Indonesian Journal of Electrical Engineering and Computer Science
Vol. 18, No. 2, May 2020, pp. 1004~1014
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v18.i2.pp1004-1014

Weighted inverse document frequency and vector space model for hadith search engine

Septya Egho Pratama[1], Wahyudin Darmalaksana[2], Dian Sa'adillah Maylawati[3], Hamdan Sugilar[4], Teddy Mantoro[5], Muhammad Ali Ramdhani[6]
[1,3,6]Department of Informatics, UIN Sunan Gunung Djati Bandung, Indonesia
[2]Department of llmu Hadits, UIN Sunan Gunung Djati Bandung, Indonesia
[4]Department of Mathematic Education, UIN Sunan Gunung Djati Bandung, Indonesia
[5]Department of Computer Science, Sampoerna University, Indonesia

Article Info

Article history:

Received Aug 24, 2019
Revised Oct 25, 2019
Accepted Nov 11, 2019

Keywords:

Classification
Convolutional neural network
Deep learning
Glove
Indonesian language process
Natural language processing
Text mining

ABSTRACT

Hadith is the second source of Islamic law after Qur'an which make many types and references of hadith need to be studied. However, there are not many Muslims know about it and many even have difficulties in studying hadiths. This study aims to build a hadith search engine from reliable source by utilizing Information Retrieval techniques. The structured representation of the text that used is Bag of Word (1-term) with the Weighted Inverse Document Frequency (WIDF) method to calculate the frequency of occurrence of each term before being converted in vector form with the Vector Space Model (VSM). Based on the experiment results using 380 texts of hadith, the recall value of WIDF and VSM is 96%, while precision value is just around 35.46%. This is because the structured representation for text that used is bag of words (1-gram) that can not maintain the meaning of text well).

Corresponding Author:

Dian Sa'adillah Maylawati,
Department of Informatics,
UIN Sunan Gunung Djati Bandung,
Jl. A.H. Nasution 105, Bandung, 40614, Indonesia
Email: diansm@uinsgd.ac.id

1. INTRODUCTION

Hadith are all the words, deeds, decrees and approvals of the Prophet Muhammad which are made provisions or laws in Islam. Hadith is used as a source of law in Islam besides the Qur'an, *Ijma'* (the agreement of the scholars in establishing a legal law in religion based on the Qur'an and Hadith in a case that occurred) and *Qiyas* (establish a law for a new case that does not exist yet), where in this case, the position of the hadith is the second source of law after the Qur'an [1-5]. Studying and practicing the contents of the hadith content in daily life is highly important for Muslims [6]. However, many fake hadiths that appear, it is necessary to have a selective in studying hadith. Many weak and fake hadiths are circulating among Muslims because of the lack of selective nature in hearing the hadith, as a result there are irregularities in social life. It is necessary to study the hadith required by a more expert to explain the hadith and references that have been guaranteed correct.

Search engine technology as one of Information Technology implementation is a computer program that designed to search spesific data based on input keywords [7-9]. Most of the search engines that already exist and are widely used today provide the results of data acquisition that has been sorted based on the level of relevance of the keywords we input. Today, search engine technology is more than database query. To increase the level of relevance of data, search engines can not be separated from the Information Retrieval (IR) and Text Mining (TM). IR is related with TM method, either text classification or text clasterizationto find the best result based on input keywords [8, 10, 11]. Even, Google Search Engine [12], Google Scholar [13],

and another bussiness and marketing using Search Engine Optimization (SEO) [9, 14, 15], continues to grow and increasingly sophisticated. Not only use text data, but also use another unstructure data such as image and sound as a keywords.

Producing the best result from search engine is very related with the algorithm that used. In the previous result. There are so many IR technique for search engine research with various algorithms, such as Principal Component Analysis [16], Naive Bayes Classification [17], and the widely used is Vector Space Model (VSM) [18-23]. This algorithm is used to measure the similarity between a document and a query by weighting the words. The document is seen as a vector that has distance and direction [24]. In the VSM, a term is represented by a vector space dimension. The similarity presentation results from calculating the match between the vector of a query and a document that has previously undergone a process of TM and weighting of the word first. Word weighting is related to the workings of the VSM algorithm. Word weight is obtained from the calculation of the number of words contained in the document divided by the number of documents containing the word searched.

Commonly, word weighting and frequency is counted using Term Frequency and Inverse Document Frequency (TF-IDF) method. However, this research try to used Weighted Inverse Document Frequency (WIDF) instead TF-IDF. WIDF weighting is a development of TF-IDF, where the weakness of TF-IDF method is that all documents containing certain terms are treated the same as binary calculations, while the WIDF method adds frequency features and document collections [25]. WIDF word weight calculation is considered more specific because it counts all existing document collections while TF-IDF treats all existing documents with binary calculations (0 and 1) regardless of the number of times a document appears. Therefore, this research aims to utilize search engine technology using IR with WIDF as document frequency algorithm and VSM as vectorization method to search hadith document based on input keywords.

2 RESEARCH METHOD

Research activity of this research that provided in the Figure 1 is begin from preparing data collection of Hadith, implementing IR and TM technique that preparing text data which is unstructured into structured representation in text pre-processing process, then counting WIDF and conducting VSM algoritm. Next, the performance of WIDF and VSM algorithm is tested with some scenarios and the result is evaluated using Recall, Precision, and Accuracy value.

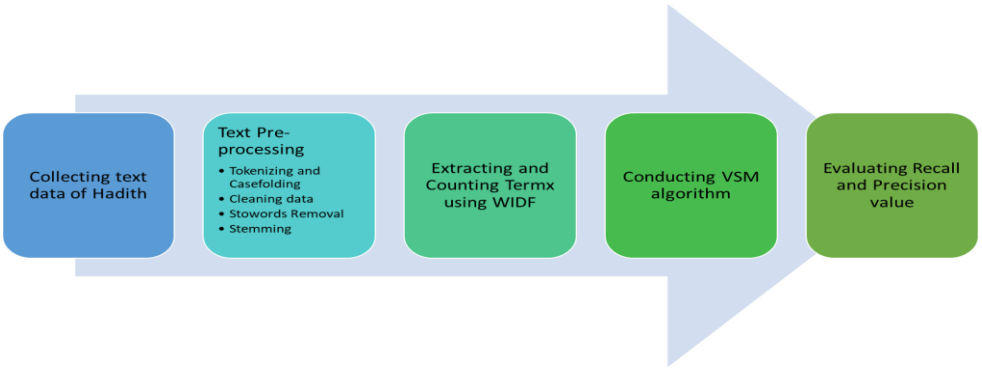

Figure 1. Research activities

2.1. Information Retrieval and Text Mining

Information Retrieval (IR) is a technique for finding relevant information according to the keywords entered [26-29]. While Text Mining (TM) is a technique for finding insight knowledge or important information from a collection of text documents [30-32]. Actually, IR and TM are very related, it can be said that IR is a part of TM, but IR is not yet TM. Because IR does not always implement a Data Mining (DM) technique such as classification or clustering [10]. However, the IR certainly does several TM stages, especially at the Pre-processing stage.

2.2. Weighted Inverse Document Frequency (WIDF)

In IR techniques that apply the concept of Text Mining, all the words that searched do not have the same weight. Giving a weight to a word is done by giving the frequency value of a word as a weight. The greater the appearance of words in the document will provide greater value relevance. The weighting method used in this study is Weighted Inverse Document Frequency (WIDF). The WIDF method is a development of the Term Frequency and Inverse Document Frequency (TF-IDF) method where the weakness

of the TF-IDF method is that all documents containing certain terms are treated the same as binary calculations, while the WIDF method adds frequency features and document collections [25, 27, 33, 34]. The formula for the WIDF method is shown in (1).

$$WIDF(d,t) = \frac{TF(d,t)}{\sum_i TF(i,t)} \quad (1)$$

Where *d* is a document collection, *t* is word or term, *i* is a related document. Then, *TF (d, t)* is the appearance of a word *(t)* in a document divided by *TF (i, t)*, which is the total number of words *(t)* in the related document *(i)*.

2.3. Vector Space Model (VSM)

Similarities between documents that defined based on *bag-of-words* representations in this research are converted to a vector space model. This model was introduced by Salton and has been used widely [35]. In VSM, each document in the database and input keywords are represented by a multi-dimensional vector or vector space dimension, where the dimensions correspond to the number of words in the document involved [23, 35, 36]. The document is represented as a vector that has distance and direction. In the VSM, a term is represented by a vector space dimension. A d_j document and a q query are represented as t-dimensional vectors as shown in Figure 2.

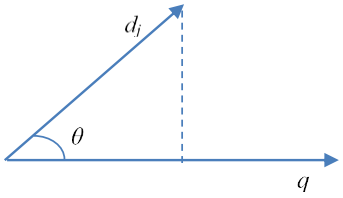

Figure 2. Vector representation

The VSM calculation process goes through the word weighting calculation stages, usually using the TF-IDF method. TF (Term Frequency) is the number of words appearing in a document while the IDF (Inverse Document Frequency) is the number of documents taken by the system where the term appears in it. However, in this research use WIDF. After that, calculate the length of each word weight in the query and document. Calculation of the length of the query and document weights using (2) and (3).

$$|q| = \sqrt{\sum_{j=1}^{t}(W_{iq})^2} \quad (2)$$

$$|d_j| = \sqrt{\sum_{t=1}^{t}(W_{ij})^2} \quad (3)$$

With $|q|$ is the length of the query, and W_{iq} is the i^{th} query weight of the document, so the length of the query ($|q|$) is calculated to get the length of the query from the document query weight (W_{iq}) called by the system. The length of a query can be calculated by the root equation of the number of squares of the query. With $|d_j|$ is the length of the document, and W_{ij} is the weight of the i document, then the length of the document ($|d_j|$) is calculated to get the length of the document from the weight of the document (W_{ij}) called by the system. The length of a document can be calculated by the root equation of the number of squares of the document. Calculation of measurement of the similarity of query documents (inner product), using (4). Similarity between query and document or Sim(q, d_j) is directly proportional to the number of query weights (q) multiplied by document weight (d_j) and inversely proportional to the root of the number of squares q ($|q|$) multiplied by the root of the number of squares of the document ($|d_j|$). Similarity calculations produce document weights that are close to value 1 or produce document weights that are greater than the values generated from inner product calculations.

$$Sim(q,d_j) = \frac{q.d_j}{|q|.|d_j|} = \frac{\sum_{i=1}^{t} W_{iq}.W_{ij}}{\sqrt{\sum_{j=1}^{t}(W_{iq})^2}.\sqrt{\sum_{t=1}^{t}(W_{ij})^2}} \quad (4)$$

2.4. Recall and Precision Evaluation

Recall and Precision evaluation aims to obtain information on search results obtained by the system. Precision is the level of accuracy or relevancy between the information that requested with input keywords and the informations that given by the system, while the Recall value is the level of success of the system in finding back information [37-39].

$$R = \frac{Number\ of\ relevant\ items\ retrieved}{Total\ number\ of\ relevant\ items\ in\ collection} \quad (5)$$

$$P = \frac{Number\ of\ relevant\ items\ retrieved}{Total\ number\ of\ items\ retrieved} \quad (6)$$

Based on (5) and (6), R being Recall, where R value is obtained by comparing the Number of relevant items retrieved with the Total number of relevant items in the collection. Recall is a document that is called from the system according to user requests that follow the pattern of the system. The greater Recall value cannot be said of a good system or not. Then, P being Precision, where P value is obtained by comparing the Number of relevant items retrieved with the Total number of items retrieved. Precision is the number of documents that are called from the relevant database after being assessed by the user with the required information. The greater the value of precision, the system can be said to be good.

3 RESULT AND ANALYSIS

In this section, the results of this study based on the research activities is presented as it shown in Figure 1, and the comprehensive analysis is also discussed.

3.1. Collecting Text Data of Hadith

The data used in this research is in the form of Indonesian translation hadith text data that obtained from the Book of Bulughul Maram [40], which contains about 1,596 hadiths (but in this research only use 380 hadiths). The book of Bulughul Maram is a thematic hadith book containing the hadiths which are used as sources of Islamic law making by *fiqh* experts, especially from the *Imam Shafi'i* and written by *Ibn Hajar Al-Aqsalani* based on his memorization without looking at the original book. This book includes the book of fiqh which received global recognition and is also widely translated throughout the world in the form of books and ebooks. So far, the Book of Bulughul Maram is only in the book or ebook version, there is no information system that can collect the data of the hadith, making it difficult for us to search for the hadiths that we want based on certain keywords.

3.2. Text Pre-processing

Test pre-processing is an important phase in Text Mining to prepare text data well befor conducting the mining process [41, 42], among others tokenizing, casefolding, cleaning text data, stopwords removal and stemming. Tokenizing and casefolding prepare text data to be easy to change into structured representation with specific and uniform term. Stopwords removal can reduce the dimension of text data with remove all an unimportant words. While stemming process is also important pre-processing phase. For Indonesian language, stemming process can be maintain the meaning of text well, because the word with affixes is a verbs that contain the maning of text [43, 44]. Even tough, in several research in text mining, the stemming process does not give a big effect in accuracy [45]. Stemming process is depend on the language, from many Indonesian stemming algorithm [46-49], this research use an improved Porter algorthm that modified based on Indonesian language [50]. The example result of text pre-processing is available in Table 2 which is pre-processing result from the text hadits example from Table 1 that provided in Indonesian Language.

3.3. Analysis of Weighted Inverse Document Frequency

After text pre-processing, as bag-of-words representation, every word is a term. Using (1), frequency of each term is calculated. The example result of WIDF for hadith text is available in the Table 3 which is counted from the terms in from the result of text pre-processing in Table 2. In addition to documents, word weight calculations are also performed on keywords, of course before the text mining process is performed first. Because in keywords, the word used is already in a simple form, then just calculate the weight of the word. For example, the keyword is "sedekah", with the same process the WIDF value of "sedekah" is 0.125.

Table 1. The Example of Hadith Data Text

Number	Hadith Text
1	Dari Abu Hurairah Radliyallaahu ‘anhu bahwa Rasulullah Shallallaahu ‘alaihi wa Sallam pernah ditanya: Wahai Rasulullah Shallallaahu ‘alaihi wa Sallam, sedekah apakah yang paling mulia? Beliau menjawab: “Sedekah orang yang tak punya, dan mulailah memberi sedekah atas orang yang banyak tanggungannya. Dikeluarkan oleh Ahmad dan Abu Dawud.
2	Dari Abu Said Al-Khudry Radliyallaahu ‘anhu bahwa Zainab, istri Abu Mas’ud, bertanya: Wahai Rasulullah, baginda telah memerintahkan untuk bersedekah hari ini, dan aku mempunyai perhiasan urge yang hendak saya sedekahkan, namun Ibnu Mas’ud menganggap bahwa dirinya dan anaknya lebih berhak untuk aku beri sedekah. Lalu Nabi Shallallaahu ‘alaihi wa Sallam bersabda: “Ibnu Mas’ud memang benar, suamimu dan anakmu adalah orang yang lebih berhak untuk engkau beri sedekah.” Riwayat Bukhari.
3	Dari Samurah Ibnu Jundab Radliyallaahu ‘anhu bahwa Rasulullah Shallallaahu ‘alaihi wa Sallam bersabda: “Meminta-minta adalah cakaran seseorang terhadap mukanya sendiri, kecuali meminta kepada penguasa atau karena suatu hal yang amat perlu.” Hadits shahih riwayat Tirmidzi.
4	Dari Abu Hurairah Radliyallaahu ‘anhu bahwa Rasulullah Shallallaahu ‘alaihi wa Sallam bersabda: “Umrah ke umrah menghapus dosa antara keduanya, dan tidak ada pahala bagi haji mabrur kecuali urge.” Muttafaq Alaihi.
5	Dari Uqbah Ibnu Amir bahwa dia mendengar Rasulullah Shallallaahu ‘alaihi wa Sallam bersabda: “Setiap orang bernaung di bawah sedekahnya sehingga ia diputuskan amal perbuatannya antara manusia.” Riwayat Ibnu Hibban dan Hakim.

Table 2. The Example of Text Pre-Processing Result from the Hadith Text in Table 1

Document 1 (*d1*)	Document 2 (*d2*)	Document 3 (*d3*)	Document 4 (*d4*)	Document 5 (*d1*)
abu	abu	samurah	abu	uqbah
hurairah	said	ibnu	hurairah	ibnu
radiyallaahuanhu	alkhudry	jundab	radiyallaahuanhu	amir
rasulullah	radiyallaahuanhu	radiyallaahuanhu	rasulullah	dengar
shallallaahualaihi	zainab	rasulullah	shallallaahualaihi	rasulullah
sallam	istri	shallallaahualaihi	sallam	shallallaahualaihi
...	...	...	...	...
jawab	sedekah	diri	tidak	amal
sedekah	hari	kecuali	ada	buat
orang	punya	minta	pahala	manusia
punya	perhiasan	penguasa	haji	riwayat
mulai	sedekah	hal	mabrur	ibnu

Table 3. The Example of WIDF Calculation for Hadith Text

Word (*t*)	*TF(d,t)*					*TF(i,t)*	*TF(d,t)/TF(i,t)*				
	d1	*d2*	*d3*	*d4*	*d5*		*d1*	*d2*	*d3*	*d4*	*d5*
abu	2	2	0	1	0	5	0,4	0,4	0	0,2	0
hurairah	1	0	0	1	0	2	0,5	0	0	0,5	0
radiyallaahuanhu	1	1	1	1	0	4	0,25	0,25	0,25	0,25	0
rasulullah	2	1	1	1	1	6	0,333	0,167	0,167	0,167	0,167
shallallaahualaihi	2	1	1	1	1	6	0,333	0,167	0,167	0,167	0,167
sallam	2	0	1	1	1	5	0,4	0	0,2	0,2	0,2
tanya	1	1	0	0	0	2	0,5	0,5	0	0	0
sedekah	**3**	**4**	**0**	**0**	**1**	**8**	**0,375**	**0,5**	**0**	**0**	**0,125**
...	...	...	...	...	...	...	...	...	...	...	...
mulia	1	0	0	0	0	1	1	0	0	0	0
jawab	1	0	0	0	0	1	1	0	0	0	0
orang	2	1	1	0	1	5	0,4	0,2	0,2	0	0,2
punya	1	1	0	0	0	2	0,5	0,5	0	0	0
mulai	1	0	0	0	0	1	1	0	0	0	0
beri	1	2	0	0	0	3	0,333	0,667	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...
hadits	1	0	1	0	0	2	0,5	0	0,5	0	0
shahih	1	0	1	0	0	2	0,5	0	0,5	0	0
ibnu	1	2	1	0	2	6	0,167	0,333	0,167	0	0,333
hibban	1	0	0	0	1	2	0,5	0	0	0	0,5
hakim	1	0	0	0	1	2	0,5	0	0	0	0,5

3.4. Analysis of Vector Space Model

After weighting the words, the next step is to calculate the match between the keywords and documents using the VSM algorithm. The first step in this algorithm is to calculate the multiplication of the weight scale by multiplying the keyword weights with the document and then adding them up. The next step is to calculate the length of the vector by squaring all the word weights and adding them together. The final step is to find a match/similarity value between the keyword and the document by dividing the value of the

scalar weight multiplication by the length of the vector. Examples of results from scalar multiplication and vector length calculations are in Table 4 (the final result value has been rounded). The results are obtained from calculations using (2), (3), and (4). For counting similarity value between input keyword and document is calculated use Cosine Similarity with (4). The similarity value is provided in Table 5 with the highest order value is Document 2, followed by Document 1 and Document 5. So, system will be produced Document 1, 2, and 5 that related with keyword "sedekah".

Table 4. The Example of Scalar Multiplication and Vector Length in VSM

Word (*t*)	Scalar Multiplication											Vector Lenght Calculation					
	Weight *(W)*						*(d,q)*					W^2					
	d1	*d2*	*d3*	*d4*	*d5*	*q*	*d1*	*d2*	*d3*	*d4*	*d5*	*d1*	*d2*	*d3*	*d4*	*d5*	*q*
abu	0,4	0,4	0	0,2	0	0	0	0	0	0	0	0,16	0,16	0	0,04	0	0
hurairah	0,5	0	0	0,5	0	0	0	0	0	0	0	0,25	0	0	0,25	0	0
radiyallaah uanhu	0,3	0.3	0,3	0,3	0	0	0	0	0	0	0	0,06	0,06	0,06	0,06	0	0
rasulullah	0,3	0,2	0,2	0,2	0,2	0	0	0	0	0	0	0,11	0,03	0,03	0,03	0,03	0
shallallaah ualaihi	0,3	0,2	0,2	0,2	0,2	0	0	0	0	0	0	0,11	0,03	0,03	0,03	0,03	0
sallam	0,4	0	0,2	0,2	0,2	0	0	0	0	0	0	0,16	0	0,04	0,04	0,04	0
tanya	0,5	0,5	0	0	0	0	0	0	0	0	0	0,25	0,25	0	0	0	0
sedekah	**0,4**	**0,5**	**0**	**0**	**0,1**	**0,1**	**0,1**	**0,1**	**0**	**0**	**0**	**0,14**	**0,25**	**0**	**0**	**0,01**	**0**
...	...	...	...	...	...	...	...	...	..	..	...	...	...	...	...	...	...
mulia	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
jawab	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
orang	0,4	0,2	0,2	0	0,2	0	0	0	0	0	0	0,16	0,04	0,04	0	0,04	0
punya	0,5	0,5	0	0	0	0	0	0	0	0	0	0,25	0,25	0	0	0	0
mulai	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
beri	0,3	0,7	0	0	0	0	0	0	0	0	0	0,11	0,45	0	0	0	0
...	...	...	...	...	...	...	...	...	..	..	...	...	...	...	...	...	...
hadits	0,5	0	0,5	0	0	0	0	0	0	0	0	0,25	0	0,25	0	0	0
shahih	0,5	0	0,5	0	0	0	0	0	0	0	0	0,25	0	0,25	0	0	0
ibnu	0,2	0,3	0,2	0	0,3	0	0	0	0	0	0	0,03	0,11	0,03	0	0,11	0
hibban	0,5	0	0	0	0,5	0	0	0	0	0	0	0,25	0	0	0	0,25	0
hakim	0,5	0	0	0	0,5	0	0	0	0	0	0	0,25	0	0	0	0,25	0
...	...	...	...	...	...	...	...	...	..	..	...	...	...	...	...	...	...
$\sum W_{d.q}$							0,1	0,1	0	0	0						
$\sum_w \mathbf{2}$												10,794	15,047	8,396	9,761	9,936	0,016
$\sqrt{\sum_w \mathbf{2}}$												3,285	3,879	2,898	3,124	3,152	0,125

Table 5. The Similarity Value of the Example Hadith Documents

Term	$\sqrt{\sum_w 2}$	$\sum W_{d.q}$	$\frac{\sum W_{d.q}}{\sqrt{\sum(W_q)^2} \cdot \sqrt{\sum(W_d)^2}}$
d1	3,285	0,047	$\frac{0,047}{0,125 \,.\, 3,285} = 1,232$
d2	3,879	0,063	$\frac{0,063}{0,125 \,.\, 3,879} = 1,939$
d3	2,898	0	$\frac{0}{0,125 \,.\, 2,898} = 0$
d4	3,124	0	$\frac{0}{0,125 \,.\, 3,124} = 0$
d5	3,152	0,016	$\frac{0,016}{0,125 \,.\, 3,152} = 0,393$

3.5. Result of Hadith Search Engine Testing

The experiment is conducted using 380 of Hadith text data from the Book of Bulughul Maram. Figure 3 shows the implementation of Hadith search engine with web-based system. There are 5 keywords with different number of term for testing scenario that conducted to evaluate the output of system, among others "sedekah", "zakat fitrah", "zakat harta rikaz", "sedekah hutang anak yatim", and "zakat hewan ternak dan tanaman". The result of those experiments use Recall and Precision value that provides in the Table 6 and Figure 4. The analysis and evaluation of exeriment result is explained in Section 4.6.

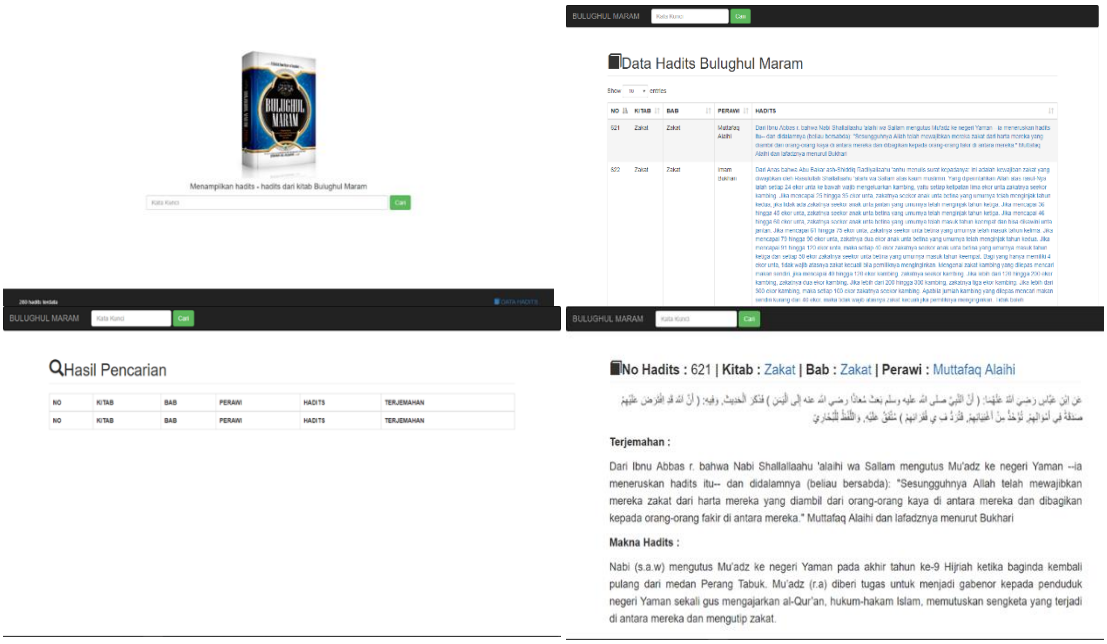

Figure 3. The example of Hadith search engine implementation

Table 6. The Experiment Result of Hadith Search Engine

Total of	Total Hadith 1st Testing "Sedekah"	2nd Testing "Zakat Fitrah"	3rd Testing "Zakat Harta Rikaz"	4th Testing "Sedekah Hutang Anak Yatim"	5th Testing "Zakat Hewan ternak dan Tanaman"
Data called	12	35	58	44	51
Relevant data that is called	12	5	3	11	15
Irrelevant data called	12	30	55	33	36
Relevant data that is not called	3	0	0	0	0
The sum of all relevant data	15	5	3	11	15
Recall	80%	100%	100%	100%	100%
Precision	100%	14,28%	8,62%	25%	29,41%

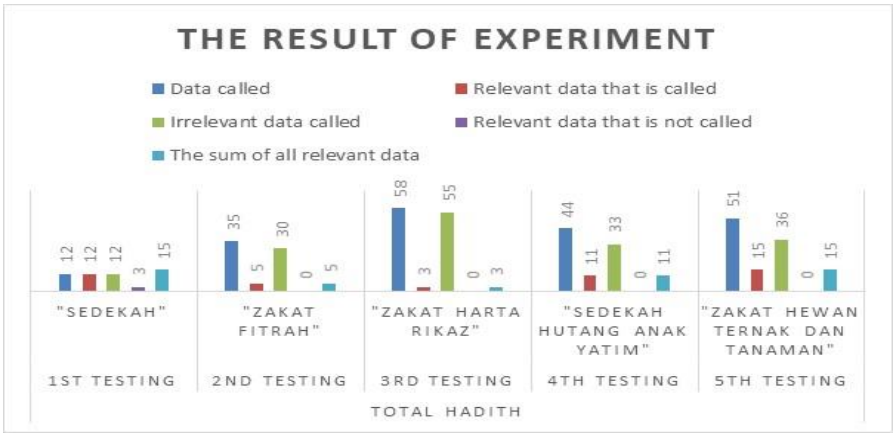

Figure 4. Graphics of experiment result

3.6. Analysis and Evaluation of Experiment and Testing Result

Based on the experiment result, this research found that:

1. Hadith search engine using WIDF and VSM success in finding the information (in this case is hadiths) in accordance with the input keywords well. It is proven by the average of Recall value which is quite high

around 96% from 5 types of experiment. While, the relevancy of hadiths that produced is low, because the average of Precision value is small, just around 35.46%.

2. The small Precision value that show that result of hadiths from search engine most of then are not relevant or not related with the purpose of the hadith that the user is looking for. This can be caused by the following things:
   a. The search engine is used bag-of-word as structured text representation. Bag-of-word representation is a structured representation from text, where text data is represented by collection of word, one word one term [51, 52]. So that, if the keyword that more than one word, it will be separated per word and not in n-gram form. This analysis is supported by the result of precision that the low precison value occurs in the input keywords with n-term/n-gram. For example, if the input keyword is “zakat fitrah”, the system will produce the hadiths that contain “zakat” and “fitrah”, whereas “zakat fitrah” is a single entity, so when there is a hadith that contains the word "fitrah" even though it is not related to zakat it will still be produced.
   b. Indonesian language is unique, still a lots of problem in Indonesian language besides punctuation, abbreviation, and character in the bracket, one of them in word with affixes [53]. The low precison value can be affected by the result of stemming process. There are letters that melt when given certain affixes, including "k", "p", "t", and "s" [54, 55]. So, if there is a word "purify yourself" when done stemming it will become "pure self", so the search engine should produce traditions related to purification. Back to the analysis in point 1, search engines will produce traditions that contain "holy" and "self", which are not all in accordance with the intent of the keywords entered, including traditions that only contain the word "self." However, if the process stemming is removed, the traditions will produce traditions containing "purify" and "self". There will be traditions related to purification that are produced and relevant, but search engines will not produce traditions related to purification that do not contain the word "purify". For example the traditions about purification but contain words other than "purify", such as containing the words "sanctified", "purified", even "holy" itself will not be displayed, so that more traditions are wasted if the stemming process is not done.
3. Poor results in terms of the precision or relevance of the information that produceed does not mean that the WIDF and VSM algorithms are wrong, However, it is because the text representation that used (bag-of word or 1-gram) is not good in maintaining the meaning of text documents. Nowadays, there are many Text Mining researches that prove and use multiple of words or n-grams that can maintain the meaning of text better [56-61]. Even Google search engines that implement IR and TM technology do not use 1-gram, because when we search for the word "Information Retrieval" (without quoting as input keywords) on Google search engines, the information that will be generated is related to "Information", "Retrieval" and "Information Retrieval".

4 CONCLUSION

This study use Weighted Inverse Document Frequency and Vector Space Model as algorithms to build search engine of Hadith which is an important source for Muslim. Based on the experiments that were conducted, this study has high Recall that means success to produce information of Hadiths in accordance with input keywords. However, the precision value is small that means most of the information that is produced are not relevant with the intent of the keywords entered. It is caused by several factors, which not due to the algorithm used, but due to the text representation used. Therefore, for the further research, besides hadith data that needs to be completed so that it is rich in information, the text representation used is also better which includes multiple of words to maintain the meaning of better and the resulting hadith information is more relevant. In addition, further research is also needed related to the effect of WIDF on search results. besides that algorithms can also be applied in addition to VSM for the search engine hadith.

ACKNOWLEDGEMENT

Authors wishing to acknowledge Research and Publication Centre of UIN Sunan Gunung Djati Bandung that supports and funds this research publication.

REFERENCES

[1] A. C. Muna, “Perkembangan Studi Hadits Kontemporer [Development of Contemporary Hadith Studies],” *Religia*, vol. 14, no. 2, 2012.
[2] Mardani, *Hukum Islam; Pengantar Ilmu Hukum Islam di Indonesia [Islamic law; Introduction to Islamic Law in Indonesia]*. Yogyakarta: Pustaka Pelajar, 2015.

[3] Rohidin, *Pengantar Hukum Islam (Dari Semenanjung Arabia Sampai Indonesia) [Introduction to Islamic Law (From the Arabian Peninsula to Indonesia)]*, 1st ed. Yogyakarta: Lintang Rasi Aksara Books, 2016.
[4] F. Djamil, *Filsafat Hukum Islam [Philosophy of Islamic Law]*. Jakarta: Logos Wacana Ilmu, 1997.
[5] 'Abd al-Wahab Khallaf, *Ilm Usul al-Fiqh*. Kairo: Dar Al-Hadith, 2003.
[6] A. Wahyudi, "Mengurai Peta Kitab-Kitab Hadits (Kajian Referensi atas Kitab-kitab Hadits) [UNDERSTANDING THE MAP OF THE BOOKS OF HADITS (Reference Study of the Books of Hadith)]," *AL-IHKAM J. Huk. Pranata Sos.*, 2015.
[7] P. W. Handayani, I. M. Wiryana, and J.-T. Milde, "Mesin Pencari Berbasiskan Semantik Untuk Bahasa Indonesia [Searching Machine Based On Semantics For Indonesian Languages]," *Jurnal Sistem Informasi*, vol. 4, no. 2. pp. 110–114, 2012.
[8] J. M. Kassim and M. Rahmany, "*Introduction to Semantic Search Engine*," 2009 Int. Conf. Electr. Eng. Informatics, vol. 02, no. August, pp. 380–386, 2009.
[9] J. B. Killoran, "How to use search engine optimization techniques to increase website visibility," *IEEE Trans. Prof. Commun.*, vol. 56, no. 1, pp. 50–66, 2013.
[10] S. M. Weiss, N. Indurkhya, T. Zhang, and F. J. Damerau, "Information Retrieval and Text Mining," *Springer Berlin Heidelb.*, no. Fundamentals of Predictive Text Mining, pp. 75–90, 2010.
[11] C. D. Manning, P. Ragahvan, and H. Schutze, *An Introduction to Information Retrieval*, no. c. 2009.
[12] P. R. Agrawal, "Google Search," 2016.
[13] C. C. Brown, "Google Scholar," *Charlest. Advis.*, 2017.
[14] A. Hassan and S. S. Dadwal, "Search Engine Marketing," in *Digital Marketing and Consumer Engagement*, 2017.
[15] A. A. Maarif, "Penerapan Algoritma TF-IDF untuk Pencarian Karya Ilmiah [Application of TF-IDF Algorithm for Scientific Work Search]," *Dok. Karya Ilm. | Tugas Akhir | Progr. Stud. Tek. Inform. - S1 | Fak. Ilmu Komput. | Univ. Dian Nuswantoro Semarang*, 2015.
[16] A. A. Okfan Rizal Ferdiansyah, Ema Utami, "Implementasi Principal Component Analysis Untuk Sistem Temu Balik Citra Digital [Implementation of Principal Component Analysis for Digital Image Retrieval Systems]," *Creat. Inf. Technol. J.*, vol. 2, no. 3, 2015.
[17] C. Slamet, R. Andrian, D. S. Maylawati, W. Darmalaksana, and M. A. Ramdhani, "Web Scraping and Naïve Bayes Classification for Job Search Engine," vol. 288, no. 1, pp. 1–7, 2018.
[18] F. Amin, "Sistem Temu Kembali Informasi dengan Pemeringkatan Metode Vector Space Model [Information Retrieval System with Vector Space Model Ranking Method]," *J. Teknol. Inf. Din.*, vol. 18, no. 2, pp. 122–129, 2013.
[19] G. Karyono, F. S. Utomo, A. Sistem, and T. Balik, "Temu Balik Informasi Pada Dokumen Teks Berbahasa Indonesia Dengan Metode Vector Space Retrieval Model [Information Retrieval in Indonesian Language Text Documents Using the Vector Space Retrieval Model]," *Semin. Nas. Teknol. Inf. dan Terap. 2012*, vol. 2012, no. Semantik, pp. 282–289, 2012.
[20] F. Sanjaya, "Pemanfaatan Sistem Temu Kembali Informasi dalam Pencarian Dokumen Menggunakan Metode Vector Space Model [Utilization of Information Retrieval System in Finding Documents Using the Vector Space Model Method]," *J. Inf. Technol.*, 2018.
[21] P. E. Mas'udia, M. D. Atmadja, and L. D. Mustafa, "Information Retrieval Tugas Akhir Dan Perhitungan Kemiripan Dokumen Mengacu Pada Abstrak Menggunakan Vector Space Model [Information Retrieval Of Final Project And Calculation Of Reflecting Documents In Abstract Using Vector Space Model]," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, 2017.
[22] I. Irmawati, "Information Retrieval in Documents using Vector Space Model," *J. Ilm. FIFO*, 2017.
[23] C. Van Gysel, M. de Rijke, and E. Kanoulas, "Learning Latent Vector Spaces for Product Search," 2016.
[24] T. Nadu, "Text Processing In Information Retrieval System Using Vector Space Model," no. 978, pp. 0–5, 2014.
[25] D. Susandi and U. Sholahudin, "Pemanfaatan Vector Space Model pada Penerapan Algoritma Nazief Adriani, KNN dan Fungsi Similarity Cosine untuk Pembobotan IDF dan WIDF pada Prototipe Sistem Klasifikasi Teks Bahasa Indonesia [Utilization of Vector Space Model in the Application of Nazief Adriani, KNN and Similarity Cosine Functions for IDF and WIDF Weighting in the Indonesian Text Classification System Prototype]," vol. 3, no. 1, pp. 22–29, 2016.
[26] C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*. 2008.
[27] A. M. Siregar and A. Puspabhuana, "*Improvement of term weight result in the information retrieval systems*," in Proceedings of 2017 4th International Conference on New Media Studies, CO*NMEDIA 2017*, 2018.
[28] F. Nadirman, A. Ridha, and A. Annisa, "Searching and Visualization of References in Research Documents," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, 2014.
[29] Y. Wang, "Design of Information Retrieval System Using Rough Fuzzy Set," *TELKOMNIKA Indones. J. Electr. Eng.*, 2014.
[30] H. Jiawei, M. Kamber, J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2006.
[31] Y. E. Zohar, "Introduction to Text Mining," *Automated Learning Group, University of Illinois*, 2002. [Online]. Available: http://www.docstoc.com/docs/25443990/Introduction-to-TextMining.
[32] I. H. Witten, "Text mining," in *The Practical Handbook of Internet Computing*, 2004.
[33] T. Tokunaga, T. Tokunaga, I. Makoto, and I. Makoto, "Text categorization based on weighted inverse document frequency," *Spec. Interes. Groups Inf. Process Soc. Japan (SIG-IPSJ*, 1994.
[34] Kurniawati and A. Syauqi, "*Term weighting based class indexes using space density for Al-Qur'an relevant meaning ranking*," in 2016 International Conference on Advanced Computer Science and Information Systems,

ICACSIS 2016, 2017.
[35] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, 1975.
[36] C. Slamet, A. R. Atmadja, D. S. Maylawati, R. S. Lestari, W. Darmalaksana, and M. A. Ramdhani, "*Automated Text Summarization for Indonesian Article Using Vector Space Model*," IOP Conf. Ser. Mater. Sci. Eng., vol. 288, no. 1, pp. 0–6, 2018.
[37] L. T. Su, "The relevance of recall and precision in user evaluation," *J. Am. Soc. Inf. Sci.*, 1994.
[38] L. Torgo and R. Ribeiro, "Precision and recall for regression," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2009.
[39] M. Junker, R. Hoch, and A. Dengel, "*On the evaluation of document analysis components by recall, precision, and accuracy*," in Proceedings of the International Conference on Document Analysis and Recognition, *ICDAR*, 1999.
[40] I. H. Al-Asqalani, *Bulughul Al-Maram, Terjemah oleh A.Hasan*. Bangil: Pustaka Tamam, 1997.
[41] S. Vijayarani, J. Ilamathi, and M. Nithya, "Preprocessing Techniques for Text Mining - An Overview," *Int. J. Comput. Sci. Commun. Networks*, vol. 5, no. 1, pp. 7–16, 2015.
[42] D. S. Maylawati, H. Aulawi, and M. A. Ramdhani, "Flexibility of Indonesian text pre-processing library," *Indones. J. Electr. Eng. Comput. Sci.*, 2019.
[43] T. Mardiana, T. Bharata Adji, and I. Hidayah, "Stemming Influence on Similarity Detection of Abstract Written in Indonesia," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, 2016.
[44] A. S. Rizki, A. Tjahyanto, and R. Trialih, "Comparison of stemming algorithms and its effect on Indonesian text processing," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, 2019.
[45] A. F. Hidayatullah, C. I. Ratnasari, and S. Wisnugroho, "Analysis of Stemming Influence on Indonesian Tweet Classification," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, 2016.
[46] J. Asian, H. E. Williams, and S. M. M. Tahaghoghi, "*Stemming Indonesian*," in Conferences in Research and Practice in Information Technology Series, 2005.
[47] M. Adriani, J. Asian, S. M. M. T. Nazief, and H. Williams, "Stemming Indonesian: A Confix-stripping approach," *ACM Trans. Asian Lang. Inf. Process.*, vol. 6, no. 1, pp. 1–33, 2007.
[48] L. Agusta, "Perbandingan Algoritma Stemming Porter Dengan Algoritma Nazief & Adriani Untuk Stemming Dokumen Teks Bahasa Indonesia," *Konf. Nas. Sist. dan Inform. 2009*, 2009.
[49] R. Setiawan, A. Kurniawan, W. Budiharto, I. H. Kartowisastro, and H. Prabowo, "Flexible Affix Classification for Stemming Indonesian Language," in *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 2016.
[50] D. S. Maylawati, W. B. Zulfikar, C. Slamet, and M. A. Ramdhani, "*An Improved of Stemming Algorithm for Mining Indonesian Text with Slang on Social Media*," in The 6th International Conference on Cyber and IT Service Management (CITSM 2018), 2018.
[51] H. M. Wallach, "Topic Modeling: Beyond Bag-of-Words," *ICML*, no. 1, pp. 977–984, 2006.
[52] D. Sa'Adillah Maylawati, M. Irfan, and W. Budiawan Zulfikar, "*Comparison between BIDE, PrefixSpan, and TRuleGrowth for Mining of Indonesian Text*," in Journal of Physics: Conference Series, 2017, vol. 801, no. 1.
[53] S. J. Putra, M. N. Gunawan, I. Khalil, and T. Mantoro, "Sentence boundary disambiguation for Indonesian language," pp. 587–590, 2018.
[54] Pusat Bahasa Kemdikbud, "Kamus Besar Bahasa Indonesia (KBBI)," *Kementeri. Pendidik. dan Budaya*, 2016.
[55] E. Setiawan, "KBBI - Kamus Besar Bahasa Indonesia [Indonesian Dictionary]," *Kamus Besar Bahasa Indonesia (KBBI)*, 2019.
[56] D. S. Maylawati and G. A. P. Saptawati, "*Set of Frequent Word Item sets as Feature Representation for Text with Indonesian Slang*," in International Conference on Computing and Applied Informatics, 2016, pp. 1–6.
[57] S. Alias, S. K. Mohammad, G. K. Hoon, and T. T. Ping, "A text representation model using Sequential Pattern-Growth method," *Pattern Anal. Appl.*, vol. 21, no. 1, pp. 233–247, 2018.
[58] H. Ahonen-Myka, "Finding All Maximal Frequent Sequences in Text," *Proc. ICML Work. Mach. Learn. Text Data Anal.*, pp. 11–17, 1999.
[59] H. Ahonen-Myka, "Discovery of Frequent Word Sequences in Text," *Proc. ESF Explor. Work. Pattern Detect. Discov.*, vol. {LNCS} (24, no. Teollisuuskatu 23, pp. 180–189, 2002.
[60] R. A. García-Hernández and Y. Ledeneva, "*Word sequence models for single text summarization*," Proc. 2nd Int. Conf. Adv. Comput. Interact. ACHI 2009, pp. 44–48, 2009.
[61] S. J. Putra, T. Mantoro, and M. N. Gunawan, "*Text mining for Indonesian translation of the Quran: A systematic review*," in *3rd* International Conference on Computing, Engineering, and Design, ICCED 2017, 2018.

BIOGRAPHIES OF AUTHORS

Septya Egho Pratama is a graduate student from Department of Informatics, UIN Sunan Gunung Djati Bandung.

Wahyudin Darmalaksana is an associate professor at Department Hadith, UIN Sunan Gunung Djati Bandung. His current research interests focus on Scientific Research and Publication in Islamic Higher Education and Hadith Science.

Dian Sa'adillah Maylawati is a lecturer in Department of Informatics at UIN Sunan Gunung Djati Bandung, Indonesia. Her current research interests focus on Software Engineering, Expert System, Text Mining, and Natural Language Processing. She takes Ph.D degree of Information and Communication Technology in Universiti Teknikal Malaysia Melaka (UTeM)

Hamdan Sugilar is a lecture in Department of Mathematic Education at UIN Sunan Gunung Djati Bandung. His current research interests focus on Education, especially Mathematic Education.

Teddy Mantoro is a Computer Science Professor in Sampoerna University, Jakarta. He obtained a PhD, an MSc and a BSc, all in Computer Science and his PhD was awarded from the School of Computer Science, the Australian National University (ANU), Canberra, Australia. He is a Senior Member of IEEE. His research interest is in information Security, pervasive/ ubiquitous computing, wireless sensor network, context aware computing, mobile computing and intelligent environment/ IoT.

Muhammad Ali Ramdhani is a Professor in Information Technology Research in Department of Informatics at the UIN Sunan Gunung Djati Bandung, Indonesia. His current research interests focus on Information System, Expert System, Decision Support System, Strategic Management, and Research Methodology.