The Study on Qur'anic surahs' Structured-ness and their Order Organization Using NLP Techniques

EHSAN KHADANGI[1]

Assistant Professor, Computer Engineering Department, Shahed University, Tehran, Iran

MOHAMMAD MOEIN FAZELI

Imam Hussein Comprehensive University, Tehran, Iran

MEHDI NAGHAVI

Faculty of Computer Engineering, Imam Hussein Comprehensive University, Tehran, Iran

ABSTRACT: Original Paper

The study of surahs' structure has attracted researchers' attention in recent years. One of the theories herein is the theory of Topic Sameness which acknowledges that each surah of Qur'an has formed on a single topic. The theory of Introduction and Explanation as one of the most important branches of Topic Sameness, proposes that the Almighty states the topic of each surah at the first section, elaborates it at different parts of the surah in the forms such as stories, signals of nature, and future predictions, and concludes from the stated contents at the final part. In this paper, we accordingly intend to study the two theories using NLP techniques for the first time. In this regard, based on the three methods of tf-idf, word2vec and roots' accompaniment in verses, the similarity of Qur'anic roots is computed. Then, the amount of similarity of the concepts within surahs to each other is calculated and compared with the random mode. The results show that the studied surahs hold the inner coherence between the concepts so that they have been formed on a single topic or a few topics related to each other. In addition, the study on the similarity between the first and the body sections of each surah shows that the structure of Introduction and Explanation seems to be true for many surahs by the designed methodology.

1 . Corresponding Author. E-mail Adress: khadangi@shahed.ac.ir

At the end, by comparing the similarity of surahs to each other versus their order distance in Qur'an and their revelation time distance, we realized that the whole Qur'an is also relatively organized in terms of surah' ordering.

KEYWORDS: Natural Language Processing, Word2vec, The Qur'an, Topic Sameness; Surahs' Structured-ness; Tf-idf.

1. Introduction

The structural vision on Qur'anic surahs, effort towards the detection of the core title of surahs, organizing the surahs based on a single topic, and finally interpretation based on the structure of the surahs have seriously appealed to some recent Qur'an researchers (Khamehgar, 2006). The major basis and presumption of the analysis of the structure of Qur'anic surahs is the order of the verses in the surahs being protected and revelatory (Khamehgar, 2008). From these researchers' viewpoint, Qur'an is an integrated and organized system for the understanding of which the relationships between the elements including the concepts, verses, and the sections inside the surahs should be known. Based on this and the elaborate studies carried out by recent researchers of Qur'anic sciences such as Khamehgar and Lesani Fesharaki, Qur'an is an integrated and accurate system the elements of which (surahs, sections, and verses) also have a revelatory organization (Khamehgar, 2002b). The analysis of this structured system can lead to many gains such as the discovery of new horizons of Qur'anic miracles, extraction of the rich Qur'anic knowledge, exploitation of the major goal and core topic of the surahs, and the order system of surahs. Based on these presumptions, many researchers have dealt with the study of the structure of many surahs, such as structural study on the surah al-Māʾidah (Q.5) by Aram and Layeqi (2017), the Structure of the surah al-Kahf (Q.18) by Fatahizadeh and Zakeri (2016), the structure of the surah al-Inshiqāq (Q.84) by Dehghani Farsani (2008), and the Structure of the surah al-Infiṭār (Q.82) by Jigareh and Sadeghi (2017). In addition, some individuals such as Khamehgar (2006) has dealt with the translation of Qur'an based on the structured-ness of surahs.

In spite of different signs discovered by believers of the structured-ness of surahs, some orientalists, based on signs such as the style of Qur'an's speech, and on the presumption of the prophet's close friends having affected the order of verses and even Qur'an not being revelatory, have concluded that the content of Qur'an is disintegrated and without logical connection.

Richard Bell 1953), a European Qur'an researcher, for instance, has stated at the introduction of his accredited Qur'an translation into English that one of the original attributes of Qur'an's style is that it is disintegrated and it is rarely possible to see coherence through a major section of a surah. Artour John Arberry (1996) has also written somewhere in the introduction of his Qur'an translation that Qur'an is far away from whatever integration related to the order of its descent and also from the logical coherence. Qur'an's reader would definitely get astonished by the apparently disordered status of many surahs especially if limited to one translation only, although the translation is linguistically accurate. These researchers' emphasis on disintegration and disorder of Qur'an's verses reminds the reader about the note that Qur'an has not stayed away from humane manipulations and at least the order of the verses is not revelatory. This is while numerous sensible reasons based on the study of the structure of the surahs as well as historical documents explicitly state that the order of the verses is revelatory and has stayed the same over time (Khamehgar, 2008).

Although different works have been done about surahs' organization and the structure of some surahs have been studied by Qur'anic Sciences researchers (Khamehgar, 2006; Fatahizadeh and Zakeri, 2016), it seems that none of the work has utilized text-mining and natural languages processing algorithms.

In the present paper, we intend to study Qur'an's system in an integrated way and to study the surahs' structured-ness in terms of their both intra- and inter-surah status by the NLP techniques and algorithms. On the one hand, the intra-structures are examined in terms of topic sameness and Introduction and Explanation theories. On the other hand, the inter-structure is examined based on the order of surahs. In this regard, the current research deals with two major questions: 1. Are Qur'anic surahs around a single topic? 2. Is the order of Qur'anic surahs organized? The rest of the paper has been organized as follows. Section 2 presents some related works. Fundamental definitions are briefly explained in section 3. Section 4 contains materials and methods. Section 5 deals with pre-processing. In section 6, we explain our method and evaluation measure in more detailed. Then we present the results of our experiments in section 7. Finally, the conclusion will be presented in section 8.

2. Related works

Besides the works by Qur'anic sciences researchers as well as orientalists already mentioned, computer sciences' researchers have also carried out many works on Qur'an analysis. Due to the significance of semantic search

in Qur'an, many works have looked for new methods of semantic search. Among these Yauri et al. (2013), Khan et al. (2013), Shoaib et al. (2009), as well as Alhawarat (2015) have presented methods based on ontology, word-net, and topic modeling, respectively, for Qur'anic semantic search. Different works have also dealt with building different Qur'an ontologies, most of which are focused on a particular field (Ismail et al., 2016). Iqbal et al. (2013) have highlighted the weaknesses of the existent Qur'an ontologies and have developed a new ontology. Safee et al. (2016) have also studied different methods of verse retrieval and have presented their weaknesses and strengths. Their findings show that there is the need for learning and building new Qur'an ontologies for correcting the contradiction between the existent ontologies.

A set of other works have dealt with the presentation of corpora suitable for analyzing Qur'an. Among these, Dukes and Buckwalter (2010) and Atwell and Sharaf (2009) have exploited the treebank of Qur'an's verses with regard to Arabic grammar and have shown it by dependency graphs. Sharaf and Atwell (2012a) have presented a corpus which connects the verses which are conceptually similar. They named this corpus QurSim. This corpus could be used for different applications such as Qur'an translation. The Corpus QurAna has also tagged Qur'an's personal pronouns based on their referents (Sharaf and Atwell, 2012b). Sherif and Ngonga Ngomo (2015) have extracted a dataset based on RDF from Qur'an translation into 43 different languages, which could be used for different applications in natural language processing. Besides the presented datasets, tools for searching and analyzing Qur'an's corpora have been presented so far. Alfaifi and Atwell (2016) have examined and compared these tools.

The approach of some researches is also the analysis of Qur'an for different applications such as developing the Qur'anic question answering system (Hamed and Aziz, 2016) and verses' classification. For instance, Sharaf, and Atwell, (2012a) looked for the classification of Qur'anic surahs into the two classes of in-Mecca and in-Medina by decision tree classifier. They utilized some features such as the length of the surah, the words and phrases used in the surahs, and prostratation verses[1].

3. Fundamental definitions

3.1. Surahs' Topic Sameness

Some Qur'anic science researchers emphasize on the structured-ness of

---

1. There are some verses in the Qur'an, which are believed they must be followed by the reader's prostration.

Qur'anic surahs and have proposed theories by assessing the organization of different surahs, the most important of which is the theory of Topic Sameness. Based on this theory, each surah holds a core topic and all the verses and discussions stated in the surah relate to that topic.

3.2. Introduction and Explanation theory

The Introduction and Explanation theory states that the Almighty defines the core topic in the beginning verses, then in different sections of the surah, explains it by analogies, anecdotes, and examples, finally concludes based on the core topic (Khamehgar, 2004; Khamehgar, 2002a). Based on this theory, the core topic of each surah is introduced in the beginning section of the surah, to be called introduction. Similarly, the final section of the surah which is somehow a conclusion of the discussions included in the surah is referred to as surah conclusion.

3.3. Section

In this paper, Qur'an's verses together, which talk about a particular topic, are called section.

3.4. Vector space models

There are different methods for representing texts and concepts in vectors. In vector space models, text is shown as a vector each component of which is related to the estimated significance of the word in the text (Soucy and Mineau, 2005). The method bag of words and its extension N-gram are one of the most applicable methods to represent texts, which, despite simplicity, act suitably for many text mining applications (Zhang et al., 2010).

4. Material and methods

In this paper, the structured-ness of Qur'anic surahs is examined based on the theory of Topic Sameness. The methodology of the current research comprises seven parts including pre-processing and preparing data, surahs' partitioning into sections, calculating the similarity of Qur'anic roots, calculating the similarity of sections and surahs, study on the relationship between surahs' title and their content, study on the topic sameness of surahs, and finally study on the structured-ness of Qur'an in terms of surahs' order.

Based on this, the corpus was initially prepared and cleaned for later processes. Then the similarity between different Qur'anic roots was

calculated by applying different NLP techniques to Qur'an corpus. The amount of relationship between the surah's title and the words within the surah was also studied. Afterwards, topic sameness of surahs was studied. For this, the similarity of intra-surah concepts was gained and compared with the random mode. Then, for examining the introduction and explanation theory the similarity of the first section to other sections of the surah and also the first section to surahs' conclusion of different surahs were calculated and the result was compared with the random mode. At the end, the organization of Qur'an in terms of surahs' order was studied in such a way that the similarity of different surahs was measured and the relationship between the order distance of surahs as well as their descent time distance and the amount of surahs' similarity was studied.

4.1. Dataset

Since the number of Qur'an's distinct words is very high, we focused on Qur'anic roots rather than derivatives. The first data were a table at each line of which there were Qur'anic words fully voweled[2], the surah locating the word, the verse related to the word, and finally the root corresponding to the considered word. At the data preparation phase, the prepositions and conjunctions as well as the vowels were initially removed from the dataset. Then, the words were replaced with Qur'anic roots, and the Qur'anic roots were numbered. Besides this data set, another dataset was built, in which the Qur'anic roots related to each verse were saved for each verse. In addition, the data sets of the number of roots' repetition and also the order number of surahs were created.

5. Pre-processing

Figure 1 shows the frequency distribution of the roots in the whole Qur'an. Since the distribution of the number of roots in log-log scale is almost linear, the number of Qur'anic roots in the whole Qur'an follows the power-law distribution. It is possible that the frequency distribution fit into lognormal (Mitzenmacher, 2004) or power-law with exponential cutoff (Clauset et al., 2009) distributions, but the diagram shown in figure 1 is distinct from both of the mentioned distributions. The distinction reason of this distribution with the lognormal is the high number of roots with very low frequency. In addition, the tail of the distribution is long enough so that it is not necessary to fit it to power-law with exponential cutoff.

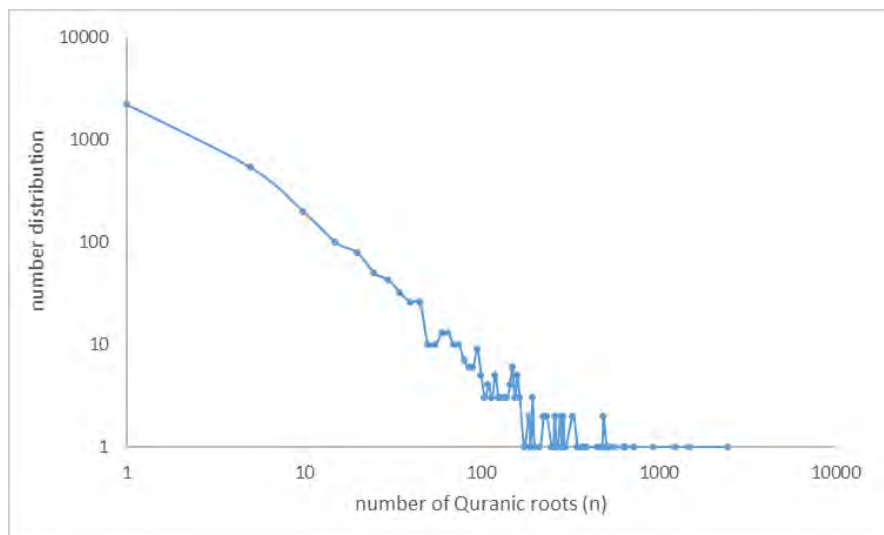2. In Arabic, vowels could be written or not.

Figure 1: The frequency distribution of the number of Qur'anic roots in the whole Qur'an

Figure 2 shows the number of roots in some large surahs of Qur'an such as al-Baqarah (Q.2), Āli ʿImrān (Q.3), al-Nisāʾ (Q.4), al-Māʾidah (Q.5), al-Anʿām (Q.6), al-Aʿrāf (Q.7), al-Anfāl (Q.8), al-Tawbah (Q.9) and Yūnus (Q.10). As seen herein, frequency distributions follow the semi-power-law distribution almost for all large surahs. The frequency distribution of roots in the whole Qur'an shows that many roots have low frequency in Qur'an, that is, around 50% of the roots have a frequency equal to or less than 3.
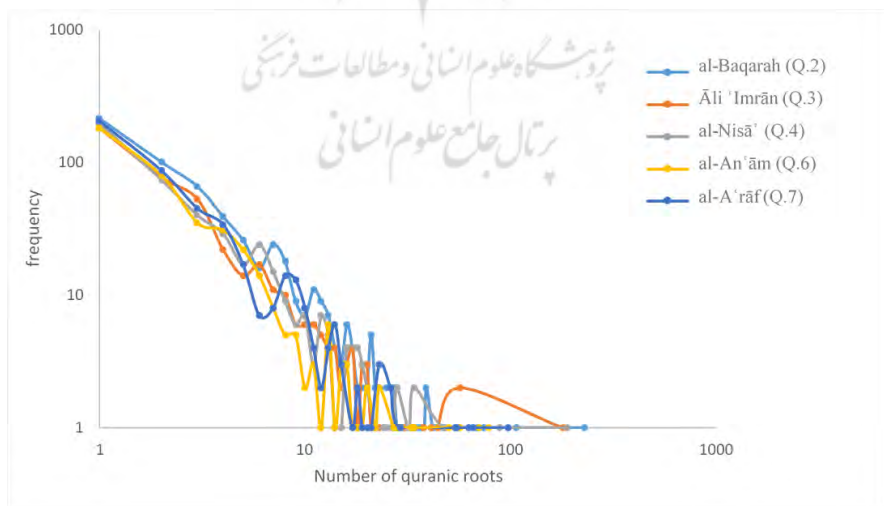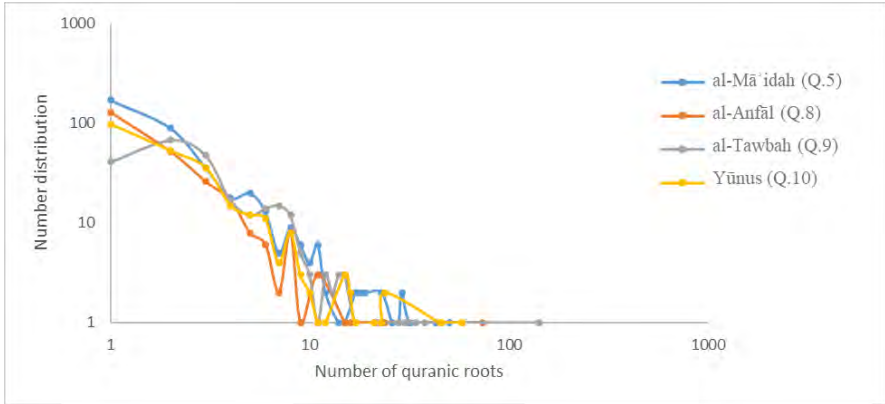
Figure 2: Frequency distributions of the number of Qur'anic roots in the large Qur'anic surahs

In addition, there are few roots such as 'ALH' and 'RBB' which are repeated many time in different surahs and are present in the whole Qur'an. The same issue is true for different surahs as well, that is, some roots are repeated much in particular surahs. The distinction of some of these roots shows that surahs' important topics and concepts could be exploited by algorithms such as tf-idf and etc.

6. The proposed method

After preparing the data, different surahs were initially partitioned. In this paper, surahs' sections proposed by Tabataba'i, (1996) has been employed with some amendments. Here on, based on the proposed methods below, the similarities of Qur'anic roots, sections, and surahs are obtained.

6.1. Calculating Qur'anic roots' similarity

Tf-idf can be calculated by the combination of term frequency in the document and the inverse document frequency. The frequency of the term t in the document d shown by $tf_{t,d}$ is the weight assigned to the term in proportion to the number of the occurrence of t in d. The inverse frequency of the document is also gained as below, where N is the total number of documents in the dataset and df_t equals the number of documents from the dataset which contain the phrase t.

$$idf_t = log\frac{N}{df_t}$$

Based on this, tf-idf of the term t in the document d is calculated according to the equation below (Larson, 2010).

$$tf\_idf_{t,d} = tf_{t,d} \times idf_t$$

For calculating the amount of similarity between two roots based on tf-idf, it is only needed to gain tf-idf vector of the roots based on the gained weight for different surahs. The cosine similarity of tf-idf vector of the roots can be a suitable measure for the similarity of roots.

Word2vec which was presented by Mikolov et al. (2013) in Google is a novel model to compute continuous vector representations of words, When the goal is representing larger elements, the generalized word2vec named sent2vec is usable (Le and Mikolov, 2014).

The other method used in this paper is the Roots' Accompaniment in verses (in short RA method). This method focuses on the accompaniment of roots in verses and is based on this presumption that if two roots in a verse are placed beside each other, they are related to each other, and the more the proportion of the two roots' accompaniment becomes, the more the amount of their relationship will be. Based on this presumption, the similarity of two roots can be calculated according to the following equation.

$$S_{ij} = \frac{N_{ij}}{\sqrt{N_i N_j}}$$

where N_{ij} is the number of the accompaniment of the roots i and j. The bottom of the fraction is also the geometric mean of the frequency of the two roots i and j in Qur'an. Then, the resulting matrix is normalized again so that the sum of elements of the similarity matrix at each row would equal 1.

$$W_{ij} = \frac{S_{ij}}{C_i}$$

$$\sum_i W_{ij} = 1$$

The value of C_i equals the sum of elements at each row of the matrix. This normalization is in order that the sum of the value of each root's similarity to other Qur'anic roots equal 1.

6.2. Calculation of similarity of sections and surahs

In this paper, the similarity of two sections, or two surahs, i and j is defined by averaging the similarity of all the roots which are in section i with all the roots in section j two by two. Therefore, the similarity of two

sections can be calculated by the equations below.

$$sim_{i,j} = \frac{\sum_{m \in i, n \in j} M_{sim}[m, n]}{l_i l_j}$$

,

$$sim_{i,j} = \sqrt[l_i l_j]{\prod_{m \in i, n \in j} M_{sim}[m, n]}$$

The first is the arithmetic mean and the second is the geometric mean. In these two formula, i and j respectively show the surah or section i and j, l_i and l_j show the number of the roots in the two sections, and M_{sim} shows the similarity matrix of roots, which is gained according to the methods mentioned above. It must be noticed that the arithmetic mean has been used in this paper.

The other solution to measure the similarity of two sections is the number of the same roots and its generalization cosine similarity. In this method, a vector is initialized for each surah as big as the number of Qur'anic roots. The elements of the vector are initialized by the number of Qur'anic roots existent in the surah. Based on this, the cosine similarity between the two vectors can use as the similarity between the two surahs. In some experiments of this paper, the simplified version of this method is used, i.e. the number or the ratio of common roots.

6.3. Evaluation

To assess different methods, the results of similarities (such as between surah title and content, between different surahs, between the first and last section of a surah, and between the concepts within a surah) have been compared to the random mode. In this paper, for selecting the random parts, we employ the selection of Qur'anic roots based on the probability of each root's occurrence in Qur'an, as below:

1) Calculate the frequency of different roots in Qur'an.

2) Take step 3 as many as the value of the length part.

3) Select a root based on the frequency of roots in Qur'an.

Therefore, it is more probable to select the roots with more frequency. In this method, to calculate the similarity with the random mode, 100 random couples have been selected and their similarities have been calculated and averaged.

6.4. Examining the structured-ness of Qur'anic surahs

In this paper, three sets of experiments have been designed for answering the first major question. The first set intends to calculate the similarity between surahs' title and within contents. The second set intends to assess topic sameness in Qur'anic surahs and studies the similarity between the concepts within a surah. The third and the most important set intends to assess the amount of relationship between the first section as an important section and the next sections.

For this, by the methods of Qur'anic roots' Accompaniment in verses and word2vec, the similarity of different Qur'anic roots was calculated and saved in a matrix called similarity matrix. Then, to calculate the similarity of the concepts within the surah, some Qur'anic surahs were selected in the way that surahs with different sizes exist among them. In addition, when the similarity between the sections of surahs is considered, small surahs and the surahs in the 30th section of Qur'an were not selected. It must be noticed that since the Qur'anic acronyms (*al- ḥurūf al-muqaṭṭaʿah*) have repeated only once in Qur'an, surahs such as YāSīn[3] (Q.36) and Qāf[4] (Q.50), the name of which has been adopted from the Qur'anic acronyms, were not selected either. As a summary, the following similarities were calculated for the selected surahs.

1) Similarity between the surah's title and the Qur'anic roots within the surah

2) Similarity between the Qur'anic roots within the surah

3) Similarity between the first and the last sections of the surah

4) Mean similarity between the first section and different sections of the surah

It must be noticed that since very frequent roots such as Allah and 'RBB' have repeated in different sections and make fewer distinctions, and that the presence of these roots in a section caused a rise in the similarity of the section with other sections, the very frequent roots were removed for RA method so as to calculate the similarity between sections.

The similarities were initially calculated by removing the roots with the frequency of above 800, and at the next stage by removing those with the frequency of above 600, 400, and 200. Then, the calculated similarities were compared with the random mode and the surahs' structured-ness was accordingly assessed in terms of topic sameness.

---

3 . يس
4 . ق

6.5. Study of Introduction and Explanation theory

Each of the Qur'anic studies' researchers has presumed a particular structure for surahs based on their own viewpoint and ideological background and each is seeking signals for proving their own claim in their own way (Fatahizadeh and Zakeri, 2016; Jigareh and Sadeghi, 2017). Based on the commonest theory of surahs' structured-ness, the Almighty proposes the topic and main idea of the surah at the first section, then explains that in different sections, and finally presents the conclusion. This structure, in this paper, has been called the structure of Introduction & Explanation. In this section, we intend to examine the structured-ness of surahs in terms of Introduction and Explanation theory.

To study this theory, we initially calculate the similarity of the first section which contains the main topic based on the theory, and the final section, which concludes from the proposed topics within the surah, then compare it to random mode. Then we calculate and show the similarity between the first section and the different sections of the surah, which are an explanation of the first section according to the theory of Introduction & Explanation.

6.6. Study of the organization of Qur'anic surahs' order

To answer the second question about the organization of surahs' order, an experiment was designed as follows.

1) The similarity of Qur'anic surahs was initially measured two by two and saved in surahs' similarity matrix. For measuring the similarity of surahs, tf-idf and RA similarity matrices were used.

2) Based on the order number of surahs, the place of each surah was defined. Therefore, the place of the surahs al-Fātiḥah (Q.1), al-Baqarah (Q.2) and al-Nās (Q.114) were considered 1, 2, and 114, respectively. Based on this, the place distance of two surahs was computed as below.

$$PD_{s_1,s_2} = \left| P_{s_1} - P_{s_2} \right|$$

, where P_{s_1} is the place of the surah s_1 and P_{s_2} is the place of the surah s_2.

3) Based on the revelation order of different surahs, the time distance of surahs s_1 and s_2 was computed as follows.

$$TD_{s_1,s_2} = \left| T_{s_1} - T_{s_2} \right|$$

, where T_{s_1} and T_{s_2} are the descent time of the surahs s_1 and s_2.

4) At this stage, the average similarity of surahs with place distance $0 <$

$pd < 114$ was calculated as follows.

$$AveDisSim(pd) = \frac{\sum_{PD_{s_1,s_2}=pd} Sim(s_1, s_2)}{114 - pd}$$

, where $114 - pd$ is the number of surahs the distance of which is equal to pd.

Similarly, the average similarity of surahs with the time distance $0 < td < 114$ was calculated as follows.

$$AveDisSim(td) = \frac{\sum_{TD_{s_1,s_2}=td} Sim(s_1, s_2)}{114 - td}$$

We plotted the diagram of surahs' similarity versus the place distance and also surahs' similarity versus the time distance was drawn and the change of the surahs' similarity was studied based on their place and time distance and the result thereof was informed.

7. Experimental results

The experimental results on studying topic sameness, Introduction and Explanation structure, and surahs' order are presented in this section.

7.1. The connection between the surah's title and the surah's content

Figure 3 shows the frequency of the title of surahs repeated in the surah itself versus the random mode.
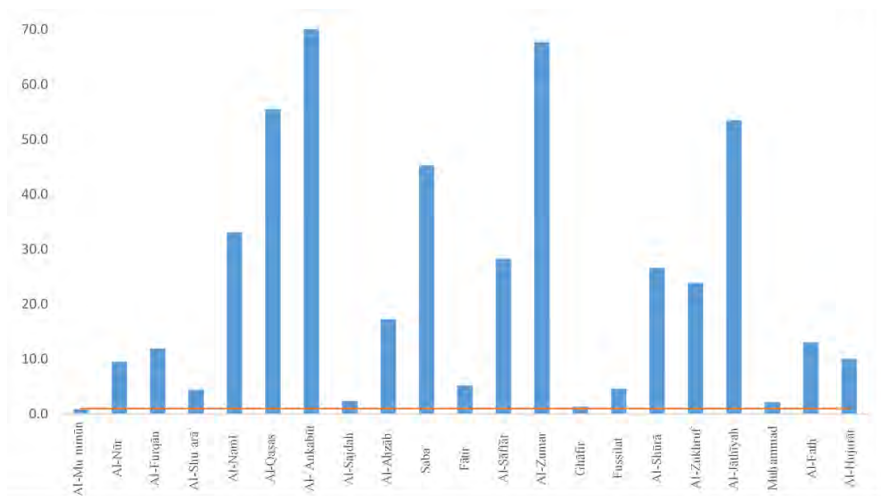
Figure 3: The frequency of titles of surahs in comparison to the random mode. The red line shows the random mode.

As seen, in most surahs, the frequency of the title within the surah is much higher than the random mode. This fact, however, is not true about all surahs and the frequency of the title is also very low for some surahs. For instance, while a majority of the surah al-Anbīyāʾ (Q.21) is about prophets, the related root[5] has not been repeated at all. However, the names of different prophets have been mentioned in this surah, such as Idrīs, Noah, Abraham, Ismāʾīl, Isaac, Jacob, Lūṭ, Yūnus, Moses, Aaron, David, Solomon, Zechariah, and Yaḥyā. Therefore, it could be said that it is possible that the title of the surah be low-frequency in the surah, but concepts similar or related to the title be repeated in the surah over and over.

To solve the above problem, the mean similarity of the surah's title with the concepts within the surah was take into study. Figure 4 shows this similarity based on the RA method versus the random mode.

In figure 4, on the contrary to Figure 3, the similarity and relationship between surahs' titles and inner concepts is much higher than the random mode for all surahs. For example, this similarity for the surah al-Anbīyāʾ (Q.21) is 7 times that of the random mode.
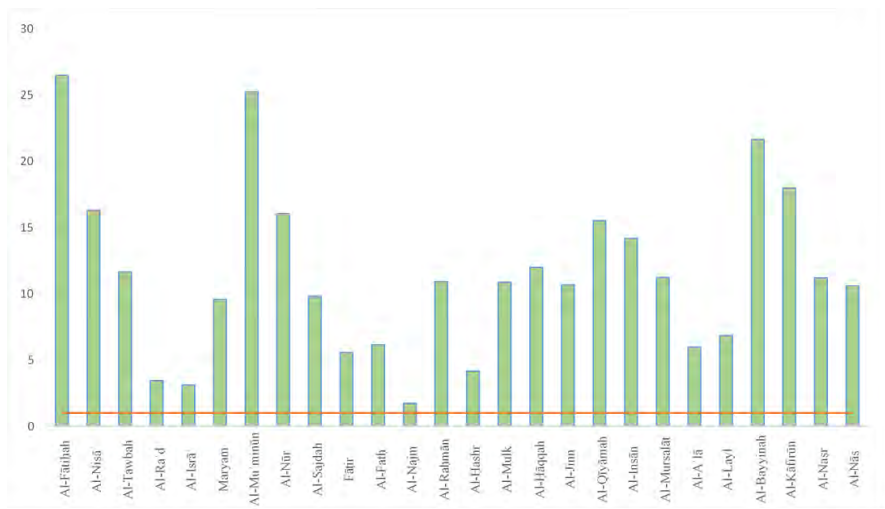
---

5. ʾNBWʾ which means message.

Figure 4: The similarity of surahs' titles with the concepts within surahs based on RA similarity

If we use the algorithm word2vec for measuring the similarity, the similarity between the surah's title and concepts therein will be as Figure 5.
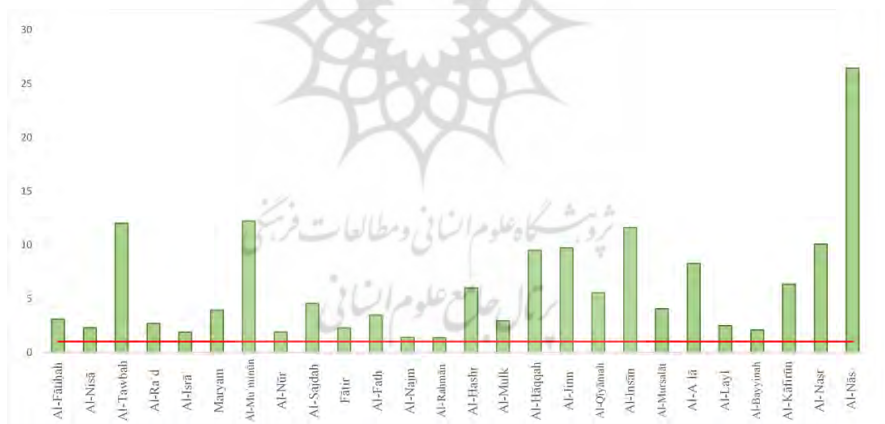

Figure 5: The similarity between the surahs' title and concepts therein based on the algorithm word2vec

According to figures 4 and 5, it could be said that the surah's title has been similar and tightly related to the inner concepts of the surah for almost all surahs. Therefore, the selection of the surah's title has been a logical issue, and cannot have come up based on the ordinary public's selection merely. However, the similarity gained by word2vec is lower, which seems

to be due to the small training data set (Qur'an).

7.2. Experimental results on topic sameness theory

After examining the surah's titles, topic sameness or, in other words, the structured-ness of surahs' inner concepts was studied. Figure 6 presents the similarity of intra-surah concepts versus the random mode.
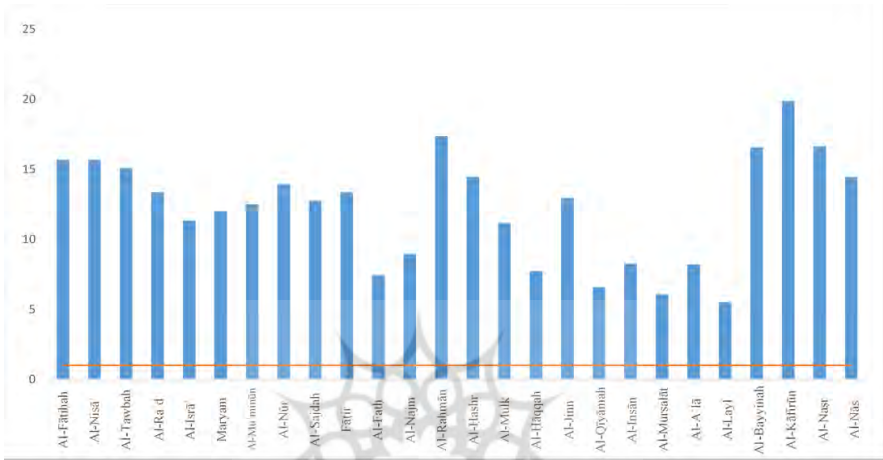

Figure 6: The amount of similarity between intra-surah concepts versus the random mode

As seen above, the similarity of the intra-surah concepts for all examined surahs is much higher than the random mode. On the average, the similarity of these concepts is above 12 times that of the random mode. This shows that the intra-surah words in all examined surahs are coherent to each other. This observation shows that each surah has formed around an explicitly single topic or several interrelated topics, although not explicitly supporting a major topic.

It must be noticed that although the results presented in this paper are related to 26 surahs, i.e. a quarter of Qur'an, they can be, for two reasons, true for the whole Qur'an except some special surahs. First, it was tried to select the surahs in a way that surahs with different sizes be studied so that if the size of the surah influences the result of the calculations, it would be recognized. Second, more than 26 surahs were studied in this paper, where the same results were also true about some other surahs, but they were not included herein due to space shortage. However, it must be noticed that the surahs the name of which has derived from the Qur'anic acronyms or the surahs the title of which has low frequency in Qur'an are exceptions to this result. This is simply because there is not adequate knowledge about their

titles, so it is not possible to calculate the similarity of these surahs' titles to other roots correctly by the existent NLP methods.

7.3. Experimental results on Introduction & Explanation theory

Figure 7 shows the proportion of the average similarity of the first and the last sections of surahs to that of the random mode. For this diagram, RA method was used. For example, RAccomp800 shows RA without consideration of roots with the frequency higher than 800.
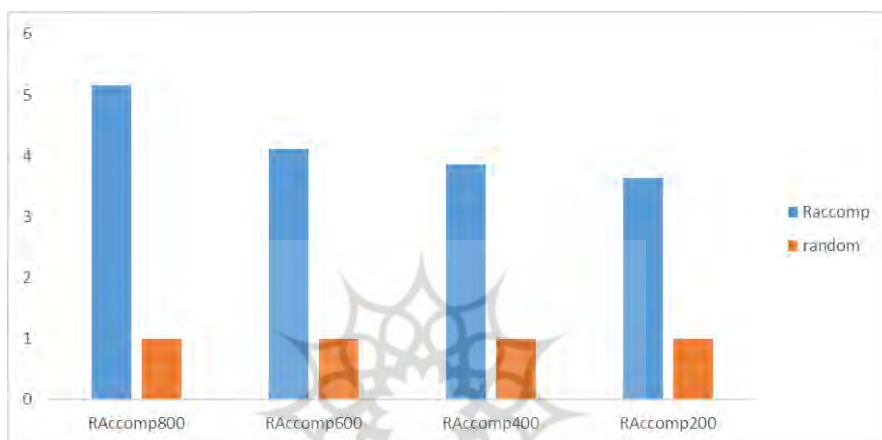

Figure 7: Comparing the average similarity of the first and the last sections of surahs with the random mode

As seen in figure 7, the average similarity of the first and last sections of surahs is much higher than the random mode. The average similarity is more than 4 times based on RA method. Although the average similarity is higher than that of the random mode, its value is not considerable enough to be able to conclude that the structure of all Qur'anic surahs is conforming to the theory of Introduction and Explanation. It seems that this issue is due to the averaging over all surahs and since it is possible that some surahs may not follow Introduction and Explanation, the final result is less than the prediction. Therefore, the similarity of the sections for different surahs should be studied separately.

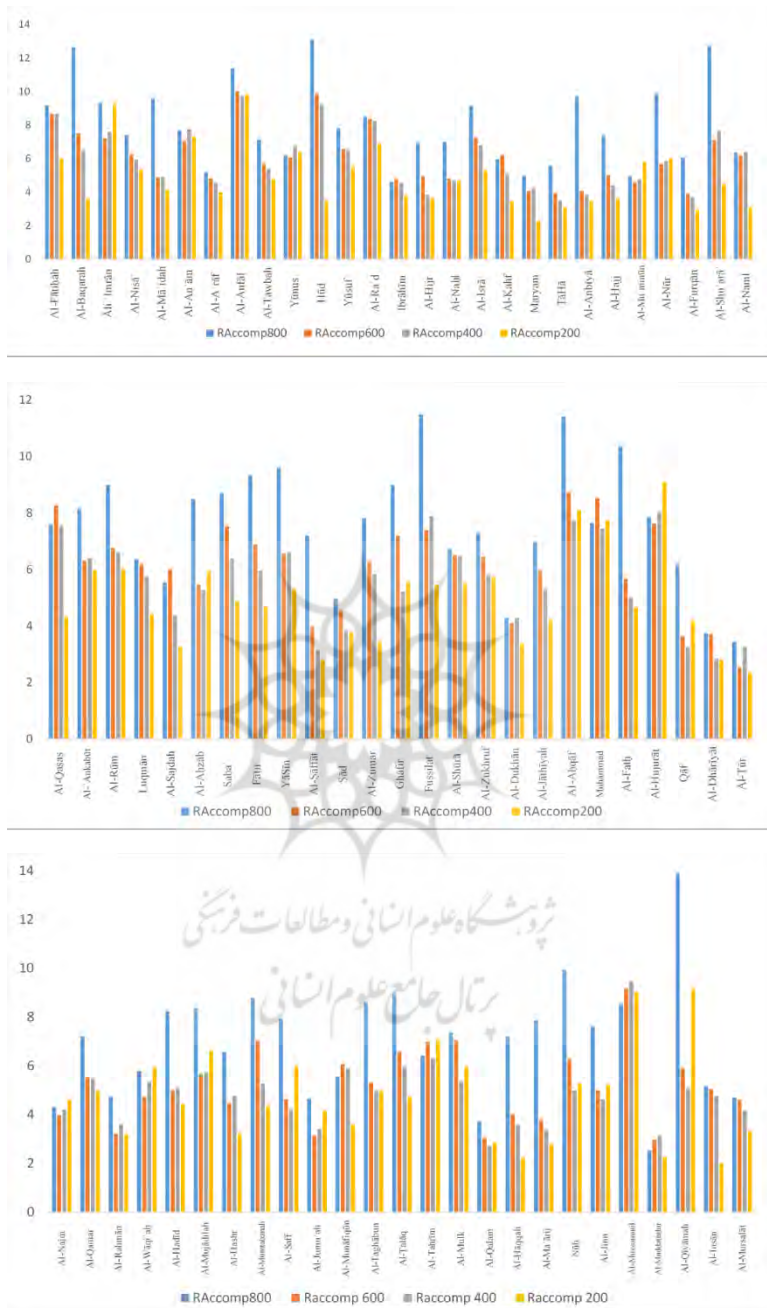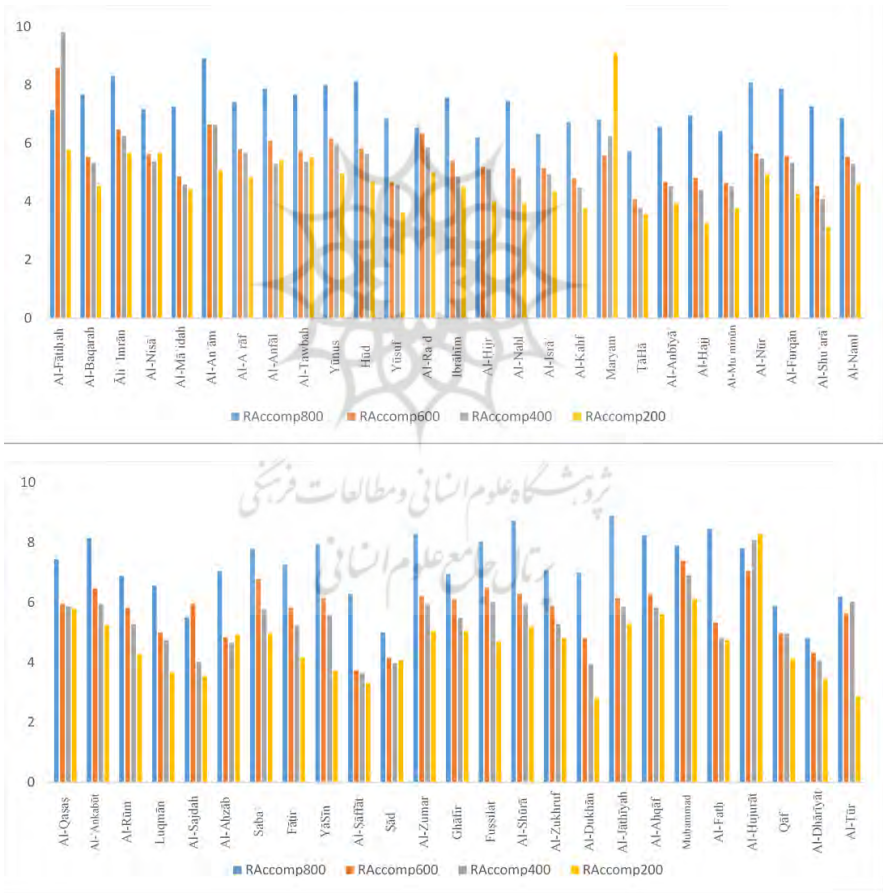Figure 8 shows the similarity of the first and last sections of each surah based on RA method.

Figure 8: The proportion of the similarity of the first section and last sections of surahs to the random mode

As seen in the figure, the similarity of the first and last sections in different surahs is much more than the random mode. Obviously, the similarly of different surahs to that of the random mode is different. Herein, except for RAcompp800 mode, the surah al-Anfāl held the most similarity. As per the RAcompp800 mode, the surahs al-Qīyāmah and Hūd showed the most similarities. In addition, with regard to the average of all modes, the surahs al-Anfāl, al-Muzzammil, al-Aḥqāf, and Hūd held the most similarity of the first and last sections respectively.

After studying the similarity between the first and last sections, we studied the presence of the first section concepts throughout the whole surah. Figure 9 shows the average similarity between the first section and all sections of the surahs al-Fātiḥah to al-Mursalāt.
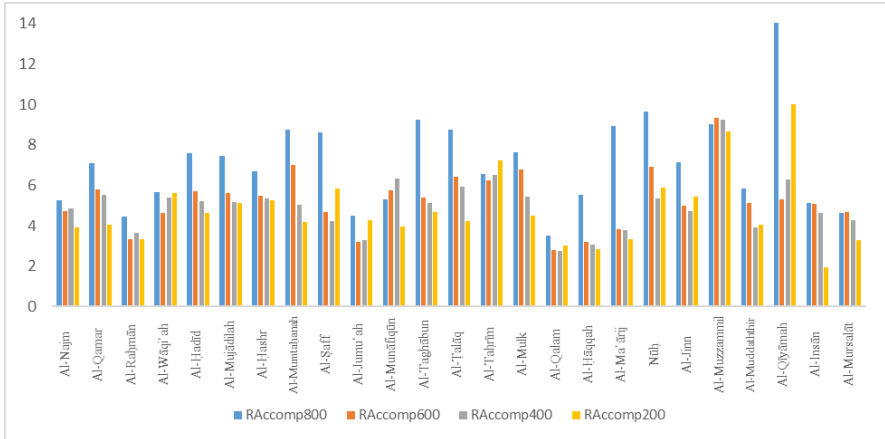
Figure 9: The proportion of the similarity of the first section and all other sections of surahs to the random mode

As seen in the figure, almost in all Qur'anic surahs, the similarity of the first section to the other sections is much more than the random mode. For the lowest case, which is related to al-Qalam, the similarity is more than 3 times that of the random, and above 9 times, for al-Muzzammil, as the highest case. The surahs al-Qīyāmah, al-Ḥujurāt, and Muḥammad are located next. This shows that the concepts stated in the introduction of the surahs are running throughout each of the surahs to some extent, being explained. Although this is more or less true for different surahs, it strengthens the Introduction and explanation theory, while more study is required for the surahs where the similarity is lower than that of other surahs.

7.4. Experimental results on surahs' order

Some researchers of Qur'anic studies regard the prophet's friends as one of the factors of surahs' order of Qur'an, so do not approve any logical or special order for surahs. Some others believe in organization of Qur'anic surahs' ordering, to be either logical or occasionally revelatory. However, organization of surahs' order is a complex problem for which the logical relationships between adjacent clusters of surahs in Qur'an must be studied using different methodologies, what is not studied in this paper. In this section, we compare the similarity of close surahs in terms of their order in Qur'an to close surahs in terms of their order of revelation with resolutions 1, 3, 5, 10, and 20. What is meant by resolution r is the size of the window in which the average similarity of surahs is calculated for those with the distance less than or equal to r. for instance, resolution 1 calculates the average similarity of adjacent surahs and resolution 3 calculates the average

similarity of the surahs the maximum distance of which is 3.

Figure 10 shows the similarity of Qur'anic surahs versus their place distance from each other for different resolutions.
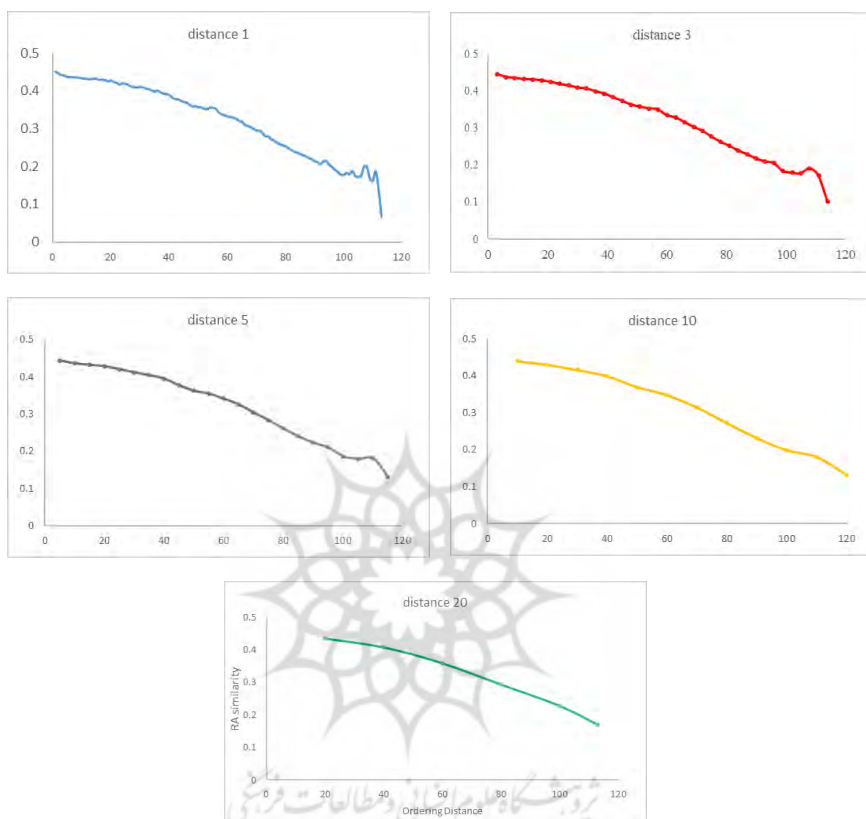

Figure 10: Similarity of different surahs versus their distance from each other based on RA similarity

According to Figure 10, similarity of different surahs reduces almost linearly by increasing their distance, as the average similarity of adjacent surahs is 0.45 and that of the surah with the most distance from each other is 0.06. The same fact is true for other resolutions.

Figure 11 also presents the diagram for the similarity of surahs based on the tf-idf versus surahs' distance.
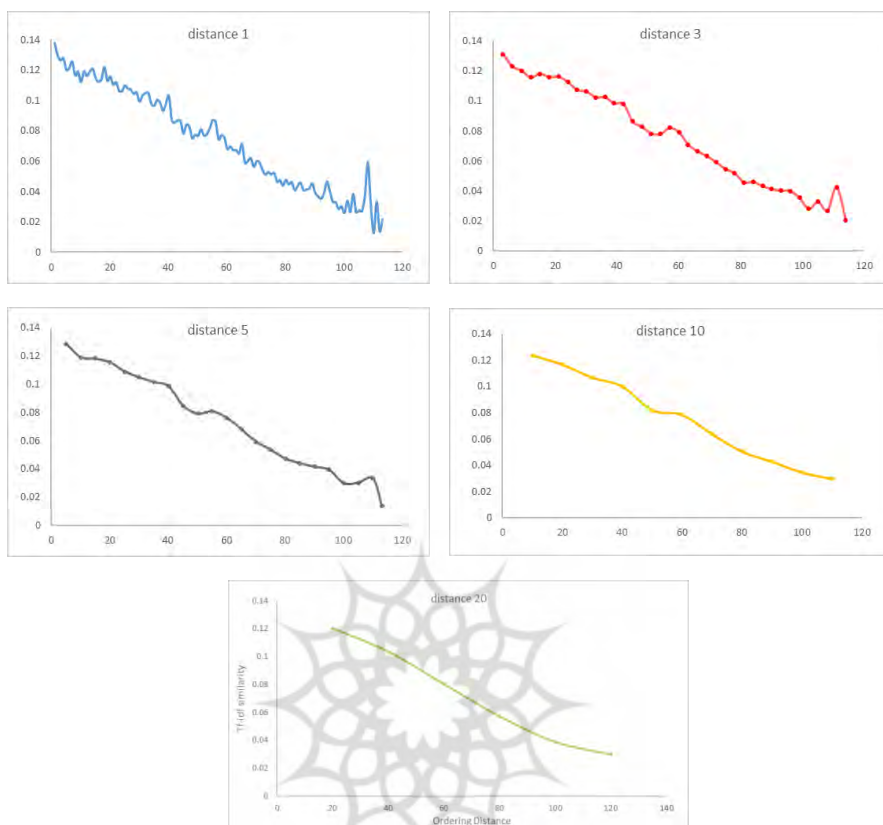
Figure 11: Similarity of different surahs versus their distance from each other based on tf-idf similarity

Based on this figure, the surah's similarity by tf-idf also shows a descending trend by increasing ordering distance.

It can be concluded from Figures 10 and 11 that surahs' distance and their similarity are correlated at least at the macroscopic level, so that the more the distance between the surahs are, the less their average similarity will be. This finding shows kind of macroscopic organization of surahs ordering beside each other. Due to short space, we postpone the microscopic analysis of surahs similarity to another paper and consider it enough to only mention that two hypotheses are imaginable based on this result: that most of adjacent surahs are conceptually related, and that Qur'anic surahs are in the form of different interrelated clusters and categories the surahs of each of which are tightly interrelated, and the category of the related surahs in Qur'an are also located beside each other.

For a more detailed study, we compared the gained results to Qur'anic surahs' similarity versus the time distance of their revelation order. Figure 12 presents the surahs' similarity versus the time distance of their revelation.
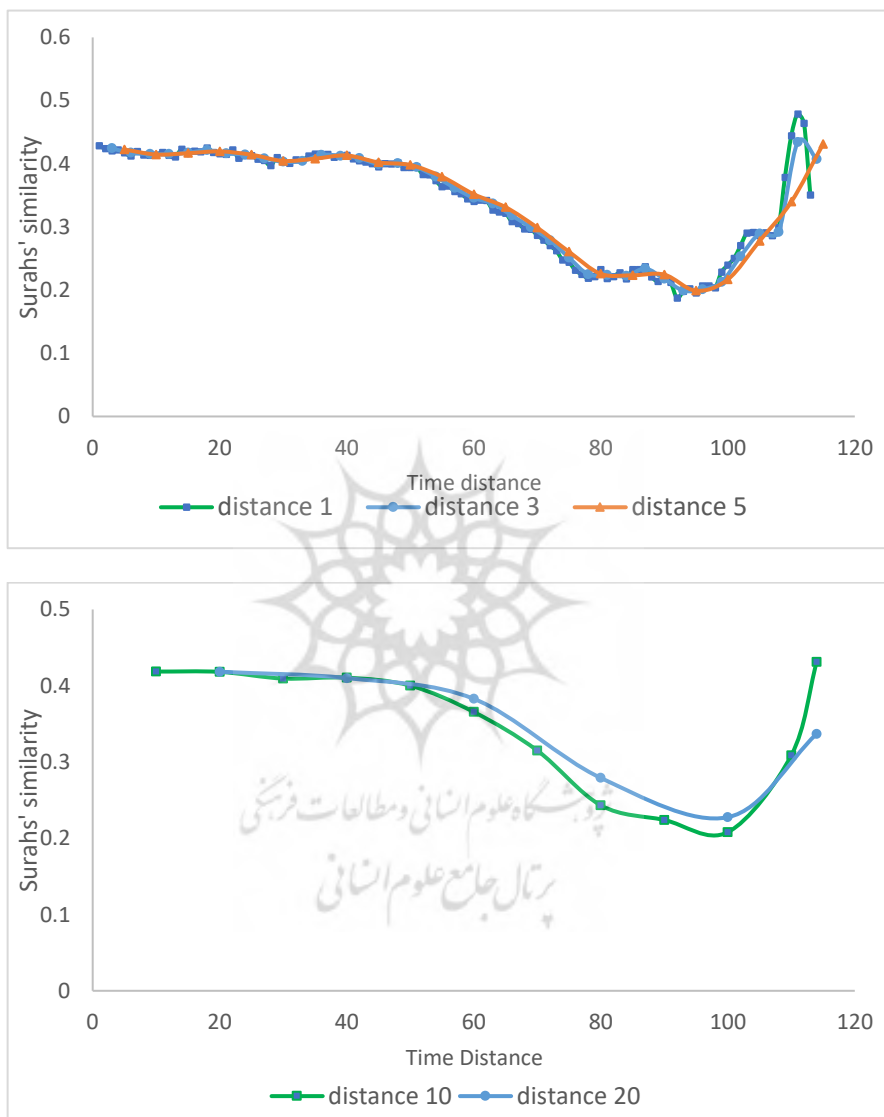

Figure 12: Similarity of different surahs versus their revelation ordering distance from each other based on RS similarity

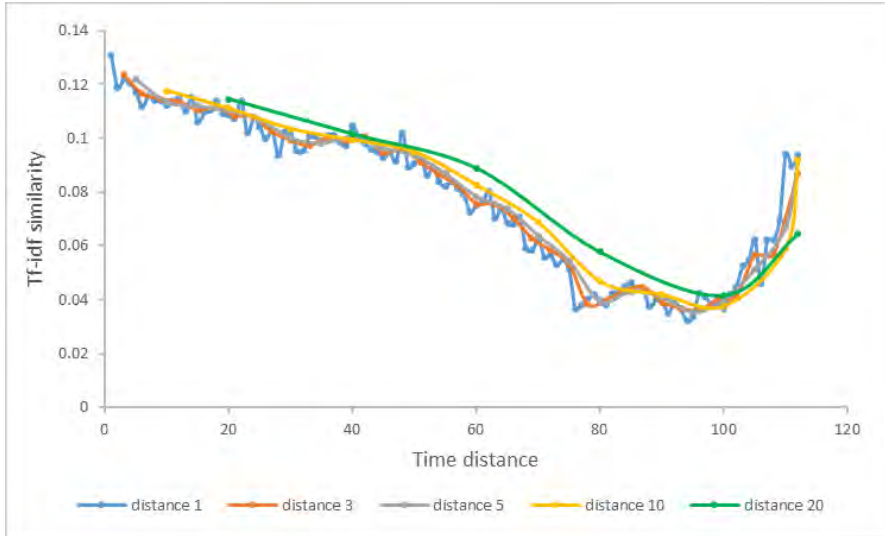This diagram will be as follows for the tf-idf similarity.

Figure 13: Similarity of different surahs versus their revelation ordering distance from each other based on RS similarity

As seen, surahs' similarity versus the order of their revelation especially based on RA similarity does not follow a special trend. In RA method, the similarity does not initially change much by the increase in the two surahs' distance of revelation order. The similarity declines for the distances between 60 and 100, then increases. The same description is true with a less intensity for tf-idf.

Accordingly, it could be concluded that the order distance of surahs in Qur'an is related to their similarity so that those closer to each other in terms of ordering are also more similar with relatively high probability. Opposite to the order distance of surahs, such organization is not true for the time distance of surahs' revelation. Therefore, it could be concluded that the surahs' ordering in Qur'an follows a logical organization, which requires more accurate and detailed study. In addition, it is accordingly recognizable why Qur'anic surahs are not ordered by their revelation order. Numerous annotating results would be gained upon more accurate study of this issue, results such as why the first surah is at the beginning of Qur'an named al-Fātiḥah that means "The Book's Opener", or what relationship there is between the close clusters of surahs.

8. Conclusion

This research was carried out with two purposes: examining each surah's

inner organization according to theories Topic Sameness and Introduction and Explanation, as well as surahs' ordering in whole Qur'an. In this regard, the Qur'anic data were initially prepared and cleared. Then by applying tf-idf, word2vec, and roots' accompaniment in verses, the similarity of Qur'anic roots was gained. By the calculated similarities, the link between the surah's topic and the body content was firstly calculated. Second, Topic Sameness of each surah was studied by calculating the similarity between the inner concepts of each surah. Third, the existence of the structure of Introduction and Explanation was assessed in Qur'anic surahs. The results compared to those of the random mode showed that the similarity of both the topic of each surah to the body concepts and between concepts to each other is much more than that of the random mode. This translates that Qur'anic surahs have been mostly formed around a single topic. In addition, surahs' organization based on the Introduction and Explanation structure is examined by computing the similarity between the first and last sections and also the first section and other sections of different surahs. Finally, based on the study of the correlation between surahs' order in whole Qur'an and their revelation order with surahs' similarity, we conclude that the surahs' ordering in whole Qur'an is relatively organized as well.

Future works

As this paper is the first work on algorithmic study on Qur'an's organization, it is reasonable that some similar works be done on each individual surah in more detail. In addition, study of the similarity between the structure of different surahs or surah clusters seems to be of interest. On the other hand, as the section-definition by Tabataba'i turned out to be imperfect, it is very important for the future works to involve manual or automatic section-definition as the prerequisite for studying the organization of surahs accurately. In addition, by involving the humane expert, rather than NLP similarity algorithms, more accurate results are available in study of surah's organization. Finally, other possible future works accordingly include those based on other methods of similarity calculation, comparing Qur'an's organization to other books, and studying organization of Qur'anic clusters.

References

Alfaifi, A. and Atwell, E. (2016). Comparative evaluation of tools for Arabic corpora search and analysis. *International Journal of Speech Technology*, 19, 347-357. https://doi.org/10.1007/s10772-015-9285-5

Alhawarat, M. (2015). Extracting Topics from the Holy Qur'an Using Generative Models. *International Journal of Advanced Computer Science and Applications*, 6(12), 288-294. https://doi.org/10.14569/IJACSA.2015. 061238

Aram, M. R. and Layeqi, F. (2017). A Study of the Structure of Surah al-Ma'idah Based on the Tree Structure Approach. *Pazhouhesh Name-ye Qur'an va Hadith*, 10(19), 55-77.

Arberry, A. J. (1996). *The Koran interpreted: A translation*. New York: Simon and Schuster.

Atwell, E. and Sharaf, A. (2009). A Corpus-based Computational Model for Knowledge Representation of the Qur'an. *Proceedings of CL2009 International Conference on Corpus Linguistics*. University of Liverpool.

Bell, R. (1953), *Introduction to the Qur'an*. University Press.

Clauset, A., Shalizi, C., Newman, M., (2009). Power-law distributions in empirical data. *SIAM review*, 51(4), 661-703.

Dehghani Farsani, Y. (2008). The Structure of Surah al-Inshiqaq. *Balagh Mobin*, 14, 3-14.

Dukes, K., and Buckwalter, T. (2010). A dependency treebank of the Qur'an using traditional Arabic grammar, *2010 The 7th International Conference on Informatics and Systems (INFOS)*. Cairo. 1-7.

Fatahizadeh, F. and Zakeri, M. (2016). A Structuralist Approach toward Surah Al-Kahf. *Journal of the Holy Qur'an and Islamic Texts*, 7(25), 101-120.

Hamed, S. and Aziz, M. (2016). A Question Answering System on Holy Qur'an Translation Based on Question Expansion Technique and Neural Network Classification. *Journal of Computer Science*, 12(3), 169-177. https://doi.org/ 10.3844/jcssp.2016.169.177

Iqbal, R., Mustapha, A., Yusoff, Z. (2013). An experience of developing Qur'an ontology with contextual information support. *Multicultural Education & Technology Journal*, 7(4), 333-343.

Ismail, R., Bakar, Z. A., Rahman, N. A. (2016). Ontology Learning Framework for Qur'an. *2016 Advanced Research in Engineering and Information Technology International Conference (AREITIC)*. Bandung. Indonesia.

Jigareh, M. and Sadeghi, Z. (2017). Investigation and analysis of Sura al-Enfetar, relying on Structuralism theory. *Journal of Critique of Arabic Literature*, 7(1), 50-74.

Khamehgar, M. (2002a). An Introduction to Structural Interpretation of the Qur'an. *Qru'anic Reserches*, 8(29-30), 208-271.

Khamehgar, M. (2002b). The Geometric Structure of Qur'an's Chapters: An Introduction to the Structural Interpretation of Qur'an. *Golestan Qur'an*, 138, 9-

13.

Khamehgar, M. (2004). The Phrasing of Qur'an's Anecdotes and the Chapters' Objectives. *Golestan Qur'an*, 179, 13-17.

Khamehgar, M. (2006). A Look into the First Structural Translation of the Holy Qur'an. *Bayyenat*, 49, 278-291.

Khamehgar, M. (2008). Theory of Purposefulness of Sūras; Principles and Backgrounds. *Qru'anic Reserches*, 13(54-55), 182-213.

Khan, H. U., Saqlain, S. M., Shoaib, M., Sher, M. (2013). Ontology based semantic search in Holy Qur'an. *International Journal of Future Computer and Communication*, 2(6), 570-575. https://doi.org/10.7763/IJFCC.2013. V2.229

Larson, R. (2010). Book Review: Introduction to information retrieval. *Journal of the American Society for Information Science and Technology*, 61(4), 852-853. https://doi.org/10.1002/asi.21234

Le, Q. and Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *Proceedings of the 31st International Conference on Machine Learning*, PMLR 32(2), 1188-1196.

Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*. https://doi.org /10.48550/arxiv.1301.3781

Mitzenmacher, M. (2004). A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Mathematics*, 1(2), 226-251. https://doi.org/10.1080/15427951.2004.10129088

Safee, M., Saudi, M., Pitchay, S., Ridzuan, F. (2016). A Systematic Review Analysis for Qur'an Verses Retrieval. *Journal of Engineering and Applied Sciences*, 11(3), 629-634.

Sharaf, A. and Atwell, E. (2012a). QurSim: A corpus for evaluation of relatedness in short texts. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul. Turkey. 2295-2302.

Sharaf A. and Atwell, E. (2012b). QurAna: Corpus of the Qur'an annotated with Pronominal Anaphora. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. 130-137.

Sherif, M. and Ngonga Ngomo, A. (2015). Semantic Qur'an: A multilingual resource for natural-language processing. *Semantic Web*, 6(4), 339-345.

Shoaib, M., Yasin, M., Hikmat, U., Saeed, M., Khiyal, M. (2009). Relational WordNet model for semantic search in Holy Qur'an. *2009 International Conference on Emerging Technologies*, 29-34, https://doi.org/10.1109/ICET .2009 .5353208.

Soucy, P. and Mineau, G. (2005). Beyond TFIDF weighting for text categorization

in the vector space model. *Proceedings of the 19th international joint conference on Artificial intelligence (IJCAI'05)*. San Francisco: Morgan Kaufmann Publishers Inc., 1130–1135.

Tabataba'i, M. H. (1996). *al-Mīzān fī Tafsīr al-Qur'an*. vol. 2. Qom: Daftar enteshārāt islāmī.

Yauri, A., Kadir, R., Azman, A., Murad, M. (2013). Qur'anic verse extraction base on concepts using OWL-DL ontology. *Research Journal of Applied Sciences, Engineering and Technology*, 6(23), 4492-4498. https://doi.org/10.19026/rjaset .6.3457

Zhang, Y., Jin, R., Zhou, ZH. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1, 43–52. https://doi.org/10.1007/s13042-010-0001-0