American Journal of Software Engineering and Applications
2019; 8(1): 18-31
http://www.sciencepublishinggroup.com/j/ajsea
doi: 10.11648/j.ajsea.20190801.13
ISSN: 2327-2473 (Print); ISSN: 2327-249X (Online)

Informatian Retrieval for Popular Words in Bahasa Translation of Al Quran and Hadith Bukhori Using Enhance Confix Stripping (ECS) Stemming

Tristyanti Yusnitasari, Irfan Humaini, Lily Wulandari, Diana Ikasari

Faculty of Computer Science and Information Technology, Gunadarma University, Depok, Indonesia

Email address:
tyusnita@staff.gunadarma.ac.id (T. Yusnitasari), irfan_humaini @staff.gunadarma.ac.id (I. Humaini),
lily@staff.gunadarma.ac.id (L. Wulandari), d_ikasari@staff.gunadarma.ac.id (D. Ikasari)

To cite this article:
Tristyanti Yusnitasari, Irfan Humaini, Lily Wulandari, Diana Ikasari. Informatian Retrieval for Popular Words in Bahasa Translation of Al Quran and Hadith Bukhori Using Enhance Confix Stripping (ECS) Stemming. *American Journal of Software Engineering and Applications.* Vol. 8, No. 1, 2019, pp. 18-31. doi: 10.11648/j.ajsea.20190801.13

Received: October 11, 2018; **Accepted**: May 23, 2019; **Published**: August 15, 2019

Abstract: This paper discusses other ways that can be used to obtain information about information seeking for popular words in Qur’an and Hadith language translations. Popular words like the example of searching for the word "corruption" will be very difficult to find in translations of the Qur’an and Hadith. Information about popular words in the translation of the Qur’an and Hadith is needed so as to facilitate the search. In the research carried out was to obtain information about popular words first and then find the word synonym. This study uses Information Retrieval Technique which is a tokenizing process, stopword removal and stemming. The stemming method used in this study is ECS (Enhance Confix Stripping). Information retrieval is used to display some of the Qur'an and hadith that relate to the keywords searched for according to certain criteria. The dataset used in this study comes from the translation of the Ministry of Religion of the Republic of Indonesia.

Keywords: Informasi, Information Retrieval, Al Quran, Hadith, ECS

1. Introduction

Information retrieval has many search techniques to facilitate information retrieval. In the study described in this paper is about information seeking, with the object of research is the translation of language in the Qur'an and Hadith Bukhori. The Qur'an is the holy book of Muslims, the things contained in the Qur'an relate to faith, science, law, rules that govern the behavior and procedures of human life, stories of previous people, worship and monotheism. Hadith also includes guidelines for living in the teachings of Islam. Hadith according to the hadith provisions is what is based on the Prophet Muhammad SAW. It is the duty of Muslims to implement daily life based on the guidance of the Qur'an and Hadith.

There are many ways to learn what is in the Qur'an and Hadith, one of which is reading it, it is only human nature that wants to do things that are easy but are reluctant to read or for some reason time constraints. Because Muslims must learn or know what is contained in the Qur'an and Hadith, because in the Qur'an and Hadith it is clearly written that one of them is what is forbidden and not prohibited so that Muslims can live according to their instructions.

As a Muslim most know what is forbidden and what is permissible according to the Qur'an and Hadith. Not a few know, only hear that what is forbidden or whatever is not forbidden without knowing correctly that it is actually written in the Qur'an and Hadith. For example, almost all Muslims know that pigs are forbidden to be consumed by Muslims, but many do not know for sure that the ban is in the Qur'an and the Qur'an and the Hadith which states the prohibition on eating pork. Many other examples such as corruption, lying, gossiping (talking about other people) and so on. Limited time to find information about it is one of the reasons and difficulties in finding the desired words to be searched for in the Qur'an and Hadith because the Qur'an consists of 30 Juz (chapters), 114 Surahs and 6326 verses and the Bukhori Muslim Hadith as many as 7008 hadith numbers. Searching for popular words that are commonly used every day or in accordance with the desired theme will be very difficult. There are already several translations that index the contents of the Qur'an, and many software developers have developed digital Qur'an and digital Hadith. In some existing software, searching for information such as searching for pig words, the search results are the name of the verse and the letter about the word pig, whereas if searching for food words that are prohibited by verses or surah containing the word pig is not part of the search results.

The research conducted here is an information retrieval system with information retrieval techniques that are techniques used in finding and tracking relevant information. Information retrieval is used to find similarities between keywords entered and documents stored in the database. The database consists of Qur'anic and Hadith documents. The resulting system can accelerate the search process on Al-Qur'an and Hadith documents, and produce relevant information.

In the search for the Qur'an and Hadith, information retrieval is used to display several verses of the Qur'an and Hadith relating to the keywords that are searched according to certain criteria. The data used in this study comes from the translation of the Ministry of Religion of the Republic of Indonesia. Then pre-processing has been carried out on the data. The pre-process process includes tokenizing, filtering, stopword removal and stemming. The stemming method used in this study is the stemming of ECS (Enhanced Confix Stripping). Whereas the search process will develop one of the models to increase accuracy so that the search for information becomes more relevant.

In the study conducted by a previous researcher for recall value in an expanded query and did not get the same value of 100% [1]. As for the precision value on unexpanded demand, the accuracy value is 27%. And in expanded queries precision values can increase up to 75%. Besides this with the expansion of this query can find verses that have the same topic. In this study using text in Indonesian, the translation texts and interpretations involved must continue to use Arabic text, so as to minimize errors in Indonesian translation. In some studies that have been conducted there has been no research on popular words and no one has considered synonyms. In this study a synonym database was used to search for words in the Qur'an and Hadith.

1.1. Text Mining

Text mining is an interdisciplinary field that refers to information search, data mining, machine learning, statistics, and linguistic computing. Because most information (general estimates say more than 80%) is currently stored as text, text mining is believed to have high commercial value potential [2]. Text mining is a technique used to deal with the problems of classification, clustering, information extraction and information retrival [5].

Text mining usually involves the process of inputting text settings (usually parsing, along with the addition of several linguistic features derived and eliminating some of them, and subsequent insertion into the database), determining patterns in structured data, and finally evaluating and interpreting output. High quality in the field of text mining usually refers to several combinations of relevance, novelty, and interestingness [20]. Text mining stages in general are text preprocessing and feature selection [9].

1.1.1. Information Retrieval

The information retrieval (IR) system is used to retrieve information relevant to the user needs of an information set automatically [1, 2]. Some experts define Information Retrieval as follows:

i. Information Retrieval is the process of finding material (documents) from an unstructured environment (usually text) that meets the information needs of a large collection (usually on a computer) [13].
ii. Information Retrieval is a part of computer science that learns about data collection and retrieval of documents [4].
iii. Information Retrieval is a discipline that deals with unstructured search data, especially textual documents, in response to a query or topic statement. [7]

Information Retrieval is part of computer science that deals with the retrieval of information from documents based on the content and context of the documents themselves. Based on the reference explained that the information retrieval is a search information based on a query that is expected to meet the user's desire of the existing document. The working principle of the information retrieval system if there is a collection of documents and a user formulating a query (request or query). The answer to that question is a collection of relevant documents and discarding irrelevant documents [19].

Figure 1 describes information retrieval; Text operations/ Preprocessing is the process of transforming documents and queries into index words. In a document there are some words that have meaning more important than other words, so preprocessing the text in a collection of documents is considered necessary in determining the word to be used as index terms. In the preprocessing stage it also includes text operations such as markup removal, stopwords removal, and stemming (base word formation).

1.1.2. Tokenizing

Tokenizing is the process of decomposition of the original description of the sentences into words and eliminate the delimiter-delimiter and characters that exist in the word. The purpose of this process is to form tokens which are each word in a string.

1.1.3. Corpus

The process of information retrieval requires a database contained one or more tables as data storage to be processed during the search process. The database uses the corpus for the process of making its supporting tables. In principle, any collection of more than one text can be called a corpus [16]: The term corpus when used in the context of modern linguistics has a more specific connotation. There are four characteristics of the corpus [16].

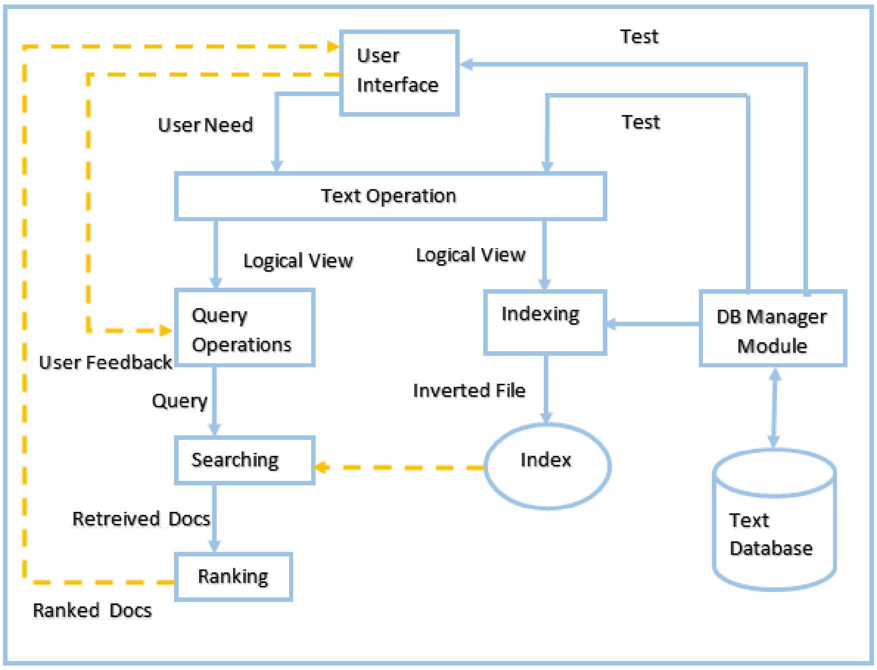

Figure 1. *IR Process (Baeza-Yates, 1999).*

1.1.4. Stopwords Removal

The words that often appear in documents in a collection (corpus) that do not represent the index to perform a search are called stopwords. The words included in this stopwords group will usually be filtered out of the group of words to be indexed. In English which is part of stopwords are articles (articles), prepositions, and conjunctions. By removing the stopwords, the size of the indexing structure can be reduced and can reduce the recall of inappropriate documents.

1.1.5. Stemming

Stem is the part of the word that remains after removing its affixes (prefix and suffix) [4]. A user often includes the word in the query but only a few of the words appear in the relevant document. Plurals, verb forms ending in -ing, and suffixes in the past are examples of syntactic types of words that would block the match between queries and related documents. This problem can be solved by word substitution with each of its original forms.

Table 1. *State Of The Art.*

	Researcher	Method	Advantages	Deficiency
1	Sistem *Information Retrieval* pencarian kesamaan ayat terjemahan Al Qur'an Berbahasa Indonesia dengan *Query Expantion* dari Tafsirnya Broto Poernomo T. P. dan Ir. Gunawan (2015)	-Stemming using the Nazief Adirani algorithm - IR Technique model Vector Space -Using *Query Expantion*	For recall values on queries that are expanding and do not get the same value that is 100%. Whereas for precision values on unexpanded queries, the precision value is 27%. And in queries that are expanded precision values can increase to 75%. Besides that with this query expansion can find verses that have the same topic	In this study using text in Indonesian, then the translation texts and interpretations involved should continue to use Arabic text. So as to minimize errors in translation into Indonesian. Synonyms and anonymous are not noticed
2	Rancang Bangun Sistem Informasi Hadist Menggunakan Teknik Temu Kembali Informasi Model Ruang Vektor Nesdi E. Rozanda, Arif Marsal, Kiki Iswanti (2014)	-Vector Space Model	Based on the results of the tests of seven queries with precision and recall calculations, the average precision 65% was obtained and the average recall was 0.97 which is almost all relevant documents taken by the system.	Database This hadith information system is still incomplete and diverse, therefore it needs to be added and always updated regularly
3	Sistem Qur'an Retrieval Terjemahan Bahasa Indonesia berbasis Web dengan reorganisasi Korpus Surya Agustian, Imelda Sukma Wulandari (2013)	-Vector Space Model	The Qur'an retrieval system built using vector space models, has given very satisfying results for some queries that are tested, have the sloping Precision-Recall graphic profile up to Recall point = 1.	Deeper testing needs to be done by parties who know more carefully regarding the content of Al-Qur'an.
4	*Question Answering* Terjemah Al qur'an Menggunakan *Named Entity Recognition* Lukman Fakih Lidimilah (2017)	-VSM, -Question Answering System (QAS) Rule Based. Named Entity Recognition (NER)	The proposed method of improving the accuracy of Rule Based is with Named Entity Recognition (NER) and document indexing improvements using ECS. the question "Who", "When", "Where" obtained accuracy of 90%, 80%, 50%	Accuracy for the "where" question is still low, The Qur'anic text document is only for Surat Al Baqarah.

	Researcher	Method	Advantages	Deficiency
5	Using Vector Space Model in Question Answering System Jovita, Linda, Andrei Hartawan, Derwin Suhartono (2015)	VSM Rule Based QAS	It has given good results recal. 547106, precision 0.662462, F-measure 0.580094	The process of searching for a long time is 29 seconds on average

1.2. ECS (Enhance Confix Stripping)

The ECS Stemmer algorithm is an improved algorithm of the Confix Stripping (CS) Stemmer algorithm. Improvements made by ECS Stemmer are improvements to some of the rules on the reference beheading table. In addition, the ECS Stemmer algorithm also adds a return step suffix in case of disappearing suffixes that should not be performed [1].

1.3. TF-IDF Method

The TF-IDF method is a term weighting method that is widely used as a comparison method for the new weighting method. In this method, the calculation of the term t weight in a document is done by multiplying the Term Frequency value with Inverse Document Frequency.

In Term Frequency (tf), there are several types of formulas that can be used [11]:

a. *binery tf*, only pay attention to whether a word exists or not in the document, if there is given a value of one, if it is not given a zero value.

b. *raw tf*, tf value is given based on the number of occurrences of a word in the document. For example, if it appears five times the word will be five.

c. tf logaritmik, this is to avoid the dominance of documents that contain few words in the query, but have a high frequency.

$$tf = 1 + log(tf) \tag{1}$$

d. tf normalization, using a comparison between the frequency of a word and the total number of words in the document.

$$tf = 0.5 + 0.5\text{x}\,[tf/max\ tf] \tag{2}$$

Inverse Document Frequency (idf) dihitung dengan menggunakan formula.

$$idfi = log(D/df)j \tag{3}$$

where, *D* is the number of all documents in the collection.

df is the number of documents containing *term* t_j.

$$\mathrm{W_{ij} = tf_{ij} \times idf_j}$$

$$Wij = tfij \times log\ (D/df)j \tag{4}$$

Where W_{ij} is weight *term* t_j to document d_i.

tf_{ij} is the number of occurrences of the *term* t_j to document d_i.

D is the number of all documents in the database.

df_j is the number of documents containing *term* t_j (there is at least one word, that is *term* t_j)

$$Wij = tfij \times log(D/df)j + 1 \tag{5}$$

2. Method

a. First step in this research is literature study about information retrieval and study of literature about Al Quran and Hadith.

b. Second Step Observing translated version of Al Quran and Hadith (Preparing Database). Figure 2.

The first thing to do is to get the Alquran translation database of Indonesian version of the Ministry of Religious Affairs of the Republic of Indonesia and Hadith Shahih Bukhori.

The second process is done Indexing (pre-process) text translated version of Alquran and Hadist consisting of tokenizing, stopword removal and stemming. Weighted tf-idf to predict document similarity to queries based on vectors formed from their constituent terms.

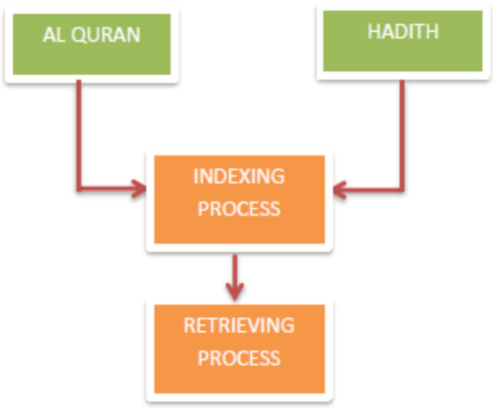

Figure 2. *Information Retrieval Processing.*

3. Result

3.1. System Framework

Figure 2 describes how the system works. Starting from first process the entire translated version of the Quran is as much as 30 juz and Muslim Hadith Bukhori as many as 7008 number of hadith. The process is tokenizing, stopword removal and stemming. Stemming using Enhanced Confix Stripping (ECS). The system framework describes that in this system synonyms are considered. In previous research, synonyms have not been considered.

Preprocessing is illustrated in Figure 4, preprocessing begins with tokenizing the Al Quran and Hadith translations. The process that is carried out is after inputting the sentence is received, the system will change all uppercase characters to lowercase. The next system will remove punctuation in the sentence. A collection of sentence words or Terms will be generated. Then read the words for the next stopword removal process or commonly known as Filtering is a process that aims to eliminate, eliminate or remove words that are not important and do not have meaning in the input sentence. The output of words from the Tokenizing process will be used as input. Then the system will compare each of these words with each stopword contained in the database. If there is a similarity of words with the words contained in the stopword database, then the word is omitted. If different, then the word will be stored and will then be used as input for the next process. The last process is to stemming to get the basic words. Stemming used in this study is ECS stemming. The basic words generated will be stored in the rootword database.

The ECS Process Stemmer Algorithm can be explained as below:

1. Words that will be stemmed are compared to the basic database. If the word is found, then the word is the basic word and the algorithm stops. If the word does not match the word in the basic dictionary, then go to step 2.
2. The word that will be stemming consists of a minimum of 3 letters, if the word consists of less than 3 letters, the word includes the base word, if not then proceed to step 3.
3. The word containing repetition will be stemmed into 1 word or no reduplication, then the word is checked with the base database, if it is found then the word is the base word, if the word is not in the basic word database then proceed to step 3.
4. Check Rule Precedence, that is, if the word has a prefix suffix "be-lah", "be-an", "me-i", "di-i", "pe-i", or "te-i" then the next stemming steps are 8, 5, 6, 7, 8, 9 but if the word input does not have the prefix-suffix pair, the stemming step goes normally, 5, 6, 7, 6, 8, 9. For example: If the word input is "merencanakan" because the word is not included in the rules that are not allowed then the next process is directly to steps 5, 6, 7, 8, 9.
5. Remove particles (P) and pronouns belonging to (PP). First remove particles (P) ("-lah", "-kah", "-tah", "-pun"). After that also remove the pronoun belongs (PP) ("ku", "mu", or "nya"). In accordance with the add-on model, it becomes: [[[DP +] DP +] DP +] Basic words [+ DS] The word "merencanakan" then the process in rule 3 does not exist because the word "-kan" belongs to Derrivation Suffixes.
6. Identification of words that contain combinations of prefixes and suffixes, if there is a word that contains a combination of prefixes and suffixes that are not allowed, then the word is considered the basic word and the algorithm stops. If no words contains a combination of prefixes and suffixes that are prohibited, then go to step 6
7. Eliminate also the suffix (DS) ("-i", "- an", and "-kan"), according to the add-on model, then become: [[[DP +] DP +] DP +] Basic words, Next to the word process "Merencanakan" will discard the word "-kan" because the word "-kan" includes derrivation suffixes or suffixes. So the word obtained is "Merencana". Because the word "Merencana" is not a basic word, the next process is step 5.
8. Elimination of prefixes (DP) ("di-", "ke-", "se-", "te-", "be-", "me-", and "pe-") follows the following steps:
   a. The algorithm will stop if:
      i. There is a combination of rules that are not allowed.
      ii. The prefix detected at this time is the same as the prefix that was previously deleted.
      iii. The word has no prefix.
   b. Identification of the prefix and suffix types, namely:
      i. If the prefix of the word is "di-", "ke-", or "se-" then.
      ii. The prefix can be immediately removed.

Delete the prefix "te-", "me", "be-", or "pe" which uses decay rules, namely the rules of Nazief Adrian's beheading, Enhanced Confix Stripping modification. Furthermore, the word "merencana" will be stemming according to the process of the fifth step which is eliminating the prefix or Derivation Prefixes. Because the word "me-" includes the start, the word "me-" will stem. Then the word "rencana" is the word obtained from this process. Then the word "rencana" is checked into the base database, because the word "rencana" includes the basic word, the stemming process stops here.

9. If all steps fail, then the prefixes that have been returned again and the word is considered the basic word. In inputting data is divided into two categories namely training data and test data. Separating the data into training data and test data is intended so that the model obtained will have good generalization ability in data classification. It is not uncommon for a classification model to classify data very well on training data, but it is very bad to do new and unprecedented data classifications, this is called overfitting. In data mining classification can be used to predict data classes from new data based on predetermined classes from existing data.

3.2. Tokenizing

The process for Tokenizing is as follows:

a. After the sentence input is received, the system will convert all uppercase characters to lowercase.
b. The next system will remove punctuation in the sentence.
c. Will be generated sentence compiler word or Terms.
d. The Tokenizing process is complete.

Tokenizing Testing

The Tokenizing process based on several verses of the Qur'an is to include a sentence to be processed:

Testing: Translated version Al Quran Surah An Nisa Verse 112.

In Bahasa: "*Dan barangsiapa yang mengerjakan kesalahan atau dosa, kemudian dituduhkannya kepada orang yang tidak bersalah, maka sesungguhnya ia telah berbuat suatu kebohongan dan dosa yang nyata.*"

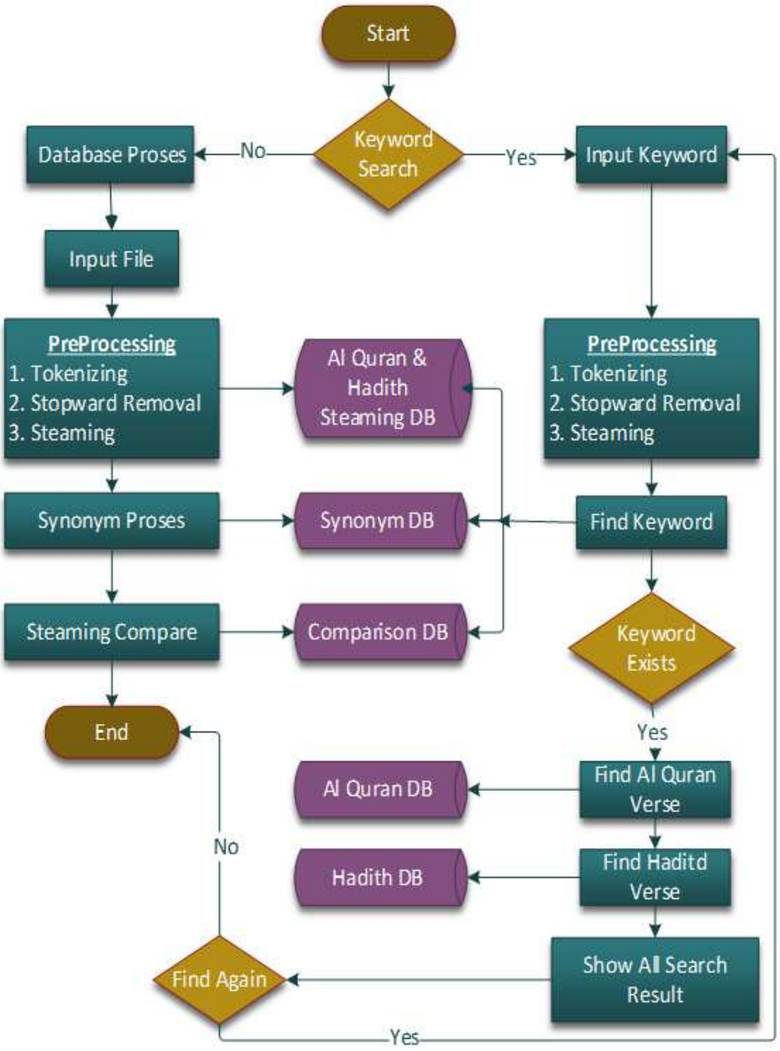

Figure 3. *System Framework.*

In English: “And whoever commits a fault or a sin, then accuses of it one innocent, he indeed takes upon himself the burden of a calumny and a manifest sin”.

Tokenizing Results:

dan barangsiapa yang mengerjakan kesalahan atau dosa kemudian dituduhkannya kepada orang yang tidak bersalah maka sesungguhnya ia telah berbuat suatu kebohongan dan dosa yang nyata.

3.3. Stopword Removal

Processes aimed at the removal, removal or disposal of unimportant and meaningless words in the input sentence. Output words from the Tokenizing process will be used as input. The system will compare each of these words with each stopword contained in the database. If there are similar words with words in the stopword database, the word is omitted. If different, then the word will be stored and for the next will be used as input on the next process.

Testing

The stopword removal process based on tokenizing result:

Testing: Translated version Al Quran Surah An Nisa Verse 112 (tokenizing result).

In Bahasa: *dan barangsiapa yang mengerjakan kesalahan atau dosa kemudian dituduhkannya kepada orang yang tidak bersalah maka sesungguhnya ia telah berbuat suatu kebohongan dan dosa yang nyata.*

Stopword Removal Result:

barangsiapa kesalahan dosa dituduhkannya bersalah berbuat kebohongan dosa nyata.

3.4. Stemming

Stemming is a process to find the root word of a word (root word). Stemming is used to change the shape of a word into a word of the word in accordance with the structure of Indonesian morphology.

Testing

The stemming process uses ECS stemmer algorithm.

Testing: Translated version Al Quran Surah An Nisa Verse 112 (stopword removal result).

In Bahasa; *barangsiapa kesalahan dosa dituduhkannya bersalah berbuat kebohongan dosa nyata.*

Stemming Results:

barangsiapa salah dosa tuduh salah buat bohong dosa nyata.

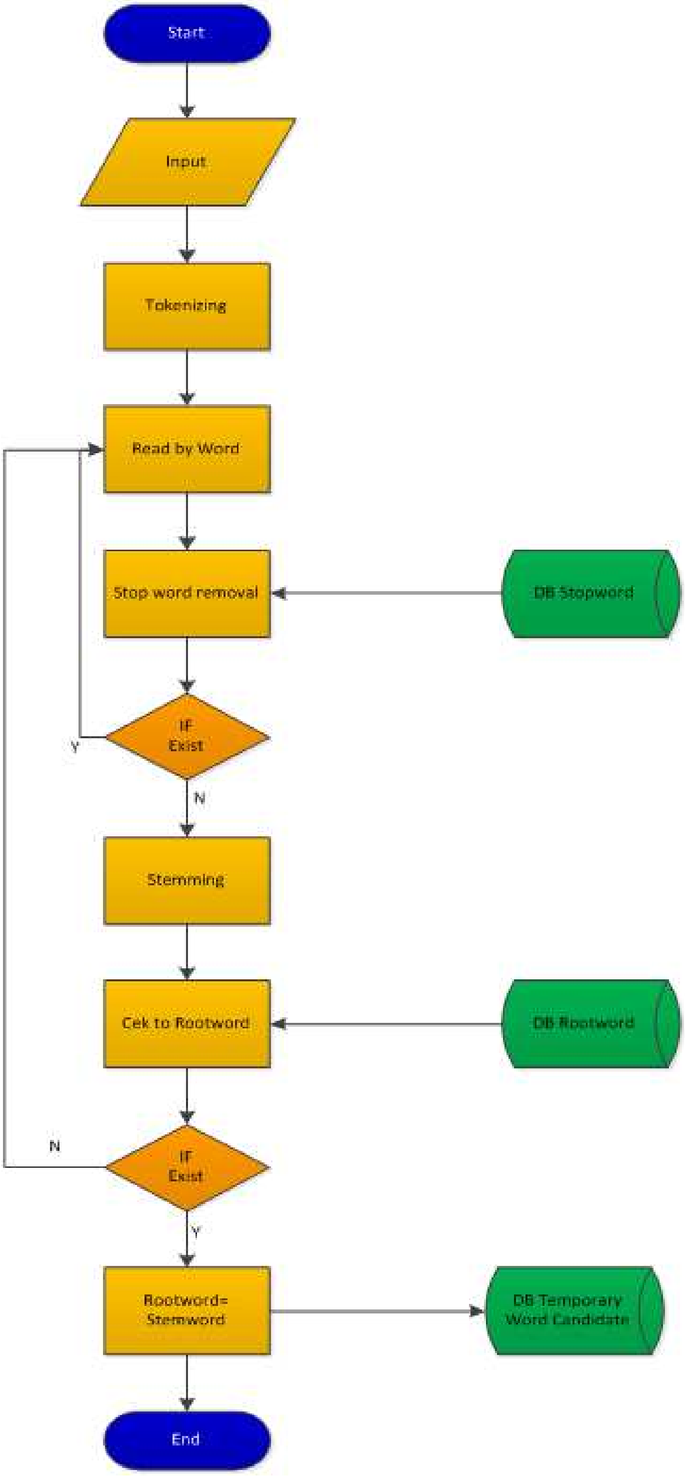

Figure 4. *Preprocessing.*

All stemming results are placed in the stemming database as shown in Figure 4. For the next will be comparable with the existing word in the synonym database The synonym database contains all of its keywords and synonyms as an example as shown in table 2. Searching information by paying attention to synonyms then search results will be more widespread and more relevant if the user searches with popular words or words that are not common. Ordinary search techniques that only use keywords usually generate words that match the keywords that are inputted only. For example testing is done by entering the keyword "Bohong", "Dusta", "Fitnah" and "Tipu". Illustration mapping of verses of Al Quran and Hadith As shown in Figure 5 shows that the words "Bohong", "Dusta", "Fitnah" and "Tipu" have the same meaning so that when done queries to the word then the results obtained will be displayed verses of the Qur'an and Hadith related to the word.

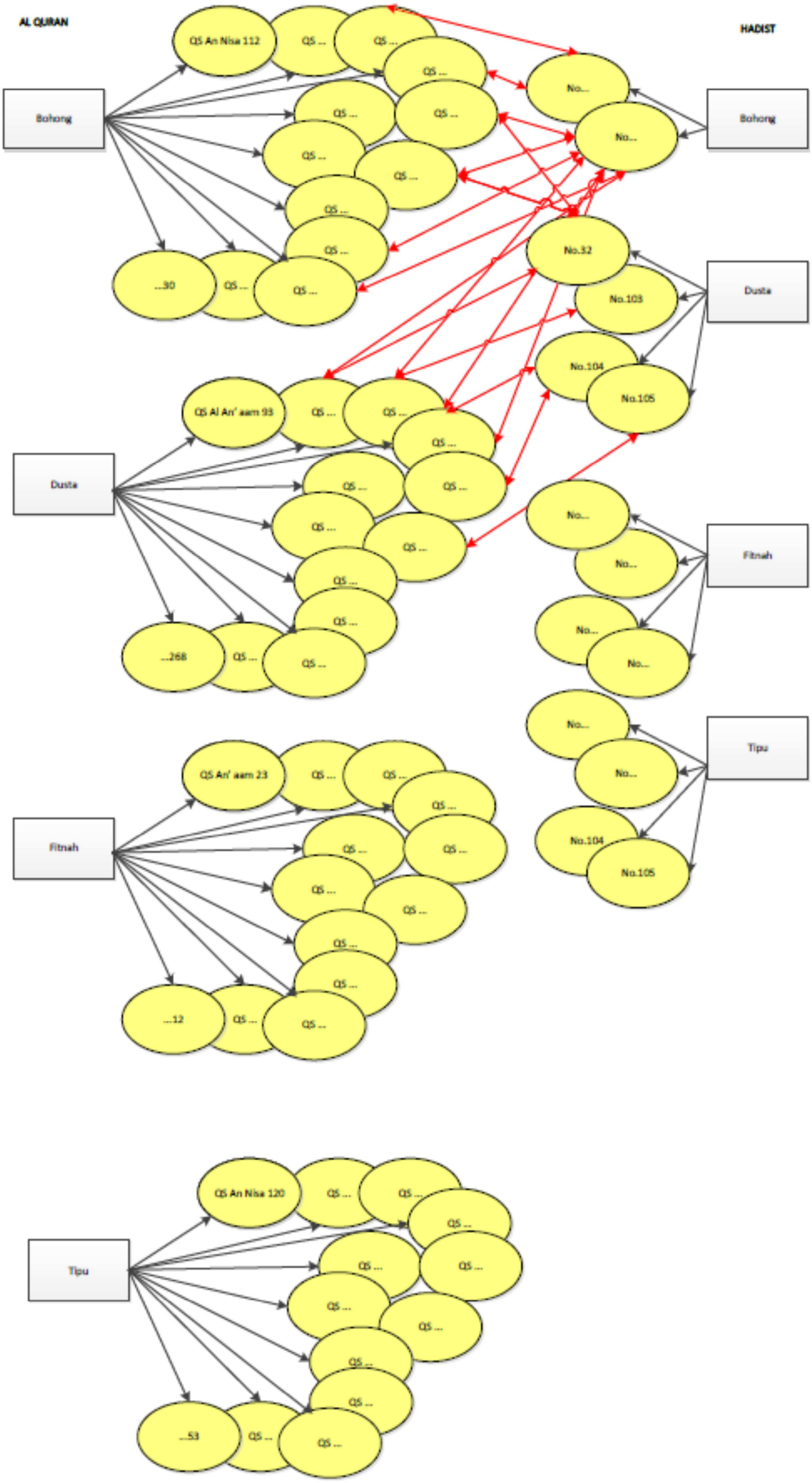

Figure 5. *Illustration of the word synonym mapping on Al Quran and Hadith.*

4. Discussion

The test results contained in Table 4 prove that the search words on keywords that have not used the database synonym only raises the text of the same document with the keyword. Using synonyms of words, the search will produce a total of 363 verses.

4.1. Testing Scenario

The trial is done with a "Korupsi" query. In Table 2, the contents of the Al-Qur'an verse which contain words that have the same meaning as "Korupsi " are described. The word "Korupsi" is not in the Qur'an and Hadith, but the meaning of the same thing as "Korupsi" is numerous, the same example of the word "Korupsi" is "Memakan Harta" As shown in table 2 the results of the information retrieval trial by testing using the keyword "Korupsi".

Table 2. *Testing IR.*

NO	Surat/ Ayat	ISI AYAT ALQURAN
1	Al Baqarah (188)	In Bahasa Dan janganlah sebahagian kamu memakan harta sebahagian yang lain di antara kamu dengan jalan yang bathil dan (janganlah) kamu membawa (urusan) harta itu kepada hakim, supaya kamu dapat memakan sebahagian daripada harta benda orang lain itu dengan (jalan berbuat) dosa, padahal kamu mengetahui. In English: And do not swallow up your property among yourselves by false means, neither seek to gain access thereby to the judges, so that you may swallow up a part of the property of men wrongfully while you know
2	An Nisa (6)	In Bahasa Dan ujilah anak yatim itu sampai mereka cukup umur untuk kawin. Kemudian jika menurut pendapatmu mereka telah cerdas (pandai memelihara harta), maka serahkanlah kepada mereka harta-hartanya. Dan janganlah kamu makan harta anak yatim lebih dari batas kepatutan dan (janganlah kamu) tergesa-gesa (membelanjakannya) sebelum mereka dewasa. Barang siapa (di antara pemelihara itu) mampu, maka hendaklah ia menahan diri (dari memakan harta anak yatim itu) dan barangsiapa yang miskin, maka bolehlah ia makan harta itu menurut yang patut. Kemudian apabila kamu menyerahkan harta kepada mereka, maka hendaklah kamu adakan saksi-saksi (tentang penyerahan itu) bagi mereka. Dan cukuplah Allah sebagai Pengawas (atas persaksian itu). In English And test the orphans until they attain puberty; then if you find in them maturity of intellect, make over to them their property, and do not consume it extravagantly and hastily, lest they attain to full age; and whoever is rich, let him abstain altogether, and whoever is poor, let him eat reasonably; then when you make over to them their property, call witnesses in their presence; and Allah is enough as a Reckoner
3	An Nisa (10)	In Bahasa Sesungguhnya orang-orang yang memakan harta anak yatim secara zalim, sebenarnya mereka itu menelan api sepenuh perutnya dan mereka akan masuk ke dalam api yang menyala-nyala (neraka). In English (As for) those who swallow the property of the orphans unjustly, surely they only swallow fire into their bellies and they shall enter burning fire
4	An Nisa (29)	In Bahasa Hai orang-orang yang beriman, janganlah kamu saling memakan harta sesamamu dengan jalan yang batil, kecuali dengan jalan perniagaan yang berlaku dengan suka sama-suka di antara kamu. Dan janganlah kamu membunuh dirimu; sesungguhnya Allah adalah Maha Penyayang kepadamu. In English O you who believe! do not devour your property among yourselves falsely, except that it be trading by your mutual consent; and do not kill your people; surely Allah is Merciful to you
5	An Nisa (161)	In Bahasa dan disebabkan mereka memakan riba, padahal sesungguhnya mereka telah dilarang daripadanya, dan karena mereka memakan harta benda orang dengan jalan yang batil. Kami telah menyediakan untuk orang-orang yang kafir di antara mereka itu siksa yang pedih. In Englih And their taking usury though indeed they were forbidden it and their devouring the property of people falsely, and We have prepared for the unbelievers from among them a painful chastisement
6	At Taubah (34)	In Bahasa Hai orang-orang yang beriman, sesungguhnya sebahagian besar dari orang-orang alim Yahudi dan rahib-rahib Nasrani benar-benar memakan harta orang dengan jalan batil dan mereka menghalang-halangi (manusia) dari jalan Allah. Dan orang-orang yang menyimpan emas dan perak dan tidak menafkahkannya pada jalan Allah, maka beritahukanlah kepada mereka, (bahwa mereka akan mendapat) siksa yang pedih, In English O you who believe! most surely many of the doctors of law and the monks eat away the property of men falsely, and turn (them) from Allah's way; and (as for) those who hoard up gold and silver and do not spend it in Allah's way, announce to them a painful chastisement
7	Al-Fajr (19)	In Bahasa dan kamu memakan harta pusaka dengan cara mencampur baurkan (yang halal dan yang bathil), In English And you eat away the heritage, devouring (everything) indiscriminately,

The test results contained in Table 2 prove that the search with the keyword "Korupsi" raises the relevant text of the document, although in the translation of Al Quran there is no word " Korupsi ", this is because a synonym database has been formed, if the synonym database does not exist then the document's text will not be found.

Testing is also done by entering the keywords "Bohong", "Dusta", "Fitnah" and "Tipu". The first is done without using a database Synonym.

The test results contained in table 3 prove that word searches on keywords that have not used synonym databases only bring up the same document text with keywords. Testing is done by entering the keyword "Bohong", "Dusta", "Fitnah" and "Tipu". Performed using a database of synonyms. The test results contained in table 4 prove that word search on keywords after using a synonym database raises all text documents not only the same as keywords, but also displays those that have the same meaning as keywords.

Table 3. *Result Of IR before using the synonym database.*

NO	KEYWORD	Number of verses
1	Bohong	30
2	Dusta	268
3	Fitnah	12
4	Tipu	53

Table 4. *Result Of IR after using the synonym database.*

NO	KEYWORD	Number of verses
1	Bohong	363
2	Dusta	363
3	Fitnah	363
4	Tipu	363

The next step is the implementation of Al Quran and Hadith information retrieval methods. Result methodology trial from Al Quran and Hadith information retrieval research to find out the results are relevant, at this stage the trial was also carried out with Alquran and Hadith experts who are academics in this field, this is to ascertain whether the output is appropriate and correct based on meaning and not scientifically wrong.

4.2. Term Weighthing

The information retrieval system is handling information retrieval that matches the query that the user wants from the document collection. The collection of documents here is a translation of Al Quran and Hadith. The collection of documents consists of documents of varying length with different content terms. Term can be in the form of words, phrases or other indexing result units in a document that can be used to determine the context of the document. Each word has a different level of importance so that each word is given term wight. Because the thing to note in information retrieval is term weighting.

Term weighting is strongly influenced by several things including Term Frequency (tf) factor, which is a factor that determines the term weight in a document based on the number of occurrences in the document. The value of the number of occurrences of a word (term frequency) is taken into account in giving weight to a word. The greater the number of occurrences of a term (high tf) in the document, the greater the weight in the document or the greater conformity value. Inverse Document Frequency (idf) factor, which is a reduction in dominance term that often appears in various documents. This is necessary because the terms that appear in various documents can be considered as common terms so that they are not important. Conversely, the term scarcity factor in the document collection must be considered in giving weight. Weighting will take into account the inverse frequency factor of the document containing a word (inverse document frequency).

The following is the application of some words from the preprocessing and weighting stages of the TFIDF method. As an example the query is done for the word "Korupsi" The word "Korupsi" is synonymous with "Memakan Harta" or "Makan Harta" As a scenario testing in this study there are 3 documents from the corpus of Al Quran translation, namely: Surat Al Baqarah verse 188, QS An Nisa verse 29 and QS An Nisa verse 10. The contents of the translation are;

Document 1 (D1) is QS Al Baqarah 188 containing: “Dan janganlah kamu makan harta diantara kamu dengan jalan yang batil dan (janganlah) kamu menyuap dengan harta itu kepada para hakim dengan maksud agar kamu dapat memakan sebagian harta orang lain itu dengan jalan dosa padahal kamu mengetahui”.

Document 2 (D2) is QS An Nisa 29 containing: “Wahai orang-orang yang beriman janganlah kamu saling memakan harta sesamamu dengan jalan yang batil (tidak benar) kecuali dalam perdagangan yang berlaku atas dasar suka sama suka diantara kamu. Dan janganlah kamu membunuh dirimu sungguh Allah maha Penyayang kepadamu”.

Document 3 (D3) is QS An Nisa 10 containing: “Sesungguhnya orang-orang yang memakan harta anak yatim secara zalim sebenarnya mereka itu menelan apa dalam perutnya dan mereka akan masuk ke dalam api yang menyala-nyala (Neraka)”.

D = Number of verse translation documents

Q = The query input by the user

The steps taken begin with preprocessing on the three documents, namely to tokenizing the three documents so that the results of tokenizing are:

Document 1 (D1) is QS Al Baqarah 188 containing: dan janganlah kamu makan harta diantara kamu dengan jalan yang batil dan janganlah kamu menyuap dengan harta itu kepada para hakim dengan maksud agar kamu dapat memakan sebagian harta orang lain itu dengan jalan dosa padahal kamu mengetahui.

Document 2 (D2) is QS An Nisa 29 containing: wahai orang-orang yang beriman janganlah kamu saling memakan harta sesamamu dengan jalan yang batil tidak benar kecuali dalam perdagangan yang berlaku atas dasar suka sama suka diantara kamu. Dan janganlah kamu membunuh dirimu sungguh allah maha penyayang kepadamu.

Document 3 (D3) is QS An Nisa 10 containing: sesungguhnya orang-orang yang memakan harta anak yatim secara zalim sebenarnya mereka itu menelan apa dalam perutnya dan mereka akan masuk ke dalam api yang menyala-nyala neraka.

The second stage was carried out in the three documents, namely to do stopword removal. Removal of stopwords, which is the removal of the terms that most often appear on documents. For this case the term experiencing stopwords is: "what, over, how, for, if, and, can, with, in, he, that, almost, if, also, times, we, then, to, when, again, then, they, by, at most, all, so that, besides, all, and, something, every, has, remains, not, for, which ". The result of the stopword removal process is;

Document 1 (D1) is QS Al Baqarah 188 containing: janganlah makan harta diantara jalan yang batil janganlah menyuap harta para hakim memakan sebagian harta orang lain jalan dosa mengetahui.

Document 2 (D2) is QS An Nisa 29 containing: orang-orang beriman janganlah memakan harta sesamamu jalan batil perdagangan berlaku atas dasar suka sama suka kamu janganlah membunuh dirimu sungguh allah maha penyayang.

Document 3 (D3) is QS An Nisa 10 containing: sesungguhnya orang-orang memakan harta anak yatim secara zalim sebenarnya menelan apa dalam perutnya masuk api menyala-nyala neraka.

The ECS Process Stemmer Algorithm;

1. Words that will be stemmed are compared to the basic database.
2. The word that will be stemming consists of a minimum of 3 letters, if the word consists of less than 3 letters, the word includes the base word, if not then proceed to step 3.

3. The word containing repetition will be stemmed into 1 word or no reduplication, then the word is checked with the base database, if it is found then the word is the base word, if the word is not in the basic word database then proceed to step 3.
4. Check Rule Precedence.
5. Remove particles (P) and pronouns belonging to (PP). First remove particles (P) ("-lah", "-kah", "-tah", "-pun"). After that also remove the pronoun belongs (PP) ("ku", "mu", or "nya").
6. Identification of words that contain combinations of prefixes and suffixes that are not allowed.
7. Eliminate also the suffix (DS).
8. Elimination of prefixes (DP) ("di-", "ke-", "se-", "te-", "be-", "me-", and "pe-").
9. If all steps fail, then the prefixes that have been returned again and the word is considered the basic word.

Stemming Result is:

Document 1 (D1) is QS Al Baqarah 188 containing: makan harta antara jalan batil suap harta para hakim makan bagian harta orang lain jalan dosa ketahui.

Document 2 (D2) is QS An Nisa 29 containing: orangorang iman makan harta sesama jalan batil dagang laku atas dasar suka sama suka. bunuh diri sungguh allah maha saying.

Document 3 (D3) is QS An Nisa 10 containing: sungguh orangorang makan harta anak yatim secara zalim benara telan perut masuk api nyalanyala neraka.

The next step is to do indexing and giving weight to the term. Each word / term in the document to be indexed must be weighted based on the number of occurrences in each document to be able to analyze the level of similarity between each paragraph. For the stage of assigning the value or weight of each term, the TF-IDF (Term Frequency - Inverse Document Frequency) algorithm is used. Giving weight to each term is defined by calculating the frequency of occurrence of documents containing a term, namely DF (document frequency); frequency calculation of the appearance of the term in the document that is TF (term frequency) and the calculation of the number of documents containing a term sought from the existing document, namely IDF (Inverse Document Frequency).

Table 5. *Frequency Term Value.*

Term	TF				
	D1	D2	D3	Df	D/ df
Makan	2	1	1	4	0,75
Harta	3	1	1	5	0,6
Jalan	2	1	0	3	1
Batil	1	1	0	2	1,5
Suap	1	0	0	1	3
Para	1	0	0	1	3
Hakim	1	0	0	1	3
Bagian	1	0	0	1	3
Orang	1	1	1	3	1
Lain	1	0	0	1	3
Dosa	1	0	0	1	3
Ketahui	1	0	0	1	3
Iman	0	1	0	1	3
Sesama	1	0	0	1	3
Dagang	0	1	0	1	3
Laku	0	1	0	1	3
Atas	0	1	0	1	3
Dasar	0	1	0	1	3
Suka	0	2	0	2	1,5
Bunuh	0	1	0	1	3
Diri	0	1	0	1	3
Sungguh	0	1	1	2	1,5
Allah	0	1	0	1	3
Maha	0	1	0	1	3
Sayang	0	1	0	1	3
anak	0	0	1	1	3
yatim	0	0	1	1	3
secara	0	0	1	1	3
zalim	0	0	1	1	3
telan	0	0	1	1	3
Api	0	0	1	1	3
perut	0	0	1	1	3
masuk	0	0	1	1	3
nyala	0	0	1	1	3
neraka	0	0	1	1	3

Table 5 explains the TF value of 3 documents. The df value is the value obtained from the sum of document terms because df is a document that contains frequency terms. As in Figure 6 is an illustration for the IDF TF value in 3 documents.

Figure 6 explains that from 3 documents there is the word "vanity" in document 1 and document 2. The word "batil" in document 1 has only one, as well as document 2 there is only 1 word "batil" and in document 3 there is no word "batil". Calculation of values such as the illustration in Figure 7.

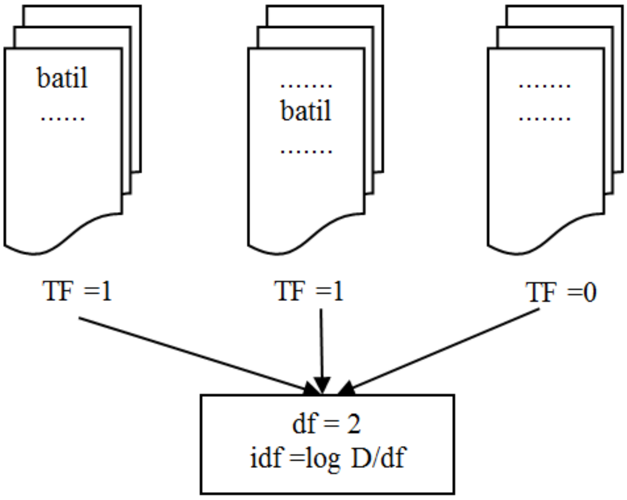

Figure 6. *Illustration of TF-IDF.*

Table 6. *Value of TF IDF.*

Term	TF					IDF =Log D/df
	D1	D2	D3	df	D/df	
Makan	2	1	1	4	0,75	-0.1249387366
Harta	3	1	1	5	0,6	-0.22184874961
Jalan	2	1	0	3	1	0
Batil	1	1	0	2	1,5	0,17609
suap	1	0	0	1	3	0.47715
para	1	0	0	1	3	0.47715
hakim	1	0	0	1	3	0.47715
bagian	1	0	0	1	3	0.47715
orang	1	1	1	3	1	0
Lain	1	0	0	1	3	0.47715
Dosa	1	0	0	1	3	0.47715
Ketahui	1	0	0	1	3	0.47715
Iman	0	1	0	1	3	0.47715
Sesama	1	0	0	1	3	0.47715
Dagang	0	1	0	1	3	0.47715
Laku	0	1	0	1	3	0.47715
Atas	0	1	0	1	3	0.47715
Dasar	0	1	0	1	3	0.47715
Suka	0	2	0	2	1,5	0,17609
Bunuh	0	1	0	1	3	0.47715
Diri	0	1	0	1	3	0.47715
Sungguh	0	1	1	2	1,5	0,17609
Allah	0	1	0	1	3	0.47715
Maha	0	1	0	1	3	0.47715
Sayang	0	1	0	1	3	0.47715
Anak	0	0	1	1	3	0.47715
Yatim	0	0	1	1	3	0.47715
Secara	0	0	1	1	3	0.47715
Zalim	0	0	1	1	3	0.47715
Telan	0	0	1	1	3	0.47715
Api	0	0	1	1	3	0.47715
Perut	0	0	1	1	3	0.47715
Masuk	0	0	1	1	3	0.47715
Nyala	0	0	1	1	3	0.47715
Neraka	0	0	1	1	3	0.47715

Table 6 is the calculation of the TFIDF value in the term of each document. Next is the calculation of the weighting of the terms in each document as in table 7.

Table 7. *Calculation of Weighting TF-IDF Term Queries in Each Document.*

Term	TF					IDF =Log D/ df	W = TF * (IDF +1)			
	D1	D2	D3	Df	D/df		IDF+1	D1	D2	D3
makan	2	1	1	4	0,75	-0.1249387366	0,87507	1,75014	0,87507	0,87507
harta	3	1	1	5	0,6	-0.22184874961	0,77815	2,33445	0,77815	0,77815
jalan	2	1	0	3	1	0	1	2	1	0
batil	1	1	0	2	1,5	0,17609	1,17609	1,17609	1,17609	
suap	1	0	0	1	3	0.47715	1,47715	1,47715	0	0
para	1	0	0	1	3	0.47715	1,47715	1,47715	0	0
hakim	1	0	0	1	3	0.47715	1,47715	1,47715	0	0
bagian	1	0	0	1	3	0.47715	1,47715	1,47715	0	0
orang	1	1	1	3	1	0	1	1	1	1
lain	1	0	0	1	3	0.47715	1,47715	1,47715	0	0
dosa	1	0	0	1	3	0.47715	1,47715	1,47715	0	0
ketahui	1	0	0	1	3	0.47715	1,47715	1,47715	0	0
iman	0	1	0	1	3	0.47715	1,47715	0	1,47715	0
sesama	1	0	0	1	3	0.47715	1,47715	1,47715	0	0
dagang	0	1	0	1	3	0.47715	1,47715	0	1,47715	0
laku	0	1	0	1	3	0.47715	1,47715	0	1,47715	0
atas	0	1	0	1	3	0.47715	1,47715	0	1,47715	0
dasar	0	1	0	1	3	0.47715	1,47715	0	1,47715	0
suka	0	2	0	2	1,5	0,17609	1,17609	0	2,35218	0
bunuh	0	1	0	1	3	0.47715	1,47715	0	1,47715	0
diri	0	1	0	1	3	0.47715	1,47715	0	1,47715	0

Term	**TF**					**IDF =Log D/ df**	**W = TF * (IDF +1)**			
	D1	**D2**	**D3**	**Df**	**D/df**		**IDF+1**	**D1**	**D2**	**D3**
sungguh	0	1	1	2	1,5	0,17609	1,47715	0	1,47715	1,47715
Allah	0	1	0	1	3	0.47715	1,47715	0	1,47715	0
maha	0	1	0	1	3	0.47715	1,47715	0	1,47715	0
sayang	0	1	0	1	3	0.47715	1,47715	0	1,47715	0
anak	0	0	1	1	3	0.47715	1,47715	0	0	1,47715
yatim	0	0	1	1	3	0.47715	1,47715	0	0	1,47715
secara	0	0	1	1	3	0.47715	1,47715	0	0	1,47715
zalim	0	0	1	1	3	0.47715	1,47715	0	0	1,47715
telan	0	0	1	1	3	0.47715	1,47715	0	0	1,47715
api	0	0	1	1	3	0.47715	1,47715	0	0	1,47715
perut	0	0	1	1	3	0.47715	1,47715	0	0	1,47715
masuk	0	0	1	1	3	0.47715	1,47715	0	0	1,47715
nyala	0	0	1	1	3	0.47715	1,47715	0	0	1,47715
neraka	0	0	1	1	3	0.47715	1,47715	0	0	1,47715
Term weighting								20,07788	23,43014	18,90187

5. Conclusion

Thus we can obtain the weight value (w) for each term in the query in each document. After the weight of each document is known, the document ranking process is carried out based on the level of relevance of the document to the query, where the greater the weight value of the document to the query, the greater the level of similarity of the document to the query being searched. Thus, a list of relevant documents can be generated based on the value of similarity between the document and query input that will then be given to the user. From the results of weighting and ranking it can be seen that document 2 (D2) has the highest level of relevance then document 1 (D1) then document 3 (D1).

Aknowledgements

The author would like to thank Gunadarma University for its support in this research. The authors also thank resource persons, Al Quran experts and Hadith Experts. The authors also thank colleagues for their input and support, and we also thank resource persons for their time in providing information related to this research.

References

[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] Akram Roshdi, Akram Roohparvar. Review: Information Retrieval Techniques and Applications, International Journal of Computer Networks and Communications Security, VOL. 3, NO. 9, 373-377, September 2015.

[4] Baeza R. Y., Neto R., Modern Information Retrieval, Addison Wesley-Pearson international edition, Boston. US. USA, 1999.

[5] Berry, M. W. & Kogan, J. 2010. Text Mining Aplication and theory.

[6] 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.

[7] Bridge, C. 2011. Unstructured Data and the 80 Percent Rule.

[8] Fatkhul Amin, Information Retrieval System with Vector Space Model Method, Journal of Business Information Systems 02, 2012.

[9] Feldman, R & Sanger, J. 2007. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press: New York.

[10] Jasman Pardede, Mira M Barmawi, Wildan D Pramono, Implementation of Generalized Vector Space Model Method In Information Retrieval Applications, No.1, Vol. 4, ISSN: 2008-5266, January - April 2013.

[11] Jovita, Linda, Andrei Hartawan, 2015, Using Vector Space Model in Question Answering System, International Conference on Computer Science and Computational Intelligence (ICCSCI 2015).

[12] Kendall, J. E. & Kendall, K. E. 2010. Analisis dan Perancangan Sistem. Jakarta: Indeks.

[13] Lukman Fakih Lidimilah, 2017, Question Answering Terjemah Al qur'an Menggunaka Named Entity Recognition, Jurnal Ilmiah Informatika Volume 2 No. 2.

[14] Mandala, Rila dan Hendra Setiawan. Peningkatan Performansi Sistem Temu KembaliInformasi dengan Perluasan Query Secara Otomatis, Laboratorium Keahlian Informatika teori Department Teknik Informatika, Institut Teknologi Bandung, 2006.

[15] Manning, Christopher D., Prabhakar Raghavan,. Introduction to Information Retrieval. Cambridge University Press, Cambridge, England, 2009.

[16] McEnery, A. M., Wilson, A. 2001. Corpus Linguistics. Edinburgh: Edinburgh University Press.

[17] Moral, C., Antonio, A., Imbert, R., Rmirez J.: A survey of stemming algorithms in information retrieval. Inf. Res.: Int Electron. J. 19 (1), 2014).

[18] 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.

[19] Salton G, Buckley C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24 (5), 513-523. https://doi.org/10.1016/0306-4573 (88) 90021-0

[20] Saraswati, N. W, 2011. Text Mining dengan Metode Naive Bayes Classifier dan Support Vector Machines untuk Sentiment Analysis. Universitas UDAYANA

[21] Subari, Ferdinandus, Health Information Retrieval System For Medical Treatment Using Space Vector Method (VSM) Method Based on WebGis, ISSN 2089-1083, Snatika 2015.

[22] 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.

[23] Tala, Fadillah Z. 2003. A Study of Stemming Efects on Information Retrieval in Bahasa Indonesia.