PEMANFAATAN SISTEM TEMU KEMBALI INFORMASI DALAM PENCARIAN DOKUMEN MENGGUNAKAN VECTOR SPACE MODEL

Taufik Ihsan Maulana[1)], Ainur Rafiq Abdillah[2)]
[1,2]Pendidikan Teknik Informatika dan Komputer, Fakultas Teknik, Universitas Negeri Jakarta / Jl. Rawamangun Muka Raya No. 11, RT. 11/RW. 14, Kel. Rawamangun, Kec. Pulogadung, Kota Jakarta Timur, Prov. DKI Jakarta, Kode Pos 13320, (021) 4898486
email: TaufikIhsanMaulana_1512620003@mhs.unj.ac.id, AinurRafiqAbdillah_1512620079@mhs.unj.ac.id.

Abstract

With the development of a very modern technological era, many documents in the form of files are used. But the document files used are always increasing every day and finding information from the contents of the files becomes more difficult. For this reason, it is necessary to implement the scientific method of document retrieval, namely information retrieval. One method of information retrieval is the Vector Space Model. In the VSM (Vector Space Model) method, a document search will be carried out with an indexing process, namely separating the contents of the text from the documents into indexing terms. Indexing term will be used for document search. The stages in the formation of indexing terms include parsing, text preprocessing, weighting and document similarity measurement.

Keywords: *document search, information retrieval, vector space model.*

Abstrak

Dengan berkembangnya zaman teknologi yang sangat modern hingga banyaknya dokumen dalam bentuk file yang digunakan. Tetapi file-file dokumen yang digunakan selalu bertambah setiap harinya dan mencari informasi dari isi file-file menjadi lebih sulit. Untuk itu perlu diimplementasikan metode ilmu pencarian dokumen yaitu temu kembali informasi *(information retrieval)*. Salah satu metode dalam temu kembali informasi adalah *Vector Space Model*. Pada metode VSM *(Vector Space Model)* akan dilakukan pencarian dokumen dengan proses *indexing* yaitu memisahkan antara isi teks dengan dokumen-dokumen menjadi *indexing term*. *Indexing term* akan digunakan untuk pencarian dokumen. Tahapan dalam pembentukan indexing term antara lain *parsing, text preprocessing, weighting* (pembobotan) dan pengukuran kesamaan dokumen *(similarity measure)*.

Kata Kunci: pencarian dokumen, *information retrieval, vector space model*.

1. PENDAHULUAN

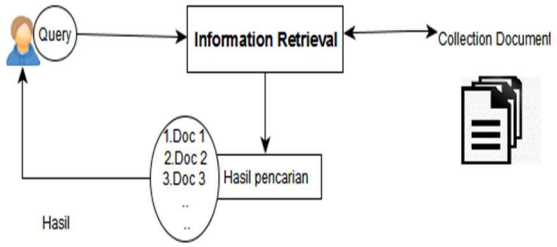
Dengan berkembangnya zaman teknologi yang sangat modern hingga saat ini yang berdampak terhadap kehidupan sehari-hari. Salah satu yang berubah adalah cara yang menggunakan data sebagai informasi pada era saat ini. Dengan semakin banyaknya jumlah dokumen yang beredar menimbulkan sebuah permasalahan dalam melakukan pencarian data yang diinginkan dengan cepat dan akurat secara *online* atau internet ataupun *offline* dengan sistem penyimpanan pada komputer. Saat ini beberapa *e-library* menggunakan algoritma untuk pencariannya seperti algoritma *boolean search* namun belum cukup kuat untuk proses pencarian karena tidak dapat mengenali dokumen yang cukup relevan.

Salah satu metode dalam temu balik informasi *(Information Retrieval)* untuk mendapatkan dokumen yang relevan berdasarkan query adalah *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 vektor yang memiliki *magnitude* (jarak) dan *direction* (arah). Metode *Vector Space Model* ini mengidentifikasi suatu dokumen dan query dalam sebuah bentuk vektor. Relevansi sebuah dokumen ke sebuah *query* didasarkan pada similaritas antara vektor dokumen dan vektor

query. (Yates, 1999). Dalam mengidentifikasi sebuah vektor dibutuhkan adanya bobot term dari sebuah dokumen atau *query*. Term dapat berupa kata, frasa, ataupun hasil dari indexing lain dalam sebuah dokumen sebagai gambaran dari isi setiap dokumen tersebut.

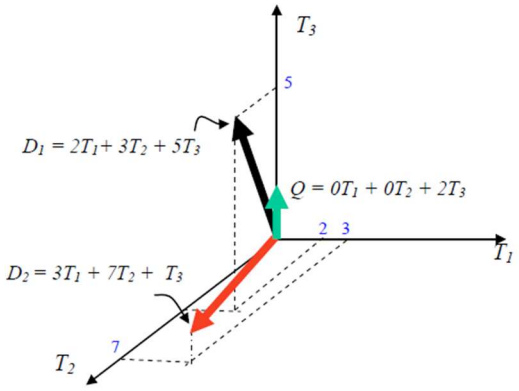
Dalam menentukan *term* pada sebuah dokumen ataupun *query* diperlukannya beberapa tahapan antara lain *filtering, stemming* dan *tokenizing*. Setiap *term* memiliki tingkat suatu kepentingan yang berbeda dalam dokumen untuk diperlukan *term weighting* (pembobotan *term*). Metode pembobotan umumnya digunakan dalam *Vector Space Model* yaitu *Term Frequency* dan *Inverse Document Frequency* (TF-IDF). Metode TF-IDF adalah suatu cara yang bertujuan untuk memberikan bobot hubungan suatu kata (term) terhadap suatu dokumen. Hasil dari pembobotan dengan metode TF-IDF nantinya dokumen dan query akan diidentifikasi dalam sebuah ruang vektor kemudian akan dicari dari tingkat kedekatannya dengan menggunakan pengukuran cosine similarity sehingga mendapatkan dokumen yang relevan dengan suatu *query*.

Information Retrieval System (Sistem temu kembali informasi) adalah suatu sistem yang menemukan (retrieve) informasi yang sesuai dengan kebutuhan pengguna *(user)* dari kumpulan informasi secara otomatis(Salton, 1989).



Gambar 1. Sistem Temu Kembali Informasi

Pada *Information Retrieval System* diperoleh metode yang digunakan dalam prosess pencarian salah satunya adalah dengan merepresentasikan proses pencarian adalah menggunakan model ruang vektor. Model ruang vektor dibuat dengan pemikiran bahwa bahwa isi dari dokumen ditentukan oleh kata-kata yang digunakan dalam dokumen tersebut. Model ini menentukan kemiripan *(similarity)* antara dokumen dengan query dengan cara mengidentifikasi dokumen dan query masing-masing ke dalam bentuk vektor.



Gambar 2. Ilustrasi Dokumen dan Query dalam ruang vektor

Text preprocessing atau disebut proses indexing adalah suatu tahapan awal pada proses mengidentifikasi koleksi dokumen kedalam bentuk tertentu untuk memudahkan dan mempercepat proses pencarian dan penemuan data dokumen kembali yang relevan. Diperoleh beberapa tahapan pada fase tersebut antara lain:

a. *Case Folding* dan *Tokenization*, *case folding* dilakukan untuk mengganti huruf besar dari setiap kata diganti menjadi huruf kecil dan menghilangkan karakter selain huruf seperti angka dan tanda baca (delimiter). Sedangkan, *tokenization* adalah membagi semua kalimat pada isi dokumen menjadi kata per kata.

b. *Filtering*, dilakukan dengan metode *stopwords* adalah menghapus term yang tidak memiliki arti atau tidak relevan. Proses ini dilakukan pada saat proses tokenisasi. Proses *Filtering* menggunakan daftar *stopwords* yang digunakan oleh Tala (2003), yang merupakan stopword bahasa Indonesia yang berisi kata-kata seperti; ada, yang, ke, kepada, dan lain sebagainya.

c. *Stemming*, digunakan untuk mengubah term yang masih melekat dalam term tersebut awalan, sisipan, dan akhiran. Proses *stemming* dilakukan dengan cara menghilangkan semua imbuhan (afiks) baik yang terdiri dari awalan (prefiks), sisipan (infiks), akhiran (sufiks) dan konfiks (kombinasi dari awalan dan akhiran) pada kata turunan. Stemming digunakan untuk

mengganti bentuk dari suatu kata menjadi kata dasar dari kata tersebut yang sesuai dengan struktur morfologi bahasa Indonesia yang benar (Tala, 2003).

Pada metode TF-IDF perhitungan bobot term t dalam sebuah dokumen dilakukan dengan mengalikan nilai *Term Frequency* dengan *Inverse Document Frequency*.

$W = tf^{ij} \times idf^{j}$

$W = tf^{ij} \times \log(D/df^{i})$

Keterangan:

W = bobot term tj terhadap dokumen di

tf^{ij} = jumlah kemunculan term tj dalam dokumen di

D = jumlah semua dokumen yang ada

df^{i} = jumlah dokumen yang mengandung term tj (minimal ada satu kata yaitu term tj)

Vector Space Model dan pembobotan TF-IDF digunakan untuk mengidentifikasi suatu nilai angka numerik pada dokumen sehingga dapat dihitung kedekatan antar dokumen. Kemiripan antar dokumen dihitung dari suatu fungsi ukuran kemiripan *(similarity measure)*. Ukuran ini memungkinkan urutan dokumen sesuai dengan kemiripan relevansinya terhadap query.

$$Sim(\vec{d_j}, \vec{q}) == \frac{\sum_{i=1}^{t}(W_{ij} \times W_{iq})}{\sqrt{\sum_{i=1}^{t}(W_{ij})^2 \times \sum_{i=1}^{t}(W_{iq})^2}}$$

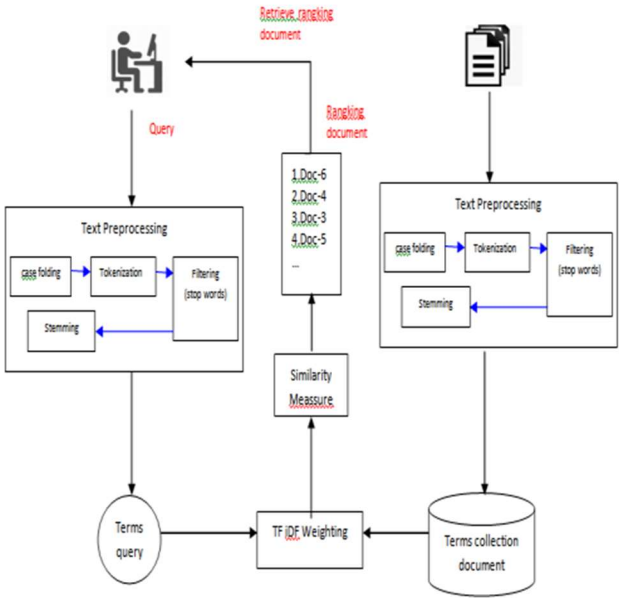
Keterangan:

D_j	:Dokumen ke *j*
Q	:*query* user
$\sum_{i=1}^{t} W_{ij}$	:jumlah bobot kata *i* pada dokumen *j*
$\sum_{i=1}^{t} W_{iq}$	:jumlah bobot kata *i* pada *query*

2. RANCANGAN SISTEM

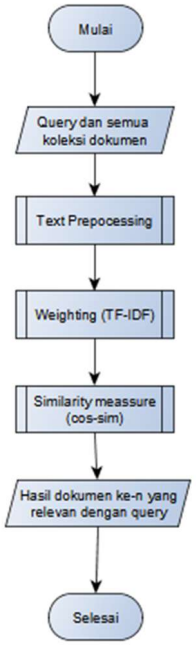
A. Analisa Data

Proses pencarian dalam sistem temu kembali informasi pada penelitian ini diterapkan pada sistem penyimpanan dokumen .**pdf* berupa jurnal dengan topik Teknologi Informasi. Dari masing-masing dokumen dilakukan indexing berdasarkan semua isi dokumen.

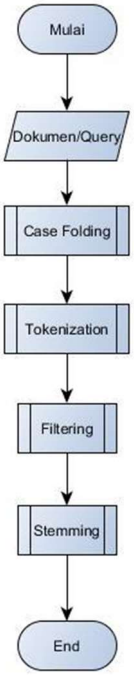


Gambar 3. Arsitektur Sistem Temu Kembali Informasi

Pada sistem temu kembali dokumen dan *query user* akan dilakukan *text preprocessing* dan pembobotan *(Weighting)* untuk memperoleh nilai kemiripannya. *Text preprocessing* pada dokumen disimpan dalam *database* berupa *index term*. Sedangkan, pada perancangan proses dilakukan untuk menjelaskan proses yang dikerjakan sistem dalam melakukan pencarian dokumen terhadap *query* dari *user*.



Gambar 4. Flowchart Perancangan Sistem



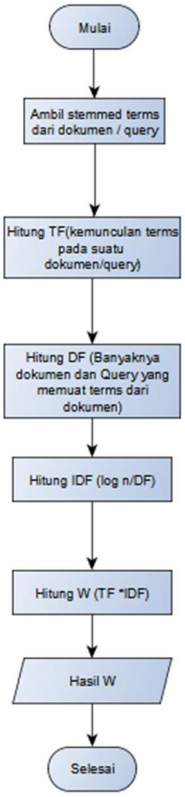
Gambar 5. Flowchart Prosedur Text Preprocessing

B. Text Preprocessing

Pada tahap *text preprocessing* dilakukan untuk mencapai *index terms* dari dokumen maupun query yang digunakan untuk pembobotan. Langkah-langkah dari tahapan adalah *case folding, tokenization, filtering, stemming.*

C. Pembobotan (Weighting)

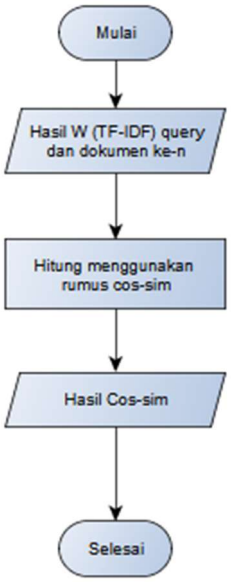
Pada tahap pembobotan *(weighting)* dilakukan dengan metode TF-IDF. Dengan cara menghitung frekuensi kemunculan suatu kata di dalam seluruh dokumen.



Gambar 6. Flowchart Pembobotan (Weighting)

D. Similarity Measure

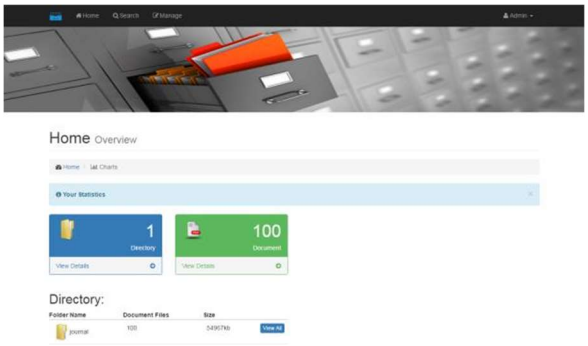
Ukuran kemiripan dokumen *(similarity measure)* menggunakan algoritma *Cosine similarity*. *Cosine similarity* adalah suatu algoritma digunakan untuk mengukur kedekatan antara dua vektor yaitu vektor query dan dokumen.



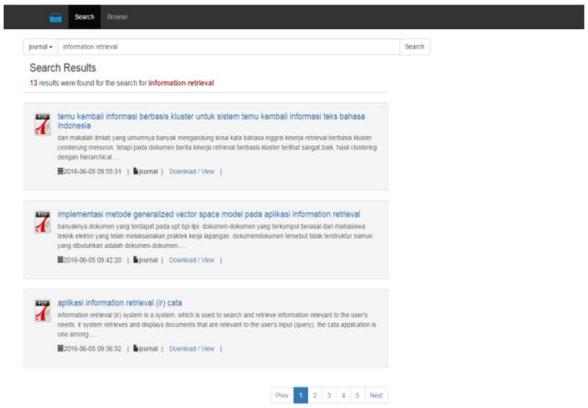
Gambar 7. Flowchart Similarity Measure

3. IMPLEMENTASI DAN PEMBAHASAN

Dalam implementasi pembuatan program menggunakan bahasa pemrograman interpreter PHP version 5.4. Selain itu, dibutuhkan library pendukung lain untuk implementasi VSM.



Gambar 8. Interface Sistem Temu Kembali Informasi



Gambar 9. Hasil Pencarian keyword "modal investasi"

Hasil Uji Recall dan Precision

Untuk mengetahui hasil uji *recall* dan *precision* sistem akan dilakukan uji terhadap 6 *query* acak dengan data dokumen pada sistem yaitu 100 dokumen dan setelah dilakukan proses *indexing* menghasilkan 13103 *terms* pada *database*. Berikut ini merupakan daftar query yang dilakukan dalam pengujian.

Query	Recall	Precision
Metode fuzzy (Q1)	1	0,5
Fuzzy (Q2)	1	0,61
Information Retrieval (Q3)	1	0,53
Sistem Informasi Geografi (Q4)	1	0,125
Implementasi Mikrokontroler Arduino (Q5)	1	1

4. PENUTUP

Kesimpulan

1. Pengembangan teknologi modern telah mengubah cara kita mengakses dan menggunakan informasi. Informasi sekarang dapat dicari dengan cepat dan akurat menggunakan sistem temu kembali informasi.
2. Vector Space Model (VSM) adalah metode yang digunakan dalam temu kembali informasi untuk mencari dokumen yang relevan berdasarkan query. Dokumen dan query direpresentasikan sebagai vektor dan relevansinya diukur menggunakan *cosine similarity*.
3. Proses *text preprocessing*, seperti *case folding*, *tokenization*, *filtering*, dan *stemming*, digunakan untuk mengidentifikasi term dalam dokumen query, sehingga memudahkan dalam pembobotan term.
4. Metode pembobotan yang umum digunakan adalah Term Frequency-Inverse Document Frequency (TF-IDF), yang memberikan bobot pada kata berdasarkan frekuensi kemunculannya dalam dokumen dan kebalikan jumlah dokumen yang mengandung kata tersebut.
5. Sistem temu kembali informasi menggunakan model ruang vektor dan pembobotan TF-IDF untuk mengidentifikasi nilai numerik yang menggambarkan kedekatan antara dokumen.
6. Metode perhitungan kemiripan dokumen yang umum digunakan adalah *cosine similarity*, yang mengukur sudut kosinus antara vektor dokumen dan vektor query.
7. Implementasi sistem temu kembali informasi pada jurnal-jurnal teknologi informasi dilakukan dengan melakukan proses *indexing* dan *pembobotan* menggunakan TF-IDF.
8. Pengujian sistem menggunakan *recall* dan *precision* untuk mengukur kinerja sistem dalam menemukan dokumen yang relevan dengan query yang diberikan.

Saran

1. Melakukan pengembangan lebih lanjut pada algoritma pencarian untuk meningkatkan akurasi dan kecepatan temu kembali informasi.
2. Menerapkan teknik pengolahan bahasa alami yang lebih canggih untuk meningkatkan pemahaman sistem terhadap query pengguna.

3. Memperluas sumber data yang diindeks untuk mencakup berbagai jenis dokumen dan sumber informasi lainnya.
4. Memperhatikan penggunaan *stopwords* dalam proses filtering, sehingga dapat menghindari penghapusan kata-kata penting dalam pencarian.
5. Menggabungkan metode pencarian *fuzzy* atau *fuzzy logic* untuk meningkatkan ketepatan pencarian dengan mempertimbangkan kemungkinan kata-kata yang mirip atau terkait.
6. Menerapkan teknik pengindeksan yang lebih efisien dan scalable untuk mengatasi jumlah dokumen yang besar dan mempercepat proses pencarian.
7. Melakukan evaluasi secara teratur terhadap kinerja sistem menggunakan metrik seperti *recall*, *precision*, dan metrik lainnya, untuk terus meningkatkan kualitas dan efektivitas sistem temu kembali informasi.

5. REFERENSI

[1] Baeza, Yates., Ribeiro, Neto. Modern Information Retrieval, Harlow, Addison-Wesley, 1999.
[2] Fatkhul, A. "Sistem Temu Kembali Informasi dengan Vector Space Model". Jurnal Fakultas Teknologi Informasi. Universitas Stikubank. Semarang. 2012.
[3] Handojo. A, dkk. "Document Searching Engine Using Term Similarity Vector Space Model on English and Indonesian Document," *Jurnal Informatics Engineering Department Faculty of Industrial Technology*, Petra Christian University Surabaya, 2014.
[4] Salton, G., "Automatic Text Processing, The Transformation, Analysis, and Retrieval of information by computer". Addison - Wesley Publishing Company, Inc. USA. 2003.
[5] Tala, F.Z., *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*. Institut for logic, Language and Computation Universiteit van Amsterdam The Netherlands, 2003.
[6] Yates, R.B, *Modern Information Retrieval, Addison Wesley*-Pearson international edition, Boston, USA. 1999.