JIK. 2021;10 (2): 132 - 135

JURNAL ILMU KOMPUTER
(COMPUTER SCIENCE JOURNAL)

http://jik.htp.ac.id

INFORMATION RETRIEVAL PADA DATA JUDUL SKRIPSI BERBASIS TEXT MENGGUNAKAN VECTOR SPACE MODEL

Eka Sabna[1], Yuda Irawan[2]
[1,2]STMIK Hang Tuah Pekanbaru, Teknik Informatika
Email :
[1]es3jelita@yahoo.com, [2]yudairawan89@gmail.com

Abstract

Data storage for student thesis titles is increasing and will continue to grow. To find information from the title of the thesis will be difficult. For this reason, a search method called information retrieval was developed. Information retrieval methods have been known for a long time, one of the most widely used methods because of its ease of implementation is the Space Vector Model (SVM). The purpose of this study is to provide an explanation of the process of searching for digital documents using the Vector Space Model method. In this model, the token and indexing process is carried out so that the maximum results are found in the thesis title data using keywords, so that a search is carried out according to keywords and will be compared with the data contained in the thesis title document file, so that it can produce correct information .

Keywords: *Information Retrieval, SVM, title, thesis, search .*

Abstrak

Penyimpanan data judul skripsi mahasiswa semakin banyak dan akan terus bertambah. Untuk mencari informasi dari judul skripsi tersebut akan menjadi sulit. Untuk itu dikembangkanlah metode pencarian yang disebut dengan temu-kembali informasi (information retrieval). Metode-metode temu-kembali informasi sudah dikenal sejak lama, salah satu dari metode tersebut yang paling banyak digunakan karena kemudahan implementasinya adalah Space Vector Model (SVM). Tujuan penelitian ini adalah memberikan paparan tentang proses pencarian dokumen digital dengan metode Vektor Space Model. Pada model ini dilakukan dengan proses token dan indexing sehingga ditemukan hasil dari maksimal terdapat dalam data judul skripsi menggunakan kata kunci, sehingga di lakukan pencarian sesuai dengan kata kunci dan akan dibandingkan dengan data yang terdapat pada file dokumen judul skripsi, sehingga dapat menghasilkan informasi yang benar.

Keywords: Information Retrieval, SVM, judul, skripsi, pencarian.

https://doi.org/10.33060/JIK/2021/Vol10.Iss2.230

E-ISSN : 2579 - 3918 | P-ISSN : 2302 - 710X

PENDAHULUAN

Informasi tumbuh dengan sangat pesat dalam berbagai basis content seperti teks, image, video, visual, audio dan sebagainya Begitu pula informasi tentang judul-judul skripsi mahasiswa semakin hari semakin bertambah. Data-data judul skripsi ini tersimpan di simpan dalam bentuk file di Perpustakaan Perguruan Tinggi. Kebutuhan saat ini adalah mahasiswa atau Dosen memerlukan Informasi judul skripsi mahasiswa . Dalam proses pencarian judul skripsi tersebut menggunaka proses pencarian manual , sehingga memerlukan waktu yang lama dalam pencarian nya. Oleh karena itu, perpustakaan perguruan tinggi membutuhkan sistem temu kembali informasi (information retrieval).

Pengambilan informasi menunjukkan proses pencarian informasi yang diperlukan (Zhou, Liu, & Liu, 2012) Information retrieval (IR) umumnya berkaitan dengan pencarian dan pengambilan informasi berbasis pengetahuan (Sharma & Patel, 2013) Salah satu penerapan prinsip relevansi yang sejak dahulu digunakan dalam pengembanngan sistem (Lestari, 2016) Information Retrieval System menemukan informasi yang biasanya dalam bentuk dokumen dari sebuah data yang tidak terstruktur dalam bentuk teks untuk memenuhi kebutuhan informasi dari koleksi data yang sangat besar umumnya tersimpan dalam database computer (Amin & Purwatiningtyas, 2015) . Information Retrieval dapat mencari judul-judul skripsi tersebut secara lebih cepat dan mudah serta menghasilkan informasi yang relevan. Penelitian ini menggunakan vektor space model yang merupakan salah satu metode informasi retrival yang bertujuan untuk mempermudah dalam proses temu kembali informasi pada dokumen berbasis text digital. Tujuan dari penelitian ini adalah menggunakan vektor space model yang merupakan salah satu metode information retrival yang bertujuan untuk mempermudah dalam proses temu kembali informasi pada dokumen berbasis text digital.

METODE PENELITIAN

Pengumpulan data dilakukan dengan cara mempelajari buku dan jurnal yang mendukung pada penelitian ini, termasuk di dalamnya literatur tentang penulisan dan mengenai hal-hal yang mendukung implementasi sistem temu kembali pada aplikasi. Metadata koleksi dokumen skripsi yang digunakan antara tahun 2021. Data tersebut tidak berurutkan, Dari hasil penelusuran informasi, dihasilkan 6 dokumen skripsi yang sering dilihat, pada tahap selanjutnya penelitian ini mengambil dari enam dokumen skripsi sebagai sempel pada penelitian ini. Metode yang digunakan dalam penelitian ini menggunakan metode Vector Space Model. Berikut langkah metode vector space model :

Proses perhitungan VSM melalui beberapa tahapan , berikut tahapan nya :

1.Perhitungan term frequency (tf) menggunakan persamaan:

$$tf = tf_{ij}$$

Dengan tf adalah term frequency, dan tfi,j adalah banyaknya kemunculan term ti dalam dokumen dj, Term frequency (tf) dihitung dengan menghitung banyaknya kemunculan term ti dalam dokumen dj.

2.Perhitungan Inverse Document Frequency (idf), menggunakan persamaan:

$$idf_i = log\frac{N}{df_i}$$

Dengan idfi adalah inverse document frequency, N adalah jumlah dokumen yang terambil oleh sistem, dan dfi adalah banyaknya dokumen dalam koleksi dimana term ti muncul di dalamnya, maka perhitungan idfi digunakan untuk mengetahui banyaknya term yang dicari (dfi) yang muncul dalam dokumen lain yang ada pada database (korpus).

3.Perhitungan term frequency Inverse Document Frequency (tfidf), menggunakan persamaan:

$$W_{ij} = tf_{ij} \cdot log\frac{N}{df_i}$$

Dengan Wij adalah bobot dokumen, N adalah jumlah dokumen yang terambil oleh sistem, tfi,j adalah banyaknya kemunculan term ti pada dokumen dj, dan dfi adalah banyaknya dokumen dalam koleksi dimana term ti muncul didalamnya.

4.Bobot dokumen (Wij) dihitung untuk didapatkannya suatu bobot hasil perkalian atau kombinasi antara term frequency (tfi,j) dan Inverse Document Frequency (dfi).

$$|d_j| = \sqrt{\sum_{i=1}^{t}(W_{ij})^2}$$

Dengan |dj| adalah jarak dokumen, dan Wij adalah bobot dokumen ke-i, maka jarak dokumen (|dj|) dihitung untuk didapatkan jarak dokumen dari bobot dokumen (Wij) yang terambil oleh sistem.

5.Jarak dokumen bisa dihitung dengan persamaan akar jumlah kuadrat dari dokumen.

$$|q| = \sqrt{\sum_{j=1}^{t}(W_{iq})^2}$$

Dengan |q| adalah jarak kueri, dan Wiq adalah bobot kueri dokumen ke-i, maka jarak kueri (|q|) dihitung untuk didapatkan jarak kueri dari bobot kueri dokumen (Wiq) yang terambil oleh sistem. Jarak kueri bisa dihitung dengan persamaan akar jumlah kuadrat dari query.

6.Perhitungan pengukuran similaritas query document (inner product), menggunakan persamaan

$$Sim(q, d_j) = \sum_{i=1}^{t} W_{iq}.W_{ij}$$

Dengan Wij adalah bobot term dalam dokumen, Wiq adalah bobot kueri, dan sim (q, dj) adalah similaritas antara kueri dan dokumen. Similaritas antara kueri dan dokumen atau inner product / Sim (q, dj) digunakan untuk mendapatkan bobot dengan didasarkan pada bobot term dalam dokumen (Wij) dan bobot query (Wiq) atau dengan cara menjumlah bobot q dikalikan dengan bobot dokumen.

HASIL & PEMBAHASAN

Sesuai dengan tahapan metode Vector Space Model maka diimplementasikan dalam data berikut sebagai sampel data sebanyak 4 judul skripsi .

D1: Penerapan Data Mining untuk pengelompokan penyakit
D2: Program berbasis web untuk pelayanan Desa
D3: Analisa Data Mining untuk promosi
D4: Klustering untuk data pasien

Proses Vector Space Model untuk sampel di atas adalah :

1.Perhitungan term frequency (tf) menggunakan persamaan:

terms(t)	tf			
	d1	d2	d3	d4
penerapan	1	0	0	0
data	1	0	1	1
mining	1	0	1	0
untuk	1	1	1	1
pengelompokan	1	0	0	0
penyakit	1	0	0	0
program	0	1	0	0
berbasis	0	1	0	0
web	0	1	0	0
pelayanan	0	1	0	0
desa	0	1	0	0
analisa	0	0	1	0
promosi	0	0	1	0
klustering	0	0	0	1
pasien	0	0	0	1

2. Perhitungan Inverse Document Frequency (idf),

terms(t)	tf				dfi	logN/dfi
	d1	d2	d3	d4		
penerapan	1	0	0	0	1	0,60206
data	1	0	1	1	3	0,124939
mining	1	0	1	0	2	0,30103
untuk	1	1	1	1	4	0
pengelompokan	1	0	0	0	1	0,60206
penyakit	1	0	0	0	1	0,60206
program	0	1	0	0	1	0,60206
berbasis	0	1	0	0	1	0,60206
web	0	1	0	0	1	0,60206
pelayanan	0	1	0	0	1	0,60206
desa	0	1	0	0	1	0,60206
analisa	0	0	1	0	1	0,60206
promosi	0	0	1	0	1	0,60206
klustering	0	0	0	1	1	0,60206
pasien	0	0	0	1	1	0,60206

3. Perhitungan Term Frequency Invers Document Frequency (tfidf)

terms(t)	tf				dfi	logN/dfi	Wij			
	d1	d2	d3	d4			d1	d2	d3	d4
penerapan	1	0	0	0	1	0,60206	0,60206	0	0	0
data	1	0	1	1	3	0,124939	0,124939	0	0,124939	0,124939
mining	1	0	1	0	2	0,30103	0,30103	0	0,30103	0
untuk	1	1	1	1	4	0	0	0	0	0
pengelompokan	1	0	0	0	1	0,60206	0,60206	0	0	0
penyakit	1	0	0	0	1	0,60206	0,60206	0	0	0
program	0	1	0	0	1	0,60206	0	0,60206	0	0
berbasis	0	1	0	0	1	0,60206	0	0,60206	0	0
web	0	1	0	0	1	0,60206	0	0,60206	0	0
pelayanan	0	1	0	0	1	0,60206	0	0,60206	0	0
desa	0	1	0	0	1	0,60206	0	0,60206	0	0
analisa	0	0	1	0	1	0,60206	0	0	0,60206	0
promosi	0	0	1	0	1	0,60206	0	0	0,60206	0
klustering	0	0	0	1	1	0,60206	0	0	0	0,60206
pasien	0	0	0	1	1	0,60206	0	0	0	0,60206

4. Perhitungan Jarak Dokumen

terms(t)	logN/dfi	Wij			
		d1	d2	d3	d4
penerapan	0,60206	0,602059991	0	0	0
data	0,124939	0,124938737	0	0,124939	0,124939
mining	0,30103	0,301029996	0	0,30103	0
untuk	0	0	0	0	0
pengelompokan	0,60206	0,602059991	0	0	0
penyakit	0,60206	0,602059991	0	0	0
program	0,60206	0	0,60206	0	0
berbasis	0,60206	0	0,60206	0	0
web	0,60206	0	0,60206	0	0
pelayanan	0,60206	0	0,60206	0	0
desa	0,60206	0	0,60206	0	0
analisa	0,60206	0	0	0,60206	0
promosi	0,60206	0	0	0,60206	0
klustering	0,60206	0	0	0	0,60206
pasien	0,60206	0	0	0	0,60206
		1,193657446	1,812381	0,831181	0,740562
		1,092546313	1,346247	0,911691	0,860559

5. Perhitungan Jarak Query

terms(t)	tf				dfi	logN/dfi	Query	
	d1	d2	d3	d4			tf	w
penerapan	1	0	0	0	1	0,60206	0	0
data	1	0	1	1	3	0,124939	1	0,124939
mining	1	0	1	0	2	0,30103	1	0,30103
untuk	1	1	1	1	4	0	1	0
pengelompokan	1	0	0	0	1	0,60206	0	0
penyakit	1	0	0	0	1	0,60206	0	0
program	0	1	0	0	1	0,60206	0	0
berbasis	0	1	0	0	1	0,60206	0	0
web	0	1	0	0	1	0,60206	0	0
pelayanan	0	1	0	0	1	0,60206	0	0
desa	0	1	0	0	1	0,60206	0	0
analisa	0	0	1	0	1	0,60206	1	0,60206
promosi	0	0	1	0	1	0,60206	1	0,60206
klustering	0	0	0	1	1	0,60206	0	0
pasien	0	0	0	1	1	0,60206	0	0
								0,831181
								0,911691

6. Perhitungan pengukuran similaritas query document

terms(t)	Wij				Query					
	d1	d2	d3	d4	tf	w	d1	d2	d3	d4
penerapan	0,60206	0	0	0	0	0	0	0	0	0
data	0,124939	0	0,124939	0,124939	1	0,124939	0,01561	0	0,01561	0,01561
mining	0,30103	0	0,30103	0	1	0,30103	0,090619	0	0,090619	0
untuk	0	0	0	0	1	0	0	0	0	0
pengelompokan	0,60206	0	0	0	0	0	0	0	0	0
penyakit	0,60206	0	0	0	0	0	0	0	0	0
program	0	0,60206	0	0	0	0	0	0	0	0
berbasis	0	0,60206	0	0	0	0	0	0	0	0
web	0	0,60206	0	0	0	0	0	0	0	0
pelayanan	0	0,60206	0	0	0	0	0	0	0	0
desa	0	0,60206	0	0	0	0	0	0	0	0
analisa	0	0	0,60206	0	1	0,60206	0	0	0,362476	0
promosi	0	0	0,60206	0	1	0,60206	0	0	0,362476	0
klustering	0	0	0	0,60206	0	0	0	0	0	0
pasien	0	0	0	0,60206	0	0	0	0	0	0
	2,232149	3,0103	1,630089	1,329059		1,630089	0,106229	0	0,831181	0,01561
	1,494038	1,735022	1,276749	1,152848		1,276749	1,907512	2,215188	1,630089	1,471898
							0,05569	0	0,509899	0,010605

Hasil dari langkah 6 ini kemudian di ranking, hasil nya adalah dokumen 3 (0,509899).

KESIMPULAN

Setelah menyelesaikan tahapan-tahapan penelitian sistem temu balik informasi, dapat diambil beberapa kesimpulan, yaitu :

1. Berdasarkan penelitian yang telah dilakukan, model ruang vektor yang digunakan memberikan hasil yang baik.
2. Berdasarkan penelitian yang telah dilakukan, penerapan stemming pada query dan dokumen meningkatkan hasil pencarian terhadap dokumen yang relevan.
3. Hasil pencarian suatu informasi dari sejumlah dokumen dengan metode SVM lebih cepat dan proses ini juga melakukan perangkingan dokumen dari kata kunci yang digunakan sebagai query pencarian.

DAFTAR PUSTAKA

Amin, F. 2012. Sistem Temu Kembali Informasi dengan Metode Vector Space Model, 02, 78– 83

Amin, F., & Purwatiningtyas. 2015. Rancang Bangun Information Retrieval System (IRS) Bahasa Jawa Ngoko pada Palintangan Penjebar Semangad dengan Metode Vector Space Model (VSM). Jurnal Teknologi Informasi DINAMIK, 20(1), 25–35.

Irmawati. 2017. Sistem Temu Kembali Informasi Pada Dokumen Dengan Metode Vector Space Model. Jurnal Ilmiah Fifo, IX(1), 74–80.

Jaya, Hendra., 2007. "Perbandingan Performansi Word Indexing dan Phrase Indexing dalam Sistem Temu Balik Informasi dengan Menggunakan Model Probabilistik." Skripsi Terpublikasi. Bandung : Institut Teknologi Bandung

Linarta, A., & Nurhadi, N. (2019). Aplikasi Bel Sekolah Otomatis Berbasis Arduino Dilengkapi Dengan Output Suara. Informatika, 10(2), 1-7.

Ramadhany., Taufik. 2008. "Implementasi Kombinasi Model Ruang Vektor dan Model Probabilistik Pada Sistem Temu Balik Informasi." Skripsi Terpublikasi. Bandung : Institut Teknologi Bandung.

Rahmalisa, U., Febriani, A., & Irawan, Y. (2021). Detector Leakage Gas Lpg Based On Telegram Notification Using Wemos D1 and Mq-6 Sensor. Journal of Robotics and Control (JRC), 2(4), 287-291.

Singh, J. N. 2012. Analysis of Vector Space Model in Information Retrieval, 14–18.

Sjaeful Afandi; Firman Ardiansya; Blasius Soedarsono. 2015. Pengembangan Sistem Temu Kembali Informasi Digital Fulltext Artikel Jurnal Di Pdii – Lipi. Baca: Jurnal Dokumentasi Dan Informasi, 36(1), 65–76. https://doi.org/http://dx.doi.org/10.14203/j.baca.v36i1.203

Sabna, E. (2019). Analisis Data Mahasiswa Dengan Algoritma K-Mean Untuk Mendukung Strategi Promosi Stmik Hang Tuah Pekanbaru. Jurnal Ilmu Komputer, 8(1), 113-117

Sabna, E. (2020). PENERAPAN TEXT MINING UNTUK PENGELOMPOKAN PENELITIAN DOSEN. Jurnal Ilmu Komputer, 9(2), 161-164.

Sabna, E., & Muhardi, M. (2016). Penerapan Data Mining Untuk Memprediksi Prestasi Akademik Mahasiswa Berdasarkan Dosen, Motivasi, Kedisiplinan, Ekonomi, Dan Hasil Belajar. Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer Dan Teknologi Informasi, 2(2), 41-44.

Susilo, J., Febriani, A., Rahmalisa, U., & Irawan, Y. (2021). Car Parking Distance Controller Using Ultrasonic Sensors Based on Arduino Uno. Journal of Robotics and Control (JRC), 2(5), 353-356.