Aplikasi *Information Retrieval* (IR) CATA Dengan Metode *Generalized Vector Space Model*

Hendra Bunyamin, Chathalea Puspa Negara

Jurusan Teknik Informatika
Fakultas Teknologi Informasi, Universitas Kristen Maranatha
. Prof. Drg. Suria Sumantri No. 65 Bandung 40164
Email: hendra.bunyamin@eng.maranatha.edu

Abstract

Information retrieval (IR) system is a system, which is used to search and retrieve information relevant to the user's needs. IR system retrieves and displays documents that are relevant to the user's input (query). The Cata application is one among Information Retrieval Systems. This application has features such as to add and change a document in document collections. There is also a feature to search the information in document collections by using Generalized Vector Space Model algorithm. Before applying this algorithm, the query which is entered by the user will be process first. The processing of words includes the disposal of stopwords and stemming. This application performs searching the documents which are relevant to the queries, based on the similarities. The searching result which is ordered based on the highest of the similarity value.

Keywords : Information Retrieval system, Generalized Vector Space Model

I. Pendahuluan

Pada saat kita melakukan pencarian melalui *search engine* (*google.com*, dan *yahoo.com*), kita bisa mendapatkan beberapa hasil, yang berupa dokumen-dokumen yang sama atau hampir sesuai dengan kata atau *query* yang kita masukkan. Demikian pula jika kita melakukan pencarian dalam aplikasi sistem informasi, seperti halnya sistem pencarian dalam perpustakaan. Aplikasi yang dibuat adalah aplikasi yang menggunakan algoritma IR (*Information Retrieval*) dengan metode sistem *Generalized Vector Space*. *Information Retrieval* (IR) merupakan suatu sistem yang membantu pengguna dalam mencari informasi di dalam kumpulan dokumen. Beberapa sistem yang menggunakan IR *system* adalah aplikasi *search engine*, seperti *google.com* dan aplikasi sistem informasi, seperti perpustakaan.

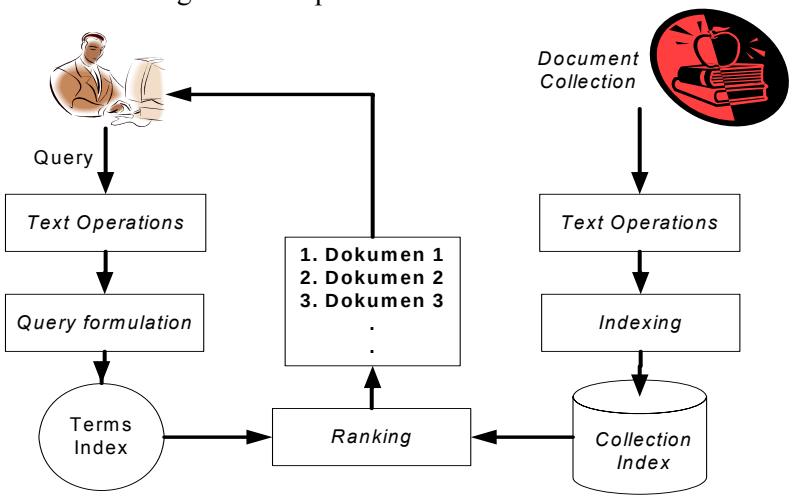
Algoritma *Generalized Vector Space Model* yang dibahas menggunakan konsep ruang vektor. Masukan dari pengguna dan kumpulan dokumen diterjemahkan menjadi vektor-vektor. Kemudian vektor-vektor tersebut dikenakan operasi perkalian titik dan hasilnya menjadi acuan dalam menentukan relevansi masukan pengguna (*query*) terhadap kumpulan dokumen.

II. *Information Retrieval System* dan *Generalized Vector Space Model*

Sistem *information retrieval* (IR) *system* adalah system yang digunakan untuk menemukan kembali (*retrieve*) informasi-informasi yang relevan terhadap kebutuhan pengguna dari suatu kumpulan informasi secara otomatis [Bunyamin, 2005].

Sistem *IR* terutama berhubungan dengan pencarian informasi yang isinya tidak memiliki struktur. Demikian pula ekspresi kebutuhan pengguna yang disebut *query*, juga tidak memiliki struktur. Hal ini yang membedakan sistem *IR* dengan sistem basis data. Dokumen adalah contoh informasi yang tidak terstruktur. Isi dari suatu dokumen sangat tergantung pada pembuat dokumen tersebut.

Sebagai suatu sistem, sistem *IR* memiliki beberapa bagian yang membangun sistem secara keseluruhan. Gambaran bagian-bagian yang terdapat pada suatu sistem *IR* digambarkan pada Gambar 1



Gambar 1 Bagian – bagian *information* sistem *retrieval* (IR)

Dari gambar 1, terlihat bahwa terdapat dua proses operasi dalam sistem *IR*. Proses pertama dimulai dari koleksi dokumen dan proses kedua dimulai dari *query* pengguna. Proses pertama yaitu pemrosesan terhadap koleksi dokumen menjadi basis data indeks tidak ada ketergantungan dengan proses kedua. Sedangkan proses kedua tergantung dari keberadaan basis data indeks yang dihasilkan pada proses pertama.

Bagian-bagian dari sistem IR menurut gambar 1 meliputi :

(1) *Text Operations* (operasi terhadap teks) yang meliputi pemilihan kata-kata dalam query maupun dokumen (*term selection*) dalam pentransformasian dokumen atau query menjadi *term index* (indeks dari kata-kata).
(2) *Query formulation* (formulasi terhadap *query*) yaitu memberi bobot pada indeks kata-kata *query*.
(3) *Ranking* (perangkingan), mencari dokumen-dokumen yang relevan terhadap *query* dan mengurutkan dokumen tersebut berdasarkan kesesuaiannya dengan *query*.

(4) *Indexing* (pengindeksan), membangun basis data indeks dari koleksi dokumen. Dilakukan terlebih dahulu sebelum pencarian dokumen dilakukan.

Sistem IR menerima query dari pengguna, kemudian melakukan perangkingan terhadap dokumen pada koleksi berdasarkan kesesuaiannya dengan *query*. Hasil perangkingan yang diberikan kepada pengguna merupakan dokumen yang menurut sistem relevan dengan *query*. Namun relevansi dokumen terhadap suatu *query* merupakan penilaian pengguna yang subjektif dan dipengaruhi banyak faktor seperti topik, pewaktuan, sumber informasi maupun tujuan pengguna.

Salah satu model sistem IR adalah model vektor. Beberapa karakteristik dalam sistem *IR* adalah :

1. Model vektor berdasarkan *index term*
2. Model vektor mendukung *partial matching* dan penentuan peringkat dokumen
3. Prinsip dasar vektor model adalah sebagai berikut :
   (a) dokumen direpresentasikan dengan menggunakan vektor *index term*
   (b) Ruang dimensi ditentukan oleh *index term*
   (c) *Query* direpresentasikan dengan menggunakan vektor *index term*
   (d) Kesamaan *document-query* dihitung berdasarkan hasil kali titik (*cross product*) antara vektor – vektor tersebut
4. Model vektor memerlukan :
   (a) Bobot *index term* untuk vektor dokumen
   (b) Bobot *index term* untuk *query*
   (c) Perhitungan *cross product* untuk vektor *document-query*
5. Kinerja
   1. Efisien
   2. Mudah dalam representasi
   3. Dapat diimplementasikan pada *document-matching*

Ada beberapa langkah atau proses untuk mendapatkan hasil dari *query* yang dimasukkan, yang disebut algoritma *Generalized Vector Space Model* [Baeza, 1999]:

1. Membuang kata depan dan kata penghubung.
2. Menggunakan *stemmer* pada kumpulan dokumen dan *query*, yaitu aplikasi yang digunakan untuk menghilangkan imbuhan (awalan, akhiran). Contoh : keagungan → agung, keabadian → abadi.
3. Menentukan *minterm* untuk menentukan kemungkinan pola frekuensi kata. Panjang *minterm* ini didasarkan pada banyak kata yang diinput pada *query*. Kemudian diubah menjadi vektor ortogonal sesuai dengan pola *minterm* yang muncul. Kemungkinan pola yang akan muncul adalah :

$$m_1 = (0,0,0,...)$$
$$m_2 = (1,0,0,...)$$
$$\cdots$$
$$m_{2t} = (1,1,1,...)$$

4. Menghitung banyaknya frekuensi atau kemunculan kata dalam kumpulan dokumen yang sesuai dengan *query*
5. Menghitung *index term* yang dapat dinyatakan dengan :

$$\vec{k}_i = \frac{\sum_{\forall r, g_i(m_r)=1} c_{i,r} \vec{m}_r}{\sqrt{\sum_{\forall r, g_i(m_r)=1} c_{i,r}^{\,2}}}$$

Dimana :

$\vec{k}_i$: *index term* ke-i

$\vec{m}_r$: vektor ortogonal sesuai pola *minterm* yang terpakai

$c_{i,r}$: faktor korelasi antara *index term* i dengan *minterm* r

Sedangkan faktor korelasi sebagai berikut :

$$c_{i,r} = \sum_{d_j | g_i(\vec{d}_j) = g_i(m_r)} w_{i,j}$$

Dimana :

$c_{i,r}$: faktor korelasi antara *index term* i dengan *minterm* r

$w_{i,j}$: berat *index term* i pada dokumen j

$g_i(m_r)$: bobot *index term* k_i dalam minterm m_r

6. Mengubah dokumen dan *query* menjadi vektor

$$\vec{d_j} = \sum_{i=1}^{n} w_{ij} \times \vec{k_i} \qquad \vec{q} = \sum_{i=1}^{n} q_i \times \vec{k_i}$$

Dimana :

$\vec{d_j}$: vektor dokumen ke-j

$\vec{q}$: vektor *query*

$w_{i,j}$: berat *index term* i pada dokumen j

q_i : berat *index term* pada *query* i

$\vec{k_i}$: *index term*

n : jumlah *index term*

7. Mengurutkan dokumen berdasarkan similaritas, dengan menghitung perkalian vektor

$$sim(\vec{dj}, \vec{q}) = \frac{\vec{dj} \bullet \vec{q}}{\left|\vec{dj}\right| \left|\vec{q}\right|}$$

Dimana :
$\overrightarrow{d_j}$: vektor dokumen j
$\vec{q}$: vektor *query*

III. Aplikasi IR dengan *Generalized Vector Space Model*

Sebagai contoh, terdapat sebuah *query* (Q), dan 3 buah dokumen yaitu dokumen 1 (D1), dokumen 2 (D2), dan dokumen 3 (D3) sebagai berikut:

Q : penyelesaian konflik Aceh
Judul D1 : Gus Dur Tak Mungkin Dijatuhkan
Judul D2 : Bondan: Bukan Saya Nggak Doyan Duit
Judul D3 : AS Dukung Kesepakatan GAM-RI
Contoh tersebut dapat diproses sesuai dengan langkah – langkah yang telah dijelaskan pada algoritma sebelumnya. Langkah – langkah tersebut antara lain :

1. Membuang kata depan dan kata penghubung. Namun dalam query tidak terdapat kata depan maupun kata penghubung. Maka proses ini tidak dilakukan.
2. Menghilangkan imbuhan (awalan, akhiran).
   Q : selesai konflik Aceh
3. Menentukan *minterms* berdasarkan banyak kata yang diinput pada *query* dan kemungkinan pola yang muncul. Berdasarkan *query* tersebut, *minterm* yang dipakai adalah m_8,m_6,m_7.
4. Menghitung frekuensi kata dalam koleksi dokumen yang sesuai dengan *query* dan menentukan vektor orthogonal sesuai dengan minterm yang dipakai

	selesai	konflik	aceh	vektor orthogonal
D_1	2	3	1	$\overrightarrow{m_1}$
D_2	1	0	4	$\overrightarrow{m_2}$
D_3	0	3	4	$\overrightarrow{m_3}$
q	1	1	1	

5. Menghitung korelasi setiap *terms*
   $C_{1,1} = 2$ $C_{2,1} = 3$ $C_{3,1} = 1$
   $C_{1,2} = 1$ $C_{2,2} = 0$ $C_{3,2} = 4$
   $C_{1,3} = 0$ $C_{2,3} = 3$ $C_{3,3} = 4$

6. Menghitung *index terms*

$$\vec{k}_1 = \frac{c_{1,1}\overrightarrow{m}_1 + c_{1,2}\overrightarrow{m}_2 + c_{1,3}\overrightarrow{m}_3}{\sqrt{{c_{1,1}}^2 + {c_{1,2}}^2 + {c_{1,3}}^2}} = \frac{2\overrightarrow{m}_1 + 1\overrightarrow{m}_2 + 0\overrightarrow{m}_3}{\sqrt{2^2 + 1^2 + 0^2}} = \frac{2\overrightarrow{m}_1 + \overrightarrow{m}_2}{\sqrt{5}}$$

$$\vec{k}_2 = \frac{c_{2,1}\vec{m}_1 + c_{2,2}\vec{m}_2 + c_{2,3}\vec{m}_3}{\sqrt{{c_{2,1}}^2 + {c_{2,2}}^2 + {c_{2,3}}^2}} = \frac{3\vec{m}_1 + 0\vec{m}_2 + 3\vec{m}_3}{\sqrt{3^2 + 0^2 + 3^2}} = \frac{3\vec{m}_1 + 3\vec{m}_3}{\sqrt{18}}$$

$$\vec{k}_3 = \frac{c_{3,1}\vec{m}_1 + c_{3,2}\vec{m}_2 + c_{3,3}\vec{m}_3}{\sqrt{{c_{3,1}}^2 + {c_{3,2}}^2 + {c_{3,3}}^2}} = \frac{1\vec{m}_1 + 4\vec{m}_2 + 4\vec{m}_3}{\sqrt{1^2 + 4^2 + 4^2}} = \frac{\vec{m}_1 + 4\vec{m}_2 + 4\vec{m}_3}{\sqrt{33}}$$

7. Mengubah dokumen dan query kedalam bentuk vektor

$$\vec{d}_1 = 2\vec{k}_1 + 3\vec{k}_2 + \vec{k}_3$$
$$= 2\left(\frac{2}{\sqrt{5}}\vec{m}_1 + \frac{1}{\sqrt{5}}\vec{m}_2\right) + 3\left(\frac{3}{\sqrt{18}}\vec{m}_1 + \frac{3}{\sqrt{18}}\vec{m}_3\right) + \left(\frac{1}{\sqrt{33}}\vec{m}_1 + \frac{4}{\sqrt{33}}\vec{m}_2 + \frac{4}{\sqrt{33}}\vec{m}_3\right)$$
$$= 1{,}7889\vec{m}_1 + 0{,}8944\vec{m}_2 + 2{,}1213\vec{m}_1 + 2{,}1213\vec{m}_3 + 0{,}1741\vec{m}_1 + 0{,}6963\vec{m}_2 + 0{,}6963\vec{m}_3$$
$$= 4{,}0843\vec{m}_1 + 1{,}5907\vec{m}_2 + 2{,}8176\vec{m}_3$$

$$\vec{d}_2 = \vec{k}_1 + 4\vec{k}_3$$
$$= \left(\frac{2}{\sqrt{5}}\vec{m}_1 + \frac{1}{\sqrt{5}}\vec{m}_2\right) + 4\left(\frac{1}{\sqrt{33}}\vec{m}_1 + \frac{4}{\sqrt{33}}\vec{m}_2 + \frac{4}{\sqrt{33}}\vec{m}_3\right)$$
$$= 0{,}8944\vec{m}_1 + 0{,}4472\vec{m}_2 + 0{,}6963\vec{m}_1 + 2{,}7852\vec{m}_2 + 2{,}7852\vec{m}_3$$
$$= 1{,}5907\vec{m}_1 + 3{,}2324\vec{m}_2 + 2{,}7852\vec{m}_3$$

$$\vec{d}_3 = 3\vec{k}_2 + 4\vec{k}_3$$
$$= 3\left(\frac{3}{\sqrt{18}}\vec{m}_1 + \frac{3}{\sqrt{18}}\vec{m}_3\right) + 4\left(\frac{1}{\sqrt{33}}\vec{m}_1 + \frac{4}{\sqrt{33}}\vec{m}_2 + \frac{4}{\sqrt{33}}\vec{m}_3\right)$$
$$= 2{,}1213\vec{m}_1 + 2{,}1213\vec{m}_3 + 0{,}6963\vec{m}_1 + 2{,}7852\vec{m}_2 + 2{,}7852\vec{m}_3$$
$$= 2{,}8176\vec{m}_1 + 2{,}7852\vec{m}_2 + 4{,}9065\vec{m}_3$$

$$\vec{q} = \vec{k}_1 + \vec{k}_2 + \vec{k}_3$$
$$= \left(\frac{2}{\sqrt{5}}\vec{m}_1 + \frac{1}{\sqrt{5}}\vec{m}_2\right) + \left(\frac{3}{\sqrt{18}}\vec{m}_1 + \frac{3}{\sqrt{18}}\vec{m}_3\right) + \left(\frac{1}{\sqrt{33}}\vec{m}_1 + \frac{4}{\sqrt{33}}\vec{m}_2 + \frac{4}{\sqrt{33}}\vec{m}_3\right)$$
$$= 0{,}8944\ \vec{m}_1 + 0{,}4472\ \vec{m}_2 + 0{,}7071\ \vec{m}_1 + 0{,}7071\ \vec{m}_3 + 0{,}1741\ \vec{m}_1 + 0{,}6963\ \vec{m}_2 + 0{,}6963\ \vec{m}_3$$
$$= 1{,}7756\ \vec{m}_1 + 1{,}1435\ \vec{m}_2 + 1{,}4034\ \vec{m}_3$$

8. Menghitung similaritas dokumen dan meranking

$$sim\left(\vec{d}_1, \vec{q}\right) = \frac{(4{,}0843 \times 1{,}7756) + (1{,}5907 \times 1{,}1435) + (2{,}8176 \times 1{,}4034)}{\left(\sqrt{4{,}0843^2 + 1{,}5907^2 + 2{,}8176^2}\right)\left(\sqrt{1{,}7756^2 + 1{,}1435^2 + 1{,}4034^2}\right)}$$
$$= \frac{(7{,}2521) + (1{,}8189) + (3{,}9542)}{\left(\sqrt{27{,}1507}\right)\left(\sqrt{6{,}4299}\right)} = \frac{13{,}0252}{13{,}2127} = 0{,}9858$$

$$sim\left(\vec{d}_2, \vec{q}\right) = \frac{(1{,}5907 \times 1{,}7756) + (3{,}2324 \times 1{,}1435) + (2{,}7852 \times 1{,}4034)}{\left(\sqrt{1{,}5907^2 + 3{,}2324^2 + 2{,}7852^2}\right)\left(\sqrt{1{,}7756^2 + 1{,}1435^2 + 1{,}4034^2}\right)}$$
$$= \frac{(2{,}8244) + (3{,}6962) + (3{,}9087)}{\left(\sqrt{20{,}7361}\right)\left(\sqrt{6{,}4299}\right)} = \frac{10{,}4293}{11{,}5469} = 0{,}9032$$

$$sim\left(\overrightarrow{d_3}, \vec{q}\right) = \frac{(2{,}8176 \times 1{,}7756) + (2{,}7852 \times 1{,}1435) + (4{,}9065 \times 1{,}4034)}{\left(\sqrt{2{,}8176^{\,2} + 2{,}7852^{\,2} + 4{,}9065^{\,2}}\right)\left(\sqrt{1{,}7756^{\,2} + 1{,}1435^{\,2} + 1{,}4034^{\,2}}\right)}$$

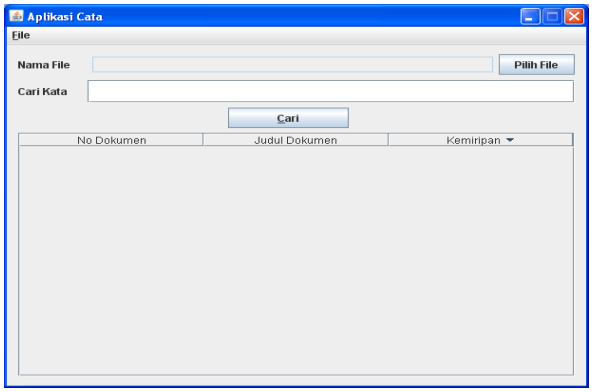
$$= \frac{(5{,}0029) + (3{,}1849) + (6{,}8858)}{\left(\sqrt{39{,}76995}\right)\left(\sqrt{6{,}4299}\right)} = \frac{15{,}0736}{15{,}9911} = 0{,}9426$$

Dari hasil similaritas pada butir (8) diatas, dapat diambil ranking yang dihasilkan adalah dokumen 1, dokumen 3, dokumen 2. Yang berarti dokumen 1 adalah dokumen yang paling relevan dengan *query*.

IV. Aplikasi Cata

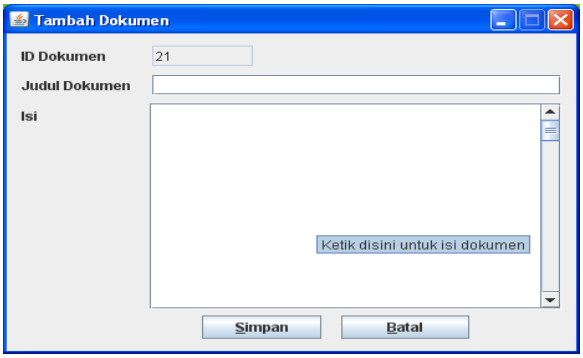
Aplikasi ini merupakan salah satu contoh IR system yang menerapkan metode vektor, yaitu *Generalized Vector Space Model*, yang selanjutnya dinamakan aplikasi Cata. Aplikasi ini berfungsi untuk mengolah *query*, serta berfungsi untuk melakukan pencocokan antara *query* dengan kata yang ada pada kumpulan dokumen. Aplikasi ini menampilkan dokumen yang relevan dengan *query* dan mengurutkannya berdasarkan kemiripan antara *query* dan dokumen yang paling tinggi. Dalam aplikasi ini, tidak ada kategori akses untuk *user*, sehingga semua *user* dapat menggunakan aplikasi ini. Aplikasi ini dibutuhkan untuk memudahkan *user* dalam mencari informasi dalam kumpulan dokumen. Terdapat pula fitur untuk menambah, mengubah dan menghapus dokumen dalam koleksi dokumen.

Berikut adalah gambar antarmuka aplikasi setiap fitur yang ada :



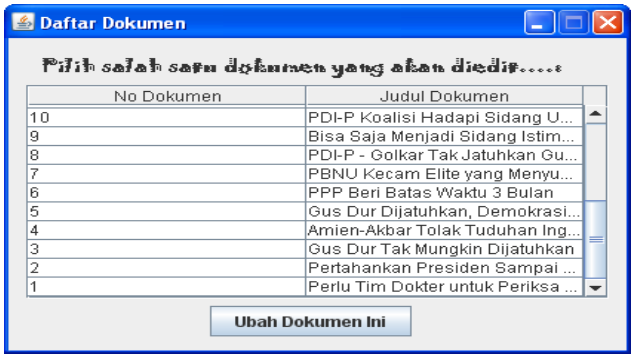
Gambar 2 *Form* Utama

Dalam *form* ini pada gambar 2, *user* dapat mencari kata yang ada pada koleksi dokumen. Jika menekan tombol cari, maka sistem akan melakukan proses pencarian. Hasilnya akan ditampilkan dalam tabel dan diurutkan berdasarkan nilai kemiripan yang paling tinggi. Dalam *form* ini, *user* dapat memilih fitur tambah, ubah, atau hapus dokumen pada menu '*File*'.



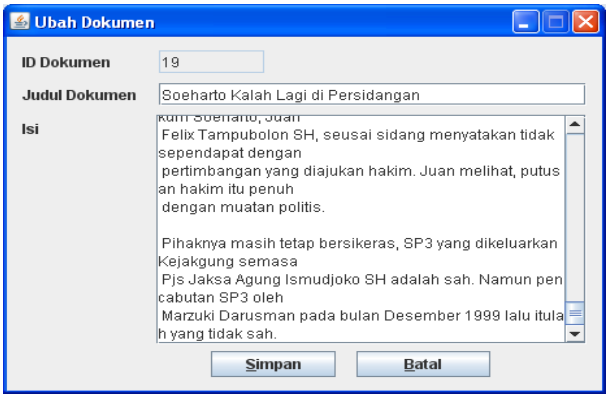
Gambar 3 *Form* Tambah Dokumen

Dalam *form* ini pada gambar 3, *user* dapat menambah dokumen. Jika menekan tombol simpan, maka sistem akan melakukan proses penambahan dokumen kedalam koleksi dokumen.



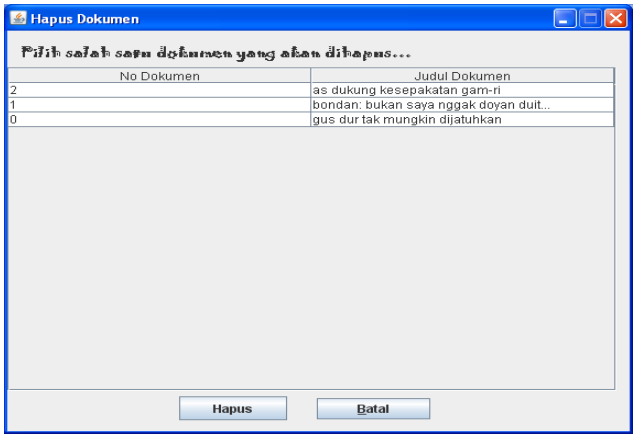
Gambar 4 *Form* Daftar Dokumen

Dalam *form* pada gambar 4, terdapat sebuah tabel yang digunakan untuk menampilkan indeks dan judul dokumen yang akan diubah oleh *user*. Tombol 'Ubah Dokumen Ini' berfungsi untuk membuka *Form* Ubah Dokumen dengan mengirimkan indeks dan judul dokumen yang dipilih.



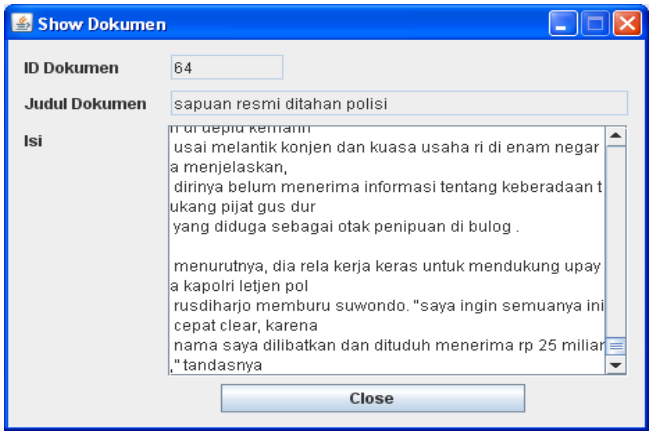
Gambar 5 *Form* Ubah Dokumen

Dalam *form* pada gambar 5, *user* dapat mengubah dokumen yang telah dipilih dalam fitur Daftar Dokumen. Jika menekan tombol simpan, maka sistem akan melakukan proses perubahan dokumen kedalam koleksi dokumen.



Gambar 6 *Form* Hapus Dokumen

Dalam *form* pada gambar 6, *user* dapat menghapus dokumen yang telah dipilih. Jika menekan tombol hapus, maka sistem akan menghapus dokumen yang telah dipilih dari koleksi dokumen.



Gambar 7 *Form* Detail Dokumen

Dalam *form* pada gambar 7, *user* dapat membaca isi dokumen dari hasil pencarian kata. Isi dokumen yang ditampilkan, tidak dapat diubah atau ditambah.

Sebelum menggunakan semua fitur yang ada, *user* diwajibkan untuk memilih sebuah file koleksi dokumen yang mempunyai extension file (.all). File ini berisi kumpulan dokumen – dokumen, yang mempunyai format sama dengan format XML. Berikut adalah contoh atau cuplikan dari format koleksi dokumen yang dipakai :

```
<documentFile>
<document>
<name>0</name>
<title>Gus Dur Tak Mungkin Dijatuhkan</title>
<content>Gus Dur Tak Mungkin Dijatuhkan
              SEMARANG- Legitimasi sosial yang begitu
kuat akan tetap mengukuhkan kedudukan Abdurrahman
Wahid sebagai Presiden RI. Walau ''digoyang'' berbagai
masalah berat sekalipun,legitimasi itu sulit untuk
digoyahkan, termasuk pada Sidang Umum Agustus
mendatang.
</content>
</document>
```

Gambar 8 Format Koleksi Dokumen

V. Kesimpulan

Hasil akhir dari seluruh proses perancangan, serta proses implementasi telah menghasilkan aplikasi yang setelah dilakukan pengujian, dinilai dapat digunakan dengan baik. Pembuatan aplikasi ini sudah mencapai tujuan utama dari aplikasi, yaitu mempermudah *user* untuk mencari informasi dalam koleksi dokumen. Kesimpulan mengenai metode yang digunakan, yaitu *Generalized Vector Space Model* adalah

1. Menggunakan bobot index term
2. Adanya vektor dokumen dan *query*
3. Perhitungan *cross product* menentukan kesamaan *query* dan dokumen

Daftar Pustaka

[Ano07] Anonymous. *IR Models*. http://www.cs.ui.ac.id/WebKuliah/TKSI/MIK/IRModels.doc. 12 Desember 2007.

[Bae99] Baeza, Ricardo, B. Ribeiro. 1999. *Modern Information Retrieval*. ACM Press. United States of America. 1999.

[Bun05] Bunyamin, Hendra. 2005. *Information Retrieval System dengan Menggunakan Metode **Latent Semantic Indexing***, Tesis S2 Magister Teknik Informatika.

[Won85] Wong, S., W Ziarko, P. Wong. 1985. Generalized Vector Space Model in Information Retrieval. http://140.122.185.120/PastCourses/2003F-InformationRetrievalandExtraction/Present_2003F/2003F_GeneralizedVectorSpaceModelInInformationRetrieval.pdf. 12 Desember 2007.

[Won87] Wong, S., W. Ziarko, V. Raghavan. 1987. On Modeling of Information Retrieval Concepts in Vector Spaces. http://delivery.acm.org/10.1145/30000/22957/p299-wong.pdf. 28 Januari 2008.