Search Applications Translation Of Translation Of Al-Qur'an Verses In Indonesian Based On Portabledocument Format (Pdf) Using ArifinSetiono Stemming Algorithm, Text Based Vector Space Text Model With Logarity Weight Frequency Words And Design Functions Dice

Dendy Febriyandra¹, Maulana Malik Ibrahim ² ,JonoAfrianto³
Faculty of Computer Science, Budi Luhur University, Jakarta, Indonesia

Abstract:

With the development of the world of information technology today that has penetrated to various sides of human life. Such developments are supported by the availability of hardware and software that increasingly powerful ability. Information Technology can not be separated with the development of the internet world today. The information presented in the internet world is very global and always cultivated ontime so that the update time of information is very fast. Through the internet world web applications are no longer limited as static information providers, but are also able to provide dynamically changed information by connecting to databases. It is not difficult to study the Qur'an, it can easily get information about the Qur'an in the virtual world. It can even download the digital Qur'an software that is widely available on some websites. The verses that exist in the Qur'an are dispersed so that it takes a long time to find a translation of Al-Qu'ran verses needed. Therefore, the authors make a web-based application search translation of Qur'anic verses 'an which will facilitate the user in searching for translations of Qur'anic verses needed with the application of ArifinSetiono Algorithm stemming, vector space-based text profiling model with frequency weighting and Dice similarity function. Based on the results of the implementation and trial of this application is able to find the translations of Qur'anic verses in Indonesian language well so that it helps users learn and get information translation of the verses of Al-Qur'an accordingly.

Keywords —Information Retrieval, Stemming ArifinSetiono, KesamaanDice , Search Engine, Al-Qur'an.

I. INTRODUCTION

1.2 Background

The Qur'an is the incomparable word of Allah, revealed to the Prophet Muhammad SAW the conclusion of the prophets and messengers, by the mediation of the Angel Gabriel a. and written on the mushafs which are then passed on to us mutawatir, and reading and studying them is worship.

With the rapid development of information technology, we can now download the digital Qur'an software that is widely available on some websites.

Perhaps we have often found a software that functions as a digital Qur'an, the software contains the verses of the Qur'an and its translations. But the verses that exist in the Qur'an are dispersed so that it takes a long time to find a translation of Al-Qu'ran verses that are needed. Therefore, the authors make a web application Al-Qur'an that will facilitate the user in search for translations Qur'an
verses needed with the application of stemming Algorithm ArifinSetiono on vector.

Space, frequency weighting and Dice similarity function that when the user input the query then this application will display the translation of Al-Qur'an from the translation pieces of the verses of Al-Qur'an in accordance or close to the query that is input by users, so it is expected to help users who want to find the translation of the verses of the Qur'an.

1.3 Problem.
By looking at the reviews from the background above, we get some problematic issues as follows:

a. The verses that exist in the Qur'an are dispersed so that it takes a long time to search for a translation of the required Al-Qur'an verses.

b. How to get the Al-Qur'an verses easily translated.

How to get translations of Al-Qur'an verses as needed.

1.4 Research Objectives.
The purpose of this research is to make the search application of translation of Al-Qur'an verses in Bahasa Indonesia based on Portable Document Format (PDF) using ArifinSetiono stemming and vector space based text modeling with weighted word logarithm and Dice similarity function. Application search translation of verses Al-Qur'an is expected to facilitate the user in search translation of Qur'anic verses needed. So that it can ease in learning and searching Al-Qur'anic verses, and helping Muslims to understand the translation of the verses of the Qur'an.

II. THEORETICAL BASIS

2.1 Concept of Information Acquisition
Information retrieval system (IRS) is the process of finding materials (usually documents) of a structured (typically text) that satisfy the information needs of large collections (usually stored on a computer). IRS is a system used to retrieve (retrieve) documents relevant to the user needs of a collection of information based on keywords or keywords or queries from the user. In addition to finding documents relevant to the query, the IRS also reviews the search results. A document that has a higher rank than other documents will be considered more relevant to the query [1].

Information retrieval is part of computer science that deals with the retrieval of information from documents based on the content and context of the document itself. Information retrieval (information retrieval) is an information search (usually a document) based on a query (user input) that is expected to satisfy the user's wishes from the existing document. While the query definition in the information retrieval is a formula used to find the information required by the user, in the simplest form, a query is a keyword (keyword) and a document containing the keyword is a document sought in the information retrieval system [2].

2.2 Stemming ArifinSetiono
Stemming is a process to find the basic word of a word. By eliminating all affixes consisting of prefixes, infixes, suffixes and confixes in derived words. Stemming is used to change the shape of a word into the root word of the word in accordance with the structure of Indonesian morphology is good and right [3].

The stemming algorithm for one language is different from the stemming algorithm for other languages. For example English has a different morphology than Indonesian so the stemming algorithm for both languages is also different. The stemming process in Indonesian text is more complicated or complex because there are variations of affixes that must be removed to get the root word of a word [4]. In this research, stemming algorithm used is stemming algorithm ArifinSetiono. The algorithm is preceded by reading each word of the sample file. So the input of this algorithm is a word which then checks all possible word forms. Each word is assumed to have 2 prefixes (3 prefixes) and 3 suffixes, so its shape becomes:

\[
\text{Prefiks} \, 1 + \text{Prefiks} \, 1 + \text{Kata dasar} + \text{Suffixes} \, 3 + \text{Suffixes} \, 2 + \text{Suffixes} \, 1
\]

Cutting is done in sequence as follows: AW:
AW (Prefix); AK: AK (Suffix); KD: KD (Basic Word); AW I, the result is stored in p1 (prefix 1); AW II, the result is stored in p2 (prefix 2); AK I, the result is stored in s1 (suffixes 1); AK II, the result is stored in s2 (suffixes 2); AK III, the result is stored on s3 (suffixes 3)

At each stage of the above cutting followed by a check in the dictionary whether the cuts are already in basic form. If the examination is successful then the process is declared completed and no need to continue the process of cutting other affixes.

However, if it comes to AK III cuts, not yet found in the dictionary, then a combination process is done. KD which is generated in combination with affixes in the following 12 configurations :

KD;KD+AK III; KD+AK III+AK II;KD+AKIII+AK II+AK I; AWI+AWII+KD;AW I + AW II + KD + AK III; AW I + AW II + KD + AK III + AK II; AW I + AW II + KD + AK III + AK II + AK I

Of course, if the result of a combination check is 'there', then checks on other combinations are no longer needed. An examination of these 12 combinations is necessary, due to the overstemming phenomenon of the affixed cutting algorithm. This weakness results in the actual part of the word cut that belongs to the base itself, which happens to be similar to one of the types of affixes. With the 12 combinations, the already existing cuts can be returned according to their position. [5].

2.3. Using ProfilDocument

a. Vector Space Model (VSM)

In obtaining information, there are various techniques and approaches that can be used, ranging from vector space model method, Boolean matching and other methods, but the most widely used now is the vector space model. In an information grabber system, the similarity between documents is defined by bag-of-word representation and converted into a vector space model (VSM) [6]. This model was introduced by Salton and has been widely used. In VSM, every document in the database and user query is represented by a multi-dimensional vector according to the number of terms in the document involved. In this model:

a) The Vocabulary is a collection of all the different terms remaining of the document after pre-processing and containing the term index. These terms form a vector space.

b) Each term i in a document or query j, is assigned a weight of real value w4

c) Documents and queries are expressed as vector t dimensions dj (w1, w2 ... wn) and there are n documents in the collection, ie j = 1, 2, ... n).

b. Model Weighting

The model used in the weighting of a word by giving the value of the number of occurrences of a word as a weight. The greater the number of occurrences that in a document will be given a value to find the greatest conformity value. In calculating the weighting should note the normalization factor to the length of the document. Documents in the document collection possess varying length characteristics. Inequality occurs because long documents will tend to have the required normalization factor in weighting.

The piercing factor for each word in the document is defined as the combination of the term frequency and inverse document frequency. To calculate the word weight value used formula:

\[
\text{Win} = \text{The weighted value of a term i in document n} \\
\text{fin= The term frequency value in document n} \\
\text{log = Logarithm function}
\]

Information:
Win = The weighted value of a term i in document n
fin= The term frequency value in document n
log = Logarithm function

In the above formula, the weighting will be performed based on the logarithm calculation of term frequency i in document n. Since log (0) is not defined then it is agreed that if in document n there is no term i eat weights given the value 0, whereas if found term i then the weight value is 1 + log (fin)
because to eliminate the value 0 from \( \log (i) \) [7].

c. Measurement of Equality

The quality of a search engine can be seen from the relevance of the data obtained. Search engines use various methods to rank search results in order to provide maximum value.

To know the similarity or level of equality between documents entered with documents that have been defined category. If the similarity of the document entered is closer to the similarity of the predefined document then the document is a suitable document for entry into one of the preceding categories [8]. Similarity used in this research is similarity dice by using formula:

\[
\sin(k,j) = \frac{2 \sum f}{\sum Xjk + \sum Xjt}
\]

Information :

\( \sin(k,j) \) = size of similarity (similarity) between queries and documents

\( Xjk \) = the weight of the query term in input

\( Xjt \) = term weight of the document query

d. Measurement of the Effectiveness of Information Acquisition

1) Accuracy Testing

Precision is a level of accuracy of results against documents relevant to the document found. The formula for calculating precision is:

\[
\text{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Number of documents retrieved in search}}
\]

2) Fittings Testing

Recall is a success rate for recognizing a document of all documents that should be recognized. The formula for recall counting is:

\[
\text{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Number of documents retrieved in search}}
\]

III. PROBLEM ANALYSIS AND PROGRAM DESIGN

3.1 Activity Diagram

Activity Diagram Search

Activity diagram search translation of verses of Al-Qur'an is as follows:

![Activity Diagram Search](image)

Picture 1: Activity Translation Seek Charts Al-Qur'an verses

3.2. Data dan BasisData

a. Logical Record Structure (LRS)

The LRS form is as follows:

![Logical Record Structure](image)

Picture 2: Logical Record Structure

b. PersistentData

Persistent data is a data that will remain even if the system is not active. Persistent data can be stored in a database (database). The following are the table structures used in database design for this application:
3.3. Search Screen Design

1). Search Screen Design

In the search form translations Qur'an verses users can input the keywords you want to search in the search text box and click the search button to search for keywords input.

2). Screen Design of Search Results Form

In the search results form of the keywords that are inputted in the search form, the data shown include no, juz, letter name, verse, appear, similarity. The user can input the keyword back in the search text box and click the search button.

3.4 Interaction Components

The interaction component of the admin is as follows:
3.5 Statechart Search Diagram

Here is a statechart search diagram.

![Statechart Search Diagram](image)

3.6 Flowchart Search

Flowchart below illustrates a search process whose function displays the translation of Qur'anic verses.

![Flowchart Proses Search](image)

3.7 Search Process Algorithm

Here is a search process algorithm:

1. Entering Query
2. If text box blank = true then
3. Results not found
4. Else
5. process of punctuation
6. deletion matching process and removing stop word process
7. stemming
8. log weights for query
9. calculate similarity
   \[
   \sin(a) = \frac{\sum q \cdot w}{\sum q \cdot \sum w}
   \]
10. Search results appear
11. End if

3.8 Implementation And Analysis Of Trial Program Tests

3.9 Display Screen

Is the initial display menu when the user wants to do a search translation of verses of the Qur'an, here the user enter the keyword in accordance with the translation of the verses are sought.

![Search Results Display Screen](image)

3.10 Application Performance Measurements

a. Search result

The following is the result of searching document equations using measurements Similarity Dice on the data with query “Have faith in God” (Q1)

![Search Results Display Screen](image)

Table 6: Precision dan Recall from data Query Q1
the following is the result of searching document equations using measurements Similarity Dice from data with query “Charge your zakat” : (Q2)

<table>
<thead>
<tr>
<th>Number of documents retrieved</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of relevant documents</td>
<td>40</td>
</tr>
<tr>
<td>Recall</td>
<td>1</td>
</tr>
<tr>
<td>Precision</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Table 7: Precision and Recall data from Query Q2

<table>
<thead>
<tr>
<th>Number of documents retrieved</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of relevant documents</td>
<td>5</td>
</tr>
<tr>
<td>Recall</td>
<td>1</td>
</tr>
<tr>
<td>Precision</td>
<td>0.7</td>
</tr>
</tbody>
</table>

b. Performance Response

1. Document search time with query “Have faith in God” as follows:
   Waktu awal = 141791177,4603
   Waktu akhir = 141791177,2903
   Waktu pencarian = Waktu akhir - Waktu awal = 141791177,2903 - 141791177,4603 = 0.8300 detik

2. Document search time with query “Charge your zakat” as follows:
   Waktu awal = 141948973,2858
   Waktu akhir = 141948973,3213
   Waktu pencarian = Waktu akhir - Waktu awal = 141948973,3213 - 141948973,2858 = 0.0355 detik

3.11 Program evaluation

Evaluation of the program is one of the things that need to be done in every application development to analyze and know the results that have been achieved by the applications that have been made in this study. Likewise in this developed application, evaluated the program to analyze the results achieved in this application. In this application found some advantages and disadvantages among others:

a. Program effectiveness

From the trial results of the program by using 2 queries to represent each translation of Qur’anic verses that have obtained the result of recall and precision value as follows:

Table 8: The average value of precision and recall

<table>
<thead>
<tr>
<th>Query</th>
<th>Q1</th>
<th>Q2</th>
<th>Average value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Precision</td>
<td>0.2</td>
<td>0.7</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Q1: value recall and precision from the first query
Q2: value recall and precision from the second query kedua.

From the data in table 8 it can be concluded that the average recall process is 100% and the average precision process is 45%.

b. Program efficiency

From the test results of the program using 2 queries to represent each translation of Qur’anic verses that have obtained the results of calculation of response time as follows:

Table 9: Average value of response time

<table>
<thead>
<tr>
<th>Query</th>
<th>Q1</th>
<th>Q2</th>
<th>Average value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waktu tanggap</td>
<td>0,8300</td>
<td>1,0355</td>
<td>0,93275</td>
</tr>
</tbody>
</table>

Q1: The response time value of the first query.
Q2: The response time value of the second query

From the data in table 9 it can be concluded that the average calculation of response time is 0.93275 seconds.

3.12 Disadvantages and Advantages of the Program

a. Excess Program

1) Search by translation of the verses of Al-Qur’an Bahasa Indonesia.
2) Stemming the basic words follow the Indonesian grammar.
3) Web-based applications that can be accessed anywhere and anytime.

b. Disadvantages of the Program

1) In the search results can still bring up irrelevant documents.
2) Documents are limited to PDF only.
3) Translation of new Qur’anic verses uploaded as much as 10 juz.
4) The application can not distinguish more relevant files if they have the same similarity value.

IV. CLOSING

4.1 Conclusion

a. This application can facilitate the user in search translation of verses of Al-Qur’an in language Indonesia.
b. Can facilitate user in learning translation of verses of Al-Qur’an.
c. Helps users to easily obtain translated Al-
Qu’ran verses.

4.2 Suggestions
To minimize the number of results from the
search, it is recommended to set the value
the smallest similarity to be displayed.

a. Expand the number of basic words and
stop word so that there is not much
formed new basic word assertions are
entered into the database.

b. Expand the number of verses from Al-
Qur’an to Juz 30.

REFERENCES
[1] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze
2009, Introduction to Information Retrieval, Cambridge
University Press.

[2] www.kajianpustaka.com/2012/10/information-retrieval-system-
irs.html dilihat 5 Oktober 2014

asi Afiks Stemming untuk Bahasa Indonesia, FMIPA
Universitas Syiah Kuala Banda Aceh.

Implementasi question answering system
Untuk dokumen Tekstbaraanberbahasa
Indonesia dengan menggunakan algoritma Rabin-Karp dan stemming
Arifin Setiono oleh, Universitas Brawijaya Malang.

[5] www.asligresik.wordpress.com/201/01/06/mencari-kata-dasar-
dengan-stemming-arifin-setiono/dilihat 1 Oktober 2014

Sistem Aplikasi Pencarian Dokumen dengan Penerapan Profil
Trigram pada Vector Space Model
dengan Pembobotan Frekuensi Setiap Trigram
Berbanding Terbalik Logaritma Jumlah Trigram dalam Dokumen
dan Fungsi Kesetaraan Dice, Universitas Budi Luhur.

etode pembobotan kombinasi TF-IDF
untuk pencarian dokumen berbahasa Indonesia.

[8] Astri Juisna Saleppang, Yanuar Firdaus A., Ema Rachmawati,
2010, Analisis dan implementasikoesfesien dice pada cross
language dan belief network pada information retrieval
Institut Teknologi Telkom Bandung.

[9] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze
2009, Introduction to Information Retrieval, Cambridge
University Press.

[10] www.kajianpustaka.com/2012/10/information-retrieval-system-
irs.html dilihat 5 Oktober 2014

sasi Afiks Stemming untuk Bahasa
Indonesia, FMIPA Universitas Syiah Kuala Banda Aceh.

[12] Shandika Aprella Yola Agustira, Indriati, Achmad Ridok, 2013,
Implementasi question answering system
Untuk dokumen Tekstbaraanberbahasa
Indonesia dengan menggunakan algoritma Rabin-Karp dan stemming
Arifin Setiono oleh, Universitas Brawijaya Malang.

[13] www.asligresik.wordpress.com/201/01/06/mencari-kata-dasar-
dengan-stemming-arifin-setiono/dilihat 1 Oktober 2014

Sistem Aplikasi Pencarian Dokumen dengan Penerapan Profil
Trigram pada Vector Space Model
(dengan Pembobotan Frekuensi Setiap Trigram
Berbanding Terbalik Logaritma Jumlah Trigram dalam Dokumen
dan Fungsi Kesetaraan Dice, Universitas Budi Luhur.

Meisy. F. 2013, Perancangan sistem membalik informasi dengan
etode pembobotan kombinasi TF-IDF
untuk pencarian dokumen berbahasa Indonesia.

Astri Juisna Saleppang, Yanuar Firdaus A., Ema Rachmawati,
2010, Analisis dan implementasikoesfesien dice pada cross
language dan belief network pada information retrieval
Institut Teknologi Telkom Bandung.