RESEARCH ARTICLE OPEN ACCESS

Information Retrieval on Hadith Sahih Bukhari Using Vector Space Model Method and Expansion Query

[1]Dede Irawan, [2]Imam Halim Mursyidin, [3]Doni Prasetyo
Faculty of Engineering *Sheikh Yusuf Islamic University, Tangerang, Banten, Indonesia*
E-mail : [1]*dedeirawan@unis.ac.id*, [2]*imamhalim@unis.ac.id*, [3]*doniprasetyo@unis.ac.id*

Abstract:

According to the big Indonesian dictionary the hadiths are words, deeds, takrir (decrees) of the Prophet Muhammad sallallaahu ‘alayhi wa sallam. narrated or told by friends to explain and establish Islamic law. There are many hadith narrators who become a reference for Muslims, some of them are Muslim Imam, Imam Ahmad and Imam Bukhari. For Muslims, Shahih Bukhari is the main reference to deepen As-Sunnah as the second source of Islamic law after the Koran. As the development of digital technology, Hadith is packaged into digital form. Currently the Hadith can be read through computers, smartphones and other digital devices. On the smartphone there is a Hadith Sahih Bukhari v.3.6 application created by Islamic Developers. The Hadith application has a search feature but the search results are still less relevant. Searching can only be done if the keyword phrase (Query) entered exactly matches the phrase in the digital text document in Indonesian. This is because the search is still conventional. For example, we tried to find hadiths with the keyword "usury" but in the search results the keyword "geriba" also appeared so it does not match what we are looking for. The problem that occurs is that of the many hadith narrated by Imam Bukhari, the author is difficult to find relevant traditions in accordance with the search. Therefore we need an Information Retrieval to support the search, to make an Information Retrieval the writer will use the Vector Space Model and Expansion Query methods. The combination of these methods is expected to provide relevant hadith information in accordance with the search.

Keywords — **Pseudo Relevance Feedback, Hadis Online, Cosine Similarity**

I. INTRODUCTION

As the development of digital technology, Hadith is packaged into digital form. Currently the Hadith can be read through computers, smartphones and other digital devices. On the smartphone there is a Hadith Sahih Bukhari v.3.6 application created by Islamic Developers and writers obtained from Playstore. There is a search feature but the search results are still less relevant. Searching can only be done if the keyword phrase (Query) entered exactly matches the phrase in the digital text document in Indonesian. This is because the search is still conventional. For example, we tried to find hadiths with the keyword "usury" but in the search results the keyword "geriba" also appeared so it does not match what we are looking for. This problem has been discussed in research conducted by (Sanjaya, 2016) but the method of decomposition of sentences has not yet yielded an optimal level of relevance. So we need an Information Retrieval to support the search. Therefore to make Information Retrieval we will use the Vector Space Model and Expansion Query methods. Because the research conducted by (Karyono, Utomo, 2012) vector space retrieval model method can give good results and also on research conducted by (Pamungkas and Ridok, 2015) Query Expansion method can add accuracy value to get relevant results. . The combination of these methods is expected to provide relevant hadith information in accordance with the search.

II. THEORITICAL BASIS

The information retrieval system is a system that finds information that fits the needs of users from a collection of information (Salton, 1989; Wibowo, 2012). The main functions of Information Retrieval as stated by (Lancaster, 1979) and (Kent, 1971) are as follows:

1. Identify sources of information relevant to the interests of the targeted user community.
2. Analyzing the contents of information sources (documents)

3. Represent the contents of the information source in a certain way that allows to be reconciled with questions (queries) users.
4. Representing users' queries in a certain way that makes it possible to bring together information sources contained in the database.
5. Bringing together statements
6. Search with data stored in a database.
7. Finding relevant information.
8. Improve system performance based on feedback given by users.

2.1. PREPROCESSING

Preprocessing is a process of changing the form of unstructured data into structured data according to needs, there are several stages in preprocessing such as case folding, tokenizing, filtering, stemming and indexing.

2.2. VECTOR SPACE MODEL

Vector Space Model (VSM) is a method to see the level of closeness or similarity of terms by weighting terms (Baeza-Yates and Ribeiro-Neto, 1999). The document is seen as a vector that has magnitude (distance) and direction (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 the similarity between the document vector and the query vector (Amin, 2012).

2.3. TERM FREQUENCY

Term Frequency is the frequency of occurrence of a term in the relevant document. The greater the number of occurrences of a term in the document, the greater the weight or the greater the value of conformity, as in the following equation:

$$tf = tf_{if}$$

Information:

tf	=	*Term frequency*
tf_{if}	=	*The number of times the term tf appears in the if document*

2.4. INVERSE DOCUMENT FREQUENCY

Inverse Document Frequency is a calculation of how the term is widely distributed in the collection of documents concerned. Inverse Document Frequency shows the availability relationship of a term in all documents. The fewer the number of documents containing the term in question, the greater the value of the Inverse Document Frequency, as in the following equation:

$$idf = log \frac{N}{Df_i}$$

Information :

idf	=	*Inverse document frequency*
N	=	*Total number of documents*
Df_i	=	*The number of documents in the collection where the term appears*

2.5. WEIGHT TERM FREQUENCY

Weight Term Frequency is the process of calculating weights on each term after getting results from the TF and its IDF. Thus the general formula for Weight Term Frequency is the merging of the raw TF calculation formula with the IDF formula by multiplying the TF value with the IDF value, as in the following equation:

$$W_{tf} = tf_{if} \, X \, idf_{fi}$$

Information :

W_{tf}	=	*Document weight*
Tf_{if}	=	*The number of times the term tf appears in the if document*
Idf_{fi}	=	*The results of the calculation of IDF on term fi*

2.6. COSINE SIMILARITY

One measure of text similarity used in VSM to sort document similarity is cosine similarity, which calculates the cosine values of angles between vectors, as in Equation 2.4.

$$\cos(\theta_{ij}) = \frac{A \cdot B}{|A||B|} = \frac{\sum_k (d_{ik} \, d_{jk})}{\sqrt{\sum_k d_{ik}^2} \sqrt{\sum_k d_{jk}^2}} \quad (2.4)$$

Information :

A	=	*Vector A*
B	=	*Vector B*
A • B	=	*Dot product between vector A and vector B*
|A|	=	*Length of vector A*
|B|	=	*Length of vector B*
|A||B|	=	*Cross product between | A | and | B |*

2.7. PSEUDO RELEVANCE FEEDBACK

Determination of the terms to be combined in the old query manually is that from the results of the first search the user will give feedback to the system then from some documents returned to the user system will determine the new term chosen from the relevant documents returned to the system. While the determination of new terms automatically is the system will determine the Top

N documents from the initial search results then the system will identify all the terms that are in the Top N documents. For the new term to be used, it is taken from words that have the maximum appearance value of the term. Then the term will be entered into the old user's query (Cios, et al, 2007).

III. METHODOLOGY

The research methodology to be carried out in this study in general can be shown by the flow chart in Figure 3.

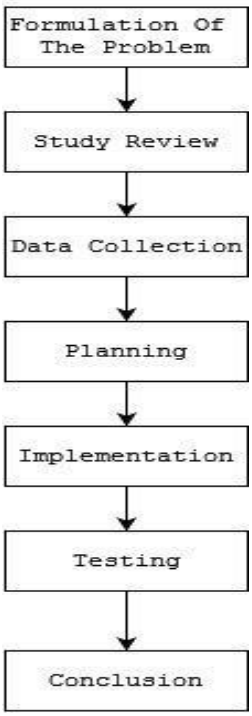

Figure 3.1 System Flowchart

3.1. DESIGN

This system starts with input in the form of a query from the user, then text preprocessing and weighting of the query will be done. Next, rank the document with the Vector Space Model that is relevant to the user's query. If the user chooses to expand, then the document with the top ranking is then taken to do a query reformulation with the Query Expansion method so that it gets a new query and then reprocesses. Figure 3.2 shows the design flow chart.

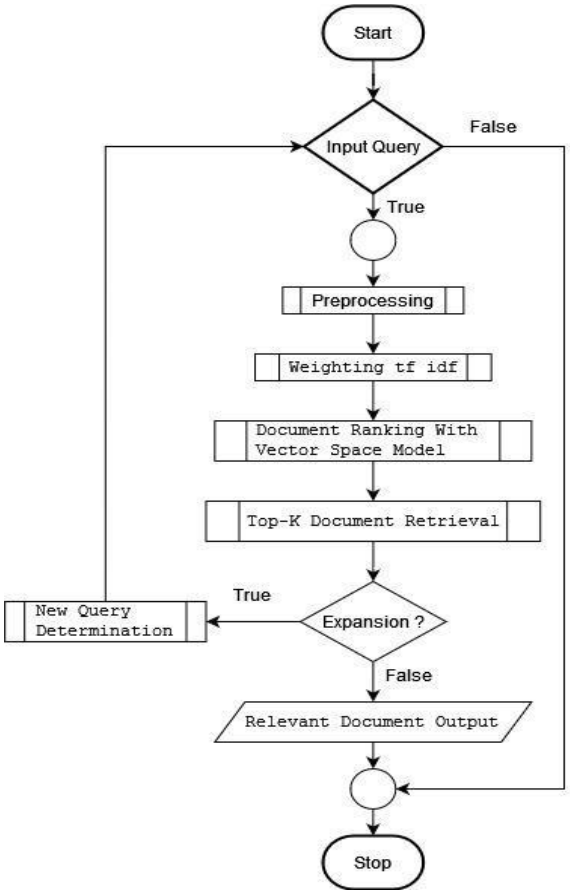

Figure 3.2 Flow Chart Design

3.2. IMPLEMENTATION

The Information Retrieval application in the Bukhari Hadith consists of several main processes, namely preprocessing, ranking documents and determining new queries. Data in the form of hadith will be processed by preprocessing to obtain a unique root word called term, then determine the relevant document rankings using the Vector Space Model, then the determination of new queries using the Pseudo Relevance feedback method. The stages of Information Retrieval using the Vector Space Model and Expansion Query methods are as follows:

3.3. FORMULATION OF THE PROBLEM

There are many hadith narrators who become a reference for Muslims, some of them are Muslim Imam, Imam Ahmad and Imam Bukhari. Imam Bukhari narrated all his traditions in the book which he called Sahih Bukhari. For Muslims, Shahih Bukhari is the main reference to deepen As-Sunnah as the second source of Islamic law after the Koran. The best compilation book and the most beautiful writer, the most correct and the least wrong, the most widely used and the most faidah and the easiest to store, the most acceptable to friends

and foes, and the most memorable in the hearts of special people and lay people are the book Shahih Al-Bukhari by Abu Abdillah Muhammad bin Isma'il Al-Bukhari then the work of Muslims bin Al-Hajjaj An-Naisaburi (Az-zabidi, 2016). As the development of digital technology, Hadith is packaged into digital form. Currently the Hadith can be read through computers, smartphones and other digital devices. On the smartphone there is a Hadith Sahih Bukhari v.3.6 application created by Islamic Developers. The application has a search feature but the search results are still less relevant because they still use conventional queries. We tried to find hadiths with the keyword "usury" but in the search results the keyword "geriba" also appeared so it does not match what we are looking for. This problem has been discussed in a previous study entitled Query Optimization for Data Search Using Sentence Decomposition (Sanjaya, 2016) but the method of decomposition of sentences still did not produce an optimal level of relevance, so that Information Retrieval was needed to support the search.

IV. RESULTS AND DISCUSSION

Query : Mahar pernikahan

In table 3.1 shows that without expansion the number of documents returned that are relevant to the query (tp) is 2 documents, while the documents that are not relevant (fp) are 0 documents. And for the number of documents that are not returned that is relevant to the query (fn) as many as 1 document, as many documents that are not relevant are (tn) 31 documents.

Whereas with the expansion of the number of documents returned that are relevant to the query (tp) as many as 26 documents, while the documents that are not relevant (fp) are as many as 0 documents. And for the number of documents that are not returned that are relevant to the query (fn) as many as 0 documents, as many as irrelevant documents are (tn) 7 documents. Then, the values of precision, recall, accuracy and F-measure for query 4 are:

Table 3.1 *Query* "Mahar pernikahan"

		VSM		*VSM + Expansion*	
		+	-	+	-
Retrieved	+	2 (TP)	0 (FP)	26 (TP)	0 (FP)
	-	1 (FN)	31 (TN)	0 (FN)	7 (TN)

***Without Expansion* :**

Precision = tp / (tp + fp)
= 2 / (2 + 0) = 2 / 2 = 1
= 1 / 1 * 100 = 100 %

Recall = tp / (tp + fn)
= 2/ (2 + 1) = 2 / 3 = 0.66
= 0.66 / 1 * 100 = 66 %

Akurasi = tp + tn / (tp + fp + tn + fn)
= 2 + 31/ (2 + 0 + 31 + 1) = 33 / 34 = 0.97
= 0.97 / 1 * 100 = 97 %

F-Measure = 2x *Recall* x *Precision* / *Recall* + *Precision*
= (2x0.66) x 1 / 0.66 + 1 = 1.32 / 1.66 = 0.79
= 0.79 / 1 * 100 = 79 %

***With Expansion* :**

Precision = tp / (tp + fp)
= 26 / (26 + 0) = 26 / 26 = 1
= 1 / 1 * 100 = 100 %

Recall = tp / (tp + fn)
= 26 / (26 + 0) = 26 / 26 = 1
= 1 / 1 * 100 = 100 %

Akurasi = tp + tn / (tp + fp + tn + fn)
= 26 + 7/ (26 + 0 + 7 + 0) = 33 / 33 = 1
= 1 / 1 * 100 = 100 %

F-Measure = 2x *Recall* x *Precision* / *Recall* + *Precision*
= (2 x 1) x 1 / 1 + 1 = 2 / 2 = 1
= 1 / 1 * 100 = 100 %

V. CONCLUSIONS

Based on the formulation of the problem, study review and research methodology making prototype of Information Retrieval application in the Hadith Sahih Bukhari with Vector Space Model Method and Expansion Query can help Muslims in finding relevant traditions in accordance with the search, the search results of the hadith search on digital text is more optimal compared to conventional query. Search results without using query expansion have good enough results with a minimum F-Measure value of 79%.

REFERENCES

[1] Adisantoso, J., Ridha, A. dan Agusetyawan, A. W. (2002) "RELEVANCE FEEDBACK PADA TEMU-KEMBALI TEKS BERBAHASA INDONESIA DENGAN METODE IDE-DEC-HI DAN IDE-REGULAR," hal. 1–8.

[2] Alfian Sukma1, Badrus Zaman, E. P. (2002) "Klasifikafi Dokumen Temu Kembali Informasi dengan K-Nearest Neghbour Information Retrieval Document Classified with K-Nearest Neighbor," *Mycological Research*, 106(11), hal. 1323–1330.

[3] Amin, F. (2012) "Sistem Temu Kembali Informasi dengan Metode Vector Space Model," *Jurnal Sistem Informasi Bisnis*, 2(2), hal. 78–83. doi: 10.21456/vol2iss2pp078-083.

[4] Az-zabidi (2016) "*Mukhtasar Shahih Bukhari*." Diedit oleh Ummul Qura. 2016.Baeza-Yates, R. dan Ribeiro-Neto, B. (1999) "*Modern Information System*. Pearson in." Boston. USA: Eddison Wesley (1718).

[5] Hidayat, W. (2013) "Indexing and Retrieval Engine untuk Dokumen Berbahasa Indonesia dengan Menggunakan Inverted Index," *Seminar Nasional Informatika dan Aplikasinya (SNIA) 2015*, (October). Tersedia pada: https://www.researchgate.net/publication/284492748.

[6] Karyono, G., Utomo, F. S., Sistem, A. dan Balik, T. (2012) "Temu Balik Informasi Pada Dokumen Teks Berbahasa Indonesia Dengan Metode Vector Space Retrieval Model," *Seminar Nasional Teknologi Informasi dan Terapan 2012*, 2012(Semantik), hal. 282–289.

[7] Kent, A. (1971) "*Information Analysis and Retrieval*. 3rd Editio." New York: Becker and Heys.

[8] Lancaster, F. W. (1979) "*Information Retrieval System: Characteristics, Testing, and Evaluation*. 2nd Editio." Diedit oleh J. Wiley. New York.

[9] Lubis, J. H. (2017) "Analisa Performansi Query Pada Database Smell," *Jurnal Manajemen dan Informatika Pelita Nusantara*, 21(1), hal. 42–49.

[10] Monica, M., Winarno, W. W. dan Sunyoto, A. (2016) "Information Retrieval Dokumen Tesis Untuk Mengetahui Kemiripan dengan Penelitian Yang Telah Ada," *Jurnal Transformasi*, 12(2), hal. 105–115.

[11] Nurul Justina Mahardianingroem, A. S. (2018) "PENERAPAN KAMUS DASAR PADA ALGORITMA PORTER UNTUK MENGURANGI KESALAHAN STEMMING BAHASA INDONESIA," 10(2), hal. 103–112.

[12] Pamungkas, Z. Y. dan Ridok, A. (2015) "QUERY EXPANSION PADA SISTEM TEMU KEMBALI INFORMASI DOKUMEN BERBAHASA INDONESIA MENGGUNAKAN PSEUDO RELEVANCE FEEDBACK Studi kasus: Perpustakaan Universitas Brawijaya."

[13] Salton, G. (1989) "*Automatic Text Processing: TheTransformation, Analysis, and Retrieval of Information by Computer*." USA: Addison-Wesley Longman Publishing Co., Inc.

[14] Sanjaya, A. (2016) "OPTIMASI QUERY UNTUK PENCARIAN DATA MENGGUNAKAN PENGURAIAN KALIMAT," hal. 6–7.

[15] Wibowo, A. (2012) "Peningkatan Performasi Sistem Temu Balik Informasi dengan Metode Phrasal Translation dan Query Expansion," *Batam: Politeknik Negeri Batam*. Tersedia pada: https://scholar.google.com/citations?user=QB4WJb4AAAAJ&hl=en#d=gs_md_cita-d&u=%2Fcitations%3Fview_op%3Dview_citation%26hl%3Den%26user%3DQB4WJb4AAAAJ%26citation_for_view%3DQB4WJb4AAAAJ%3A9yKSN-GCB0IC%26tzom%3D-420.

[16] Zain, M. Y. dan Suswati (2016) "Information Retrieval System Pada Pencarian File Dokumen Berbasis Teks Dengan Metode Vector Space Model Dan Algoritma ECS Stemmer," *Jurnal Insand Comtech*, 1(1), hal. 30–37