Relationship Analysis of Keyword and Chapter in Malay-Translated Tafseer of Al-Quran

S.Chua[1], P.N.E.Nohuddin[2]
[1]*Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak.*
[2]*Institute of Visual Informatics, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor.*
chlstephanie@unimas.my

***Abstract*—A number of studies have gained popularity to study the unseen knowledge categories and relationship of subject matters discussed in the Al-Quran or the Tafseer. This research investigates the relationships between verses and chapters at the keyword level in a Malay translated Tafseer. A combination technique of text mining and network analysis is developed to discover non-trivial patterns and relationships of verses and chapters in the Tafseer. This is achieved through keyword extraction, keyword-chapter relationship discovery and keyword- chapter network analysis. A total of 130 keywords were extracted from six chapters in the Tafseer. The keywords and their relative importance to a chapter are computed using term weighting. A network analysis map was generated to visualize and analyze the relationship between keyword and chapter in the Tafseer. The relationship between the verses and chapters at the keyword level are successfully portrayed through the combination technique of text mining and network analysis. The novelty of this approach lies in the discovery of the relationships between verses and chapters that is useful for grouping related chapters together.**

***Index Terms*—Text Mining; Network Analysis; Al-Quran Tafseer.**

I. INTRODUCTION

Al-Quran is the principle book of Muslims faith and practice. According to [1], 1.6 million Muslims worldwide use the Al-Quran as their reference book and therefore, it is beneficial for Muslims in general and Islamic scholars to be able to gain information from it. Islamic scholars or even any Muslim always recite or refer to any verses or chapters to support formal and informal practices. Al-Quran has been translated into various languages, normally done by individual Islamic nation, to facilitate the understanding of the content of Al-Quran by its people. Many verses in Arabic Al-Quran are difficult to understand. This leads to difficulty in understanding the knowledge categories and relationship of subject matters discussed in the Al-Quran. During the process of translating the Al-Quran, many aspects in Al-Quran's syntax, rhetoric, word and verbal similarity are involved [2], [3]. One of the significant issue is that in translating the Al-Quran, there is a lack of similarity or the absence of the equivalent Islamic terms [4]. Thus, it can be a challenge to have a complete understanding on Al-Quran and its translation for non-Arabic speaker Muslims. This serves as a motivation to explore the relationship of subject matters discussed in the Al-Quran. In this research, the Malay-translated Tafseer of Al-Quran is used. Data Mining is a process of discovering patterns in large databases. It includes processes of cleaning and analyzing data with the aim to discovering hidden knowledge [5]. Similarly, text mining is the process of discovering information in textual data. Natural language text is unstructured, formless and relatively difficult to deal with in comparison with other qualitative type of data such as numerical, nominal and decimal data. Data mining techniques have since been adopted in text mining to discover patterns in unstructured textual data. Text mining techniques often process documents to categorize content, classify documents according to keywords, making links be- tween otherwise unconnected documents and providing visual maps. Furthermore, an effective visualization technique is vital to ensure that the discovered knowledge is properly conveyed. The aim of this study is to discover the relationships between keywords and chapters using a combination technique of text mining and network analysis to discover non-trivial patterns and relationships of verses and chapters at keyword level in a Malay translated Tafseer. As mentioned earlier, it can be difficult to understand the content and important knowledge in the Tafseer. In this study, a text mining technique is introduced to transform verses and chapters into keyword-chapters vector, and after which, to use network analysis method to discover the relationships between the keywords and chapters. This can assist Muslims to comprehend the Tafseer based on thematic (keyword) approach, which is more effective to grasp [6].

The rest of the paper is organized as follows. Section II discusses the background of several topics that are related to the keyword-chapter relationship analysis. Then Section III provides a description of the modules of the proposed framework for keyword-chapter relationship analysis. This is followed by Section IV which presents the results and discussion. Finally, in Section V, the paper is concluded with a brief summary and future research work.

II. BACKGROUND AND RELATED WORK

A. Al-Quran and Tafseer

The Al-Quran contains a unity of subjects or themes which are described in 144 chapters (surahs) and each chapter consists of a number of verses (ayat). It contains the extraordinary words of Allah discovered by Prophet Muhammad through the angel Jibrail. The Quranic revelation is regarded by Muslims as the exact words of Allah. The longest chapter in Al-Quran has 286 verses (Surah Al-Baqarah) while the shortest chapter has only three verses (Surah Ul-Kawthar). The rest of the chapters have different numbers of verses that fall in between. The length of the

chapters may vary and likewise, the verses in each chapter have different lengths too. The chapters of Al- Quran are not arranged in chronological order. It starts with Surah Al-Fatihah and ends with Surah An-Nas.

There are several subjects covered in the Al-Quran that is essential knowledge in the Islamic code of practice. Verses in the Al-Quran are in Arabic. However, it has been said that a proper Muslim is required to accurately understand the content of the Al-Quran. Therefore, the Al-Quran is translated into many languages so that Muslims around the world can understand its content.

Several researches had been done on analyzing the contents of the Al-Quran. [7] proposed a statistical classifier software tool to group the Al-Quran corpus, in Arabic, and a theme table is generated to illustrate the verses and chapters classification. Frequent patterns were also mined from the Al-Quran, both in Arabic [8] and Malay [9] to discover association rules. More recently, [10] automatically categorize the Tafseer of verses of Holy Quran using the KNN algorithm. [11] used a semantic interpreter and a supervised learning method for classification of candidate answers to questions based on Quranic ontology of concepts. [1] used text mining approaches to discover the characteristics of the Holy Quran using term frequencies.

Despite the text mining approaches to mine the Al-Quran, there is no known text mining approach to discover the relationship of chapters in the Al-Quran or Tafseer. Therefore, the proposed framework of keyword-chapter relationship analysis offers an option for discovering groups of related chapters in the Malay translated Tafseer at the keyword level.

B. *Keyword Extraction using the Term Frequency-Inverse Document Frequency statistics*

A keyword is defined as a term selected from a document, which is deemed important in relation to that document. Keywords are usually identified in text mining tasks as they are the basic units to work with in a document. The importance of a candidate keyword can be calculated using a number of term weighting functions, commonly used in text mining. One of the popular functions to calculate how important a term is to a document is the term frequency-inverse document frequency (TF-IDF) statistics. Term frequency is simply the number of times a term occurs in the document. On the other hand, the inverse document frequency measures the commonness of a term across all documents. This is calculated by using the logarithm of the total number of documents in a collection divided by the number of documents containing the term. TF- IDF can be formulated as follows:

$$TF - IDF\,(t, d, D) = tf(t, d) \cdot log(|D|/|d(t)) \qquad (1)$$

where, t is a term in a document, d and D is a collection of documents. Further explanation of this function can be found in the review by [12].

C. *Network of Document Relationships*

Finding documents which are similar or related in terms of subject matters can assist researchers, academicians and leisure readers to understand more about the collection of documents. This can be done using combination of techniques in text mining, information retrieval and document cluster analysis.

Document cluster analysis is one of the clustering techniques used to group similar themes or subjects of documents together based on terms or keywords. [13] introduced a network analysis learning model to classify documents which have been labeled as positive and negative documents based on the subjectivity ratings. Then, the model forms communities of documents to show the similarity and relationship between them. More recently, [14] performed text summarization on Martin Luther King Jr's speech using text mining and social network analysis approach. TF-IDF algorithm and weighted closeness centrality were used to compute importance of a sentence in a social network of sentences.

In general, one of the challenges for the researchers is how to represent and signify the identified relationships of the keywords and chapters. Most of the visualization applications calculate the proximity or distance between the keywords and chapters which indicate that the further apart the chapters are in relation to distance, the less related they are to one another. Thus, in this study, we are interested in using a network analysis map to form clusters of chapters using keywords as the basic unit.

III. METHODS

Figure 1 presents the framework for Keyword and Chapter Relationship Analysis (KCRA) which consists of three modules: (i) Keyword extraction module (ii) Keyword-chapter relationship discovery module and (iii) Keyword-chapter network analysis module. The framework was demonstrated using a collection of Al-Quran chapters of Malay translated Tafseer as its text collection.

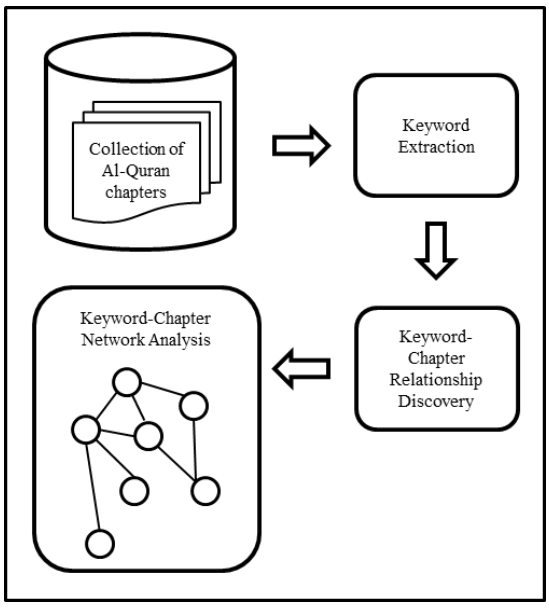

Figure 1: The framework for keyword and chapter relationship analysis (KCRA)

A. *Keyword Extraction*

In the keyword extraction module, there are two sub-modules. They are document preprocessing and keyword iden- tification as shown in Figure 2. The documents, in this case, chapters of the Tafseer texts, need to be prepared so that they are in a suitable format for further processing. In document preprocessing, these documents will first be

preprocessed by removing stop words, symbols and punctuation marks. Then, all the terms in the documents will be converted into lower case for standardization. Stop words are terms that are deemed to be unimportant and do not add value to the content of a document. These include terms like articles and prepositions. A partial list of Malay stop words obtained from the research of [15] was used in this research. This list contains 35 most frequently occurring words, which are mostly articles and prepositions. An additional word is added to the stop words list, which is "lagi", as it is a high frequency word with no semantic value to the context of this research. This stop word list of 36 words is currently sufficient for the purpose of this research. The 36 stop words are listed in Table I.

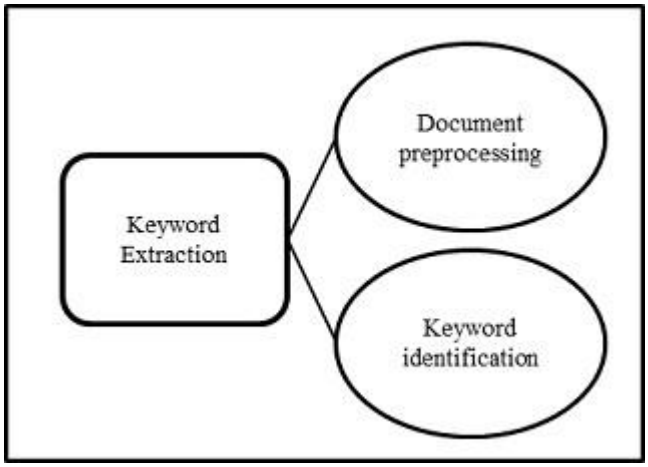

Figure 2: Sub-modules in keyword extraction

Table 1
36 Malay Stop Words

dan	kita	juga	tetapi	dunia
yang	dalam	pada	seperti	berkata
di	dari	bagi	makanan	ke
ini	untuk	pertanian	negara	daripada
dengan	halal	akan	oleh	beliau
itu	kepada	umat	rakyat	bukan
tidak	mereka	telah	ada	boleh
				lagi

After the documents are preprocessed, keywords are identified using the term-weighting statistics TF-IDF. Each term is assigned a weight that signifies its importance to the document. The higher the weight assigned, the more important the term. Therefore, the list of terms in a document can be sorted in order of importance. A subset of the most important terms can then be selected as keywords.

B. Keyword-Chapter Relationship Discovery

There are two sub-modules in the keyword-chapter relation- ship discovery module as shown in Figure 3. As discovered in the previous module, sets of keywords in each chapter are bound to be similar to sets of keywords in other relevant chapters in the Tafseer. As such, the list of terms is arranged in a keyword matrix to show which chapters contain these terms. As shown in Table II, a list of terms can be found in more than one chapter. For example, Term1 may appears n times in m of chapters. n is describing the frequency counts of the terms and $n \geq 0$. The list of x terms can then be selected based on their weightings or according to a theme of interest. The maximum number of chapters is $1 \leq m \leq 114$.

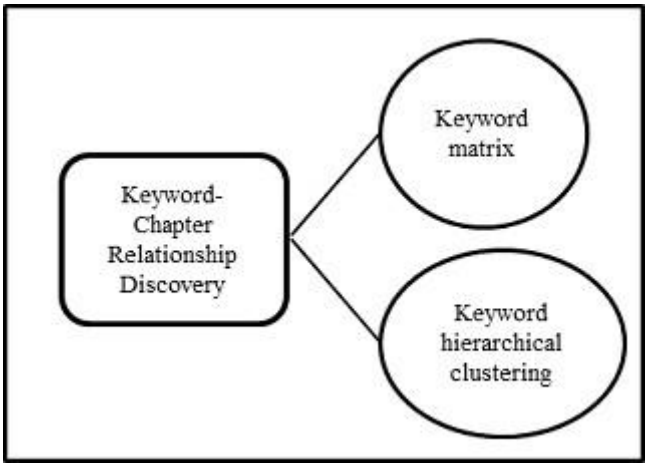

Figure 3: Sub-modules in keyword-chapter relationship discovery

Table 2
Keyword-Document Matrix

Keyword	Chapter (Surah)					
	S_1	S_2	S_3	S_4	…	S_m
Term 1	n	n	n	n	…	n
Term 2	n	n	n	n	…	n
Term 3	n	n	n	n	…	n
…	…	…	…	…	…	…
Term x	n	n	n	n	…	n

From the keyword-document matrix, relationships between keyword and chapter can be discovered. Keyword hierarchical clustering is established based on the hierarchical clustering technique to form groups of chapters in relation to keywords that link them together.

C. Keyword-Chapter Network Analysis

From Figure 1, it is shown that the end result of the KCRA framework will be shown in a network analysis map. Many studies are conducted on the multitude of document relation-ships such as citation and bibliography relationships [16, 17]. These previous studies serve as the motivation to identify multiple connections between keywords and chapters in the Tafseer. The keyword-chapter relationship discovery analysis of the Tafseer is therefore illustrated in a network analysis map so that the importance and richness of the keyword-chapter relationships can be identified as communities of related keywords and chapters.

The relationships of keyword and chapter are normally based on similarities of theme or subjects. In this module, the keywords and chapters are clustered using the hierarchical clustering technique. A hierarchical clustering mechanism, founded on the Newman method [18] for identifying clusters in network data, was applied. Newman proceeds in the standard iterative manner on which hierarchical clustering algorithms are founded. The process starts with a number of clusters equivalent to the number of nodes. The two clusters (nodes) with the greatest similarity are then combined to form a merged cluster. The process continues until a best cluster configuration is arrived at or all nodes are merged into a single cluster.

IV. RESULTS AND DISCUSSION

A. Keyword Extraction

Six short chapters are selected for this experiment: (i) Surah 109 Al Kafirun, (ii) Surah 110 An Nashr, (iii) Surah 111 Al Lahab, (iv) Surah 112 Al Ikhlas, (v) Surah 113 Al Falaq and

(vi) Surah 114 An Nas. These chapters are very short. The longest chapter out of the six is Al Kafirun, which contains six verses. The chapters are prepared in plain text document which is preprocessed initially to remove the stop words, symbols and punctuation marks in the paragraphs.

After preprocessing the chapter sets, the terms left are extracted as keywords from these chapters. There are 130 keywords which appeared at least once in any one of the selected chapters. Table 3 provides a part sample of the keywords extracted from the six chapters and the frequencies of their occurrences. It can be observed that each keyword appeared in more than one chapter. Note that however, these chapters may not necessarily have any relationship (similarity) in terms of subject of discussion even though they use the same keywords in their paragraphs.

Table 3
Keyword-Chapter Matrix (Term Frequency)

Keyword	Chapter (Surah)					
	109	110	111	112	113	114
aku	4	0	0	0	1	1
Allah	1	2	0	3	1	1
apabila	0	1	0	0	2	0
berlindung	0	0	0	0	1	1
kejahatan	0	0	0	0	2	1
makhluk	0	0	0	1	5	0
…	…	...	...	...	...	...
Muhammad	1	1	0	1	1	1
orang	4	0	0	0	1	0

B. Keyword-Chapter Relationshop Discovery

From Table 3, the frequency of each keyword in each chapter is shown. The distribution of the co-occurrence of each keyword in each chapter can also be viewed. However, in interpreting the semantic relation between keywords, meaningful phrases may need to be formed. The use of phrases will be investigated as part of the future works for the KCRA framework.

The TF-IDF weighting is then calculated and assigned to each of the keyword. A heavier weighting, noted by a bigger value, shows that a particular keyword is more importantly related to a chapter. Table 4 shows the part sample of the same keywords and their TF-IDF values.

Table 4
Keyword-Chapter Matrix (TF-IDF)

Keyword	Chapter (Surah)					
	109	110	111	112	113	114
aku	0.03	0	0	0	0.028	0.029
Allah	0.007	0.015	0	0.029	0.007	0.008
apabila	0	0.044	0	0	0.088	0
berlindung	0	0	0	0	0.044	0.045
kejahatan	0	0	0	0	0.088	0.045
makhluk	0	0	0	0.059	0.220	0
…	…	...	...	...	...	...
Muhammad	0.007	0.007	0	0.010	0.007	0.008
orang	0.163	0	0	0	0.044	0

C. Keyword-Chapter Network Analysis

The keyword-chapter matrix can be a very long list and thus, is difficult to view the connection between keyword and chapter in the Tafseer of the Al-Quran. The matrix is therefore converted into the network analysis map to illustrate the islands of keywords and chapters that are linked to one another.

Figure 4 shows the network of the six chapters with all its keywords. Note that the keywords extracted from Surah 111 Al Lahab however, were not linked to any of the other keywords linked to the other five chapters, and thus, is an island by itself. This is because the subject domain of Al Lahab does not share the same set of keywords as the rest of the five chapters.

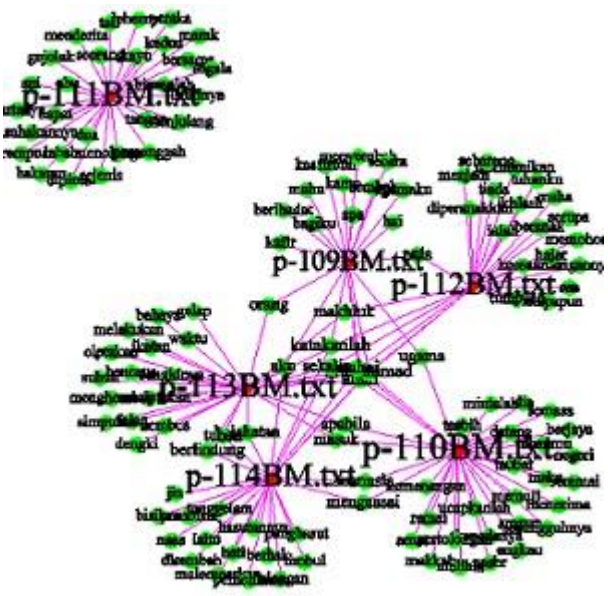

Figure 4: Network analysis map (all keywords)

It is assumed that if a large number of keywords and chapters were processed, more islands or communities of keywords and chapters could be identified. These communities represent the relationship or connections between keywords and chapters of the Tafseer of the Al-Quran in relation to a certain subject content matter or theme. An island of keywords would signify that the particular chapter is a subject or theme of its own, having its own set of keywords. A cluster of keywords that linked chapters together would signify that those chapters share similar subject content matter or theme.

Figure 5 shows the network of the six chapters for 17 keywords. A smaller set of keywords here enable a closer inspection of the network. It can be seen that these keywords link to more than one chapter. For example, “kejahatan” appeared in two chapters, Al Falaq and An Nas. Some of the keywords such as “Allah”, “berlindung” and “menguasai” are connected to three chapters, namely, An Nashr, Al Falaq and An Nas. The same group of keywords that co-occurred in these chapters indicated that they are related to a similar subject content matter or theme.

V. CONCLUSION

In this paper, a combination of text mining and network analysis approach is introduced to extract keywords and identify relationships between keywords and chapters. The KCRA framework is proposed, which consists of (i) Keyword extraction module (ii) Keyword-chapter relationship discovery module and (iii) Keyword-chapter network analysis module. The applicability of this approach is demonstrated using the Malay translated Tafseer of the Al-Quran. The novelty of this approach is shown in the discovery of the relationships between verses and chapters at keyword

level that is useful for finding and grouping related chapters together based on shared subject content matter or theme. This approach can contribute to applications such as document clustering and text summarization. For future works, the research will be extended by including the use of phrases. The use of phrases can preserve the semantic relations of individual keywords to give a greater representation of the subject matter, meaning of keywords or theme in the chapters.

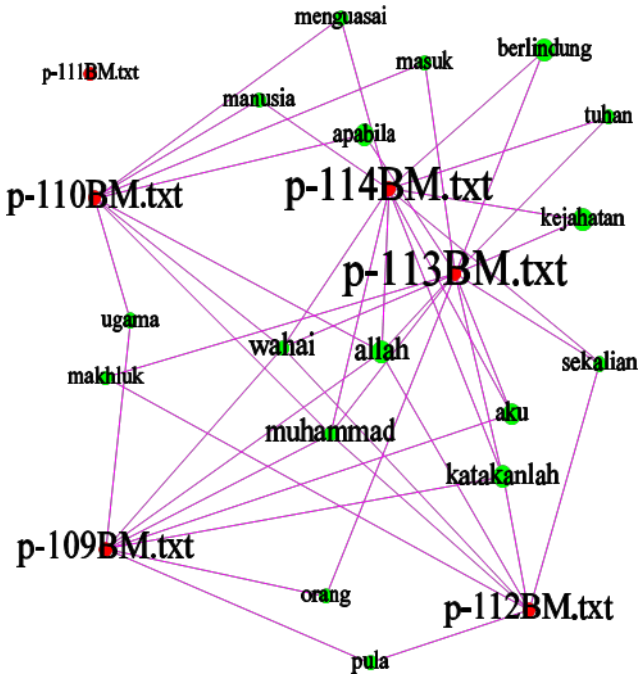

Figure 5: Network analysis map (17 keywords)

ACKNOWLEDGMENT

The authors would like to thank Universiti Malaysia Sarawak (UNIMAS) for the financial support in this research. This research was supported by the Research Acculturation Grant Scheme (RAGS/ICT07(2)/1049/2013(16)).

REFERENCES

[1] M. Hegazi, A. Hilal and M. Alhawarat. Processing the Text of the Holy Quran: A Text Mining Study. International Journal of Advanced Computer Science and Applications, 6(2), 262-267, 2015.

[2] N. M. Abdelaal, and S. M. Rashid. Grammar-Related Semantic Losses in the Translation of the Holy Quran, with Special Reference to Surah Al Aaraf (The Heights). SAGE Open, 6(3), 2016.

[3] M. Abdelwali. The Loss in Translation of the Qurn, in Translation Journal: Religious Translation, 11(2), 2007.

[4] A. Ali, M. A. Brakhw, M. Z. F. Nordin, and S. F. S. Ismail. Some Linguistic Difficulties in Translating the Holy Quran from Arabic into English. International Journal of Social Science and Humanity, 2(6), 2012.

[5] J. Han, M. Kamber and J. Pei. Data Mining: Concepts and Techniques (3rd ed.) Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2011.

[6] S. Saeed. Thematic Tafsir Methodology. url: http://quranica.com/thematic-tafsir-methodology/ [Accessed on 13 April 2017]

[7] M. N. Al-Kabi, M. G. Kanaan, R. Al- Shalabi, K. M. O. Nahar and B. M. Bani-Ismail. Statistical Classifier of the Holy Quran Verses (Fatiha and Yaseen Chapters). Journal of Applied Sciences, 5, 580-583, 2005.

[8] 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, 6(3), 127-134, 2012.

[9] S. Chua and P. N. E. Nohuddin. Frequent pattern extraction in the Tafseer of Al-Quran. In Proceedings of the Information and Communication Technology for The Muslim World, 1-5, 2014.

[10] G. S. Hassan, S. K. Mohammad and F. M. Alwan. Categorization of Holy Quran-Tafseer using K-Nearest Neighbor Algorithm. International Journal of Computer Applications11/2015, 129(12), 1-6, 2015.

[11] R. Mohamed, M. Ragab, H. Abdelnasser, N. M. El-Makky and M. Torki. AlBayan: A knowledge-based system for Arabic answer selection. In Proceedings of the 9th International Workshop on Semantic Evaluation, Colorado, USA, 2015.

[12] F. Sebastiani. Machine learning in automated text categorization. ACM Computing Surveys, 4(1), 1-47, 2002.

[13] M. Kim, B. Zhang and J. Lee. Subjective Document Classification Using Network Analysis. In Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining, IEEE Computer Society, Washington, DC, USA, 365-369, 2010.

[14] S. G. Cho and S. B. Kim. Summarization of documents by finding key sentences based on social network analysis. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer-Verlag, 285292, 2015.

[15] F. Sidi, M. A. Jabar, M. N. Selamat, A. A. A. Ghani, M. N. Sulaiman and S. Baharom. Malay Interrogative Knowledge Corpus. American Journal of Economics and Business Administration, 3(1), 171-176, 2011.

[16] L. Egghe and R. Rousseau. Co-citation, bibliographic coupling and a characterization of lattice citation networks. Scientometrics: An Interna- tional Journal for All Quantitative Aspects of the Science of Science and Science policy, 55(3), 349-361, 2002.

[17] M. Eto. Spread co-citation relationship as a measure for document retrieval. In Proceedings of the fifth ACM workshop on research advances in large digital book repositories and complementary media, ACM, New York, USA, 7-8, 2012.

[18] M. E. J. Newman. Fast Algorithms for Detecting Community Structure in Networks. Phys. Rev. E 69, 066113, 1-5, 2004.