Journal of Engineering and Applied Sciences 11 (12): 2809-2814, 2016
ISSN: 1816-949X

Data Selection and Collection for Printed Quranic Documents

[1]Ruziana Mohamad Rasli, [2]Mime Azrina Jaafar, [3]Faudziah Ahmad and [3]Siti Sakira Kamaruddin
[1]Department of Information Technology and Communication,
Tuanku Syed Sirajuddin Polytechnic, Pauh Putra, Arau, Perlis, Malaysia
[2]Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia
[3]School of Computing, College of Arts and Sciences,
Universiti Utara Malaysia, Sintok, Kedah, Malaysia

Abstract: The purpose of this study is to explain on the process of data selection and collection for Quranic knowledge. The data used in this research is from the Holy Al-Quran and Hadiths. Basically, there are five sources used which is the Holy Al-Quran, Hadith Sahih Bukhari, Hadith Sahih Muslim, Hadith Sunan Abu Dawud and Hadith Sunan Ibn Majah. In order to minimize the resources, only Zakat topics is extracted. From all the data collection and selection processes, it is concluded that the total of 339 sentences are extracted from 72 Surah in the Holy Al-Quran and a total of 589 sentences are extracted from four Hadiths (Hadith Sahih Bukhari, Hadith Sahih Muslim, Hadith Sunan Abu Dawud and Hadith Sunan Ibn Majah). The total of sentences extracted from these two sources are 928 sentences. These two processes took 8 month to be completed. Most of the times are allocated in the process of proof reading the documents after OCR conversion. The output of this paper is 928 sentences from the Al-Quran and Hadiths that are focuses to Zakat topic. These ayats will be used for the next phase which is data processing and TF-IDF calculation.

Key words: Data selection, data collection process, quranic documents, Majah, sources

INTRODUCTION

The main purpose of this study is to show the process of data selection and collection of Quranic documents that will be used in the research that covers text mining and graphical knowledge representation. Basically the full methodology of this research covers 12 main phases starting from preliminary study to the last phase which is the evaluation phase as in Fig. 1.

According to Fig. 1, this research starts with preliminary study and followed by data selection and collection in the second phase. To start this research, a preliminary study had been made to find the most suitable topics to be used in this research on Quranic documents. Preliminary study involves the process of giving out questionnaire to random respondents aged between 20-55 years old. Once the questionnaires had been collected, the data is analyze based on the issues in Zakat. The result generated from the questionnaire is that Zakat will be chosen as the domain topics since most of the respondents have basic knowledge of Zakat, its characteristics and importance.

MATERIALS AND METHODS

Not only that Zakat had been chosen as the scope for this research because of two things. First, Zakat is the third pillars of Islam and Zakat is mentioned repeatedly in the Holy Al-Quran and ahadith (ahadith is the plural of hadith). Second, Zakat is considered as the main alternatives to solve problem in poverty which is really good to cater the world's financial crisis (Johari *et al*., 2013). Zakat is a "financial" form of ibadah that makes the computation of Zakat crucial in fulfilling this obligation. However, there are a lot of issues in understanding Zakat such as the way Zakat being paid, how to calculate Zakat, types of Zakat being paid, conditions for wealth that need to be paid for Zakat, who is the Zakat recipient, who is eligible for Zakat payments and other. Adnan and Nur Barizah (2009) reveals that there is a general misconception of Zakat in the standard and guidelines of it and as a result, there is an inappropriateness on Zakat recognition, measurement and presentations.

There is a need to represent information in Zakat as easy as it could. From this topics, the next phase is data selection and followed by data collection.

Corresponding Author: Ruziana Mohamad Rasli, Department of Information Technology and Communication, Tuanku Syed Sirajuddin Polytechnic, Pauh Putra, Arau, Perlis, Malaysia

Data selection and data collection is considered as an important task since the data gathered from this phase will be used throughout the whole research and the methods that will be developed depends on it. However, there are several issues in selecting data source and also to collect related Zakat data.

In data selection, there are five sources that will be used in this research which consists of The Holy Al-Quran and 4 Hadiths (Hadith Sahih Muslim, Hadith Sahih Bukhari, Hadith Sunan Ibn Majah and Hadith Sunan Abu Dawud). The main source is the Holy Al-Quran which a considered as the most important book for Muslim's reference and guidance. Hadiths will be used as the second main source because basically, the contents of the Holy Al-Quran is in basic form. There is no detail explanation and in order to get further explanation on the ayats in the Holy Al-Quran, Hadiths is being used. Basically Hadiths is the sayings of Prophet Muhammad SAW Hadiths cover a lot of topics from economics to daily routines of a Muslim. As a Muslim, in order to be a good Muslim, there is a need of guidance and these two sources are the best main resources that being referred by Muslim thousands of years ago.

Since, the data that will be used covers religious topics, the data must be valid, authentic and genuine. This is because, the model that will be generated will be used by Muslims all over the world and therefore, the data must be in perfect condition.

Al-Quran is the main source used by Muslims as a source of guidance and a book of knowledge and wisdom (Yauri *et al*., 2013). There are 114 chapters in the Holy Al-Quran with 6236 verses covering various topics from science to economy. The Holy Al-Quran contains divine words of wisdom that helps Muslim to be in the right path (Mukhtar *et al*., 2012). Currently, with the new era of technological advancement, a lot of documents can be uploaded by anyone about everything without knowing that the data is genuine and authentic. There are several cases where the Hadiths were fabricated by irresponsible people for their own advantages and sadly, being followed by other Muslim (Bukhari and Ismaiel, 1997).

Hadith is the second most important documents after Al-Quran that is being used as a reference by the Muslim (Bukhari and Ismaiel, 1997). In this research, only four authentic Hadiths will be used which is from Hadith Sahih Bukhari, Hadith Sahih Muslim, Hadith Sunan Abu Dawud dan Hadith Sunan Ibn Majah. The most widely used Hadiths in the world are Hadith Sahih Bukhari dan Hadith Sahih Muslim. This is because, these two Hadiths provides Sahih information not like other Hadiths which will have a combination of Sahih hadiths and Da'if Hadiths.

Da'if Hadiths is weak Hadiths that resulted from the lowest ranking and it have a variation in the nature of the weaknesses associated with its reporters. However, the main problem in a hadiths is that there is several cases where Maudu' (Fabricated) Hadith being carried out in the Internet (Bukhari and Ismaiel, 1997). Fabricated hadiths are really dangerous for Muslim because the contents of the hadiths are being changed and fabricated by irresponsible people for their own usage. Usually fabricated hadiths are not the original sayings of the Prophet Muhammad SAW and had been manipulated by irresponsible people.

Before going thru data collection phase, the data must be selected. Data selection involves only selecting the genuine and authentic approved by Department of Islamic Development Malaysia (JAKIM). Once the authentic and genuine data had been prepared, the steps will move to Data Collection.

RESULTS AND DISCUSSION

The Holy Al-Quran: The Holy Al-Quran is the main reference used by Muslim all over the world that covers a lot of topics in it. The Holy Al-Quran is divided into 114 chapters with 6236 sentences. Since the topics that will be used in this research is only on Zakat, not all chapters will be referred. However, the list of Holy Al-Quran surahs are given as:

- Al-Fatihah
- Al-Baqarah
- 'Ali 'Imran
- An-Nisa'
- Al-Ma'idah
- Al-'An'Am
- Al-'A'Raf
- Al-'Anfal
- At-Tawbah
- Yunus
- Hud
- Yusuf
- Ar-Ra'D
- 'Ibrahim
- Al-Hijr
- An-Nahl
- Al-'Isra'
- Al-Kahf
- Maryam
- Taha
- Al-'Anbya'
- Al-Haj
- Al-Mu'minun

- An-Nur
- Al-Furqan
- Ash-Shu'Ara'
- An-Naml
- Al-Qasas
- Al-'Ankabut
- Ar-Rum
- Luqman
- As-Sajdah
- Al-'Ahzab
- Saba'
- Fatir
- Ya-Sin
- As-Saffat
- Sad
- Az-Zumar
- Ghafir
- Fussilat
- Ash-Shuraa
- Az-Zukhruf
- Ad-Dukhan
- Al-Jathiyah
- Al-'Ahqaf
- Muhammad
- Al-Fath
- Al-Hujurat
- Qaf
- Adh-Dhariyat
- At-Tur
- An-Najm
- Al-Qamar
- Ar-Rahman
- Al-Waqi'Ah
- Al-Hadid
- Al-Mujadila
- Al-Hashr
- Al-Mumtahanah
- As-Saf
- Al-Jumu'Ah
- Al-Munafiqun
- At-Taghabun
- At-Talaq
- At-Tahrim
- Al-Mulk
- Al-Qalam
- Al-Haqqah
- Al-Ma'Arij
- Nuh
- Al-Jinn
- Al-Muzzammil
- Al-Muddaththir
- Al-Qiyamah
- Al-'Insan
- Al-Mursalat
- An-Naba'
- An-Nazi'At
- 'Abasa
- At-Takwir
- Al-'Infitar
- Al-Mutaffifin
- Al-'Inshiqaq
- Al-Buruj
- At-Tariq
- Al-'A'La
- Al-Ghashiyah
- Al-Fajr
- Al-Balad
- Ash-Shams
- Al-Layl
- Ad-Duhaa
- Ash-Sharh
- At-Tin
- Al-'Alaq
- Al-Qadr
- Al-Bayyinah
- Az-Zalzalah
- Al-'Adiyat
- Al-Qari'Ah
- At-Takathur
- Al-'Asr
- Al-Humazah
- Al-Fil
- Quraysh
- Al-Ma'Un
- Al-Kawthar
- Al-Kafirun
- An-Nasr
- Al-Masad
- Al-'Ikhlas
- Al-Falaq
- An-Nas

From all these 114 Surahs, not all of the Surahs will be used. Due to time constraints, only 72 Surahs will be used which focusses on Zakat topic. Once the data of the Holy Al-Quran are manually analyzed and selected, a total of 339 sentences are extracted from 72 chapters. Figure 1-3 shows the chapters and its ayats' frequency distribution.

From Fig. 1-3, the second Surah, Surah Al-Baqarah contributes the highest number of ayats (sentences) with 34 ayats. The second highest ayats is from Surah 64, Al-Qalam with the total ayats of 17 ayats and the third highest is from Surah 22, Al-Hajj with 10 ayats. Altogether, from 6236 ayats in the Holy Al-Quran only 5%

Fig. 1: Research methodology

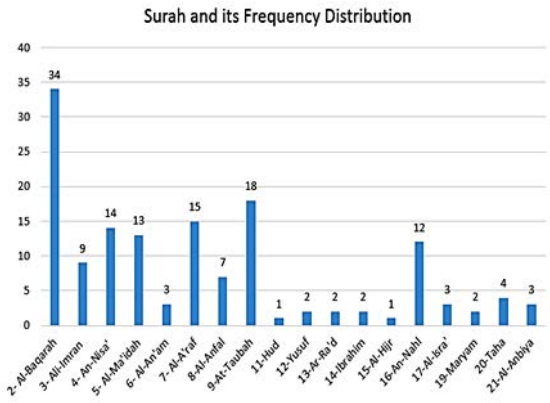

Fig. 2: Surah and its ayats' frequency distribution

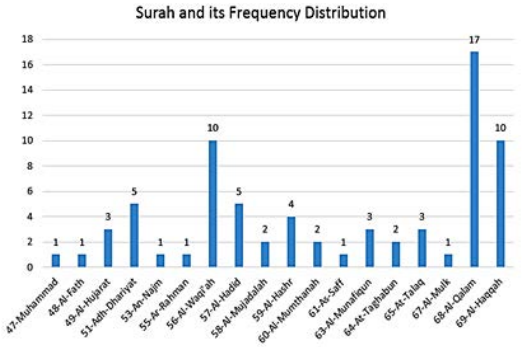

Fig. 3: Surah and its ayats' frequency distribution

of the total ayats in the Al-Quran is used in Zakat topic. The next section will discuss on four types of Hadiths focusing on the The Book of Zakat chapters.

Hadiths: As discussed in the previous part, there are four Hadiths that will be used in this research which is Hadith Sahih Muslim, Hadith Sahih Bukhari, Hadith Sunan Abu Dawud and Hadith Sunan Ibn Majah. Once selected, there are a total of 208 chapters for the whole Hadiths.

However, due to the need of authentic hadiths, only hadiths that are Sahih were selected. Basically, there are three types of Hadiths which are Sahih Hadith (Authentic Hadiths), Da'if Hadith (Weak hadith) and Maudu' Hadith (Fabricated Hadith). Sahih Hadith is the one which has a continuous isnad which made up of trustworthy reporters that preservers from similar authorities. For Sahih Hadith, the grading is given only to those which are transmitted by Al-Bukhari, Muslim and declared by other traditionists. For Da'if hadiths, the content fails to reach the ranking status of Hasan and Sahih and also it has a variation in the nature of the weaknesses associated with its reporters. For Maudu' Hadith, the contents goes against the established norms or its reporters are not a valid reports which includes a liar. Maudu' Hadith falls under eight categories of causes of fabrication. These causes of fabrication can come from political difference, factions based on issues of creed, fabrication by zanadia, fabrication by story tellers, fabrications by ignorant ascetics, prejudice in favours of town, race or particular imam, inventions for personal motives and proverbs turned into hadiths.

Figure 4 shows the division of chapters according to the Hadiths which shows that Hadith Sahih Bukhari give the largest contributions of data with 38% (78 chapters), followed by Hadith Sahih Muslim with 27% (56 chapters).

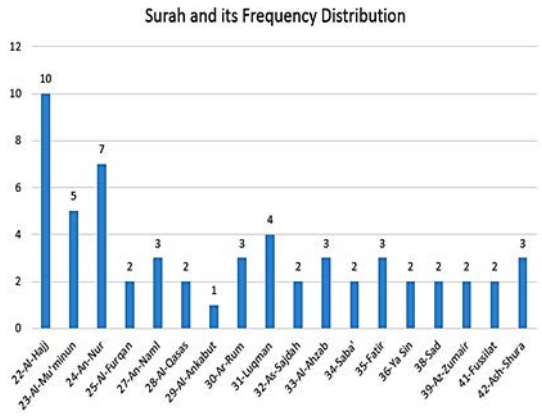

Fig. 4: Surah and its ayats' frequency distribution

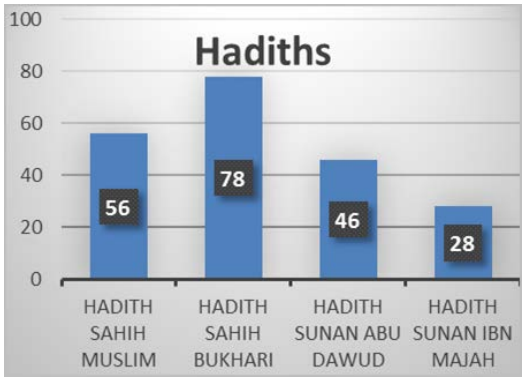

Fig. 5: Distribution of Hadith's Chapter on Zakat

Hadith Abu Dawud and Hadith Ibn Majah have the lowest contribution of chapters by 22 and 13% with 46 and 28 chapters, respectively. From these 208 chapters, a total of 589 sentences are extracted.

The difference between Hadith Sahih Muslim and Hadith Sahih Bukhari is 85 ayats and the difference of Hadith Ibn Majah and Hadith Abu Dawud is only 2 ayats. The lowest number of ayats if from Hadith Ibn Majah with 62 ayats which make it contributes to only 13% of the whole data.

Now that we have all the information needed based on the Holy Al-Quran and Hadiths, we will try to convert from printed version to softcopy version. The data will not be downloaded from the Internet since it is not proven to be a valid and authentic version. In this research, only the sources that is validated from Department of Islamic Studies Malaysia (JAKIM) will be used. Once selected, the researcher gathered the sources in hardcopy version. In order to convert the data from hardcopy to softcopy, there are several steps that need to be done as in next part which is scanning, OCR conversion and proofread.

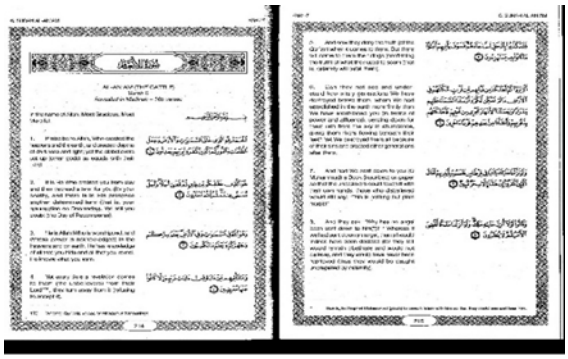

Fig. 6: Example of scanned document

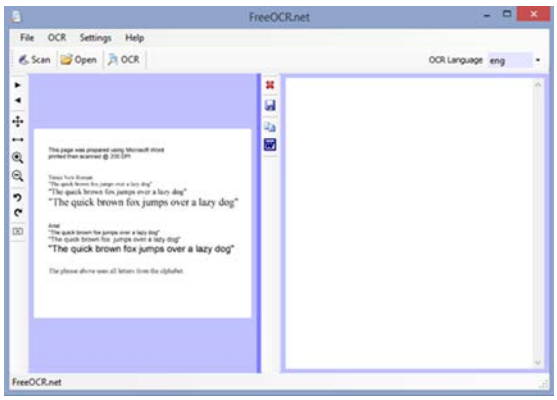

Fig. 7: Interface of FreeOCR.net

Scanning: A scanner is used to scan the related data pages of the Holy Al-Quran and Hadiths which is then automatically converted it to .jpg image file format. The main drawback of this image file is that the contents cannot be copied straight away from the image file. Figure 5 shows an example of image file from the Holy Al-Quran.

According to Fig. 6, only the English translated section is being selected because the main scope of this research is the English translated Al-Quran and Hadiths. To select this part, additional tool called FreeOCR.net is used to convert the images to text. The detailed main function of the tool will be discussed in the next phase.

Ocr conversion: Optical Character Recognition (OCR) is a process of converting images with text into text documents using a tool equipped with automated computer algorithms. Basically, the process of converting the images are based on two sources. The first source is image or PDF files obtained from any types of scanner and the second source is photos taken with digital cameras or mobile phones. For this research, a tool called FreeOCR.net is used to convert all the hardcopy documents into a softcopy documents. Figure 7 shows the interface of the tool.

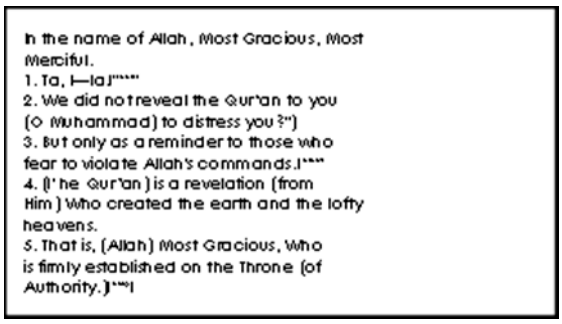
In the name of Allah, Most Gracious, Most Merciful.
1. Ta, H—la J"""
2. We did not reveal the Qur'an to you (O Muhammad) to distress you ?")
3. But only as a reminder to those who fear to violate Allah's commands.I'""
4. (I' he Qur'an) is a revelation (from Him) Who created the earth and the lofty heavens.
5. That is, (Allah) Most Gracious, Who is firmly established on the Throne (of Authority.)'""I

Fig. 8: Example of text generated

Proofread: OCR Conversion, the output of the text will be in .txt format. The main drawback of this process is the text generated from this tools is not 100% accurate. There are some symbols that are mistakenly recognized as other symbol. Majority of these mistaken symbols have similar characteristics with the other symbols. Example of original text generated is listed in Fig. 8.

According to Fig. 8, there is a need to compare the scanned text with the original hardcopy version. These process is called proof read. Proof read takes a lot of time to finish since the proof reader need to check words by words in the data sources. This is the most time consuming process which takes several months to be completed.

The output from this process will be used in the next phase which is the pre-processing phase. In this next phase, the processed words are then further analysed using TF-IDF calculation to categorize the terms to get the concept words of these data.

CONCLUSION

From all these processes, it is concluded that the total of 339 sentences are extracted from 72 Surah in the Holy Al-Quran and a total of 208 chapters are extracted from four Hadiths (Qazwini, 2007; Qadhi, 2008; Al-Khattab, 2007).

These two processes took 8 months to be completed. Most of the times are allocated in the proof reading the documents after OCR conversion. Although the domain is Zakat, there is also other topics related to Zakat. Therefore, these data need to be simplified only to cover one single topic.

This study is one of the parts in the research on developing graphical knowledge representation techniques. The outcome of this research will help Muslim all over the world to understand Zakat easily using graphical model.

ACKNOWLEDGEMENT

Researchesr are grateful for the Fundamental Research Grant Scheme provided by Ministry of Higher Education Malaysia.

REFERENCES

Adnan, M.A. and A.B. Nur Barizah, 2009. Accounting treatment for corporate Zakat: A critical review. Int. J. Islamic Middle Eastern Finance Manage., 2: 32-45.

Al-Khattab, N., 2007. English Translation of Sahih Muslim. 1st Edn., Darussalam Publisher and Distributes, Riyadh, Saudi Arabia.

Bukhari, A. and M. Ismaiel, 1997. The Translation of the Meaning of Sahih Al-Bukhari. Darussalam Publisher, Riyadh, Saudi Arabia,.

Johari, F., A.M.R. Aziz, M.F. Ibrahim and A.F.M. Ali, 2013. The roles of islamic social welfare assistant (Zakat) for the economic development of new convert. Middle East J. Sci. Res., 18: 330-339.

Mukhtar, T., H. Afzal and A. Majeed, 2012. Vocabulary of quranic concepts: A semi-automatically created terminology of Holy Quran. Proceedings of the 15th International Conference on Multitopic Conference (INMIC), December 13-15, 2012, IEEE, Islamabad, Pakistan, ISBN:978-1-4673-2249-2, pp: 43-46.

Qadhi, Y., 2008. English Translation of Sunan Abu Dawud. Darussalam, Riyadh.

Qazwini, M.Y.M.A., 2007. English Translation of Sunan Ibn Majah. Darussalam Publisher, Riyadh, Saudi Arabia,.

Yauri, A.R., R.A. Kadir, A. Azman and M.A.A. Murad, 2013. Ontology semantic approach to extraction of knowledge from Holy Quran. Proceedings of the 5th International Conference on Computer Science and Information Technology, March 27-28, 2013, IEEE, Amman, Jordan, pp: 19-23.