Analysis and implementation of computer-based system development of stemming algorithm for finding Arabic root word

F E Zamani*, K Umam, W D I Azis and W S Abdillah
UIN Sunan Gunung Djati Bandung, Jl. AH Nasution No. 105 Bandung, West Java, Indonesia
*fadliemsa@gmail.com

Abstract. At present many experts in the field of information technology have designed and developed algorithms to solve stemming problems, especially in Arabic. But of the many stemming analyses in Arabic, there is no standardization of a good stemming algorithm in analysing the accuracy of the text in the Koran. The development of stemming in the Koran is significant to work because it supports the Sharaf classification in the Koran to understand the meaning of every word in the Qur'an. One stemmer or stemming an algorithm to find the primary form of an Arabic word is the Khoja Stemmer algorithm. The way of working from Khoja Stemmer is to try to find the root of an Arabic word by removing the longest prefix and the longest suffix of a word, then try to determine the root of the remaining words using the root word dictionary. In this study, the Khoja Stemmer built was able to calculate the average stemming of the Koran by 95.295%. But the root words produced by Khoja Stemmer if manually checked, there are still several errors. Thus, an Al-Quran dictionary is needed to analyse each stemming result conducted by Khoja stemmer in stemming the Koran.

1. Introduction
In text mining research Arabic has a unique structure because it has morphology and grammar that are different from other languages [1], problems that are often encountered in stemming are excessive cutting or over stemming and under stemming research for stemming or root word search very few are found [2], one of the algorithms developed is the Khoja algorithm [3]. In the research that has been done to process stemming in Arabic text, there is still some incorrect accuracy due to the manual weighting or stemming process using novel root which still has weaknesses in it [4]. In this case, the author tries explicitly to examine the stemming process because stemming is an essential step in processing words in Arabic, especially for the Koran. Besides that, the right algorithm is also needed to get the accuracy of good stemming results to make it easier for people to learn the Koran. The Khoja Stemmer algorithm is one of
the most popular and widely used Arabic language stemmers. The way the Khoja Stemmer algorithm works is to eliminate the longest suffix and the longest prefix of a word. Then match the remaining words with verbal patterns and nouns (noun), to get the root of the word. Therefore, researchers built a stemming application on the Qur'an using the Shereen Khoja Stemmer algorithm to find out whether the algorithm suitable for stemming the Al-Quran making it easier for students to understand morphology or Sharaf in the Koran compared to algorithms that have been developed previously from accuracy, accuracy and stemming results [5,6] or a stemming algorithm that must change to another form first [7].

2. Materials and method
The system built aims to find the primary form of a word in every word in Al-Qur'an. The input is an Al-Qur'an dataset consisting of 30 juz which is loaded into the application. Then the app will read the input dataset, and the tokenization process is performed on each input line. After all the input lines are formed into a token, then applies will conduct stemming in each token to produce a root or root of each token. A root is generated from the associated root word dictionary dataset with the application. After that, the output is generated in the form of stemming from each word in the Qur'an and the accuracy of the system in the process of stemming.

The following is a flowchart about an overview of the system being built.

Figure 1. Flowchart.
2.1. Read dataset input
Input is an Al-Quran dataset in a .txt format consisting of 30 Juz. The data is done twice as divided into two parts, i.e., 1-15 in the first part and juz 16-30 in the second part to minimize execution time. At this stage, the application will read and display the results of the Al-Quran dataset input. Then the app will process stem. The following process will be elaborated stem in building stemming applications of the Koran using the Khoja Stemmer algorithm.

2.2. Tokenize
At this stage, the input is spelled out in a row for the tokenization process; one line consists of one verse of the Koran. The system performs the tokenization process of the Al-Quran dataset input by breaking the input sentence into words for each input. The following is an example of tokenization in verse in the Koran, which is applied to Khoja Stemmer.

![Figure 2. Example of tokenization in verse in the Koran.](image)

2.3. Find stem / Root
After all the input sentences are changed to per word in the previous stage, the system will look for the root or root of the word. Following are the steps [8,9]:

- Removing all diacritics or harakat from the results of the word which have been typed.
- Determine each waqof sign into Non-Letter Words.
- Normalize the word that has been typed, which consists of:
  - Change ﴾٣﴿ to ﴾١﴿
  - Change ﴾٣﴿ to ﴾١﴿
  - Change ﴾٣﴿ to ﴾١﴿
  - Change ﴾٣﴿ to ﴾١﴿
  - Change ﴾٣﴿ to ﴾١﴿
- Tokenization results match each word with the root word dictionary dataset to take root or roots of his word by removing any prefix (prefix of the word) and the suffix (suffix word) the longest contained in words.
- Then it is determined the removal process of affixes (prefix and suffix).

After all the prefixes and suffixes are omitted, the remaining words will be matched with the root word dictionary dataset to determine the root word. Following are the root search rules found in the stemmer Khoja algorithm:

- If the root generated consists of 3 letters, the system would output that has been determined in the Tri roots dataset.
• If the root generated consists of 4 letters, the system would output the specified number on the Quad roots dataset.
• If the system can find the root or root of a word, the system will issue stemmed Words i.e. the output of all the words that succeed in stemming. If the system

3. Results and discussion
The results of the trial use Khoja Stemming with the output of fi’il madhi in the table below, by inputting the word fi’il mudhori ‘which is in Al-Quran Juz 30. The results of the trials can be seen in table 1.

3.1. Testing Khoja for stemming

| Table 1. Tests using Khoja with fi’il madhi output. |
|-----------------------------------------------|
| **Data** | **Fi’il Mudhori** | **Fi’il Madhi** (Khoja) | **Accuracy** | **Basic Word** |
|-----------|--------------------|------------------------|-------------|---------------|
| 1.        | ﷐ ﷑ ﷑ ﷗ ﷘ ﷠ ﷣ | ﷗ ﷗ ﷗ | Right | ﷗ ﷗ ﷗ |
| 2.        | ﷐ ﷑ ﷑ ﷗ ﷘ ﷠ ﷣ | ﷗ ﷗ ﷗ | Right | ﷗ ﷗ ﷗ |
| 3.        | ﷐ ﷑ ﷑ ﷗ ﷘ ﷠ ﷣ | ﷗ ﷗ ﷗ | less precise | ﷗ ﷗ ﷗ |
| 4.        | ﷐ ﷑ ﷑ ﷗ ﷘ ﷠ ﷣ | ﷗ ﷗ ﷗ | less precise | ﷗ ﷗ ﷗ |
| 5.        | ﷐ ﷑ ﷑ ﷗ ﷘ ﷠ ﷣ | ﷗ ﷗ ﷗ | less precise | ﷗ ﷗ ﷗ |
| 6.        | ﷐ ﷑ ﷑ ﷗ ﷘ ﷠ ﷣ | ﷗ ﷗ ﷗ | less precise | ﷗ ﷗ ﷗ |
| 7.        | ﷐ ﷑ ﷑ ﷗ ﷘ ﷠ ﷣ | ﷗ ﷗ ﷗ | less precise | ﷗ ﷗ ﷗ |
| 8.        | ﷐ ﷑ ﷑ ﷗ ﷘ ﷠ ﷣ | ﷗ ﷗ ﷗ | Right | ﷗ ﷗ ﷗ |
| 9.        | ﷐ ﷑ ﷑ ﷗ ﷘ ﷠ ﷣ | ﷗ ﷗ ﷗ | less precise | ﷗ ﷗ ﷗ |

In the experiments that have been carried out, there are words that are correct in cutting, but overall, the method of speech for stemming has quite a high accuracy. From one sentence, almost everything is true in cutting the word.

3.2. System accuracy
The following is a table of details in numbers from the results of application testing.

| Table 2. Experiment result. |
|----------------------------|
| **Data** | **Stemmed Words** | **Words Not Stemmed** | **Stop words** | **Non Letter Words (Waqof Sign)** | **Total Words** | **Accuracy (%)** |
|-----------|--------------------|-----------------------|---------------|-----------------------------------|----------------|------------------|
| Juz 1-15  | 23.278             | 1.898                 | 13.754        | 2.556                             | 41.486         | 95.42            |
| Juz 16-30 | 24.191             | 1.984                 | 13.140        | 1.823                             | 41.138         | 95.17            |

95.295

From the table above, the test data is divided into two parts to minimize application execution time [10]. In the first test data, namely juz 1-15 stemmed words were 23.278, words not derived totalled 1,898, stop
words amounted to 13,754 words, and non-letter words numbered 2,556 so that all the words processed in the first test data amounted to 41,486. The accuracy of stemming produced in the first test data is 95.42%.

In the second test data, namely juz 16-30 produced stemmed words totalling 24,191, words not stemmed numbered 1,984, stop words numbered 13,140 words, and non-letter words totalled 1,823 so that all the words processed in the second test data amounted to 41,138. The accuracy of stemming produced in the second test data is 95.17%. So the total stemming accuracy generated from all test data is 95.295%.

4. Conclusion
Although the Khoja Stemmer Algorithm has the accuracy of stemming the Al-Quran by 95.295%, the root generated by stemmer when checked manually and compared to the Al-Quran dictionary, it turns out there are still many mistakes. So that this stemmer is not suitable to be applied to stemming applications for the Koran.

References
[1] Ababneh M, Al-Shalabi R, Canaan G and Al-Nobani A 2012 Building an effective rule-based light stemmer for Arabic language to improve search effectiveness International Arab Journal of Information Technology (IAJIT) 9 368-372
[2] Jaffar A, Masnizah M and Ghassan K 2013 Enhanced Arabic information retrieval: Light stemming and stop words In Soft computing applications and intelligent systems (pp 219-228) Springer, Berlin
[3] Firmanto A, Widowati S and Rakhmatsyah A 2011 Implementation of Principal Component Analysis and Back propagation Neural Network in Classifying the Observation of the Verses of Knowledge in the Koran
[4] Al-Kabia M N, Kazakzehb S A, Atab B M A, Al-Rababahc S A and Alsmadid I M 2015 A novel root based Arabic stemmer Journal of King Saud University - Computer and Information Sciences 27 (2) 94-103
[5] Zitouni A, Damankesh A, Barakati F, Atari M, Watfa M and Oroumchian F 2010 Corpus-based Arabic stemming using N-grams Proceedings of the 6th Asia Information Retrieval Society Conference (AIRS2010) 6458 280-289
[6] Thabet N 2004 Stemming the Qur’an Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages 85-88
[7] Sawalha M and Atwell E 2008 Comparative Evaluation of Morphological Arabic Language Analysts and Stemmers White Horse Research Online 107-110
[8] Khoja S and Garside R 1999 Stemming Arabic text, technical report (Lancaster: Lancaster University Computing Department)
[9] Al-Shammari E and Lin J 2008a Towards an error-free Arabic stemming Proceedings of the 2nd ACM Workshop on Improving non-English web searching 9-16
[10] Froud H, Lachkar A and Alauoi Ouatik S 2012 A comparative study of root-based and stem-based approaches for measuring the similarity between Arabic words for Arabic text mining applications Journal of Advanced Computing (ACIJ) 3 55-67