American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN : 2320-0936
Volume-10, Issue-8, pp: 254-258
www.ajer.org
Research Paper Open Access

Developing Information Retrieval System For Indonesian Quran Translation Using Vector Space Model (VSM)

Rina Candra Noor Santi, Sri Mulyani, Sri Eniyati
[1,2,3] *(Informationof Technology Faculty, Stikubank University, Indonesia*
Corresponding Author: Rina Candra Noor Santi

ABSTRACT: *Word search is a part of Information Retrieval, one of the models is Vector space model. VSM has effectiveness in word search because the search results are based on similarity of vectorquery and document vector. This study implements the VSM Algorithm in stages: preproccessing text using 4 stages, weighting terms using the TF-IDF method, and ranking using the Cosine Similarity method, testing the system with Recall & Precision.*
KEYWORDS :*Vector Space Model* (VSM), *Information Retrieval,* TF-IDF, *Cosine Similarity, Recall&Precision,*

Date of Submission: 04-08-2021 Date of acceptance: 17-08-2021

I. INTRODUCTION

Word search is one part of Information Retrieval that deals with retrieving information from documents based on the content and context of the documents themselves. Information Retrieval is the science of searching information on documents, searching for documents themselves, searching in databases for text, sound, images, or other data.

There are three models used in information retrieval, the first is Probabilistic model, the example of this model is the application of Bayes theorem in a probabilistic model, the second is Set-theoretic models, the example of this model is Standard Boolean and the last is Algebratic model, the model represents documents and queries as vector similarity between vector documents and vector queries. Examples of this model are Vector space models. From three models of information retrieval, Algebratic models with examples of models Vector space model is the simplest model in word search, has proven to have effectiveness in word search by displaying search results based on similarity of vector queries and vector documents. Vector Space Model is an IR model that represents documents and queries in the form of a dimensional vector. The basic concept of VSM is to calculate the distance between documents and then sort them based on their proximity. The smaller the distance between documents, the more similar they [4]. While according to Amin [2], the VSM method was chosen because the way this model works is efficient, easy in representation and can be implemented in document matching. Based on this information, the author uses the Vector Space Model (VSM) approach as a word search model in the Al-Qur'an translation.

The Qur'an has a total of 30 chapters, 114 letters and 6,236 verses. In this case the author conducted research as many as 13 juz, with 15 letters and 1,803 verses, namely from Surah Al-Fatihah verse 1 - Surah Al-Hijr verse 1 which has 28% content of the Al-Qur'an. Vocabulary in Al-Qur'an experiences many repetitions, especially in Surah Al-Baqarah which has 80% of Al-Qur'an vocabulary and 20% in other surahs. In this study, the author uses the Al-Qur'an text source and the translation from the Tanzil.net website. Tanzil is an international project, which is not incorporated into a sect, organization, or country that was launched in early 2007 to produce Unicode-verified Al-Qur'an text that functions in the Al-Qur'an website and application.

The hope of this study is the achievement of a search for the entire vocabulary in the Al-Qur'an which is supported by the Recall & Precision test which serves as a measure of the accuracy of the system in generating a suitable search.

II. LITERATURE REVIEW

2.1. Information Retrieval

System According to Salton [2], the information retrieval system is a system that retrieves information that is appropriate to the user's needs of the information collection automatically. The working principle of the information retrieval system if there is a document and a the user who formulates a question (request or query). The answer to that question is a collection of relevant documents and discarding irrelevant documents. The information retrieval system will take one of these possibilities. The information retrieval system is divided into two main components, namely the indexing system to generate a system database and retrieval is a combination of user interface and look-up table.The information retrieval system is designed to find documents or information needed by the user.

There are three models used in information retrieval, the first is Probabilistic model, the example of this model is the application of Bayes theorem in a probabilistic model, the second is Set-theoretic models, the example is Standard Boolean and the last is Algebratic model, the model represents the document and query as vector similarity between vectors and vector queries. Examples of this model are Vector space models [4].

2.2. Vector Space Model (VSM)

According to Baeza [2] Vector Space Model (VSM) is a method to see the level of closeness or similarity of terms by means of weighting terms. Documents are seen as a vector that has magnitude and direction. In the Vector Space Model, a term is represented by a dimension of vector space. The relevance of a document to a query is based on similarity between document vectors and query vectors.

VSM provides a partial matching framework is possible. This is achieved by assigning non-binary weights to index terms in queries and documents. The term weight is finally used to calculate the level of similarity between each document stored in the system and the user's request. Documents taken are sorted in a sequence that has similarities, the vector model takes into account the consideration of documents relevant to user requests. The result is a collection of documents that are taken far more accurately (in the sense that they are in accordance with the information needed by the user) [2].

The stages for VSM are:

1. Term Weighting

The first step in calculating VSM is Weighting Term Frequency Inverse Document Frequency (TFIDF), which combines two concepts for calculating weights, the frequency of occurrence of a word in a document (TF) and Inverse.

$$tf = tf_{i,j} \qquad (1)$$

2. Calculating Cosine Similarity

Through VSM and TFF weighting, a representation of the numerical value of the document will be obtained so that the proximity between documents can be calculated. The closer two vectors are in a VSM, the more similar the two documents are represented by the two vectors. Similarities between documents can be calculated using a similarity measure function. This size allows the ranking of documents according to their similarity or relevance to the query. Cosine Similarity or Sim (q, dj) is used to evaluate the level of similarity or similarity of documents (dj). related to query (q) as a correlation between vector dj and q [2].

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

***3.** Recall & Precision*

Precision is the proportion of documents taken by the system that are relevant. Recall value is the proportion of relevant documents taken by the system.

Recall = (number of relevant documents found) / (number of relevant documents in the collection)(3)

The highest recall value is 1, which means that all documents in the collection have been found. Precision is defined by finding only relevant documents in the collection. [7]

Precision = (number of relevant documents found) / (number of documents found) (4)

The highest precision value is 1, which means all documents found are relevant. [7]

III. METHODOLOGY

3.1. Requirements Planning Phase In this phase the author takes the following steps:

a. Identify the purpose of making the application

b. Analyze application requirements.

c. Make application features that will be created.

3.2. Design Process Phase (Workshop Design) In this phase the author takes the following steps:

a. Designing a database, is the design of tables needed in data processing. These tables will be implemented into databases, which use the MySQL program.

b. Design the system using UML. Yaitain design about what processes are needed by the system. Then change it to an algorithm that can be implemented into the program.

c. Designing an interface, which is the display that will be shown to the user and is expected to be easy to use.

3.3. Implementation (Implementation System) In this phase the author takes the following steps:

a. Make programming and implement VSM into programming with the best algorithms.

b. Test the system to check whether the program is running well or there are still errors in the coding or the system.

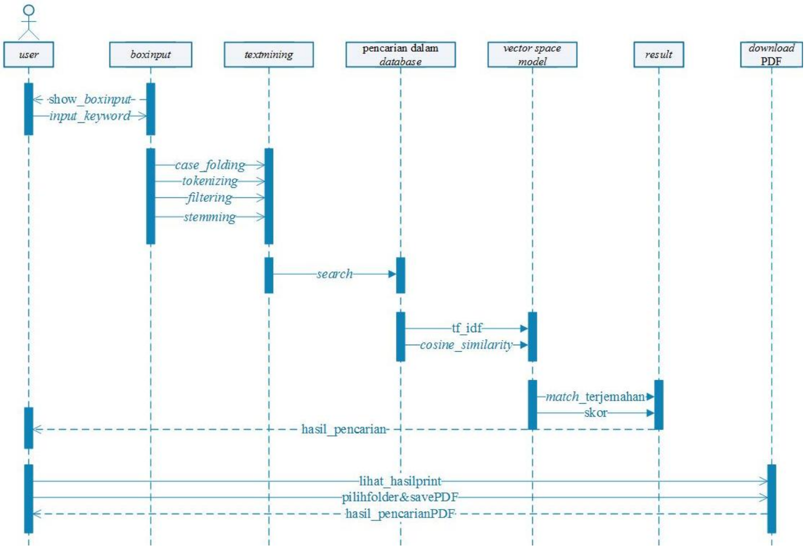

Fig. 1 Sequence Diagram

IV. DISCUSSION AND RESULT

4.1. Translated Search Application al-Qur'an juz 1-13

In running this application, the first step that must be done is textpreprocessing in the database id_indonesian that is a database containing al-qur'an translation from juz 1-13 whose results are stored in the tm_id_indonesian database. This process is carried out with 4 stages, namely Case Folding, Tokenizing, Stopword Removal, and Stemming.

4.2. TermWeighting & Vector Space Model (VSM)

In this process, there are several calculations with different formulas. The first step in this calculation is to calculate the distance of the Query and each document (paragraph). The trick is to calculate | q | * | aj |, q as the query (input from the user) and aj is the paragraph containing the query.

No.	*Term*	W^2							
		Q^2	$A1^2$	$A2^2$	$A3^2$	$A4^2$	$A5^2$	$A6^2$	$A7^2$
1	alam			0.714					
2	allah	0.293	0.293	0.293					
3	balas					0.714			
4	beri								0.714
5	engkau						1.183		0.293
6	hari					0.714			
7	jalan							0.293	2.663
8	kuasa					0.714			
9	lurus							0.714	
10	maha		1.183		1.183				
11	minta						0.714		
12	murah		0.293		0.293				
13	murka								0.714
14	nama		0.714						
15	nikmat								0.714
No.	*Term*	W^2							
		Q^2	$A1^2$	$A2^2$	$A3^2$	$A4^2$	$A5^2$	$A6^2$	$A7^2$
16	orang								2.856
17	puji			0.714					
18	sayang		0.293		0.293				
19	sebut		0.714						
20	sembah						0.714		
21	semesta			0.714					
22	sesat								0.714
23	tolong						0.714		
24	tuhan			0.714					
25	tunjuk							0.714	
SUM		0.293	3.496	3.151	1.773	2.142	3.325	1.723	8,67
SQRT		0.544	1.869	1.775	1.331	1.463	1.823	1.312	2.944

Fig. 2 Example Calculation of CosSim-distance Query

4.3. Recall and Precision

Testing is done with 10 keywords in the search process just limited to juz 1. Using keywords with 1 term (hereafter, shaitan, reward, prayer), 2 terms (inhabitants of hell, Banu Israel, recipients of repentance, Baitul Maqdis), and 3 terms (creators earth sky, jewish and Christian). The basis for query retrieval is the Convenience Sampling method, where the researcher assumes that the words at the top are often used, while the bottom is rarely used.

Table 1. Recall and Precision Result

No	*Query*	*Recall*	*Precision*
1	Syaitan	100%	100%
2	Shalat	100%	100%
3	Pahala	100%	100%
4	Bani Israil	100%	100%
5	Yahudi dan Nasrani	100%	100%
6	Akhirat	100%	83,3%
7	Baitul Maqdis	100%	50%
8	Penghuni Neraka	100%	42,8%
9	Penerima Taubat	100%	37,5%
10	Pencipta Langit Bumi	100%	22,2%

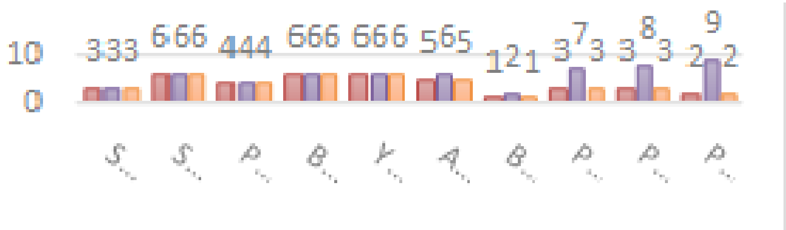

Fig. 3. Recall and Precision Chart

To calculate the recall by dividing the number of relevant documents found with the number of relevant documents in the collection. And to calculate precision by dividing the number of relevant documents found with the number of documents found.

V. CONCLUSION

The conclusion that the implementation of VSM for word search on the Al-Qur'an translation through the first stage is preproccessing prostextext which uses 4 stages, namely case folding, tokenizing, filtering, stemming (porter stemmer algorithm) and term weighting (TF-IDF method) which serves to maximize search result. And the process of similarity (cosine similarity) serves to get the match of the paragraph with the query and the distance for sorting.

The application system is able to search the translation of the Qur'an and display the search results accompanied by the results of calculations using the VSM Algorithm and sort them by ranking. The results of recall & presicion testing shows that the search results have an average 100% recall of the system can produce a complete paragraph that has a query .

REFERENCES

[1]. A.S., R., & Shalahuddin, M. (2014). *Rekayasa Perangkat Lunak, Terstruktur dan Berorientasi Objek.* Bandung: Informatika, 2014.

[2]. Amin, F. (2012). *Sistem Temu Kembali Informasi dengan Metode Vector Space Model*. Semarang: Universitas Stikubank, Fakultas Teknologi Informasi, Program Studi Sistem Informasi Bisnis, 2012.

[3]. Amin, F., Purwatiningtyas, Wicaksono, A., & Setiawan, D, *Rancang Bangun Information Retrieval System (IRS) Bahasa Jawa Ngoko pada Palintangan Penjebar Semangad dengan Metode Vector Space Model (VSM).* Semarang: Fakultas Teknolohi Informasi, Universitas Stikubank (UNISBANK), 2015.

[4]. Bari, A., & Saputra, R. H, *Penerapan Pencarian Kata dengan Vector Space Model pada Aplikasi Terjemahan Juz Amma berbasis Java ME.* Palembang: Prodi Teknik Informatika, STMIK GI MDP, 2011.

[5]. Brata, D., & Hetami, A. (2015). *Perancangan Information Retrieval (IR) Untuk Pencarian Ide Pokok Teks Artikel Berbahasa Inggris Dengan Pembobotan Vector Space Model*. Stmik Asia Malang, 2015.

[6]. Chu, H, *Information Representation and Retrieval in the Digital Age, 2nd Edition.* New Jersey: Information Today, 2010.

[7]. Fitri, M, *Perancangan Sistem Temu Balik Informasi dengan Metode Pembobotan Kombinasi TF-IDF untuk Pencarian Dokumen Berbahasa Indonesia*. Pontianak: Prodi Teknik Informatika, Jurusan Teknik Elektro Fakultas Teknik, Universitas Tanjungpura, 2013.

[8]. Guritno, S., & Sudaryono, U., *Theory and Application of IT Research, Metodologi Penelitian Teknologi Informasi.* Yogyakarta: Andi, 2011.

[9]. Harjanto, D. S., Endah, S. N., & Bahtiar, N, *Sistem Temu Kembali Informasi pada Dokumen Teks Menggunakan Metode Term Frequency Inverse Document Frequency (TF-IDF).* Semarang: Universitas Diponegoro, Fakultas Sains dan Matematika, Jurusan Matematika & Jurusan Ilmu Komputer/Informatika, 2012.

[10]. K.L. Sumathy, M, *Text Mining : Concepts, Applications, Tools and Issues - An Overview.* College Thanjavur, 2013.

[11]. Karyono, G., & Utomo, F. S, *Temu Balik Informasi pada Dokumen Teks Berbahasa Indonesia dengan Metode Vector Space Retrieval Model*. Purwokerto: Prodi Teknik Informatika & Sistem Informasi, STMIK AMIKOM, 2012.

[12]. Mustaqhfiri, M., Abidin, Z., & Kusumawati, R, *Peringkasan Teks Otomatis Berita Berbahasa Indonesia Menggunakan Metode Maximum Marginal Relevance*. Malang: Jurusan TI Fakultas SainTek - UIN Malik Ibrahim, Malang, 2012.

[13]. Nasir, M., *Mukjizat Rasm Al-Qur'an (Telaah Atas Tulisan Mushaf Usmani).* Yogyakarta: Jurusan Tafsir dan Hadis Fakultas ushuluddin - UIN Sunan Kalijaga Yogyakarta, 2008.

[14]. Nofriansyah, D., *Konsep Data Mining VS. Sistem Pendukung Keputusan.* Yogyakarta: Deepublish, 2014.

[15]. Pudjiantoro, T. H. , *Analisa Kompetensi Calon Pegawai Menggunakan Metode TF-IDF.* Bandung, 2013.

[16]. Rizqi, D. Y., *Implementasi Algoritma Porter Stemmer dan Algoritma Boyer Moore pada Pencarian Ensiklopedia Etika Islam berdasarkan Al-Qur'an dan Hadits*. Jakarta: Prodi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Syarif Hidayatullah, 2014.

[17]. Salamah, *Penerapan Tokenisasi Kalimat dan Metode TF (Term Frequency) Pada Peringkas Teks Otomatis Artikel Berita Berbahasa Indonesia.* Jakarta: Prodi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Syarif Hidayatullah, 2014.

[18]. Solihin, A. , *Pemrograman Web dengan PHP dan MYSQL.* Budi Luhur, 2016.

[19]. Tala, F. Z., *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*. Belanda: Institute for Logic, Language and Computation - Universiteit van Amsterdam, 2003.