Processing the Text of the Holy Quran: a Text Mining Study

Mohammad Alhawarat
Department of Computer Science
College of Computer
Engineering and Sciences
Salman Bin Abdulaziz University
Al-Kharj, Kingdom of Saudi Arabia

Mohamed Hegazi
Department of Computer Science
College of Computer
Engineering and Sciences
Salman Bin Abdulaziz University
Al-Kharj, Kingdom of Saudi Arabia

Anwer Hilal
Department of Computer Science
Preparatory Year Deanship
Salman Bin Abdulaziz University
Al-Kharj, Kingdom of Saudi Arabia

***Abstract*—The Holy Quran is the reference book for more than 1.6 billion of Muslims all around the world Extracting information and knowledge from the Holy Quran is of high benefit for both specialized people in Islamic studies as well as non-specialized people. This paper initiates a series of research studies that aim to serve the Holy Quran and provide helpful and accurate information and knowledge to the all human beings. Also, the planned research studies aim to lay out a framework that will be used by researchers in the field of Arabic natural language processing by providing a "Golden Dataset" along with useful techniques and information that will advance this field further. The aim of this paper is to find an approach for analyzing Arabic text and then providing statistical information which might be helpful for the people in this research area. In this paper the holly Quran text is preprocessed and then different text mining operations are applied to it to reveal simple facts about the terms of the holy Quran. The results show a variety of characteristics of the Holy Quran such as its most important words, its wordcloud and chapters with high term frequencies. All these results are based on term frequencies that are calculated using both Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) methods.**

Keywords*—Holy Quran; Text Mining; Arabic Natural Language Processing

I. Introduction

The Holy Quran is the reference book for more than 1.6 billion of Muslims all around the world. Extracting information and knowledge from the Holy Quran is of high benefit for both specialized people in Islamic studies as well as none-specialized people. The Holy Quran is the word of God and hence needs careful handling when processed by automated methods of machine learning, natural language processing and artificial intelligence. The language of the Holy Quran is Arabic which is known to be one of the challenging natural languages in the field of natural language processing and machine learning. This is due to some of its special characteristics such as diacritic, multiple derivations of words, complicated Diglossia and others [1], [2], [3], [4]. These make dealing with Arabic language a challenging task when applying machine learning and artificial intelligence techniques.

Few research studies have considered the Arabic text of Quran [5], [6], [7], [8], instead many studies deal with the translations of the meaning of the words of the holy Quran [9], [10], [11], [12], [13], [14]. Kais and his colleagues have created an open source Quranic corpus [15] using both arabic words as well as translations of these words.

To the best of our knowledge, there is no research study that analyzed the Arabic text of the holy Quran using text mining techniques the way it is done in this paper. The aim of this paper is to find an approach for analyzing Arabic text and then providing statistical information might be helpful for the people in this research area. Also, this study aims at providing a framework for future studies in this field of study. The paper used the holly Quran to achieve these aims; first the holly Quran text is preprocessed and then different packages of *R* has been used such as: `tm and RWeka`.

It is important here to stress that this is not a religious study, instead it is an automated study that gives statistical results. These results in no way are accepted until they are approved by Islamic scholars.

The rest of the paper is organized as the following: Section II is dedicated to explain the process of preparing the text of the Holy Quran, in section III experiments that are applied to the text of the Holy Quran are explained, in section IV the results that are obtained in the paper are discussed, and finally section V concludes the paper.

II. Preparing Text

The holy Quran has around 78 thousand words. These words are grouped into verses. A set of verses are grouped into: parts, chapters, group (Hizb)or Hizb quarter.

The text of the Holy Quran has been first downloaded from Tanzil project website[16] which represents an authentic verified source of the holy Quran text. The downloaded file includes the whole text of the Quran without diacritic. The file has been divided semi-automatically into five different set of documents:

- 114 Chapter (Sura),
- 30 Part (Juza),
- 60 Group (Hizb),
- 240 Hizb Quarter or

- 6236 Verse.

After that the encoding of the files have been converted into `CP1256` because the original encoding of the files are unreadable by *R*.

The files have then been read as a corpus and cleaned by removing stop words. *R* does not support stop word removal for Arabic language, hence a list of around 2000 stop words have been created manually and manipulated from different sources including [17]. Also, *R* does not support stemming for Arabic language, therefore simple cleaning has been applied on the corpus such as normalizing some words by replacing different shapes of the word with its normal form. For example the words:

لله ، والله ، بالله ، فالله ، تالله ، فلله ، اللهم ، آلله

have all been replaced by الله.

Also, the words:

ربنا ، ربهم ، ربكم ، ربك ، ربها ، ربه ، ربي ، برب ، ورب

have all been replaced by رب.

The variations of the previous two words are due to the some of the prefixes and suffixes of the Arabic language. Note that both الله and رب are considered stems rather than roots for the aforementioned variations for both words.

This procedure has been applied to few words because the processing of all shapes of all words (the stemming procedure) is out of the scope of this paper. Although stemming algorithms for Arabic language do exist, but their accuracy still need to be enhanced. For this reason, applying such algorithms is not suitable for the holy Quran as it is the word of God, and hence errors are not tolerated.

After that the corpus have been converted into both Term-Document and Document-Term matrices as both needed for different type of experimentations. The next section illustrate different experimentation applied to both matrices.

III. Experiments

In this section different set of experimentations are carried out on the text of the holy Quran. These experiments are based on the Term matrices that are built according to two selected partitioning methods: Chapters and Parts. These are chosen as examples because using all partitioning methods will produce numerous results and figures.

The experiments will manipulate the text of the holy Quran in order to produce its most frequent terms, wordcloud and clusters.

A. Experiments on Chapters of the holy Quran

In this subsection the text of the holy Quran is studied based on its 114 chapters. Each chapter in the holy Quran talks in general about one theme but it might include different topics. But it is considered the most coherent partitioning methodology.

1) Most Frequent Words: The term-document matrix has been used in one experimentation setup to calculate the frequency of the terms of the holy Quran. There are many frequent terms in the Holy Quran, hence figure 1 depicts the most 30 frequent words. These are calculated using TF measure. Also, the most frequent 30 terms in the holy Quran is calculated based on TF-IDF as shown in figure 2.

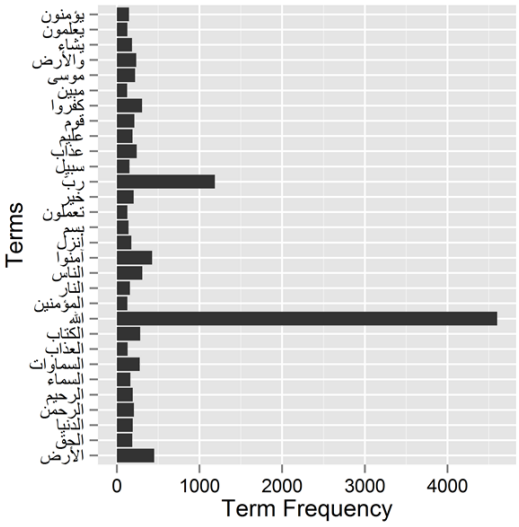

Fig. 1: Most frequent terms in the holy Quran measured by TF (Based on Chapters)

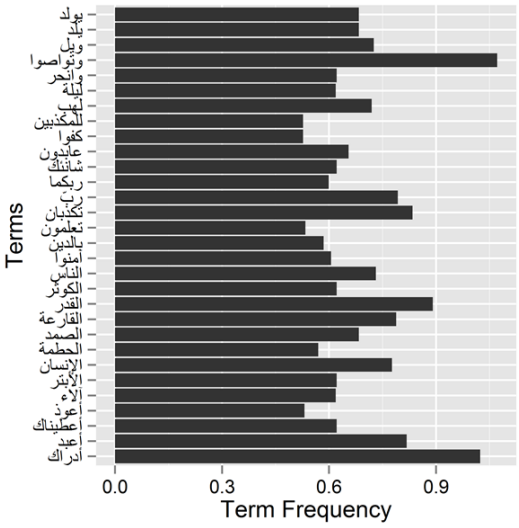

Fig. 2: Most frequent terms in the holy Quran measured by TF-IDF (Based on Chapters)

2) Word Cloud: It is important for specialized as well as non-specialized people in Islamic studies to visualize the words of the Holy Quran. Figures 3- 4 show the wordcloud of the Holy Quran for the most frequent 100 words measured using TF and TF-IDF measures respectively.

Fig. 3: Wordcloud for the most frequent 100 words in the holy Quran measured by TF (Based on Chapters)

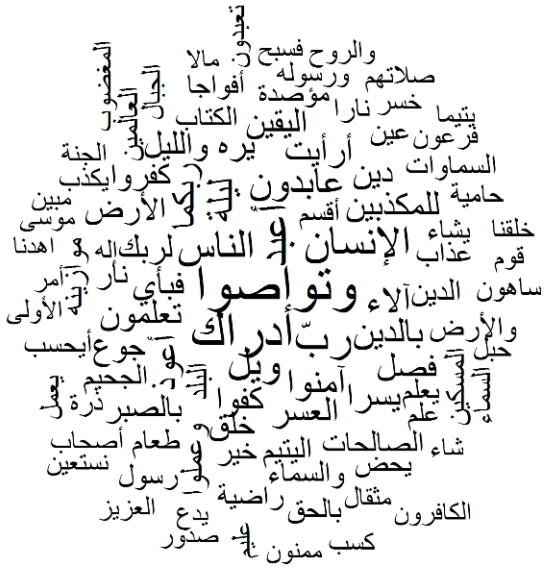

Fig. 4: Wordcloud for the most frequent 100 words in the holy Quran measured by TF-IDF (Based on Chapters)

It's also important to visualize the wordcloud of the holy Quran based on bi-gram, tri-gram and four-gram words. These appear in figures 5- 7.

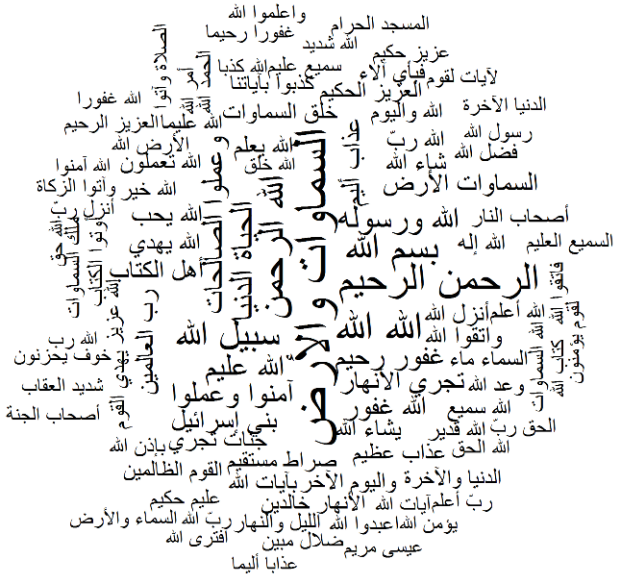

Fig. 5: Wordcloud for the most frequent 100 bi-gram terms (Based on Chapters)

Fig. 6: Wordcloud for the most frequent 100 tri-gram terms (Based on Chapters)

Fig. 7: Wordcloud for the most frequent 100 four-gram terms (Based on Chapters)

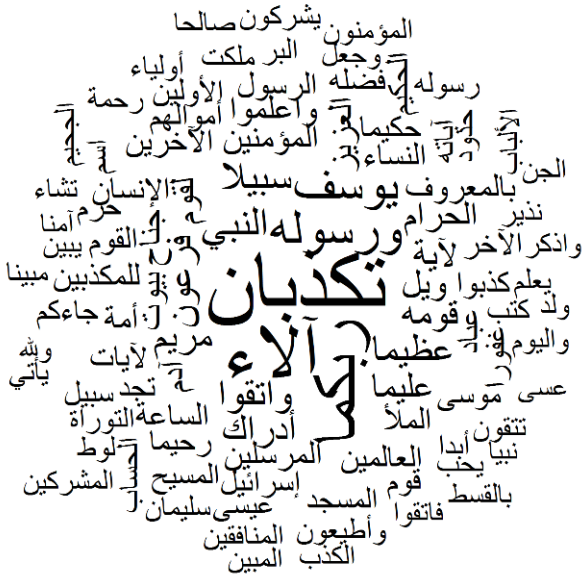

Fig. 9: Wordcloud for the most frequent 100 words in the holy Quran measured by TF-IDF (Based on Parts)

B. Experiments Based on Parts of the holy Quran

The same experimentation setups that are used for Chapters of the holy Quran are applied on the Parts of the holy Quran and samples of the results are depicted in figures 10- 11.

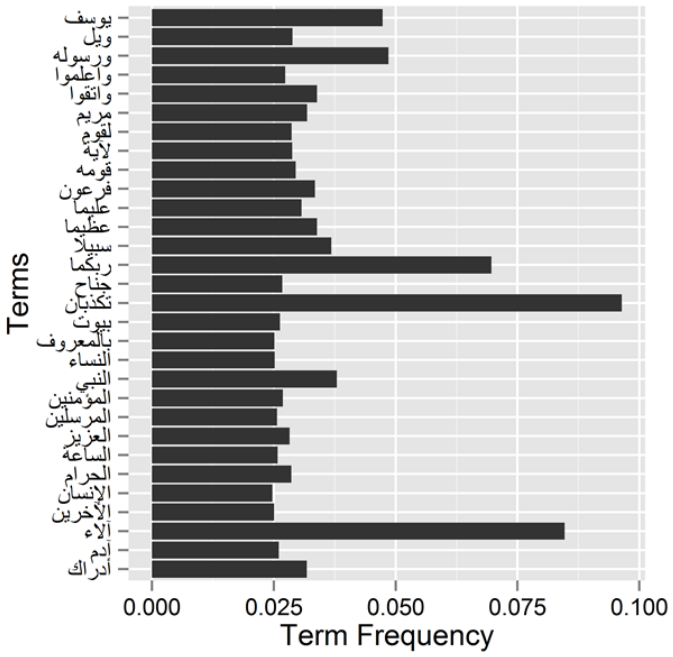

Fig. 8: Most frequent terms in the holy Quran measured by TF-IDF (Based on Parts)

C. Experiments Based on Document-Term Matrix

In this subsection a different experimentation setup is carried out using document-term matrix based on both chapters and parts of the holy Quran.

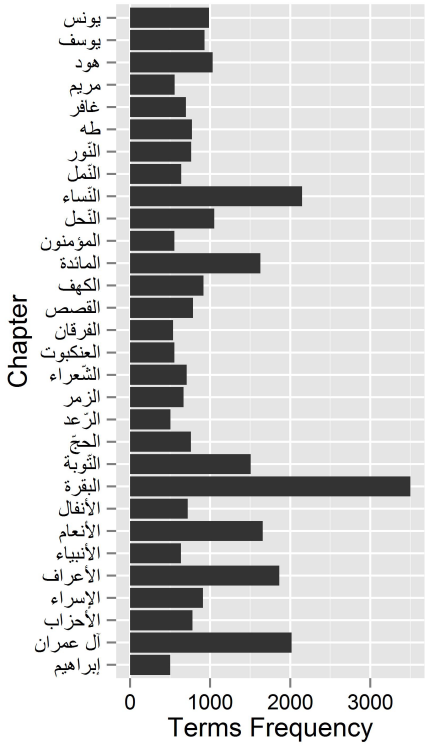

Fig. 10: Term Frequencies based on the Chapters of the holy Quran with Term Frequency larger than 500

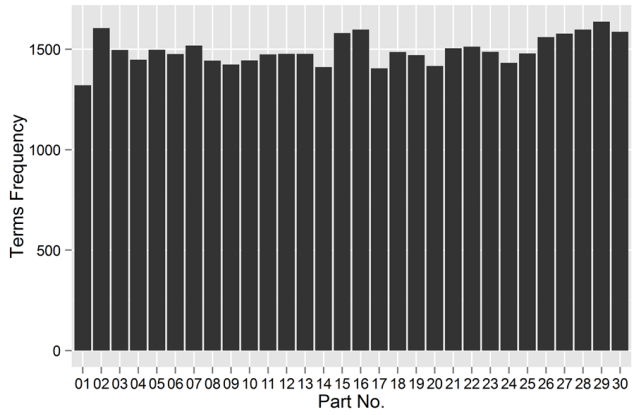

Fig. 11: Term Frequencies based on the Parts of the holy Quran

IV. Discussion

The results that are obtained based on chapters are different from those obtained when parts are used as documents. This is clear when comparing figure 2 with figure 8 where the most frequent terms are different. This is because TF-IDF measure is used; which depends on the ratio of number of all documents to the number of documents a specific term appears on. However, when TF measure is used then there are no differences. That is due to the fact that term-document matrix is used and TF-IDF measure is based on documents compared to TF measure which depends on terms solely.

The bi-grams, tri-grams and four-grams that appear in figures 5- 7 reveal the most frequent n-gram terms in the holy Quran. These terms are of great benefit for future work that is related to semantic search. Notice that these terms are important because stop words are removed before they are extracted.

One important result in this paper is that the partitioning methods affect the distribution of the terms of the holy Quran and there frequencies. This appears when figure 10 and figure 11 are compared. Notice that the term frequencies for Parts are almost equal where the term frequencies for Chapters dramatically vary from one chapter to another.

There are many applications for the most frequent terms obtained in this paper. For example, frequent terms presented in figure 1 and figure 2 are calculated using TF and TF-IDF measures respectively. These calculations were based on chapters of the holy Quran; however results that are shown in figure 8 were calculated based on Parts of the holy Quran. Choosing the right application depends on both the document-partitioning method as well as the calculation measure. The first might affect the precision and quality of results of a specific application; and hence need to be tested in a specific application. However, the calculation measure might affect choosing the right application, for example if semantic search or clustering applications are chosen then it is more suitable to use terms produced using TF measure. This is due to the fact that it will give more weight for terms that are repeated most in all documents. On the other hand, if topic modelling application is chosen, then it is more appropriate to use terms calculated based on TF-IDF measure. This is because using it will return terms that are more important in a specific document and less important in other documents.

The results obtained in this paper could be improved if a stemming algorithm is used. This is because using stems instead of the original words will give more accurate results as known in information retrieval field of study; where using stems will increase the precession. For Arabic language, stemming algorithms are still immature and have high error rates. For the holy Quran, error rates was calculated and it is in the range of 22-55%. For more details please see [18], [19], [20]. Such error rates are not acceptable when analyzing the text of the holy Quran because it is the word of God and errors will change the meanings of its words.

V. Conclusion

This study aims to layout a framework for future work that is related to the application of natural language processing, data mining and text mining to the text of the holy Quran. This is done by initially preprocessing the text of the holy Quran and by considering the different possible partitioning. Choosing one document portioning of the holy Quran affects the resulted Frequent Terms of the holy Quran. Moreover, if TF-IDF weighting is used, then the resulted Frequent Terms are also change. Graphical representation of the main terms of the holy Quran is depicted using term-frequencies plot and wordcloud plots including bi-grams, tri-grams and four-grams.

One important result in this paper is that frequencies of the terms of the holy Quran depend on the chosen partitioning method that is used in the analysis: If chapters are used then Term frequencies vary according to the size of each chapter; If parts are used instead then Term frequencies are almost equal.

Another important result in this study is that a list of frequent terms are suggested to be used in different applications: on one hand for, those terms calculated based on TF measure they might give good results for semantic search and clustering; on the other hand, terms that are calculated based on TF-IDF measure are more suitable for topic modelling.

This study constitutes the first phase in a large project that aims at advancing the research in the field of arabic natural language and more specifically in exploring the text of the holy Quran. Therefore, the results illustrated in this paper represent a simple sample of what could be obtained from the analysis of the text of the holy Quran.

Although the results of this study are interesting, however it is based on the original words of the holy Quran. More accurate results will be obtained if an efficient stemming algorithm is used. Unfortunately, there is no accurate known stemming algorithm exists for Arabic language due to the challenges that are faced in processing Arabic language.

Future work may include preprocessing the text of the holy Quran with efficient and accurate algorithm that might give words like stems as light stemmers algorithms do. If such algorithm is developed then further study on the text of the holy Quran will be carried out to extract knowledge and important information that is useful to all humanity using machine learning techniques.

ACKNOWLEDGMENT

This project was supported by the deanship of scientific research at Salman bin Abdulaziz University under the research project number 109/T/33.

REFERENCES

[1] M. DIAB and N. HABASH, "Arabic dialect tutorial," in *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL07)*, 2007, pp. 29–34.

[2] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," vol. 8, no. 4, pp. 14:1–14:22, Dec. 2009. [Online]. Available: http://doi.acm.org/10.1145/1644879.1644881

[3] N. Y. Habash, *Introduction to Arabic Natural Language Processing*, G. Hirst, Ed. Morgan and Claypool Publishers, 2010.

[4] M. Saad and W. Ashour, "Arabic morphological tools for text mining," in *6th International Symposium on Electrical and Electronics Engineering and Computer Science, European University of Lefke, Cyprus, 2010*, 2010, p. 112117.

[5] I. Ali, "Application of a mining algorithm to finding frequent patterns in a text corpus: A case study of the arabic," *International Journal of Software Engineering and Its Applications*, vol. 6, pp. 127–134, 2012.

[6] M. Al-Yahya, H. Al-Khalifa, A. Bahanshal, I. Al-Odah, and N. Al-Helwah, "An antological mdoel for representing semantic lexicons: An application on times nouns in the holy quran," *The Arbaian Journal for Science and Engineering*, vol. 35(2c), pp. 21–37, 2010.

[7] K. Dukes, E. Atwell, and N. Habash, "Supervised collaboration for syntactic annotation of quranic arabic," *Language Resources and Evaluation*, vol. 47, no. 1, pp. 33–62, 2013. [Online]. Available: http://dx.doi.org/10.1007/s10579-011-9167-7

[8] M. H. Panju, "Statistical extraction and visualization of topics in the qur'an corpus," Master's thesis, University of Waterloo, 2014.

[9] N. Ismail, N. Rahman, Z. Bakar, and T. Sembok, "Terms visualization for malay translated quran documents," in *International Conference on Electrical Engineering and Informatics, Bandung, Indonesia*, 2007, pp. 554–557.

[10] M. S. Hikmat Ullah Khan, Syed Muhammad Saqlain and M. Sher, "Ontology-based semantic search in holy quran," vol. 2, no. 6, pp. 562–566, 2013.

[11] N. Shahzadi, A. ur rahman, and M. J. Sawar, "Semantic network based classifier of holy quran," *International Journal of Computer Applications*, vol. 39, no. 5, pp. 43–47, February 2012.

[12] M. Shoaib, M. Nadeem Yasin, U. Hikmat, M. Saeed, and M. Khiyal, "Relational wordnet model for semantic search in holy quran," in *International Conference on Emerging Technologies, 2009. ICET 2009.*, Oct 2009, pp. 29–34.

[13] A. A. Aliyu Rufai Yauri, Rabiah Abdul Kadir and M. A. A. Murad, "Quranic verse extraction base on concepts using owl-dl ontology." vol. 6, no. 23, pp. 4492–4498, 2013.

[14] M. Yunus, R. Zainuddin, and N. Abdullah, "Semantic query for quran documents results," in *Open Systems (ICOS), 2010 IEEE Conference on*, Dec 2010, pp. 1–5.

[15] K. Dukes. (2014, Dec.) Quranic arabic corpus. [Online]. Available: http://corpus.quran.com/

[16] Tanzil.net. (2014, Oct.) Tanzil quran text download @ONLINE. [Online]. Available: http://tanzil.net/download/

[17] T. Zerrouki. (2014, Nov.) Arabic stop words @ONLINE. [Online]. Available: http://sourceforge.net/projects/arabicstopwords/

[18] M. Sawalha and E. Atwell, "Comparative evaluation of arabic language morphological analysers and stemmers," in *Coling 2008: Companion volume: Posters*. Coling 2008 Organizing Committee, 2008, pp. 107–110.

[19] N. Thabet, "Stemming the qurán," in *Workshop on Computational Approaches to Arabic Script-based Languages*. Coling 2008 Organizing Committee, 2008, pp. 28–31.

[20] R. J. R. Yusof, R. Zainuddin, M. S. Baba, and Z. M. Yusoff, "Quránic words stemming," *Arabian Journal for Science and Engineering*, vol. 35, p. 38, 2010.