords obtained from the extraction of the text in the document so that it gets sorted data from that similarity (level similaritas) most high so the search papers become more efficient as the relevant information

Kata Kunci : TF-IDF, Scientific Papers, Keyword

1. Pendahuluan

1.1. Latar Belakang

Perkembangan teknologi khususnya internet sangat berperan dalam kehidupan sehari-hari. Dengan adanya internet, informasi dapat dengan mudah disebarluaskan dan diakses oleh banyak orang. Banyaknya informasi yang beredar tentu membuat kebutuhan akan informasi yang relevan semakin meningkat. Salah satu cara yang bisa digunakan untuk mendapatkan informasi yang relevan adalah dengan menggunakan sistem temu kembali informasi *(information retrieval).*

1.2. Identifikasi Masalah

Dalam penelitian ini masalah dapat dirumuskan bagaimana menerapkan algoritma TF-IDF yang dapat digunakan untuk mencari karya ilmiah.

1.3. Tujuan Penelitan

Tujuan penelitan ini adalah untuk menerapkan algoritma TF-IDF yang dapat digunakan untuk mencari karya ilmiah sebagai pengukur tingkat similaritas antara dokumen dengan *keyword* yang didapat dari ekstraksi teks pada dokumen sehingga mendapatkan data yang terurut dari yang kemiripannya (tingkat similaritas) paling tinggi sehingga pencarian karya ilmiah menjadi lebih efisien sebagai informasi yang relevan.

1.4. Manfaat

Adapun manfaat yang diharapkan dalam penelitian ini adalah :

Mempermudah dalam pencarian informasi karya ilmiah karena dapat mempresentasikan hasil informasi secara terurut berdasarkan kemiripan antara *query* dengan informasi yang ada pada dokumen karya ilmiah.

2. Tinjauan Pustaka

2.1. Information Retrieval

Information Retrieval (IR) adalah ilmu pencarian informasi dari sejumlah data yang sudah hilang karena terlalu banyaknya data yang ada. Ilmu ini dipopulerkan oleh Vannevar Bush (1945) dan implementasinya mulai dikenalkan pada tahun 1950-an. Pada tahun 1990-an, sudah banyak teknik dan metode dari in*form*atioan retrieval yang dikembangkan dan dipakai.

2.2. Sistem Temu Kembali Informasi

Sistem temu kembali informasi merupakan kegiatan yang bertujuan untuk menyediakan dan memasok informasi bagi pemakai sebagai jawaban atas permintaan atau berdasarkan kebutuhan pemakai. Pada dasarnya sistem temu balik informasi adalah suatu proses untuk mengidentifikasi, kemudian memanggil (*retrieve*) suatu dokumen dari suatu simpanan (file), sebagai jawaban atas permintaan informasi. [5]

2.3. Text Mining

Text mining dapat didefinisikan secara luas sebagai suatu proses menggali informasi dimana seorang user berinteraksi dengan sekumpulan dokumen menggunakan *tools* analisis yang merupakan komponen-komponen dalam *data mining* yang salah satunya adalah kategorisasi. *Text mining* bisa dianggap subjek riset yang tergolong baru. *Text mining* dapat memberikan solusi dari permasalahan seperti pemrosesan, pengorganisasian / pengelompokkan dan menganalisa *unstructured text* dalam jumlah besar. [3]

2.4. Tokenisasi

Secara garis besar *tokenisasi* adalah tahap memecah sekumpulan karakter dalam suatu teks kedalam satuan kata. Sekumpulan karakter tersebut dapat berupa karakter whitespace, seperti enter, tabulasi, spasi. Namun untuk karakter petik tunggal (‘), titik (.), semikolon (;), titik dua (:) atau lainnya, juga dapat memiliki peran yang cukup banyak sebagai pemisah kata. Sebuah titik (.) biasanya untuk tanda akhir kalimat, tapi dapat juga muncul dalam singkatan, inisial orang, alamat internet,dll. Kemudian tanda hypen (-) biasanya muncul untuk menggabungkan dua token yang berbeda untuk membentuk token tunggal. Tapi dapat pula ditemukan untuk menyatakan rentang nilai, kata berulang, dsb. Atau karakter slash (/) sebagai pemisah file atau direktori atau url ataupun untuk menyatakan “dan atau”.

2.5. Filtering

Filtering yaitu proses pembuangan *stopword* yang dimaksudkan untuk mengetahui suatu kata masuk kedalam *stopword* atau tidak. Pembuangan *stopword* adalah proses pembuangan *term* yang tidak memiliki arti atau tidak relevan. *Term* yang diperoleh dari tahap *tokenisasi* dicek dalam suatu daftar *stopword*, apabila sebuah kata masuk didalam daftar *stopword* maka kata tesebut akan masuk keproses berikutnya.

2.6. TF-IDF

Metode TF-IDF merupakan metode untuk menghitung bobot setiap kata yang paling umum digunakan pada *information retrieval*. Metode ini juga terkenal efisien, mudah dan memiliki hasil yang akurat [9]. Metode ini akan menghitung nilai *Term Frequency* (*TF*) dan *Inverse Document Frequency* (*IDF*) pada setiap token (kata) di setiap dokumen dalam korpus. Metode ini akan menghitung bobot setiap token t di dokumen d dengan rumus:

$$W_{dt} = tf_{dt} * IDF_t$$

Dimana :

d : dokumen ke-d

t : kata ke-t dari kata kunci

W : bobot dokumen ke-d terhadap kata ke-t

tf : banyaknya kata yang dicari pada sebuah dokumen

IDF : *Inversed Document Frequency*

Nilai *IDF* didapatkan dari

IDF : log2 (D/df)

dimana

D : total dokumen

df : banyak dokumen yang mengandung kata yang dicari

Setelah bobot (W) masing-masing dokumen diketahui, maka dilakukan proses pengurutan dimana semakin besar nilai *W*, semakin besar tingkat similaritas dokumen tersebut terhadap kata kunci, demikian sebaliknya.

2.7. Karya Ilmiah

Karya ilmiah merupakan suatu ulisan yang diperoleh sesuai dengan sifat keilmuannya dan didasari oleh hasil pengamatan, peninjauan, penelitian dalam bidang tertentu, disusun menurut metode tertentu dengan sistematika penulisan yang bersantun bahasa dan isisnya dapat dipertanggungjawabkan kebenarannya/ keilmiahannya. Karya ilmiah adalah suatu karya dalam bidang ilmu pengetahuan (*science*) dan teknologi yang berbentuk ilmiah. Suatu karya dapat dikatakan ilmiah apabila proses perwujudannya lewat metode ilmiah. [10]

3. Langkah Penyelesaian

Tahap *tokenizing* dalam penelitian ini memecah kata berdasarkan spasi dan menghasilkan kata yaitu pengetahuan, logistic, manajemen, transaksi, logistic, pengetahuan, antar, individu, dalam, manajemen, pengetahuan, terdapat, transfer, pengetahuan, logistic.

Tahap *filtering* dalam penelitian ini membuang *stopword* yaitu kata **antar,** kata **dalam** dan **terdapat** dihapus karena termasuk *stopword* atau kata yang tidak mempunyai makna atau arti sehingga kata menjadi pengetahuan, logistic, manajemen, transaksi, individu, transfer.

Setelah dilakukan tahap tokenizing dan proses *filtering.*

Implementasi sederhana dari TF-IDF adalah sebagai berikut :

Kata Kunci (KK) : sistem komputer

D1 : Perancangan Sistem Pakar Troubleshooting Personal Computer

D2 : Aplikasi Sistem Pakar Diagnosis Pada Sistem Komputer

Jumlah dokumen (D) : 2

Tahap *tokenizing* dalam penelitian ini memecah kata berdasarkan spasi dan menghasilkan kata yaitu pengetahuan, logistic, manajemen, transaksi, logistic, pengetahuan, antar, individu, dalam, manajemen, pengetahuan, terdapat, transfer, pengetahuan, logistic.

Tahap *filtering* dalam penelitian ini membuang *stopword* yaitu kata **antar,** kata **dalam** dan **terdapat** dihapus karena termasuk *stopword* atau kata yang tidak mempunyai makna atau arti sehingga kata menjadi pengetahuan, logistic, manajemen, transaksi, individu, transfer.

Setelah dilakukan tahap *tokenizing* dan proses *filtering*, maka hasil

	TF			Df	D/df	IDF	W		
Token	**KK**	**D1**	**D2**				**KK**	**D1**	**D2**
Sistem	1	0	6	1	2	0.301	0.301	0	1.806
Komputer	1	7	3	2	1	0	0	0	0

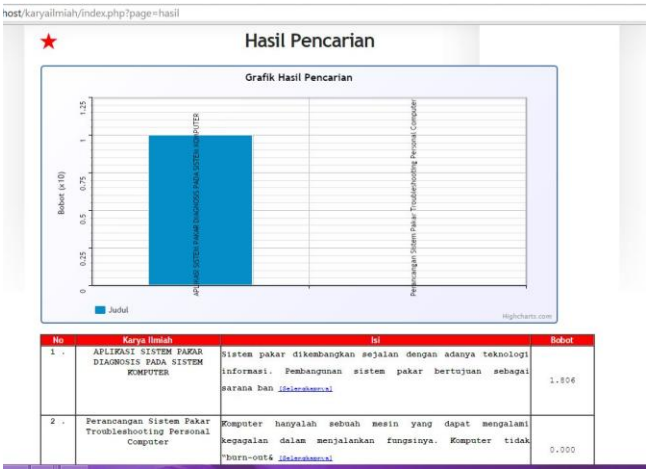
Bobot *(W)* untuk D1 = 0 + 0 = 0
Bobot *(W)* untuk D2 = 1.806 + 0 = 1.806

4. Implementasi

1. Halaman Utama Pencarian

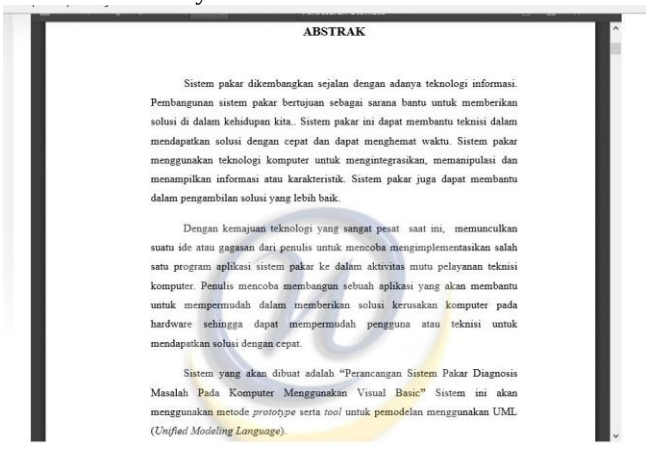
Gambar 1 Halaman Utama Pencarian

2. Hasil Pencarian



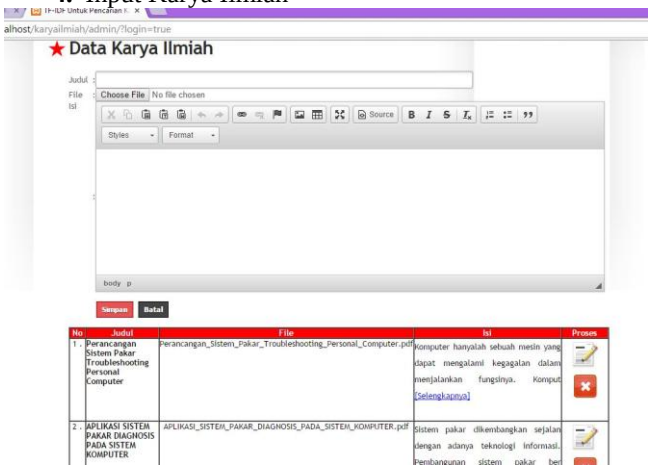
Gambar 2 Hasil pencarian

3. Detail Karya Ilmiah



Gambar 3 Detail Karya Ilmiah

4. Input Karya Ilmiah



Gambar 4 Input Karya Ilmiah

5. Kesimpulan dan Saran

5.1. Kesimpulan

Berdasarkan penelitian yang dilakukan oleh penulis dalam membuat penerapan algoritma TF-IDF untuk pencarian karya ilmiah, maka penulis dapat menarik kesimpulan sebagai berikut:

1. Telah berhasil dibuatnya sebuah aplikasi tentang pencarian karya ilmiah dan penerapan algoritma TF-IDF untuk pencarian karya ilmiah disertakan dengan hasil nilai bobot dari tiap karya ilmiah yang ditemukan.
2. Penerapan algoritma TF-IDF untuk pencarian karya ilmiah menghasilkan karya ilmiah yang dapat dilihat atau diunduh dalam bentuk format pdf.
3. Penerapan algoritma TF-IDF untuk pencarian karya ilmiah dibuat dengan menggunakan pemograman PHP, database MySQL dan perancangan sistem menggunakan use case diagram, class diagram dan activity diagram.
4. Penerapan algoritma TF-IDF untuk pencarian karya ilmiah ini dapat memverifikasi dokumen karya ilmiah dari kata kunci yang dicari dengan hasil pencarian merupakan karya ilimiah yang ditampilkan merupakan karya ilmiah yang mengandung kata kunci.

5.2. Saran

Berikut ini saran penulis terhadap pengembangan dan penerapan algoritma TF-IDF untuk pencarian karya ilmiah lebih lanjut, yaitu :

1. Diharapkan dapat dilakukan pengembangan lagi pada Penerapan algoritma TF-IDF untuk pencarian karya ilmiah dengan menggunakan metode lain.
2. Data-data yang sudah lama sebaiknya di *backup* guna untuk menghindari kehilangan data bila terjadi kerusakan pada sistem atau pada perangkat keras.
3. Perlunya dilakukan manajemen yang baik dan teratur terhadap sistem informasi yang diterapkan, hal ini dilakukan sebagai upaya pemeliharaan terhadap sistem.

DAFTAR PUSTAKA

[1] Aditya, Alan Nur. *Jago PHP & MySQL Dalam Hitungan Menit*. *Dunia Komputer*. Bekasi. 2010

[2] Akbar, Fakhreza, *Menentukan Nilai Tes Esai Online Menggunakan Algoritma Latent Semantic Analysis (Lsa) Dengan Pembobotan Term Frequency/Inverse Document Frequency*, Universitas Sumatera Utara Medan. 2011

[3] Feldman, Ronen., Sanger, Jerman., *The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data*. http://www.books24x7.com/marc.asp?bookid=23164 diakses pada tanggal 11 Januari 2015

[4] Gunadi, Suhendar Hariman, *Penggunaan Bahasa Alamiah dan Kosa Kata Terkontrol Dalam Sistem Temu Kembali Informasi Berbasis Teks*. http ://repository.usu.ac.id/bitstream/123456789/17059/... /pus-des2006-1.p diakses pada tanggal 11 Januari 2015

[5] Hasugian, Jonner, *Visual Modelling Menggunakan UML dan Rational Rose*. Informatika. Bandung. 2006

[6] Herman, Achmad, Andani, Ilham, Amil Ahmad. *Klasifikasi Dokumen Naskah Dinas Menggunakan Algoritma Term Frequency – Inversed Document Frequency Dan Vector Space Model*. Universitas Hasanudin. 2010

[7] Nugroho, Adi. *Konsep Perancangan Sistem Basis Data.* Andi. Yogyakarta. 2006

[8] Prasetyo, Didik Dwi. *Administrasi Database Server MySQL*. Elex Media Komputindo. Jakarta. 2006

[9] Robertson, Stephen, *Understanding Inverse Document Frequency: On theoretical arguments for IDF*, Journal of Documentation, Vol. 60, pp. 502–520

[10] Susilo, Eko. *Pengertian Karya Ilmiah*. https://blogueblog.wordpress.com/tag/pengertian-karya-ilmiah/ diakses pada tanggal 11 Januari 2015

[11] Herwansyah, Adhit. *Aplikasi Pengkategorian Dokumen Dan Pengukuran Tingkat Similaritas Dokumen Menggunakan Kata Kunci Pada Dokumen Penulisan Ilmiah Universitas Gunadarma*. Universitas Gunadarma. 2010