Searching Concepts and Keywords in the Holy Quran

Ahmad T. Al-Taani
Alaa M. Al-Gharaibeh
Searching Concepts and Keywords in the Holy Quran

Ahmad T. Al-Taani and Alaa M. Al-Gharaibeh
Department of Computer Science, Yarmouk University, Jordan

Abstract: The Arabic language is one of the oldest languages in the world that presents its own features and challenges while searching for Arabic-based contents. The most search systems for the Holy Quran is organized around words (contained in the target verses) rather than the concepts which those words denote. As a word denotes many concepts (polysemy) and a concept can be denoted by many words (synonyms). In this study, we present a methodology to search about concepts and keywords in the Holy Quran while improving the chances of finding all desired verses. Three approaches are applied; text-based, stem-based, and synonyms-based system. These three approaches are compared with respect to precision. The text-based approach is based on a full word, the synonyms-based approach is based on the synonyms of the word, and the stem-based approach is based on the stem of the word. We apply the light stemming algorithm to find the stem of the word.

Keywords: Information Retrieval, Stem-Based System, Text-Based System, Synonyms-Based System, The Holly Quran.

1. Introduction

The Arabic language raises several challenges to Natural Language Processing (NLP) largely due to its rich morphology. Morphological processing becomes particularly important for Information Retrieval (IR), because IR needs to determine an appropriate form of words as index [1]. Arabic language is a semantic language with a composite morphology. Arabic words are categorized as particles, nouns, or verbs. Unlike most western languages, Arabic script writing orientation is from right to left. There are 28 characters in Arabic. The characters are connected and do not start with capital letter as in English. Most of the characters differ in shape based in their position in the sentence and adjunct letters [2]. The Holy Qur'an is originally written in classical Arabic language, consists of 114 surah of varying lengths. Each surah also differs from one another in terms of the number of verses (ayat).

Orthographic variations and the use of diacritics and glyphs in the representation of the language of classical Arabic increase the difficulty of stemming [3]. Information Retrieval (IR) can be defined as the study of how to determine and retrieve from a corpus of stored information the portions which are relevant to particular information needs. Let us assume that there is a store consisting of a large collection of information on some particular topics, or combination of various topics. The information may be stored in a highly structured form or in an unstructured form, depending upon its application. A user of the store, at times, seeks certain information which he may not know to solve a problem. He therefore has to express his information need as a request for information in one form or another. Thus IR is concerned with the determining and retrieving of information that is relevant to his information need as expressed by his request and translated into a query which conforms to a specific information retrieval system (IRS) used. An IRS normally stores surrogates of the actual documents in the system to represent the documents and the information stored in them [4].

Stemming is a pre-processing step in any Information Retrieval (IR) system. It is the process of removing all affixes (prefixes, suffixes and infixes) from a word to extract its root [5].

Stemming has been widely used in several fields of natural language processing such as data mining, information retrieval, and multivariate analysis. It is a method for improving the performance of information retrieval systems.

The two approaches of stemming are light stemming and heavy stemming. The former is "the process of stripping off prefixes and suffixes to produce stem of the word"; and heavy stemming is "the process of stripping off prefixes, suffixes and infixes to produce the root of the word" [6].

2. Related Work

Kadri et al. [1] proposed a new stemming method that tries to determine the core of a word according to linguistic rules. The new method shows the best retrieval effectiveness. The linguistic-based method can better determine the semantic core of a word.

Eiman Al-Shammari [2] introduced an algorithm to normalize noisy text, which only focuses on the Arabic language. And introduced a new similarity measure to stem Arabic noisy document. The need for such a new measure stems from the fact that the common rules applied in stemming cannot be applied on noisy texts, which do not conform to the known grammatical rules and have various spelling mistakes. The proposed normalization algorithm automatically group words after applying the similarity measure.

Naglaa Thabet [3] proposed a new light stemming approach that gives better results, when applied to a rich vocalized text as the Quran. The stemmer is basically a light stemmer to remove prefixes and suffixes and is applied to a version of the Quran transliterated into western script. The use of the transliteration is highly significant for resolving the problem of diacritics in the Qur'an. Given that the transliteration of the Quran is available in western script, the problem of diacritics is resolved, since in the transliterated version of the Quran, each diacritic is translated into a letter in Roman script.
El-Khoribi et al. [5] a novel Arabic text categorization system has been developed based on statistical learning. The system uses a new method for feature extraction. The system consists of three modules, namely, the stem lookup table generation module, the training module and the classification module. In the stem lookup table generation module, words are transformed to their stems using the stemmer. A feature vector is then constructed for each stem using a corpus of text. Finally, vector quantization is used to obtain a stem-label lookup table to be used in the second and third modules. In the training module, a set of documents (word sequences) corresponding to one subject are submitted to the system. The word sequences are converted to label sequences using the stemmer then the stem lookup table. The discrete data hidden Markov model is then used to generate a model for the given documents. In the classification module, the input document is submitted and converted to a label sequence.

El Emary et al. [6] proposed an Automatic Information Retrieval System to handle the Arabic data, and presented some type of comparison between the retrieval results using the vector space model in two different indexing methods: the full-word indexing and the root indexing. The proposed Automatic Information Retrieval system was implemented and built using a traditional model technique: Vector Space Model (VSM) where the cosine measure similarity was used. It provided better retrieval performance by ranking the retrieved data descending according to the similarity which gives indication to the user which may be suited his need better. The output results showed that the root indexing improved the retrieval performance more than the full-word indexing on the Arabic documents, and it reduced the size of stored data and minimized the time of system processing.

Kanan et al. [7] proposed a new light stemming technique that used a set of rules to determine if a certain sequence of characters is a part of the original word or not and this helped us solving some ambiguity problems, and introduced a way for handling the majority of broken plural forms and reducing them to their singular form and this helped us grouping words of the same meaning in a common form.

Goweder et al. [8] proposed a light stemmer that incorporates a broken plural recognition component, and evaluated it within the context of information retrieval. The results show that identifying broken plurals and reducing them to their correct stems providing a significant improvement in the performance of information retrieval systems.

Riyad Alshalabi [9] provided a technique for extracting the trilateral Arabic root for an unvocalized Arabic corpus. It provides an efficient way to remove suffixes and prefixes from the inflected words. Then it matches the resulting word with the available patterns to find the suitable one and then extracts the three letters of the root by removing all infixes in that pattern.

Aljlayl [10] presented three methods of query translation using a bilingual dictionary for Arabic-English CLIR. First, the Every-Match (EM) method is presented. This method yields ambiguous translations since many extraneous terms are added to the original query. To disambiguate the query translation, the First-Match (FM) method is presented that considers the first match in the dictionary as the candidate term. Finally, the Two-Phase (TP) method is presented to reduce the ambiguity of the every match method. In this method; all the terms that do not retranslate to the original Arabic query word are ignored.

Noordin et al. [11] proposed block-level link, stemming, sentence completion, and other common retrieval techniques like phrase searching for Qur'anic text. This paper proposed citation analysis to determine the inclusion of items in the database/storage. Block-level link involves organizing texts in separate blocks, and each block is linked to its own list of articles and sources (stored items). Chen et al [12] identified sets of prefixes and suffixes. They identified the prefixes that must be removed: 19 three-character, 14 two-character, and 3 one character, and the suffixes: 18 two-character, 4 one character. To remove the prefixes and suffixes in the predefined sets, each algorithm proposes their own rules. For example, they applied the following rules: If the word is at least five-character long, remove the prefixes of three characters: ﺣ, ﺍ, ﺜ, ﺪ, ﺰ, ﺑ, ﺕ, ﺞ, ﺔ, ﺢ. If the word is at least four-character long, remove the first two characters: ﺨ, ﻤ, ﻦ, ﺔ, ﺢ, ﺪ, ﺞ, ﺖ, ﺑ, ﺕ, ﺪ, ﺞ, ﺖ. If the word is at least four-character long and begins with ﺔ or with ﺢ remove ﺔ or ﺢ.

Al-Shammari et al. [13] introduced a novel approach for stemming Arabic documents, and compared the performance of the new algorithm with that of the Khocha stemming algorithm using stemming weight as an evaluation criterion. They showed that the use of presently neglected Arabic stop words can be highly effective and can provide a significant improvement when processing Arabic documents. A new framework is introduced to normalize Arabic documents by overcoming the limitations of previous approaches, caused by the early removal of stop words.

3. Qur'anic Search System (QSS)

Almost all information retrieval systems work in the same way and pass several steps before retrieve the most relevant related documents in the field of some formulated queries. These steps deal with a set of documents and its text contents deal with representations of documents. The implemented system accepts easily a query (text query) which is related to the Holy Qur'an subjects and the system retrieves verses in which the query key terms are mentioned. The methodology of this research is divided into three phases discussed in next subsections and illustrated in Figure 1. These phases basically are first analyzing the Holy Qur'an (representing its keywords in database file). Second implementing the light stemming algorithm. Finally analyzing the queries in three search systems; text based system, stem based system, and synonyms based system.

3.1. Pre-Processing

Qur'anic text is pre-processed in this research. Each text file will be read verse by verse, and then each verse will be read word by word. Preprocessing in this project can be divided into four text operations:
(1) Lexical analysis of the text with the objective of treating digits, hyphens, punctuation marks.
(2) Elimination of the stop words.
(3) Stemming
(4) Selection of index term

![Diagram of QSS Processes](image)

**Figure 1: The Processes in QSS**

### 3.1.1. Lexical Analysis of the Text Files

The text of every text file is converted into a stream of words (the candidate words to be adopted as index terms). The following three cases have to be considered with care: not Arabic word, punctuation marks, digits. If the word is not an Arabic word then the word is considered as a useless word. Punctuation marks are removed entirely in the process of lexical analysis, the punctuation marks that are removed , " " . & @ # % ^ * ) ( - _ ; . If the word contains digits then the word is considered as a useless word.

### 3.1.2. Stop Words Elimination

Stop words are words which are too frequent among text files which do not carry a particular and useful meaning for IR. Elimination of stop words has an additional important benefit. It reduces the size of the indexing structure. Stopwords table is built that contains 1219 words.

The word categories that were used are:
- Adverbs
- Prepositions
- Pronouns
- Interrogative Pronouns
- Conditional Pronouns

For every word in the text if it is a stop word then the word is considered as a useless word.

### 3.1.3. Stemming

Stemming of the remaining words with objective of removing affixes (prefixes and suffixes) and allowing the retrieval of documents containing syntactic variations of query terms (e.g., بُلُوْلِيْلٌ، صلالة، صلَّنا، صَلَّنا). Light stemming algorithm is used in this project.

### 3.1.4. Index Term Selection

Index term (or keyword) a pre-selected term which can be used to refer to the content of a document [14]. In this project all the words in the text files are used as index terms. The words of each text file are inserted in a column in the table in Microsoft access before and after preprocessing, and each row represent the words of each verses also there are two columns that contain the sura name and verse ID.

### 3.2 Implementing Light Stemming Algorithm

Any light stemming algorithm has common steps of normalization and stemming, and the main difference among different light stemming algorithms is the number of prefixes and suffixes removed from each one. During the normalization process, all diacritics are removed. The light stemming algorithm is represented in details by these steps:

- Remove Diacritics.
- Normalize the Word.
- Remove Prefixes.
- Recursively Removes Suffixes.

#### Step 1: Remove Diacritics

Diacritics are commonly used in religious scripts such as the holy Qur'an. Arabic grammar it is important to understand the representation of the Arabic language in the written form in order to understand the components of a word that would actually be depicted in its written form. Short vowels and other diacritics are removed from every text file. Short vowels include the fatha ( ﻓِ), Domma ( ﻓُ), and kasra ( ﻓَ). Other diacritics such as the shadda ( ّ), sikkun ( ﻓِ), and tanween or double fatha, Domma, and kasra.

#### Step 2: Normalization of the Words

Normalization is the process of unification of different forms of the same letter as follows:

- Normalize ١ to ١.
- Normalize ٥ to ٥.
- Normalize ی to ٨.
- Normalize the ٨ and the sequence ٨ to ٨ to ٨ to ٨ to ٨.

#### Step 3: Remove Prefixes

The algorithm removes a set of prefixes as presented by many authors [1, 2, 15].

After removing these prefixes it checks if the word length is less than 3 letters, in this case these prefixes considered as a main part of the word and so the removed prefix returned back to the word. Table 1 shows a set of Arabic prefixes.

<table>
<thead>
<tr>
<th>Prefix 1</th>
<th>مولؤي</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prefix 2</td>
<td>آت إل إس ل ظرف علم ل ظرف علم</td>
</tr>
<tr>
<td>Prefix 3</td>
<td>قبل بال است قال كلأل</td>
</tr>
<tr>
<td>Prefix 4</td>
<td>قبل كلأل</td>
</tr>
</tbody>
</table>

#### Step 4: Remove Suffixes

In this step a set of suffixes are recursively removed from the tail of word. The longest suffix is removed first, then the shorter. This process is recursive because most suffixes are compound of pronouns, gender, for example the word (هَمْ) has a composite suffix (هَمْ وَهُمْ) which is made from two parts (هَمْ) for feminine plural and the pronoun (حَمْمَهُمْ). Also as done in the previous step the algorithm checks if the word length is greater than 3 letters in order to prevent remove a main part of the word.

Table 2 shows a set of Arabic suffixes as presented by...
3.3. The Proposed Approach

The proposed algorithm that is used for searching the Qur'an about keywords and concepts can be described as shown in the following steps.
1. Select the text file that contains Qur'anic text as search data.
2. Build the stop word table.
3. Build the index table
4. Read the text file verse by verse.
5. IF the word is not an Arabic word THEN consider this word as a useless word.
6. IF the word contains digits THEN consider this word as a useless word.
7. IF the word length is less than three characters THEN consider this word as a
   Useless word
8. Remove diacritics.
9. Normalize the word.
10. Apply the stop word test to check if this term is a stop word, if the term is a stop word, discard it. Otherwise, if you are using full word indexing go to step 3.5 In case of root indexing go to step 3.4.1
11. Remove prefixes and recursively remove suffixes, then go to step 3.5.
12. Gather the words of each verse
13. Insert the group of words of each verse to the index table.
14. Match the word of the query with the words in the index table.
15. Retrieve the verses which may more relevant to the user demand.

3.4 Analyzing the Queries

3.4.1 Queries in Stem Based System

This search method is based on the root of the words, each word of the user query go back to the previous phase (text files preprocessing) and do all preprocessing steps (removing stop words, Normalization, and stemming). Each root words of the user query is matched to the root word in the index table then retrieves the verses that have the same root word.

When key words tool is used, the system searches in the Qur'anic text after root word indexing.

3.4.2 Queries in Text Based System

This search method is based on the full word, each words of the user query is matched to the same word in the index table then the system retrieves the verses that have the same word. When key words tool is used, the system searches in the Qur'anic text after full word indexing.

3.4.3 Queries in Synonyms Based System

This search method is based on the synonyms of the words, each word of the user query go to the Arabic thesaurus in Microsoft word 2003 and get the synonyms of each word. Each synonyms word of the user query is matched to the same word in the index table then the system retrieves the verses that have the same word. When key words tool is used, the system searches in the Qur'anic text after full word indexing.

4. Experiments and Evaluation

This section presents an evaluation of text based system, stem based system, and synonyms based system when applying them on the Holy Qur'an to search for keywords and concepts. And determining the efficiency of light stemming algorithm which is used by stem based system. The quality of the retrieved verses is evaluated by specialist in Islamic who determine relevant verses to the test query. For the following subsections explain the evaluation measures, queries collection that have been used for evaluation, the experiment result, introduce the evaluation results and the experiments environment.

4.1 Evaluation Measure

The precision criterion in the field of information retrieval is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search. This criteria range between 0 and 1. A perfect Precision score of 1 means that every result retrieved by a search was relevant (but says nothing about whether all relevant documents were retrieved).

In this research, a precision criterion is calculated manually for each query runs in one of the three search systems. At the end of applying all selected queries, the average precision is calculated, to specify the effectiveness of three search systems.

Precision is calculated in this project depending on this formula:

\[
\text{Precision} = \frac{\text{number of relevant verses}}{\text{number of the verses retrieved}}
\]

4.2 Collecting and Applying Queries for Evaluation

This objective is accomplished by applying different queries from the Holy Qur'an, and calculating the Precision. We examined the three search systems by using 10 queries for each one when searching about words.

Our experimental results also concerned with the evaluation of retrieval efficiency and its effectiveness by comparing the results of text based system, stem based system, and synonyms based system precision measurement.

4.3 Experiments and Results

In this section, we calculate the precision for each query which runs on each one of the search systems by using this formula:

\[
\text{Precision} = \frac{\text{number of relevant verses}}{\text{number of the verses retrieved}}
\]

At the end of applying all selected queries, the average precision is calculated. Experiments and results for stem based system, synonyms based system, and text based system are presented in next subsection.
4.3.1 Experiments and Results for Stem Based System

We applied 10 queries for searching about words in the Holy Qur'an, when stem based system is selected for search it finds the stem of the word by using light stemming algorithm then match it with the same stem word that is stored in the database. We choose 10 words that are different in meaning and form. Table 3 shows the words query, the stem of the words that is founded by the system, and shows the number of relevant verses and the number of the retrieved verses, and shows the precision for each query.

Table 3: Precision for the Words Queries Using Stem Based System

<table>
<thead>
<tr>
<th>Words</th>
<th>Stem</th>
<th># of relevant verses found</th>
<th># of the retrieved verses</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>الحسنة</td>
<td>حسن</td>
<td>156</td>
<td>156</td>
<td>1</td>
</tr>
<tr>
<td>قائمة</td>
<td>قائمة</td>
<td>6</td>
<td>8</td>
<td>0.75</td>
</tr>
<tr>
<td>مهين</td>
<td>مهين</td>
<td>17</td>
<td>24</td>
<td>0.70</td>
</tr>
<tr>
<td>منكر</td>
<td>منكر</td>
<td>14</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>واقعة</td>
<td>واقعة</td>
<td>2</td>
<td>7</td>
<td>0.28</td>
</tr>
<tr>
<td>متكبر</td>
<td>متكبر</td>
<td>5</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>فاقد</td>
<td>فاقد</td>
<td>8</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>فاكهة</td>
<td>فاكهة</td>
<td>11</td>
<td>32</td>
<td>0.34</td>
</tr>
<tr>
<td>مكر</td>
<td>مكر</td>
<td>17</td>
<td>19</td>
<td>0.89</td>
</tr>
<tr>
<td>مفترون</td>
<td>فتير</td>
<td>33</td>
<td>46</td>
<td>0.78</td>
</tr>
</tbody>
</table>

The average of the precision is 0.77 when stem based system is used for searching about words.

4.3.2 Experiments and Results for Synonyms Based System

We applied 10 queries for searching about words in the Holy Qur'an, when synonyms based system is selected for search it finds the synonyms of the word by using Arabic thesaurus in Microsoft office then match each word with the same word that is stored in the database. I choose 10 words that are different in meaning and form. Table 4 shows the words query, the synonyms of the words that is founded by the system, the number of relevant verses, the number of the retrieved verses, and shows the precision for each query.

Table 4: Precision for the Words Queries Using Synonyms Based System

<table>
<thead>
<tr>
<th>Words</th>
<th>Synonyms</th>
<th># of relevant verses</th>
<th># of the retrieved verses</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>الحسنة</td>
<td>حسن</td>
<td>30</td>
<td>33</td>
<td>0.90</td>
</tr>
<tr>
<td>قائمة</td>
<td>قائمة</td>
<td>8</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>مهين</td>
<td>مهين</td>
<td>15</td>
<td>17</td>
<td>0.70</td>
</tr>
<tr>
<td>منكر</td>
<td>منكر</td>
<td>32</td>
<td>35</td>
<td>0.91</td>
</tr>
<tr>
<td>واقعة</td>
<td>واقعة</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>متكبر</td>
<td>متكبر</td>
<td>10</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>فاقد</td>
<td>فاقد</td>
<td>58</td>
<td>69</td>
<td>0.84</td>
</tr>
<tr>
<td>مكر</td>
<td>مكر</td>
<td>11</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>مفترون</td>
<td>مفترون</td>
<td>49</td>
<td>56</td>
<td>0.87</td>
</tr>
<tr>
<td>منكر</td>
<td>منكر</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The average of the precision is 0.92 when synonyms based system is used for searching about words. That is show a good result for synonyms based system.

4.3.3 Experiments and Results for Text Based System

We applied the 10 queries for searching about words in the Holy Qur'an, when text based system is selected for search it match the word with the same word that is stored in the database. We choose 10 words that are used in stem based system and synonyms based system. We choose the same sample of the words to determine which one of the search system is the best with respect to the precision. Table 5 shows the words query, the number of relevant verses, the number of the retrieved verses, and the precision for each query.

Table 5: Precision for the Words Queries Using Text Based System

<table>
<thead>
<tr>
<th>Words</th>
<th># of relevant verses found</th>
<th># of the retrieved verses</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>الحسنة</td>
<td>11</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>قائمة</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>مهين</td>
<td>16</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>منكر</td>
<td>14</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>واقعة</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>متكبر</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>مكر</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>مفترون</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The average of the precision is 0.80 when text based system is used for searching about words.

The comparison between each three systems is illustrated by Figure 2.

Figure 2: Comparison Between Three Search Systems with Respect to Precision

6. Conclusion

Three search techniques are applied Stem based system, Synonyms based system, and Text based system to determine the effectiveness of verses retrieval in Holy Qur'an by calculating the precision for each query. We compute the precision on an Arabic set of queries (10 queries) for each one of the three search techniques. Our experimental results showed that using the Synonyms based system will give the highest precision than using the Stem based system and Text based system. When searching about words, the average precision in the synonyms based system is 0.92 whereas the average precision of the text based system is 0.80, and the average precision of the stem based system is 0.77. Text based system gives the limitation in retrieval, for example when we search about (النظرية) (in Arabic) the retrieved verses is zero because the word contains the (ال) and the existed verses contain the word (فلكية) so it doesn't match with the
existed verses. In synonyms based system gives the synonyms of the word (فاكهة, فاكهة) then match the synonyms of the word with existed verses that contains the word (فاكهة) so it retrieves all verses that contain (فاكهة). In case of stem based system it find the stem of the word (فاكهة) and produce (فاك) so it retrieves all the existed verses that contain (فاك) but also retrieve irrelevant verses that contains another word that have the same stem. Therefore, in stem based system the chance of retrieve the irrelevant verses are higher than in synonyms based system.

7. Future Work

We want to develop our search system by applying the stem-synonyms based system which finds the stem of the words after removing the stop word from text and finds the synonyms of the words. The purpose of this system is to gets identical words after removing prefix and suffix and compares the results of this system with the stem based system and synonym based system. And we will try to calculate the recall criterion which defined as the number of relevant documents retrieved by a search divided by the total number of existing relevant documents (which should have been retrieved).

References:


