SISTEM INFORMATION RETRIEVAL PENCARIAN KESAMAAN AYAT TERJEMAHAN AL QURAN BERBAHASA INDONESIA DENGAN QUERY EXPANSION DARI TAFSIRNYA

Broto Poernomo T.P.[1] dan Ir. Gunawan[2]
[1]Teknik Informtika Sekolah Tinggi Manajemen Informatika dan Komputer Asia
[2]Teknik Informatika Sekolah Tinggi Teknik Surabaya
papung@gmail.com dan gunawan@stts.edu

ABSTRAK

Fokus dari penelitian ini adalah melakukan pencarian ayat-ayat dalam Al Quran yang memiliki kesamaan yang didasarkan pada teks terjemahan bahasa Indonesia dan dua teks tafsir. Yaitu teks tafsir Jalalain dan tafsir dari Departement Agama Republik Indonesia. Teks tafsir ini berfungsi sebagai query expansion terhadap query user. Penambahan teks tafsir ini juga dilakukan pada semua ayat pada koleksi dokumen. Dalam penelitian ini sistem dibagi menjadi dua sub sistem yaitu sub sistem untuk pre-processing dataset dan sub sistem untuk pencarian ayatnya. Output yang diharapkan adalah daftar ayat-ayat yang relevan yang memiliki kesamaan topik dengan query user. Sedangkan perankingannya didasarkan pada nilai kesamaan yang didapatkan.

Dataset yang digunakan pada penelitian ini bersumber dari program Holy Quran. Selanjutnya dataset tersebut di pre-processing yang meliputi tokenizing, filtering, pembentukan Inverted Index dan stemming menggunakan algoritma Nazief Adirani. Sedangkan proses pencariannya menggunakan pemodelan berbasis Vector Space Model.

Untuk nilai recall pada query yang mengalami ekspansi dan tidak didapatkan nilai yang sama yaitu 100%. Sedangkan untuk nilai precision pada query yang tidak diekspansi didapatkan nilai precision 27%. Dan pada query yang diekspansi nilai precision dapat meningkat mencapai 75%. Selain itu dengan query expansion ini dapat menemukan ayat-ayat yang memiliki kesamaan topik.

Kata Kunci: *Information Retrieval System, Algoritma Nazief Adriani, Vector Space Model, Ekspansi Query*

ABSTRACT

This study is to search the verses in the Qur'an Indonesian translation and two texts of tafseer. That is tafseer Jalalain and tafseer from Departement Religion of Republic Indonesia. This text serves as a text query expansion to a user query. In this study, the system is divided into two sub-systems, sub-systems for pre-processing datasets and sub-systems to search the verse. The expected output is a list of verses that is relevant topic to the user query. While the ranking is based on the similarity values obtained.

The dataset used in this study comes from the Holy Quran program. Furthermore, these datasets in a pre-processing which include tokenizing, filtering,

information of Inverted Index and stemming with Adirani Nazief algorithm. The search process uses the modeling-based Vector Space Model.

The recall value with expanded query and none have same value of 100%. The value of precision in non-expanded query is 27%. And the expanded query precision value can be increased to 75%. In addition to query expansion can find verses that have a common topic.

Keywords: Information Retrieval System, Nazief Adriani algorithm, Vector Space Model, Query Expansion

I. PENDAHULUAN

Sumber yang paling utama dan pertama untuk dipelajari oleh seorang muslim adalah Al Quran, karena Al Quran berfungsi sebagai petunjuk dan pedoman hidup umat Islam dalam kehidupan sehari-hari. Di dalam Al Quran mengandung hal-hal yang berhubungan dengan keimanan, ilmu pengetahuan, hukum, muamalah (peraturan-peraturan yang mengatur tingkah laku dan tata cara hidup manusia), kisah-kisah umat sebelumnya, ibadah dan tauhid (pengesaan Allah).

Dalam menerangkan hal-hal di atas, Al Quran mengemukakannya dalam dua cara. Pertama, Al Quran menerangkannya secara detil dan terperinci. Kedua, Al Quran menerangkan secara umum dan garis besarnya saja. Baik ayat yang menjelaskan secara detil maupun global, di dalamnya masih banyak mengandung makna yang luas yang memerlukan penjelasan. Sehingga untuk bisa memahami setiap ayat dalam Al Quran diperlukan tafsirnya (penjelasan terperinci). Banyak kitab-kitab tafsir yang telah ditulis oleh para ulama diantarannya adalah Tafsir Jalalain, Tafsir Al Azhar, Tafsir Ath Thabrani, Tafsir As Sa'di, tafsir yang diterbitkan Departemen Agama Republik Indonesia dan kitab tafsir yang lainnya.

Pada penelitian ini akan dibangun sistem pencarian kesamaan ayat-ayat dalam Al Quran berdasarkan teks terjemahan bahasa Indonesianya. Untuk menentukan kedekatan kesamaannya menggunakan teks tafsir dari dua kitab tafsir. Yaitu kitab tafsir Jalalain dan kitab tafsir yang diterbitkan oleh Departemen Agama Republik Indonesia. Untuk pemodelannya berbasis Vector Space Model (VSM) dengan melakukan pre-proccesing terlebih dahulu pada teks-teks yang akan digunakan.

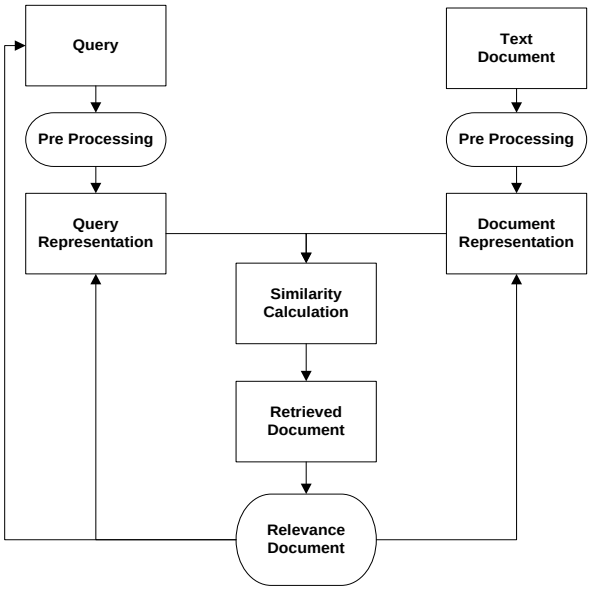
II. KAJIAN TEORI

1. Information Retrieval

Information Retrieval merupakan bagian dari computer science yang berhubungan dengan pengambilan informasi dari dokumen-dokumen yang didasarkan pada isi dan konteks dari dokumen-dokumen itu sendiri. Berdasarkan referensi dijelaskan bahwa Information Retrieval merupakan suatu pencarian informasi yang didasarkan pada suatu query yang diharapkan dapat memenuhi keinginan user dari kumpulan dokumen yang ada. Pada Sub ini dibahas tentang definisi Information Retrieval System, Arsitektur Information Retrieval System, fungsi-fungsi dari definisi Information Retrieval System dan pembahasan tentang Recall dan Precision.

2. Arsitektur IR

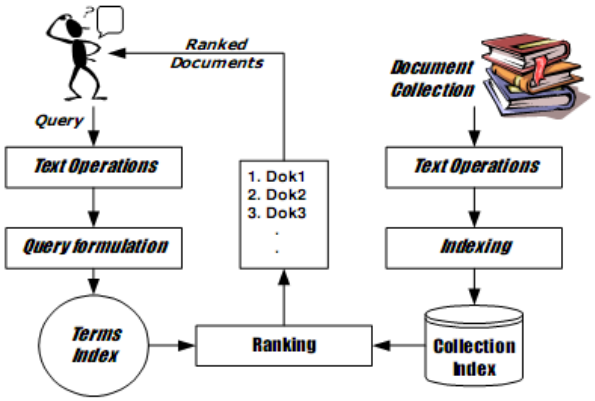
Ada dua pekerjaan utama yang dikerjakan oleh Information Retrieval, yaitu melakukan pre-processing terhadap database dan kemudian menerapkan metode tertentu untuk menghitung kedekatan (relevansi atau similarity) antara dokumen di dalam database yang telah di-preprocess dengan query pengguna. Sehingga arsitektur dari Information Retrieval. Arsitektur IR dapat dilihat pada gambar 1.



Gambar 1. Arsitektur IR

3. Query Expansion

Seringkali pengguna mengalami kesulitan dalam membentuk query yang ditujukan untuk menemukembalikan informasi hal ini dikarenakan mereka tidak mengetahui detail dari konstruksi koleksi dan lingkungan information retrieval. Padahal, jumlah dokumen relevan yang diperoleh dipengaruhi oleh jumlah kata kunci dalam query. Hal ini akan mengakibatkan hasil pencarian yang dilakukan pun menjadi kurang optimal.



Gambar 2. Arsitektur Information Retrieval System dengan Query Formulation

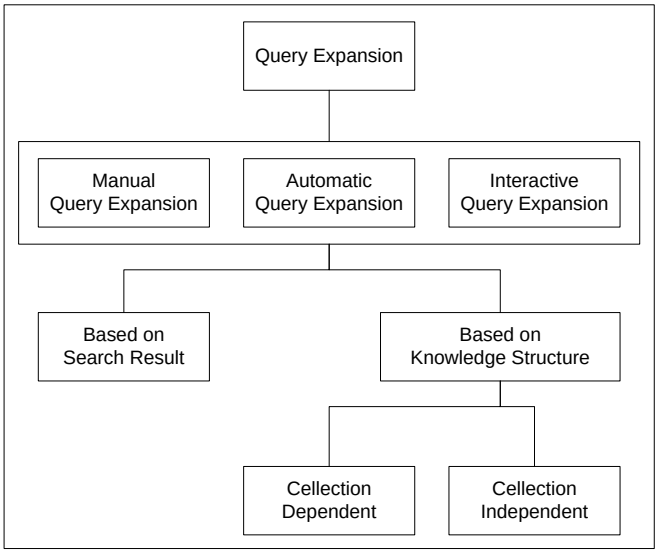
Menyusun ulang query (*query reformulation*) yang dimasukkan oleh user adalah hal yang sering dilakukan dalam information retrieval. Hal ini dilakukan untuk mengatasi ketidaksesuaian antara query yang dimasukkan oleh user dengan informasi yang ingin didapatkannya.

Query reformulation yang sering dipakai adalah dengan Query Expansion, yaitu dengan memanjangkan query yang dimasukkan oleh user dengan menambahkan beberapa term kedalamnya. Ekspansi query merupakan salah satu teknik yang dapat digunakan dalam membantu pengguna dalam memberikan query yang baik. Ekspansi query dapat berperan sebagai penghubung karena adanya vocabulary gaps antara query dan dokumen. Query yang dimasukkan oleh user pada umumnya pendek dan query expansion dapat melengkapkan informasi yang ingin dicari user.

Dari gambar diatas dapat dilihat bahwa terjadi perubahan pada teks query user sebelum query tersebut di olah. Query expansion dapat dilalkukan dengan menggunakan salah satu dari tiga metode yaitu berikut ini.

1. Manual Query Expansion (MQE)
   Pada metode ini, sistem tidak memberikan bantuan sama sekali kepada pengguna. Pengguna mengubah sendiri query secara manual ketika merasa tidak puas dengan hasil yang didapatkan.
2. Automatic Query Expansion (AQE)
   Pada metode ini, sistem menambahkan kata perluasan berdasarkan kata yang berhubungan dengan query. Modifikasi query dilakukan tanpa perlu kendali dari pengguna. Beberapa teknik yang digunakan antara lain:
   a. Analisis Global (GA)
      Teknik yang menganalisis korpus untuk memeriksa kemunculan kata dan mendapatkan hubungan kata. Analisis Global memeriksa seluruh dokumen yang ada dalam koleksi untuk membangun struktur yang menyerupai thesaurus (pseudo-doc of concept). Perluasan query menggunakan istilah-istilah dalam thesaurus dengan melihat istilah yang berhubungan erat dengan semua istilah pada query dalam ruang lingkup koleksi. Analisis global membutuhkan informasi kemunculan dari setiap pasangan kata pada koleksi yang merupakan tugas yang berat secara komputasi.
   b. Analisis Lokal (LA)
      Analisis lokal memperluas query berdasarkan informasi pada dokumen peringkat teratas yang ditemukan menggunakan query awal. Metode ini mengasumsikan bahwa dokumen-dokumen teratas tersebut relevan untuk kemudian membangkitkan sebuah query baru.
   c. Analisis Konteks Lokal (LCA)
      Analisis Konteks Lokal merupakan sebuah teknik yang mengkombinasikan kelebihan dari analisis global dan analisis lokal. Teknik analisis global yang digunakan adalah Phrasefinder, sedangkan analisis local adalah local feedback. Analisis konteks lokal mengambil ide dari analisis global seperti penggunaan konteks dan konsep tetapi menerapkannya pada dokumen hasil temu kembali menggunakan analisis lokal.
3. Interactive Query Expansion (IQE)
   Pada interactive query expansion, pengguna memilih sendiri kata-kata perluasan yang akan ditambahkan berdasarkan kata-kata yang dihasilkan oleh sistem. Ide dari interactive query expansion adalah bahwa pengguna lebih baik dalam memilih

ekpansi term daripada sistem. Teknik yang tercakup didalamnya adalah relevance feedback. Hubungan antara tiga teknik tersebut dapat dilihat pada gambar berikut.



Gambar 3. Arsitektur Query Expansion

Relevance feedback merupakan suatu teknik dalam information retrieval dimana user memberikan feedback (pengaruh) pada dokumen hasil temu kembali yang dianggap relevan. Relevance feedback adalah metode yang sudah diterima secara luas untuk meningkatkan keefektifan information retrieval secara interaktif. Sebuah pencarian awal dilakukan oleh sistem menggunakan query yang diberikan oleh user dan sebagai hasilnya dikembalikan kepada user sejumlah dokumen. User memeriksa dokumen-dokumen tersebur dan menandai dokumen-dokumen yang dianggap relevan. Sistem kemudian secara otomatis memodifikasi query berdasar penilaian relevansi user tadi. Kemudian query baru dijalankan untuk kembali menemukan dokumen-dokumen yang lebih relevan. Proses ini dapat berjalan berulang terus sampai user merasa kebutuhan informationya terpenuhi.

4. Pembobotan TF/IDF

Analyzing adalah merupakan tahap penentuan seberapa jauh keterhubungan antar kata-kata pada dokumen yang ada dengan menghitung frekwensi term pada dokumen. Tahap ini disebut juga tahap pembobotan, yaitu dijelaskan sebagai berikut:

1. Pembobotan Term

   Term adalah suatu kata atau suatu kumpulan kata yang merupakan ekspresi verbal dari suatu pengertian. Dalam information retrieval sebuah term perlu diberi bobot, karena semakin sering suatu term muncul pada suatu dokumen maka kemungkinan term tersebut semakin penting dalam dokumen.

   Dari proses pembobotan term maka akan didapatkan hasil akhir berupa Term Frequency (TF), yaitu merupakan frekwensi atau jumlah masing-masing kata. Hasil pembobotan term kemudian akan digunakan sebagai dasar perhitungan pada basis Term Frequency-Inverse Document Frequency (TF-IDF).

2. Term Frequency-Inverse Document Frequency (TF-IDF)
   Term Frequency-Inverse Document Frequency (TF-IDF) adalah cara pemberian bobot hubungan suatu kata (term) terhadap dokumen. Untuk dokumen tunggal tiap kalimat dianggap sebagai dokumen. Basis ini menggabungkan dua konsep untuk perhitungan bobot, yaitu Term Frequency (TF) merupakan frekwensi kemunculan kata (t) pada kalimat (d). Document Frequency (DF) adalah banyaknya kalimat dimana suatu kata (t) muncul. TF-IDF dapat dirumuskan sebagai berikut:

   $$TF\text{-}IDF\ (tk, dj) = TF\ (tk, dj) * IDF\ (tk) \quad ..... \quad (1)$$

   Keterangan:
   dj = Dokumen ke-j.
   tk = Term ke-k.

Dimana sebelumnya dihitung terlebih dahulu Term Frequency (TF) yaitu frekwensi kemunculan suatu term ditiap dokumen. Kemudian dihitung Inverse Document Frequency (IDF) yaitu nilai bobot suatu term dihitung dari seringnya suatu term muncul dibeberapa dokumen. Semakin sering suatu term muncul dibanyak dokumen, maka nilai IDF-nya akan kecil. Berikut rumus-rumus TF dan IDF:

$$TF\ (tk, dj) = f\ (tk, dj) \quad ..... \quad (2)$$

Keterangan:
TF = Jumlah frekwensi term.
f = Jumlah frekwensi kemunculan.
dj = Dokumen ke-j.
tk = Term ke-k.

Kemudian untuk menghitung nilai IDF bisa menggunakan persamaan sebagai berikut:

$$IDF(tk) = 1 / df\ (t)\ Atau\ IDF\ (tk) = log\ (N / df\ (t)) \quad ... \quad (3)$$

Keterangan:
IDF = Bobot term.
N = Jumlah total dokumen.
df = Jumlah kemunculan dokumen.
dj = Dokumen ke-j.
tk = Term ke-k.

Persamaan pertama hanya boleh digunakan apabila hanya terdapat satu buah dokumen saja yang diproses sedangkan persamaan kedua digunakan pada proses yang melibatkan banyak dokumen.

5. Vector Space Model

Vector Space Model adalah salah satu metode atau algoritma yang sering digunakan untuk sebuah sistem temu kembali informasi yang biasa dikenal dengan

Information Retrieval System. Algoritma ini merupakan sebuah model yang digunakan untuk mengukur kemiripan antar beberapa dokumen. Dalam Information Retrieval System, kemiripan antar dokumen didefinisikan berdasarkan representasi bag-of-words dan dikonversi ke suatu model ruang vector (Vector Space Model). Model ini diperkenalkan oleh Salton dan telah digunakan secara luas. Pada Vector Space Model, setiap dokumen di dalam database dan query pengguna direpresentasikan oleh suatu vector multi-dimensi. Dimensi sesuai dengan jumlah term dalam dokumen yang terlibat.

Vocabulary merupakan kumpulan semua term berbeda yang tersisa dari dokumen setelah preprocessing dan mengandung t term index. Kumpulan term ini membentuk suatu ruang vektor.

Setiap term i di dalam dokumen atau query j, diberikan suatu bobot (weight) bernilai real wij. Dokumen dan query diekspresikan sebagai vektor t dimensi.

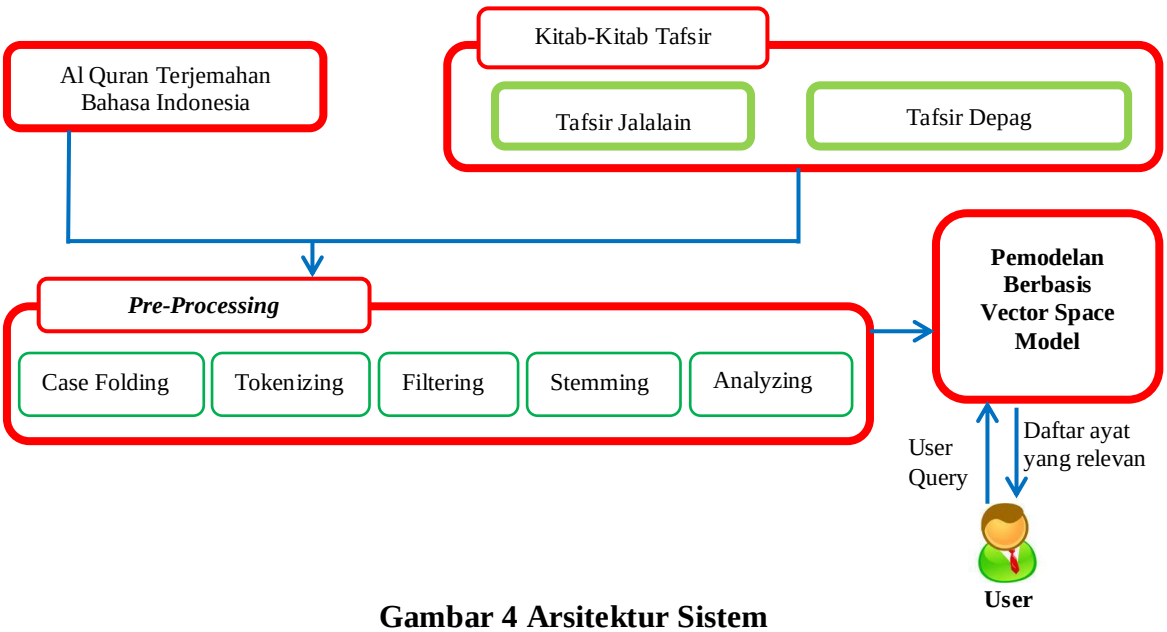
$$dj = (w1, w2, ..., wtj) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (4)$$

Dan terdapat n dokumen didalam koleksi, yaitu j = 1, 2, ..., n. Berikut merupakan contoh dari vector space model tiga dimensi untuk dua dokumen D1, dan D2, satu query pengguna Q1, dan tiga term T1, T2, T3. Representasi Vector Space Model permasalahan diatas dapat dilihat pada gambar 2.2 berikut ini.

III. PEMBAHASAN

1. Arsitektur Sistem

Arsitektur sistem digunakan untuk menggambarkan sistem kerja yang digunakan pada proses analisa dan implementasi. Dengan arsitektur sistem dapat dilihat alur sistem secara lengkap, adapun arsitektur sistem dari keseluruhan sistem ditunjukan pada gambar dibawah



Gambar 4 Arsitektur Sistem

Data yang dipakai dalam penelitian ini adalah bersumber dari program Holy Quran. Baik dari teks terjemahan maupun teks tafsirnya. Data ini terbagi menjadi 114 surat dan terdiri dari total 6.236 ayat. Dari 6.236 ayat tersebut terdapat 29 ayat yang tidak memiliki tafsir atau yang biasa disebut ayat mutasyabihat.

2. Pre_Processing

Pre-processing dilakukan setelah dataset terbentuk. Tahapan-tahapan dalam pre-processing data meliputi tokenizing, filtering dan stemming. Dan supaya hasil dari pre-processing ini dapat dilihat hasilnya maka pada tahap filtering dan stemming hasil pengolahannya akan disimpan.

IV. PENUTUP

Kesimpulan

1. Hasil stemming pada tahapan pre-processing menggunakan algoritma Nazief Adriani menghasilkan akurasi 95% untuk teks terjemahan Al Quran dan tafsir berbahasa Indonesia. Kegagalan pembentukan kata dasar pada tahap ini dikarenakan masih adanya kata dalam bahasa Arab yang tidak diterjemahkan ke bahasa Indonesia.
2. Query expansion yang dilakukan dengan penambahan teks dari dua teks tafsir yaitu tafsir jalalain dan tafsir dari Depag dapat meningkatkan jumlah ayat-ayat yang ditemukan. Hal ini dikarenakan adanya penambahan keyword paad query.
3. Didapatkan nilai recall yang sama yaitu 100% untuk query user yang di ekspansi maupun yang tidak. Sehingga dapat disimpulkan bahwa ekspansi query tidak mempengaruhi dari nilai recall. Hal ini dapat terjadi karena penambahan keyword pada query berbanding lurus dengan peningkatan jumlah ayat yang ditemukan.
4. Untuk query yang tidak diekspans didapatkan nilai precision sebesar 27%. Nilai precision ini didasarkan pada ketentuan bahwa hubungan antar ayat ditentukan oleh kesamaan topiknya dan tidak didasarkan pada banyaknya teks yang sama.
5. Untuk query yang diekspansi didapatkan nilai precision sebesar 75%. Perluasan query dengan penambahan teks tafsir dapat meningkatkan nilai precision karena dengan teks tafsir ini dapat menambahkan makna baru pada query user. Selain itu penambahan teks tafsir yang dilakukan pada semua ayat yang menjadi koleksi dokumen juga menjadi factor peningkatan nilai precision. Sehingga ayat-ayat yang ditemukan dapat memberikan penjelasan yang lebih mendetil dari ayat yang menjadi query user.

Saran

1. Pada penelitian ini menggunakan teks dalam bahasa Indonesai. Diharapkan pada penelitian selanjutnya teks terjemahan dan tafsir yang dilibatkan sebaiknya tetap menggunakan teks berbahasa Arab. Sehingga dapat meminimalkan kesalahan dalam penerjemahan ke bahasa Indonesia.
2. Pada penelitian selanjutnya diharapkan memperhatikan sinonim dan anomin kata.
3. Dalam penelitian ini semua algoritma dibangun menggunakan bahasa pemrograman PHP. Sehingga tidak menggunakan tool sama sekali. Dalam penelitian selanjutnya disarankan menggunakan tool seperti Sphinx Search dan Weka untuk lebih memaksimalkan hasilnya baik dari sisi kecepatan dan akurasinya.

VI. DAFTAR PUSTAKA

[1] Agusta, Ledy. *Perbandingan Algortima Stemming Porter Dengan Algoritma Nazief & Adriani Untuk Stemming Dokumen Teks Bahasa Indonesia*. Universitas Kristen Satya Wacana. 2009.

[2] Berry, Michael W., Castellanos, Malu. *Survey of Text Mining: Clustering, Classification, and Retrieval, Second Edition*. 2007.

[3] Carmel, David. *An Extension of the Vector Space Model for Querying XML Documents via XML Fragments*. Computer Science Dept Haifa University, Haifa. 2009.

[4] Feldman Ronen, Sanger James. *The Text Mining Handbook*. Apress. 2000.

[5] Karyono Giat, Setyo, Fandy. *Temu BalikInformasi Pada Dokumen teks Berbahasa Indonesia dengan Metode Vektor Space Model*. Seminar Nasional Teknologi Informasi & Komunikasi Terapan. 2012.

[6] Munteanu, Dan. *Vector Space Model For Document Representation In Information Retrieval*. Department of Computers and Applied Informatics Romania. 2007.

[7] Rub´ en Tous, Jaime Delgado. *A Vector Space Model for semantic Similarity Calculation and OWL Ontology Alignment*. Universitat Pompeu Fabra. 2009.

[8] Sharaf, Abdul-Baquee, Atwell Eric. *Knowledge representation of the Quran through frame semantics A corpus-based approach*. University of Leeds. 2009.