Penerapan Algoritma Tf-Idf Vector Space Model (Vsm) Pada Information Retrieval Terjemahan Al Quran Surat 1 Samai Dengan Surat 16 Berdasarkan Kesamaan Makna

Implementation of TF-IDF Vector Space Model (VSM) Algorithm in Information Retrieval of AL QURAN Translation verse 1 to verse 16 based on meaning similarity

[1]Irfan Humaini, [2]Lily Wulandari, [3]Diana Ikasari, [4] Tristyanti Yusnitasari
[1,2.3.4]Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Gunadarma Depok
[1]Irfan_Humaini@staff.gunadarma.ac.id, [2]lily@staff.gunadarma.ac.id,[3] d_ikasari@staff.gunadarma.ac.id, [4]tyusnta@staff.gunadarma.ac.id

Abstrak –*Information Retrieval (IR) merupakan pencarian informasi yang biasanya dalam suatu teks dokumen. Pada penelitian ini membahas IR terhadap Al Quran terjemahan Bahasa Indonesia Korpus sinonim (tesaurus) dibentuk untuk mendukung information retrieval agar hasil pencarian menjadi lebih luas. Metode yang digunakan adalah TF-IDF Vector Space Model (VSM) dengan pengembangan pada pembobotan keyword dan proses kueri, yaitu hasil kueri yang menjadi peringkat satu pada hasil pencarian information retrieval dijadikan kueri untuk proses pencarian selanjutnya. Cosine similarity digunakan untuk perhitungan kemiripan dokumen. Pembentukan basis data korpus sinonim (tesaurus) dilakukan dengan cara mengembangkan suatu sistem agar dapat dilakukan secara otomatis. Pengujian dilakukan dengan menguji pencarian ayat Al Quran dalamaplikasi information retrieval dan membandingakan hasil pencarian aplikasi dengan pendapat pakar Al Quran dan Hadist. Persentase keberhasilan pengujian dengan menggunakan 1 kata mencapai 100%. Keberhasilan pencarian pengujian menggunakan lebih dari 1 kata atau sebuah kalimat, pada 10 peringkat teratas dari dokumen yang ditemukan, keberhasilan mencapai 95,6%. Penelitian initelah membuktikan bahwa information retrieval dengan menggunakan korpus sinonim(tesaurus), dan penambahan bobot kata dari keyword pertama yang dicari menambah tingkat relevan, karena secara signifikan memperluas hasil pencarian dan mengeliminir dokumen yang tidak relevan*.

Kata kunci : *Al Quran, Hadist, Korpus, Information Retrieval, TF-IDF, VSM, Cosine Similarity, Machine Learning, Tesaurus*

Abstract –*Information Retrieval (IR) is an information retrieval that is usually done in every text document. This study discusses IR on the Al Qur'an in Indonesian translation. The synonym corpus (thesaurus) was made to support information retrieval so that search results become wider. The method used is TF-IDF Vector Space Model (VSM) with the development of keyword weighting and query processing, namely query results that rank first in the search results for information retrieval that is retrieved by the query for the next search process. Cosine similarity is used to calculate the similarity of documents. The establishment of a synonym corpus database (thesaurus) is done by developing a system that can be done automatically. Testing is done by trying to search verses of the Qur'an in the application of information retrieval and comparing the results of the search application with the experts of the Koran and Hadith. The percentage of successful testing using 1 word reaches 100%. The success of the test search using more than 1 word or sentence, in the top 10 ranking of documents found, reached 95.6%. This research has proven that searching for information by using a corpus of synonyms (thesaurus), and using the word weight of the searched keyword adds relevance, because it significantly adds to search results and eliminates irrelevant documents.*

1. Pendahuluan

Al Quran adalah kitab suci umat Islam, "Kitab (Al Quran) ini tidak ada keraguan padanya, petunjuk bagi mereka yang bertakwa" (Al Quran. Al-Baqarah:2). "Dan sesungguhnya Kami telah mendatangkan sebuah Kitab (Al Quran) kepada mereka yang Kami telah menjelaskannya atas dasar pengetahuan Kami, menjadi petunjuk dan rahmat bagi orang-orang yang beriman" (Al Quran. Al Araf:52). "Dan Kami turunkan dari Al-Quran suatu yang menjadi penyembuh dan rahmat bagi orang-orang yang beriman dan Al-Quran itu tidaklah menambah kepada orang-orang yang dzalim selain kerugian" (Al Quran. Al-Isra: 82). Hal-hal yang terkandung di dalam Al Qur'an berhubungan dengan keimanan, ilmu pengetahuan, hukum, peraturan-peraturan yang mengatur tingkah laku dan tata cara hidup manusia, kisah-kisah umat sebelumnya, ibadah serta tauhid (pengesaan Allah). Al Quran terdiri dari 30 Juz, 114 Surat dan 6326 Ayat sehingga untuk melakukan pencarian terhadap kata yang sesuai dengan tema yang diinginkan akan sulit sekali. Pada beberapa perangkat lunak yang ada, pencarian informasi seperti mencari kata bohong maka hasil pencarian adalah nama ayat dan surat mengenai kata bohong saja, sedangkan dalam terjemahan Al Quran dan Hadist banyak makna yang sama dengan kata bohong seperti dusta, tipu, fitnah dan lain-lain yang tidak akan ditemukan melalui perangkat lunak tersebut karena metode pencarian hanya berdasarkan kata kunci saja. Bahkan ada banyak kata-kata yang dianggap populer di masyarakat yang tidak ada terjemahannya dalam Al Quran, contohnya kata korupsi sehingga jika dilakukan pencarian dengan kata kunci korupsi maka proses pencarian tidak memberikan hasil apapun. Berdasarkan beberapa contoh tersebut maka perlu dibuat korpus Sinonim (tesaurus) untuk mendukung proses pencarian informasi (information retrieval) sehingga ha.sil pencarian menjadi lebih luas dan lebih relevan.

Penelitian yang dilakukan di sini adalah sistem pencarian informasi dengan teknik temu kembali informasi yang merupakan teknik yang digunakan dalam pencarian dan penelusuran informasi yang relevan. Teknik temu kembali informasi (information retrieval) digunakan untuk mencari kemiripan antara kata kunci yang dimasukkan dengan dokumen yang tersimpan di dalam basis data. Basis data terdiri dari dokumen Al Quran Pada pada penelitian ini teknik information retrieval yang digunakan, yaitu pemodelan TF-IDF Vector Space Model. Pada pencarian Al Quran, information retrieval digunakan untuk menampilkan seluruh ayat Al Quran yang sesuai dengan kata kunci (Keyword) yang dimasukkan. Terhadap dokumen atau ayat Al Quran yang terdeteksi berdasarkan keyword dilakukan pemeringkatan dokumen yang memiliki kedekatan paling tinggi dengan keyword, pengukuran peringkat dilakukan menggunakan metode Cosine Similarity. Pemeringkatan dokumen perlu dilakukan agar hasil yang ditampilkan mengurutkan dari dokumen yang paling besar tingkat kemiripannya sampai ke dokumen yang paling kecil tingkat kemiripannya. Setelah didapatkan dokumen yang memiliki tingkat kemiripan paling tinggi (peringkat 1), dokumen tersebut dijadikan kueri selanjutnya ditambah dengan keyword pertama yang dimasukkan beserta sinonimnya. Pada penelitian ini juga dilakukan pengklasifikasian tema, hal ini tujuan utamanya adalah agar proses pencarian dokumen pada information retrieval sudah terfokus pada dokumen-dokumen dengan tema

ISBN: 978-602-60581-1-9

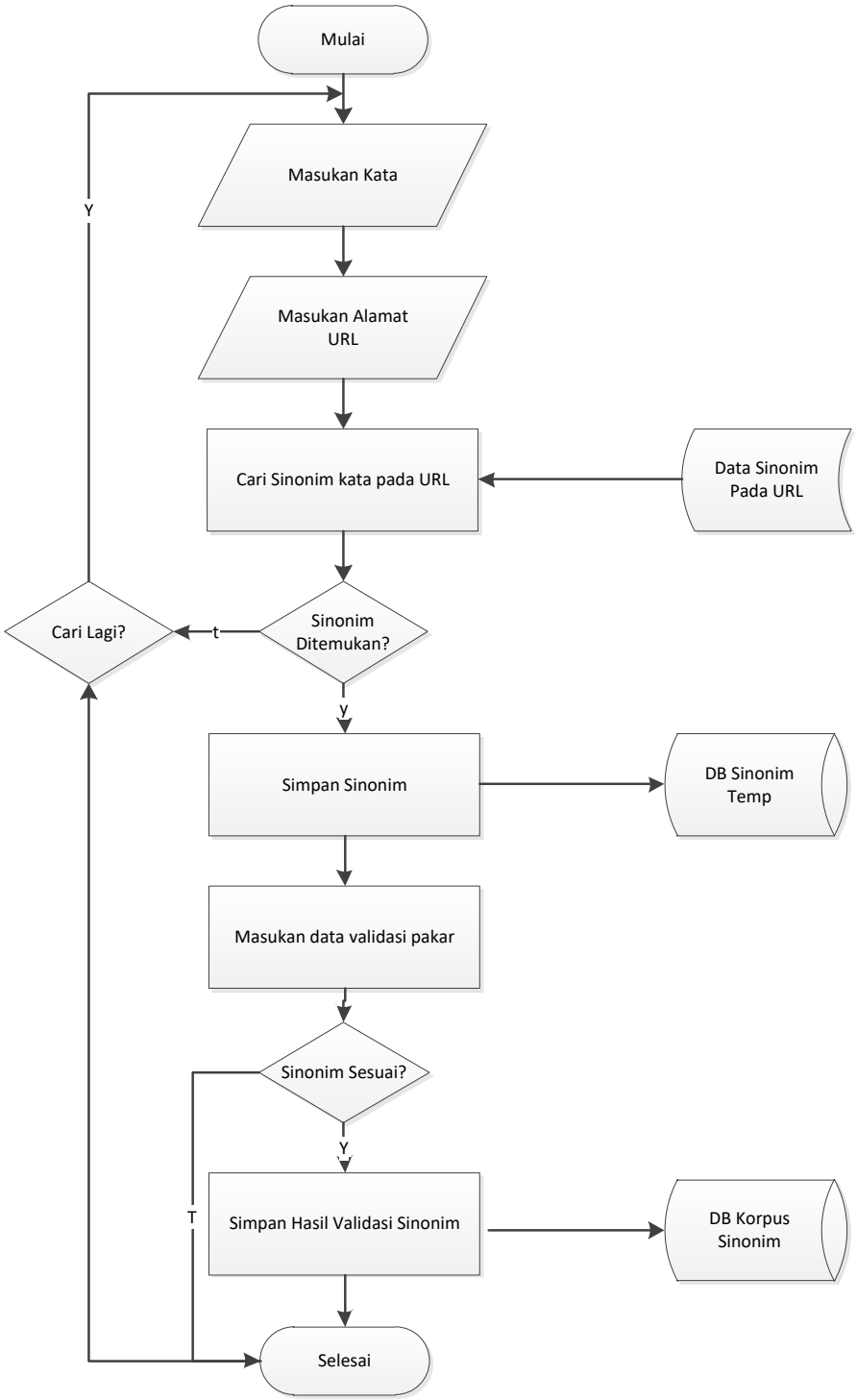
yang sesuai berdasarkan keyword, sehingga hasil yang ditampilkan lebih relevan. Tujuan penelitian ini adalah mengembangkan mekanisme information retrieval terhadap Al Quran terjemahan Bahasa Indonesia dengan pendekatan teknik TF-IDF Vector Space Model (VSM).

Information Retrieval (IR) system digunakan untuk menemukan kembali (retrieve) informasi-informasi yang relevan terhadap kebutuhan pengguna dari suatu kumpulan informasi secara otomatis Gambaran dari serangkaian dokumen sebagai vektor-vektor dalam ruang vector umum disebut dengan model ruang vektor dan merupakan dasar untuk operasi pencarian sejumlah informasi mulai dari memberi nilai dokumen pada kueri, serta klasifikasi dan pembagian dokumen (Manning, 2009). Model ruang vektor (vector space model) menyadari bobot-bobot biner yang terlalu terbatas pada model Boolean, lalu menawarkan sebuah framework yang memungkinkan adanya partial matching (sesuai sebagian) yang dapat dilakukan dengan menugaskan bobot-bobot non-biner ke dalam index terms pada kueri dan dokumen. Pada model ruang vektor, dokumen di dalam basis dan kueri pengguna direpresentasikan oleh suatu vektor multi-dimensi. Dimensi sesuai dengan jumlah term dalam dokumen yang terlibat. Pada penelitian yang dilakukan sebelumnya (Poernomo, Gunawan 2015) metode yang digunakan *Stemming* menggunakan algoritma Nazief Adirani, Teknik IR model ruang vector dan menggunakan *Query Expantion*. Pada penelitian ini menggunakan teks dalam bahasa Indonesia, selanjutnya teks terjemahan dan tafsir yang dilibatkan sebaiknya tetap menggunakan teks berbahasa Arab. Sehingga dapat meminimalkan kesalahan dalam penerjemahan ke bahasa Indonesia. Sinonim dan anonim tidak diperhatikan 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. Penelitian yang dilakukan pada Hadist (Rozanda, Iswanti, 2014) menggunakan model ruang vector, pada penelitian ini hasil uji dari tujuh *query* dengan perhitungan *precision* dan *recall*, maka didapatkan hasil rata-rata *precision* 65% dan menghasilkan *recall* rata-rata 0,97 yaitu hampir semua dokumen yang relevan terambil oleh sistem. Database Sistem informasi hadits ini masih belum lengkap dan beragam, oleh karena itu perlu ditambah dan selalu di*update* secara berkala. (Agustian, Wulandari, 2013) Sistem Qur'an *retrieval* yang dibangun dengan menggunakan model ruang vektor,telah memberikan hasil yang sangat memuaskan untuk beberapa kueri yang diuji, memiliki profil grafik *Precision-Recall* yang landai sampai titik *Recall*=1. Pengujian lebih mendalam perlu dilakukan oleh pihak-pihak yang lebih mengetahui secara seksama mengenai isi kandungan Al-Qur'an, agar hasil pengujian lebih objektif.

2. Merode Penelitian

Langkah yang dilakukan untuk menyusun basis data sinonim (thesaurus) digambarkan pada gambar 1. Alur proses pada gambar 1 di mulai dengan tahap mengumpulkan istilah atau kata populer melalui berapa cara seperti menggunakan kuisioner.

ISBN: 978-602-60581-1-9



Gambar 1 Proses Menyusun Basis Data Sinonim (Tesaurus)

ISBN: 978-602-60581-1-9

Hal ini untuk mengetahui secara langsung respon dari masyarakat. Kemudian mengamati kata-kata yang digunakan di internet seperti dari media sosial, forum masyarakat, forum keagamaan, portal berita dan lain-lain khususnya yang membahas tema tentang Islam. Hasil dari pengumpulan istilah atau kata yang dianggap populer pada masyarakat langkah selanjutnya digunakan sebagai acuan untuk mencari padanan kata atau sinonim di beberapa alamat URL. Alamat URL yang digunakan untuk membangun sinonim adalah www.tesaurus.kemendikbud .go.id, www.senonimkata.com, www.artikata.com, www.kbbi.online.com. Setelah sinonim kata ditemukan maka divalidasi oleh pakar apakah sinonim tersebut sesuai dengan kebutuhan information retrieval Al Quran dan Hadist. Pakar juga akan memberi masukan atau penambahan terhadap istilah atau kata yang tidak ditemukan padanan atau sinonimnya pada URL Jika pakar menyatakan bahwa sinonim (tesaurus) tersebut telah sesuai dengan kebutuhan information retrieval, maka sinonim dari kata-kata tersebut disimpan dalam basis data. Jika kata tersebut dianggap oleh pakar tidak sesuai atau tidak perlu dimasukkan ke dalam basis data maka kata tersebut dihapus dari data sementara. Pada proses pembentukan korpus sinonim ini juga dihasilkan kata-kata yang terkait secara makna dengan kata yang dijadikan acuan. Sebagai contoh kata adil memiliki keterkaitan makna dengan kata timbangan, neraca, setara, dan takar.

3. Hasil dan Pembahasan

Pembentukan Sinonim Korpus(Tesaurus)

Berikut ini langkah-langkah detail dari algoritma Pembentukan Korpus Sinonim (Tesaurus) yaitu:

(a) Langkah 1 adalah untuk membaca kata yang akan dicari sinonimnya.
(b) Langkah 2 memasukkan alamat URL tentang sinonim mana yang dituju.
(c) Langkah 3 mencari padanan kata pada URL.
(d) Langkah 4-5 menyimpan pada basis data sementara jika sinonim ditemukan atau selesai pencarian.
(e) Langkah 6memvalidasi hasil sinonim di basis data sementara dengan pakar, jika sesuai sinonim disimpan di basis data korpus Sinonim.
(f) Langkah 8 jika sinonim di basis data sementara menurut pakar tidak sesuai maka dihapus.

Algoritma Pembentukan Korpus Sinonim (Tesaurus)

Algoritma Pembentukan Korpus Sinonim(Tesaurus)
Input : Kata
Output : Padanan Kata (Sinonim)

Proses

1. Baca Kata
2. Crawling ke Alamat URL
3. Cari padanan kata
4. IF kata ditemukan Then simpan sinonim di db_Temp
5. Else go to 9
6. Hasil di db_temp divalidasi pakar
7. IF validasi sesuai Then Simpan di db_Korpus Sinonim
8. Else Hapus kata di db_sinonim

ISBN: 978-602-60581-1-9

9. End

Berikut ini langkah-langkah detail dari algoritma TF IDF VSM

(a) Langkah 1 adalah untuk membaca keyword.
(b) Langkah 2-5 digunakan untuk Preprocessing agar hasil akhirnya hanyaberupa kata dasar.
(c) Langkah 6-9merupakan proses pencarian sinonim terhadap kata yang menjadi keyword, jika ditemukan sinonim maka akan disertakan dalam kueri pencarian.
(d)Langkah 10-11 digunakan untuk perhitungan kemiripan dan pemeringkatan dari dokumen yang terdeteksi sesuai keyword.
(e) Langkah 12-13 digunakan untuk proses di mana dokumen dengan rangking pertama serta keyword dan sinonimnya (tesaurus) dijadikan kueri.
(f) Langkah ke 14 digunakan untuk pemberian bobot tambahan pada keyword yang digunakan dalam kueri.
(g)Langkah 15-16 digunakan untuk menghitung bobot, kemiripan dan perangkingan dari setiap dokumen.
(h) Langkah 18-19 digunakan untuk proses pencarian berdasarkan tema klasifikasi yang terkait dengan kueri pada basis data Al Quran terjemahan bahasa Indonesia dan tema Klasifikasi.
(i) Langkah 20 digunakan untuk menampilkan hasil Information Retrieval berupa ayat Al Quran beserta tema apa yang terkandung dari setiap ayat

Algoritma Information Retrieval Al Qur'an dan Hadist

Algoritma IR Al Qur'an dan Hadist
Input : Keyword [t]
Output : Ayat Al Quran[d.n] dan Hadist[d.n]

Proses

1. Baca Keyword [t]
2. Preprocessing
3. Tokenizing
4. Stopword Removal
5. Stemming
6. Cari padanan kata di Db_Sinonim
7. IF Sinonim ditemukan sertakan pada pencarian
8. Else proses keyword 1
9. End IF
10. Perhitungan TF-IDF VSM
11. Perangkingan Cosine Similarity
12. IF Rangking dokumen[d.n] = 1 then
13. Dokumen[d.n] sebagai queri / keyword baru
14. Beri bobot Tf IDF tambahan pada kata yang diinput pada keyword1
15. Perhitungan TF-IDF VSM
16. Perangkingan Conine Similarity

ISBN: 978-602-60581-1-9

17. End IF
18. Cari klasifikasi tema Al Quran dan Hadist pada Db_klasifikasi Tema
19. Cari terjemahan Al Quran dan Hadist pada Db_terjemahan Al Quran dan Hadist
20. Tampilkan Hasil IR Ayat Al Quran dan Hadist
21. End

Tabel 1 merupakan tabel hasil pengujian terhadap proses *information retrieval* terhadap 10 surat Al Quran. Tabel – tabel tersebut memiliki 7 kolom yaitu kolom nomor urut, kolom surat yang diuji, total ayat yang ditemukan oleh *pakar*, no ayat dan surat yang ditemukan oleh pakar, total ayat yang ditemukan oleh aplikasi, no ayat dan surat yang ditemukan oleh aplikasi dan prosentase ketepatan aplikasi dari masing-masing dokumen. Sebagai contoh pengujian dokumen nomor urut satu.

Tabel 1 Hasil Pengujian Pencarian Dengan *Keyword* : DOSA

No	Surat Al Quran	Total Ayat				Prosentase Ketepatan
		Pakar	Surat:Ayat	Aplikasi	Surat:Ayat	
1.	81. Surah At-Takwir (29)	1	81:9	1	81:9	100%
2.	82. Surah Al-Infitar (19)	0		0		100%
3.	83. Surah Al Mutafifin (36)	5	83:12 83:17 83:29 83:31	5	83:12 83:17 83:29 83:31	100%
4.	84. Surah Al-Insyiqaq (25)	1	84:24	1	84:24	100%
5.	85. Surah Al-Buruj (22)	3	85:10 85:12 85:4	3	85:10 85:12 85:4	100%
6.	86. Surah At-Tariq (17)	0		0		100%
7.	87. Surah Al-A'la (19)	0		0		100%
8.	88. Surah Al-Gasyiyah (26)	1	88:24	1	88:24	100%

ISBN: 978-602-60581-1-9

9.	89. Surah Al-Fajr (30)	1	89:13	1	89:13	100%
10.	90. Surah Al-Balad (20)	0		0		100%

Pengujian terhadap proses *Information Retrieval* Al Quran terjemahan bahasa Indonesia terhadap 10 surat Al Quran dengan *keyword* : Dosa dapat dilihat pada tabel 1. Tabel 1 merupakan tabel hasil pengujian proses *information retrieval* Al Quran terjemahan bahasa Indonesia terhadap 10 surat Al Quran. Pada tabel 1 memiliki 7 kolom yaitu kolom nomor urut, kolom surat yang diuji, total ayat yang ditemukan oleh *pakar*, nomor ayat dan surat yang ditemukan oleh pakar, total ayat yang ditemukan oleh aplikasi, nomor ayat dan surat yang ditemukan oleh aplikasi dan prosentase ketepatan aplikasi dari masing-masing dokumen.Sebagai contoh pengujian adalah surat 81 nama surat At-Takwir. Total ayat yang ditemukan oleh pakar adalah 1 ayat. Total Ayat yang ditemukan oleh aplikasi adalah 1 ayat. Nilai prosentase ketepatan dokumen nomor urut satu sebesar 100%. Jumlah surat yang diujikan sebanyak 10 surat Al Quran untuk pencarian dengan *keyword:* Dosa . Total ayat Al Quran hasil pencarian oleh pakar sebanyak 12 ayat Al Quran. Total ayat Al Quran hasil pencarian dengan aplikasi sebanyak 12 ayat Al Quran..

Tabel 2 Hasil Rata-Rata Keberhasilan Pengujian IR Dengan 35 Keyword

Keyword	**Prosentase**
Dosa	100%
Hamil	100%
Neraka	100%
Surga	100%
Pelit	100%
Sholat	100%
Istana	100%
Gila	100%
Alam	100%
Jiwa	100%
Gaib	100%
Bohong	100%
Tubuh	100%
Adil	100%
Durhaka	100%
Keyword	**Prosentase**
Tolong	100%
Canda	100%
Korupsi	100%
Amal	100%
Sulit	100%
Perempuan	100%
Catat	100%
Sumpah	100%
Kapok	100%
Mati	100%
Kkeluarga	100%

ISBN: 978-602-60581-1-9

Hilang	100%
Iman	100%
Takwa	100%
Gosip	100%
Orang Tidak Percaya Kepada Tuhan	87%
Hukuman Orang Tidak Adil	77.70%
Wajah Berseri-seri	
Al Quran itu tiada lain hanyalah peringatan bagi semesta alam	88.80%
Aku benar-benar bersumpah dengan kota ini (Mekah)	100%
RATA-RATA KEBERHASILAN	**98.70%**

Tingkat keberhasilan aplikasi untuk pencarian information retrieval adalah :

$$Prosentase\ Ketepatan = \frac{(m-(m-a)}{m}\ x\ 100\%$$

$$Prosentase\ Ketepatan = \frac{(12-(12-12)}{12}\ x\ 100\%$$
$$= 100\%$$

dimana
m = pakar
a = aplikasi

Pada tabel 2 adalah persentase keberhasilan rata-rata dari proses information retrieval pencarian ayat-ayat Al Quran dengan percobaan dilakukan sebanyak 35 kali dengan keyword yang berbeda, hasilnya di dapat nilai rata-rata 98.70%.

4. Kesimpulan

Telah berhasil dilakukan penerapan algoritma pembentukan Sinonim (tesaurus) untuk penyusunan istilah-istilah atau kata populer beserta sinonimnya yang bersumber dari beberapa URL dan masukan dari pakar, sehingga terbentuk korpus sinonim (tesaurus) yang baik. Ujicoba information retrieval terhadap 20 kata yang telah menggunakan korpus sinonim, sudah berhasil mendeteksi ayat-ayat dari surat Al Quran yang cukup luas dan telah memunculkan hasil dari keyword yang dimasukkan beserta sinonimnya. Hasil ujicoba Information retrieval jika tanpa mennggunakan korpus sinonim akan memunculkan ayat-ayat dari surat Al Quran yang lebih sedikit karena hanya mendeteksi kata yang sama dengan keyword. Perbandingan

ISBN: 978-602-60581-1-9

hasil information retrieval dengan menggunakan korpus sinonim hasilnya lebih signifikan dibandingkan tanpa menggunakan korpus sinonim.

Referensi

1. Adriani, M., Asian, J., Nazief, B. Tahaghoghi, S.M.M., Williams, H.E. 2007. *Stemming Indonesian*: *A Confix-Stripping Approach. Transaction on Asian Langeage Information Processing*.
2. Agusta, Ledy. Comparison of Algortima Stemming Porter With Nazief & Adriani Algorithm For Stemming Indonesian Text Document. Satya Wacana Christian University. 2009.
3. Baeza R.Y., Neto R., Modern Information Retrieval, Addison Wesley-Pearson international edition, Boston. US. USA, 1999.
4. Broto Poernomo T.P, Ir. Gunawan, Information Retrieval System Search Similarities AlQur'an Translation Version in Indonesian with Query Expansion from Tafsirnya IDeaTech, ISSN: 2089-1121, 2015.
5. Manning, Christopher D., Prabhakar Raghavan,. Introduction to Information Retrieval. Cambridge University Press, Cambridge, England, 2009.
6. Nesdi E. Rozanda, Arif Marsal, Kiki Iswanti, Design of Hadist Information Systems Using Technique of Retrieval of Vector Space Model Information, ejournal.uin-suska.ac.id, 20014.
7. Surya Agustian, Imelda Sukma Wulandari, Qur'an Retrieval System Web-based Indonesian Translation with Reorganization of Corps, KNSI 2013, ISBN 978-602-17488-0, 2013.
8. Tala, Fadillah Z. 2003. *A Study of Stemming Efects on Information Retrieval in* Bahasa Indonesia.

ISBN: 978-602-60581-1-9