Reputasi: Jurnal Rekayasa Perangkat Lunak
Volume 1 No. 1 Mei 2020
E-ISSN 0000-0000

Information Retrieval System Pada Pencarian File Dokumen Berbasis Teks Dengan Metode Vector Space Model

Astrid Noviriandini[1], Diah Ayu Ambarsari[2], Fahmi Aprian[3]
[1,2,3] Universitas Bina Sarana Informatika

e-mail: [1]astrid.asv@bsi.ac.id, [2]diah.das@bsi.ac.id, [3]17190532@bsi.ac.id

Abstrak - Pencarian informasi berdasarkan query oleh pengguna yang diharapkan dapat menemukan koleksi dokumen berdasarkan kebutuhan pengguna dikenal dengan *Information Retrieval* atau temu kembali informasi. Penelitian ini membahas tentang implementasi sistem temu kembali informasi untuk mencari dan menemukan dokumen teks berbahasa Indonesia dan bahasa Inggris menggunakan metode *Vector Space Model*. Tujuan penelitian ini untuk menyediakan solusi pada mesin pencari agar mampu menyediakan informasi dokumen teks pada database yang tepat menggunakan kata kunci tertentu. Hasil dari pencarian direpresentasikan dengan urutan/rangking kemiripan dokumen dengan query.

Kata Kunci: Informatin Retrieval, Temu Kembai Informasi, Vector Space Model

Abstract—*Information retrieval based on a query by the user, which is expected to find a collection of documents based on user requirements, known as Information Retrieval or information retrieval. This study discusses the implementation of information retrieval system to search and find the text documents in Bahasa Indonesia and English using the Vector Space Model. The purpose of this study to provide a solution in search engines to be able to provide information on a text document right database using specific keywords. Results of the search represented by the order/rangking similarity with the query document.*

Keywords: *Information Retrieval, Information Retrieval, Vector Space Model*

PENDAHULUAN

Perkembangan ilmu pengetahuan yang pesat dewasa ini telah mendorong permintaan akan kebutuhan informasi ilmu pengetahuan itu sendiri. Cara pemenuhan kebutuhan akan informasi ini dapat dilakukan dengan beraneka ragam. Mulai dari sekedar membaca Koran, majalah atau jurnal-jurnal tertulis hingga menggunakan teknologi digital yang terus berkembang. Akan tetapi semakin luas dan berkembangnya informasi yang beredar membuat masyarakat mengalami kesulitan untuk mendapatkan informasi yang dibutuhkannya dari media cetak. Lambat laun masyarakat mulai menggunakan teknologi digital untuk memudahkan mereka dalam mencari informasi yang dibutuhkan.

Perpustakaan di SMPN 2 Sepatan merupakan salah satu perpustakaan yang menyediakan berbagai informasi koleksi pustaka yang ada seperti tentang buku pelajaran dan buku pengetahuan lainnya. Dokumen tersebut terus bertambah setiap saat sehingga membuat dokumen semakin lama semakin banyak. Untuk mencari dokumen-dokumen tersebut dibutuhkan waktu yang relative lama, karena pencariannya dilakukan secara manual. Maka dari itu dibutuhkan sebuah *search engine* yang dapat mencari dokumen-dokumen tersebut secara lebih cepat, mudah serta menghasilkan informasi yang relevan.

Sistem temu kembali informasi *(Information Retrieval System)* merupakan suatu sistem yang menemukan *(retrieve)* informasi yang sesuai dengan kebutuhan *user* dari kumpulan informasi secara otomatis. Prinsip kerja sistem temu kembali informasi jika ada sebuah kumpulan dokumen dan seorang *user* yang memformulasikan sebuah pertanyaan *(request* atau *query)*. Jawaban dari pertanyaan tersebut adalah sekumpulan dokumen yang relevan dan membuang dokumen yang tidak relevan (Irmawati, 2017).

Tujuan yang ingin dicapai dari penelitian ini adalah Untuk membuat suatu aplikasi yang dapat membantu Pengguna Khususnya siswa-siswi dalam pencarian Dokumen informasi dengan menerapkan metode *Vektor Space Model* (VSM).

Adapun manfaat penelitian ini adalah :

a. Untuk membantu dan mempermudah siswa-siswi dalam mencari Dokumen informasi.
b. menghasilkan informasi yang lebih relevan dengan hasil ketepatan (*precision*) tinggi dan perolehan (*recall*) rendah.

METODE PENELITIAN

A. Metode Penelitian

Untuk memperoleh gambaran yang jelas mengenai penelitian ini, maka penulis perlu mendapatkan data yang akurat. Beberapa langkah yang dilakukan untuk mendapatkan data tersebut sebagai berikut :

1. Pengamatan Langsung (observasi)

Penulis mengadakan observasi secara langsung terhadap obyek yang diteliti, yaitu perpustakaan SMPN 2 Sepatan. Hal-hal yang diamati adalah kegiatan yang terjadi di lapangan, dan mencatat secara sistematis tentang hal-hal tertentu yang diamati. yaitu :

a. Penulis melakukan pengamatan pada proses pencarian data oleh petugas perpustakaan.
b. Penulis melakukan pengamatan pada hasil pencarian dan kesesuaian hasil yang diinginkan dengan fakta data yang ada.

2. Komunikasi Langsung atau Wawancara

Interview ini dilakukan dengan cara mengumpulkan data dan berkomunikasi secara langsung dengan objek peneliti agar mendapatkan informasi yang lebih akurat tentang permasalahan-permasalahan yang sebelumnya kurang jelas. Dalam hal ini penulis melakukan wawancara langsung kepada petugas pengelola perpustakaan SMPN 2 Sepatan, serta user pada perspustakaan.

Hal yang menjadi perhatian penulis adalah data yang tersedia serta kesesuaian pencarian yang diinginkan oleh user terhadap hasil pencarian yang telah dilakukan.

3. Studi Pustaka

Pengumpulan data juga dilakukan dengan cara mempelajari buku-buku yang mendukung pada penelitian ini, termasuk di dalamnya literatur tentang penulisan dan mengenai hal-hal yang mendukung pembuatan program aplikasi.

B. Information Retrieval System

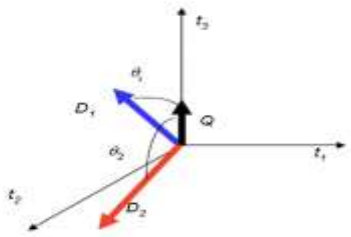
Menurut (Fadlil, 2018) sistem temu kembali adalah bagian dari ilmu komputer yang berkaitan dengan pengambilan informasi dari dokumen-dokumen berdasarkan isi dari dokumen-dokumen itu sendiri . Sedangkan menurut (Fauziah, Sulistyowati, & Asra, 2019) *Vector Space Model* (VSM) adalah metode untuk melihat tingkat kedekatan atau kesamaan *(similarity)* term dengan cara pembobotan term. Dokumen dipandang sebagai sebuah vector yang memiliki *magnitude* (jarak) dan *direction* (arah). Pada *Vector Space Model,* sebuah istilah direpresentasikan dengan sebuah dimensi dari ruang vector. Relevansi sebuah dokumen ke sebuah *query* didasarkan pada similaritas diantara vector dokumen dan vector *query*.

C. Vector Space Model

Vector Space Model (VSM) adalah metode untuk melihat tingkat kedekatan atau kesamaan *(similarity)* term dengan cara pembobotan term. Dokumen dipandang sebagai sebuah vector yang memiliki *magnitude* (jarak) dan *direction* (arah). Pada *Vector Space Model,* sebuah istilah direpresentasikan dengan sebuah dimensi dari ruang vector. Relevansi sebuah dokumen ke sebuah *query* didasarkan pada similaritas diantara vector dokumen dan vector *query* (Amin, 2011).

Menurut (Siregar, Sinaga, & Arianto, 2017) karakterisitk model ruang vector antara lain :

1. Model vector berdasarkan keyterm.
2. Model vector mendukung partial matching (sebagian sesuai) dan penentuan peringkat dokumen.
3. Prinsip dasar model vector adalah :
   a. Dokumen direpresentasikan dengan menggunakan vector keyterm.
   b. Ruang dimensi ditentukan oleh keyterms.
   c. Query direpresentasikan dengan menggunakan vector keyterm.
   d. Kesamaan dokumen keyterm dihitung berdasarkan jarak vector.
4. Model ruang vector memerlukan :
   a. Bobot keyterm untuk vector dokumen.
   b. Normalisasi keyterm untuk vector dokumen.
   c. Normalisasi keyterm untuk vector query.
   d. Perhitungan jarak untuk vector dokumen keyterm.
5. Kinerja model ruang vector :
   a. Efisien.
   b. Mudah dalam representasi.
   c. Dapat diimplementasikan pada document matching dan partial matching



Sumber : (Amin, 2011).

Gambar 1. Ilustrasi *Vector Space Model*

D. Metodologi Indexing Teks

Pada penelitian ini menggunakan metode penelitian sebagai berikut pada gambar 1.

Sumber : (Noviarindini dan Ambarsari, 2020)

Gambar 1. Metodologi Indexing Text

Pada gambar 1. data angket dari perustakaan mrupakan dataset yang digunakan dalam pencarian informasi berdasatkan kata kunci yang diberikan. Kata kunci dilakukan proses *tokenizing, filtering, stemming, Indexing* dan kemudian dilakukan proses perhitungan *TF-IDF* . Kemudian hasil yang sudah di dapatkan dilakukan perhitunagnan *Vectror Space Model* .

A. Tokenizing

Proses penghilangan tanda baca pada kalimat yang ada dalam dokumen sehingga menghasilkan kata-kata yang berdiri masing-masing.

B. Filtering

Tahap pengambilan kata-kata yang penting dari hasil tokenizing. Tahap filtering ini menggunakan daftar stoplist atau wordlist. Stoplist yaitu penyaringan terhadap kata-kata yang tidak layak untuk dijadikan sebagai pembeda atau sebagai kata kunci dalam pencarian dokumen sehingga kata-kata tersebut dapat dihilangkan dari dokumen. Sedangkan wordlist adalah daftar kata yang mungkin digunakan sebagai kata kunci dalam pencarian dokumen.

C. Stemming

Proses mengubah kata menjadi kata dasarnya dengan menghilangkan imbuhan-imbuhan pada kata dalam dokumen.

D. Indexing

Teks dokumen yang telah melalui proses tokenizing dan stemming, kemudian di-indeks ke dalam database.

HASIL DAN PEMBAHASAN

Contoh :

Query (Q) = Sejarah Perkembangan Teknologi Informasi dan Komunikasi

Dokumen 1 (D1) = Dari zaman prasejarah hingga saat ini, teknologi informasi dan komunikasi terus berkembang seiring dengan perkembangan peradaban dan kebutuhan manusia

Dokumen 2 (D2) = Sejarah perkembangan teknologi informasi dan komunikasi, khususnya komputer memang tidak bisa lepas dari sejarah perkembangan alat hitung

Dokumen 3 (D3) = Dalam rangka memenuhi kebutuhan utama manusia tersebut, manusia terus menciptakan teknologi-teknologi baru di bidang informasi dan komunikasi.

Tabel 1. Index

Index		
D1	D2	D3
Zaman	Sejarah	rangka
Prasejarah	Kembang	Penuh
Teknologi	Teknologi	Butuh
Informasi	Informasi	Utama
Komunikasi	Komunikasi	Manusia
kembang	Khusus	Tersebut
Seiring	Komputer	Manusia
kembang	Memang	Terus
adab	Bisa	Cipta
butuh	Lepas	Teknologi
manusia	sejarah	Teknologi
	Kembang	Baru
	Alat	Bidang
	hitung	Informasi
		Komunikasi

Sumber: (Noviriandini & Ambarsari, 2020)

Tabel 2. Perhitungan *tf*

Token	**Tf**				**df**
	Q	D1	D2	D3	
Zaman	0	1	0	0	1
Prasejarah	0	1	0	0	1
Teknologi	1	1	1	2	3
Informasi	1	1	1	1	3
Komunikasi	1	1	1	1	3
Kembang	1	2	2	0	2
seiring	0	1	0	0	1
adab	0	1	0	0	1
butuh	0	1	0	1	2
manusia	0	1	0	2	2
Sejarah	1	0	1	0	1
khusus	0	0	1	0	1
Komputer	0	0	1	0	1
Memang	0	0	1	0	1
Bisa	0	0	1	0	1
Lepas	0	0	1	0	1
Alat	0	0	1	0	1
hitung	0	0	1	0	1
Rangka	0	0	0	1	1
Penuh	0	0	0	1	1
Butuh	0	0	0	1	1
Utama	0	0	0	1	1
Tersebut	0	0	0	1	1
Terus	0	0	0	1	1
Cipta	0	0	0	1	1
baru	0	0	0	1	1
bidang	0	0	0	1	1

Sumber: (Noviriandini & Ambarsari, 2020)

D1, D2, D3 = Dokumen
tf = banyak kata yang dicari pada sebuah dokumen
D = total dokumen
Df = Banyak dokumen yang mengandung kata yang dicari

Tabel 3. Perhitungan *tf.idf*

idf	***Tf.idf***			
log(D/df)	Q	D1	D2	D3
0,4771	0	0,4771	0	0
0,4771	0	0,4771	0	0
0	0,4771	0,4771	0,4771	0,1761

0	0,4771	0,4771	0,4771	0,4771
0	0,4771	0,4771	0,4771	0,4771
0,1761	0,4771	0,1761	0,1761	0
0,4771	0	0,4771	0	0
0,4771	0	0,4771	0	0
0,1761	0	0,4771	0	0,4771
0,1761	0	0,4771	0	0,1761
0,4771	0,4771	0	0,4771	0
0,4771	0	0	0,4771	0
0,4771	0	0	0,4771	0
0,4771	0	0	0,4771	0
0,4771	0	0	0,4771	0
0,4771	0	0	0,4771	0
0,4771	0	0	0,4771	0
0,4771	0	0	0,4771	0
0,4771	0	0	0	0,4771
0,4771	0	0	0	0,4771
0,4771	0	0	0	0,4771
0,4771	0	0	0	0,4771
0,4771	0	0	0	0,4771
0,4771	0	0	0	0,4771
0,4771	0	0	0	0,4771
0,4771	0	0	0	0,4771
0,4771	0	0	0	0,4771

Sumber: (Noviriandini & Ambarsari, 2020)

Tabel 4. Perhitungan Jarak Q-D

Jarak Q-D			
Q	**D1**	**D2**	**D3**
0	0,2276	0	0
0	0,2276	0	0
0,2276	0,2276	0,2276	0,0310
0,2276	0,2276	0,2276	0,2276
0,2276	0,2276	0,2276	0,2276
0,2276	0,0310	0,0310	0
0	0,2276	0	0
0	0,2276	0	0
0	0,2276	0	0,2276
0	0,2276	0	0,0310
0,2276	0	0,2276	0
0	0	0,2276	0
0	0	0,2276	0
0	0	0,2276	0
0	0	0,2276	0
0	0	0,2276	0
0	0	0,2276	0
0	0	0,2276	0
0	0	0	0,2276
0	0	0	0,2276
0	0	0	0,2276
0	0	0	0,2276
0	0	0	0,2276
0	0	0	0,2276
0	0	0	0,2276
0	0	0	0,2276
0	0	0	0,2276
1,1382	2,0798	2,5351	2,7938
sqrt (Q)		sqrt (D)	
1,0669	1,4422	1,5922	1,6715

Sumber: (Noviriandini & Ambarsari, 2020)

Tabel 5. Perhitungan Dot Product

Dot Produk		
Q*D1	**Q*D2**	**Q*D3**
0	0	0
0	0	0
0,0518	0,0518	0,0071
0,0518	0,0518	0,0518
0,0518	0,0518	0,0518
0,0071	0,0071	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0,0518	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0,1625	0,2143	0,1107

Sumber: (Noviriandini & Ambarsari, 2020)

Perhitungan Similaritas

Langkah selanjutnya adalah menghitung nilai *Cosinus* sudut antara *vector Query* dengan tiap dokumen dengan rumus :

$$Cosine\ \Theta\ D_i = \frac{Q*D}{|Q|*|Di|}$$

$$D1 = \frac{0,1625}{1,0669 * 1,4422} = 0,1056$$

$$D2 = \frac{0,2143}{1,0669 * 1,5922} = 0,1262$$

$$D3 = \frac{0,1107}{1,0669 * 1,6715} = 0,0621$$

Membuat Ranking

Dari Analisa Vector Space Model diperoleh hasil untuk ketiga dokumen di atas adalah sebagai berikut.

Tabel 6. Pembuatan Ranking

D1	D2	D3
0,1056	0.1262	0.0621
Rank 2	Rank 1	Rank 3

Sumber: (Noviriandini & Ambarsari, 2020)

Hasil perhitungan *Cosine* diketahui bahwa Dokumen 2(D2) memiliki tingkat similaritas tertinggi kemudian disusul dengan D1 dan D3.

KESIMPULAN

Kesimpulan dari hasil penelitian program tugas akhir yang telah dilakukan adalah sebagai berikut:

a. *Information Retrieval System* yang dibuat dapat mencari informasi dari isi *file* dokumen yang disimpan di dalam sistem.
b. Proses peng-indeks-an dokumen di dalam aplikasi *Information Retrieval System* yang dikembangkan melalui beberapa tahapan pemrosesan teks, yaitu *tokenizing*, *filtering*, *stemming*. Sedangkan untuk proses pencariannya juga melalui beberapa tahapan proses yaitu penghitungan bobot dengan *tf-idf*, menghitung jarak tiap dokumen dan *query*, menghitung *dot product*, menghitung similaritas dan perangkingan.
c. pada penelitian ini telah berhasil mengembangkan aplikasi IR *System* dengan metode VSM untuk menemukan kembali dokumen berbahasa Indonesia dan bahasa inggris berformat *.doc, *.docx, dan *.pdf.

Adapun saran yang dapat dipertimbangkan lebih lanjut adalah Penggunaan model dari *information retrieval system* yang lainnya untuk dapat membandingkan hasil kinerja *information retrieval system* sehingga dapat ditemukan model yang paling baik dari sistem temu kembali informasi.

REFERENSI

Amin, F. (2011). IMPLEMENTASI SEARCH ENGINE (MESIN PENCARI) MENGGUNAKAN METODE VECTOR SPACE MODEL Fatkhul Amin Dosen Fakultas Teknologi Informasi Universitas Stikubank Semarang. *Dinamika Teknik*, *V*(1), 45–58.

Fadlil, A. (2018). Aplikasi Sistem Temu Kembali Angket Mahasiswa Menggunakan Application of Information Retrieval for Opinion Student. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *6*(1), 33–40. https://doi.org/10.25126/jtiik.201961184

Fauziah, S., Sulistyowati, D. N., & Asra, T. (2019). *OPTIMASI ALGORITMA VECTOR SPACE MODEL DENGAN ALGORITMA K-NEAREST NEIGHBOUR PADA PENCARIAN JUDUL ARTIKEL JURNAL*. *15*(1), 21–26.

Irmawati, I. (2017). Sistem Temu Kembali Informasi Pada Dokumen Dengan Metode Vector Space Model. *Jurnal Ilmiah FIFO*, *9*(1), 74. https://doi.org/10.22441/fifo.v9i1.1444

Siregar, R. R. A., Sinaga, F. A., & Arianto, R. (2017). Aplikasi Penentuan Dosen Penguji Skripsi Menggunakan Metode TF-IDF dan Vector Space Model. *Computatio : Journal of Computer Science and Information Systems*, *1*(2), 171. https://doi.org/10.24912/computatio.v1i2.1014