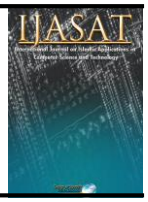

Keyword based Clustering Technique for Collections of Hadith Chapters

Puteri N. E, Nohuddin[1, a], Zuraini Zainol[2, b], Kuan Fook Chao[2, c], A. Imran Nordin[1, d], and M. Tarhamizwan A. H. James[2, e]

[1]Institute of Visual Informatics, National University of Malaysia, 46300 Bangi, Malaysia
[2]Department of Computer Science, Faculty of Defence Science and Technology, National Defense University of Malaysia, Sungai Besi Camp 57000 Kuala Lumpur, Malaysia
[a]puteri.ivi@ukm.edu.my, [b]zuraini@upnm.edu.my, [c]chao6923@gmail.com, , [d]aliimran@ukm.edu.my, [e]tarhamizwan123@outlook.com

Abstract

Hadith chapters are collections of the narrations that quote of what Prophet Muhammad (pbuh) said and preached on Islamic way of living based on the Al Quran. It covers various subjects that concern us as human beings, including wisdom, doctrine, worship and the law especially on the subject of the relationship between Allah and His creatures. In a broader application, hadith chapters are also interpreted with the deeds and acts of Prophet Muhammad (pbuh) and also reports about his companions which the prophet agreed upon. All these three categories are known as the Sunnah. This research investigates the relationships between words in the hadith chapters at the keyword level using a combination technique of text mining and Self Organizing Maps (SOM) cluster analysis to discover frequency of keywords occurred in Hadith chapters and its similarities between different hadith chapters. In this study, we used the hadith documents which were translated into English. The pre-processing steps are necessary in order to eliminate noise and to only keep the useful words. This is an effective and efficient method for Hadith chapters document clustering. The result shows the discovery of the relationships between keywords in the hadith chapters and their relevance. This may give benefits to the Muslims and Islamic scholars to make full use of the Hadith and Sunnah in their daily and also formal practices.

***Keywords*:** Hadith, Clustering, Self -Organizing Maps, SOM, Text Mining and keywords.

1. Introduction

A large number of Islamic documents have been documented and published for Muslims to be referred to and applied in their daily activities. Several researches have been conducted to introduce and provide techniques in processing these documents that can facilitate users to group and identify the documents according to their subject themes. Text mining and clustering techniques amongst others are known as popular methods used to divide a set of data into groups of similar objects. In this study, a set of Hadith Chapters are referred as document objects and the main objective of this study is to investigate the relationships between words in hadith chapters and their relevance at the keyword level using a combination of text mining and Self Organizing Maps (SOM) clustering techniques. These techniques are used to determine the frequency of keywords occurred in the content of Hadith chapters and their similarities. Whereas, SOM is used to visualize the clustered results based on their similarities on the Hadith chapters.

The paper is organized as follows: background and related work on Hadith Chapters are presented in Section 2. Section 3 presents the framework for keyword based clustering technique, while Section 4 demonstrates the data clustering analysis using a set of Hadith Chapters. Finally, in Section 5, the paper is concluded with a brief summary and future research work.

2. Background and Related Work

2.1 Hadith

One of the fundamental sources of Islamic references and guidance for the Muslims after the Holy Book, Al-Quran is Hadith. Hadith, in Arabic, refers to a report, statement, act, story, narration or discourse. Hadith chapters are collection of the narrations that quote of what Prophet Muhammad (pbuh) said and preached on Islamic way of living based on the Al Quran by reliable narrators (Rahman *et al.*, 2010). In a broader application, Hadith chapters are also interpreted with the deeds and acts of Prophet Muhammad (pbuh) and also reports about his companions which the prophet agreed upon. All these three categories are known as the Sunnah. There are several aspects of human life covered in Hadith chapters such as spiritual, economics, politics, social, judicial aspects of the life of Muslims, etc. Besides, Hadith Chapters cover the administration of inheritance of wealth of deceased Muslims, the settlement of marriage squabbles, divorce as well as paternity of the child. It also covers areas of Islamic activities such as naming of a child, the burial rites as well as settlement of disputes between two communities (Musa *et al*., 2012)

Other than the Al-Quran, Hadith is also one of the main sources of Islamic religious law or Shariah Law for the Muslims. Each hadith consists of two major aspects, a chain of narrators (transmitters) reporting the hadith (the isnad), and the text of the hadith (the mat'n). The early Muslims scholars classified the Hadith based on the degree of authenticity and reliability where each category had to meet certain criteria. The categories of Hadith are: (i)[1] Sahih refers to "correct", "true", "valid" or "sound". In other words, this category is genuine and has passed all tests; (ii) Moothaq almost like the Sahih but the narration is not as strong as sahih; (iii) Hassan: the fair traditions but not as authentic; (iv) Dha'eef is basically a weak hadith which makes it unreliable and acceptable. The two most highly respected collections of Hadith are Sahih Bukhari and Sahih Muslim while the other four collections are the Sunan of Tirmidhi, Nasa'i, Ibn Majah and Abu Da'ud. These four collections and two sahih collections are then formed as a "six books" or Al-Kutub al-Sitta.

2.2 Text Mining

Text mining (in Computer Science) is a technique used to discover new information by automatically extracting it from different written resources (V. Gupta & Lehal, 2009). The key element of this technique is the extracted information is linked together to form new facts or new hypotheses that can be explored further using conventional means of experiments. Text mining differs from web search or directory search. In the web search, people are usually looking for what is already known or has been written by someone. In comparison, the goal of the text mining is to discover something that nobody knows or not yet written down. Text mining can be identified either as intelligent text analysis, text data mining or knowledge discovery in text (KDT). Text mining can be performed with the unstructured or semi structured data set. It mines information and knowledge from a mountain of texts.

[1] www.al-islam.org

2.3 Clustering

One of the data mining techniques that have been widely applied in text mining is document clustering. Clustering is an unsupervised learning method that is aimed to group a set of data into a meaningful subclasses or clusters (Akbar, 2008; Ding & Fu, 2012). It can also be used to discover the structure and the contents of the unknown texts. Clustering of documents has been used in many applications such as in information retrieval systems, assisting users on web sites, and personalization of search engine results amongst others (Gharib *et al.*, 2012). Document clustering groups a set of documents into a number of clusters where each cluster contains documents with similar topics based on its similarity measures. Generally, there are two (2) approaches of document clustering namely partitional and hierarchical. The former constructs various partitions and then evaluates them by some criterion. The most common algorithms used in partitional methods are k-means and k-medoids. Whilst the later, is a set of nested clusters that creates a cluster hierarchy (a tree of clusters), also known as dendrogram (Berkhin, 2006). The hierarchical clustering methods can be categorized into agglomerative (bottom-up) and divisive (top-down). A number of researches have been conducted to explore the implementation of text clustering algorithms in text documents, for example, SOM (Bação *et al.*, 2005; Bakus *et al.*, 2002), k-means (Ahmed & Tiun, 2013; H. Gupta & Srivastava, 2014), Fuzzy c-means, quality threshold (QT), etc. Despite all of these, SOM has multiple benefits where it can visualize the results immediately (Chumwatana *et al.*, 2010). This makes the process of understanding the documents easier in meaningful way.

2.4 Self-Organizing Maps (SOM)

SOM is unsupervised learning rules type of Artificial Neural Network (ANN) that is capable of visualizing important relationships among the data which are hidden in the input set of data (Kohonen, 1998). It visualizes the data set in the simplest representation form (Chifu & Cenan, 2004). It clarifies how to classify input vectors according to how they are grouped in the input space. The weight vectors are firstly initialized using a random number generator. Then, the algorithm processes the input data through a record by a record. The smallest value of the distance in each record will be assigned as the "winning" node. This will be defined as a mapping from the input vector onto a two-dimensional array of nodes. SOM is also used to map the high dimensional space into some low dimensional space. In addition, SOM operates in two ways: training and mapping. Training builds the map using input example, whereas mapping automatically classifies a new input vector (Roy & Bandyopadhyay, 2014).

2.5 Related Work

Gharib *et al*., (2012), proposed a semantic text documents based on WordNet lexical categories and applied SOM to cluster the documents. In their experiments, they used three different clustering algorithms namely k-means, SOM and bisecting k-means. The results suggests that SOM obtains the highest clustering quality compared to the other two algorithms. Similarly, Chumwatana *et al.*, (2010) found that a non-segmented document clustering method using SOM has improved the efficiency of information retrieval. Their data sets consist of 50 Thai documents which are related to several categories: sport, travel, education and political news, etc. The results showed that SOM is capable to cluster all of the documents into different clusters accurately. In addition, Isa *et al*., (2009) introduced a hybrid text document classification approach by integrating the naïve Bayes classification method using SOM to cluster the vectorised text documents. These data sets were collected from four (4) categories of vehicles. The results indicated that this new approach has improved the performance of the document classification.

3. The Framework For Keyword based Clustering Technique

Figure 1 illustrates the Framework for Keyword-based Document Clustering (FKDC). It consists of the two main stages: (i) Word Pattern Analysis and (ii) SOM Document Clustering Technique. The first component consists of the pre-processing task while the second component is mainly focused on SOM. The materials (input) being processed in this framework was a collection from English translated Hadith Chapters.

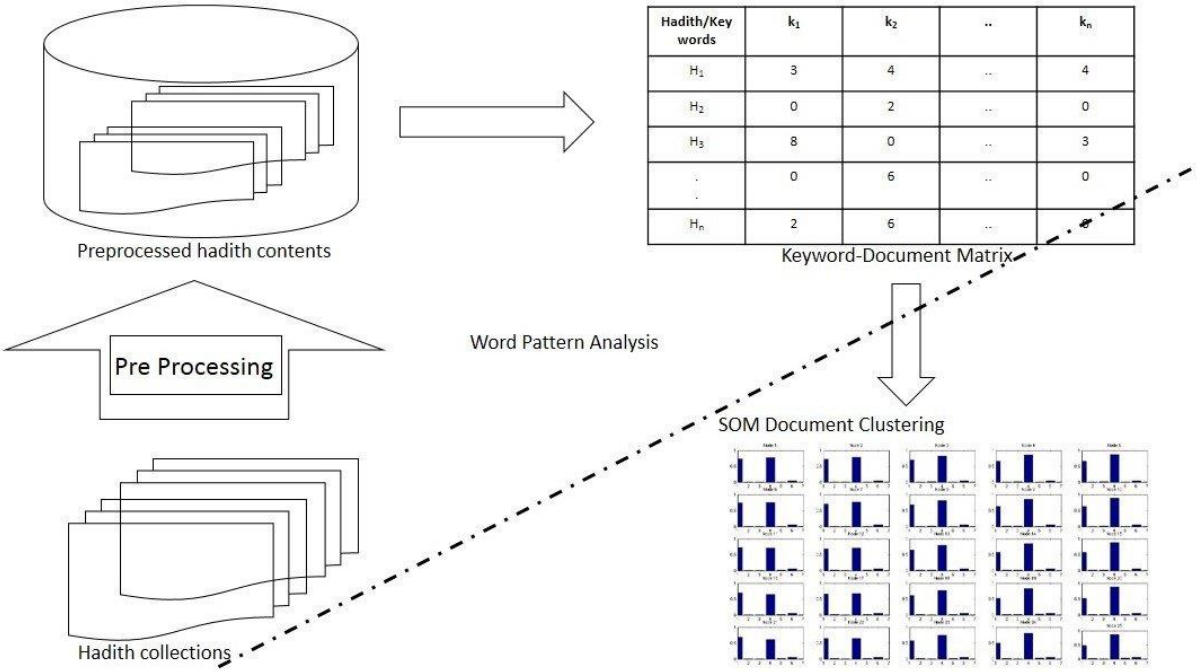

Figure. 1 The Framework for Keyword-based Document Clustering

3.1 Word Pattern Analysis

In the first stage of FKDC, there are two sub-modules that process a collection of the text of the Hadith chapters. As shown in Figure 2, the two sub-modules are document pre-processing and keyword selection. Each hadith needs to be pre-processed in a suitable format for further processing.

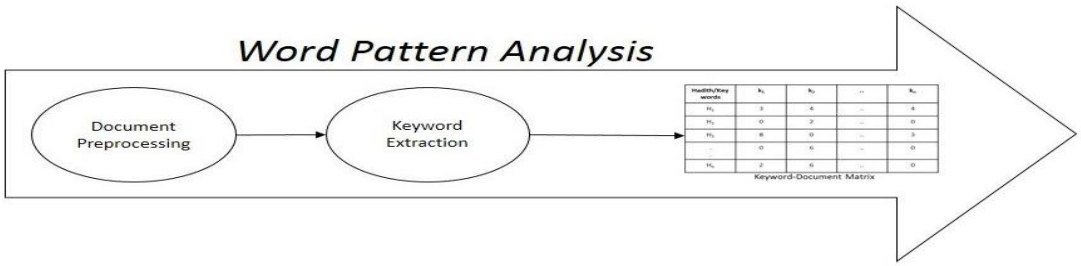

Figure. 2 Sub-modules in the Word Pattern Analysis

During document pre-processing, a set of Hadith chapters are processed to remove some stop words (e.g. the, you, is, are, etc.), symbols and punctuations. Stop words are deemed to be unnecessary in the hadith contents because these terms do not add important meaning to the contents. Moreover, during the preprocessing stage, a stemming tool is used to remove the prefix and suffix of Stemming which is the basic text processing procedure for English text. The goal is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.

Then, all the "cleaned" hadith contents are compiled in one data mart of documents for keywords extraction. During the process of keywords extraction, the frequency of each word in every chapter is recorded. Only those keywords with frequency count above a minimum *threshold*[2] are considered as "important" keywords. A 2-D matrix is constructed to present the result of keyword extraction from the hadith chapters. The matrix *n* x *m* consists of a list of distinct keywords as the column and hadith chapters as the row (see Table 1). The matrix records the frequency of keywords appeared in the hadith chapters. In the matrix, h_n represents the Hadith chapters and k_n denotes the extracted keywords. As the number of Hadith chapters increases in the compilation of documents, the chapter-keyword matrix may grow into a very long list as many distinct keywords are extracted from the Hadith chapters. Thus, it is difficult to view and analyze the connection between chapters and keywords. Therefore, a clustering technique is proposed to group similar content of Hadith chapters based on the chapter-keyword matrix.

Table 1: Chapter–Keyword Matrix

	k_1	k_2	k_3	k_4	k_5	k_6	k_7	…..	k_n
h_1	41	3	7	3	3	2	5	…..	0
h_2	6	4	5	0	0	1	4	…..	2
h_3	8	1	4	2	0	3	2	…..	1
h_4	6	1	3	1	0	0	0	…..	0
:	:	:	:	:	:	:	:	…..	:
h_n	10	1	0	0	0	0	0	…..	5

3.2 Cluster analysis of word pattern

The final stage of FKDC is the SOM document clustering. This stage aids the analysis of a long list of keywords gathered from the preprocessing stage. The clustering technique groups hadith chapters that consist of similar keywords. Hence, from those extracted keywords, it is possible to determine the topic of the grouped chapters (in a cluster). The topic of the Hadith Chapters that are being discussed could be similar but its sources could be from a single narrator or by multiple narrators. This clustering technique ensures that the number of cluster SOM is pre-set with a number of possible clusters could be generated. A number of experiments are performed to determine the best number of clusters. In this study, experiments of SOM size with 6 x 6 is selected to have a fine distribution of hadith chapters in 36 clusters. The bigger size of SOM leads to more refine distribution of the cluster members. The map was then trained and the remaining examples assigned to nodes using a distance function.

The SOM document clustering produced two (2) maps to represent (i) a cluster prototype map identified from the chapter-keyword matrix and (ii) features cluster map using bar charts. From the prototype map, a variation of bar charts is illustrated to show keywords combination for each cluster. Hadith with similar keyword combinations are grouped into the same cluster map. In the second map, it illustrates the details of each cluster and the number of graphs included in each cluster. Nohuddin *et al.* (2011) described further information about SOM clustering technique.

[2] Threshold is a point that must be exceeded to be significant.

4. Experimental Results and Analysis

This section provides the results from the experiment done using FKDC. A set of 40 Hadith Chapters are used as an input for this experiment. These Hadith Chapters are narrated by Bukhari, Muslim, Da'ud and Malik on Hajj. Firstly, Word Pattern Analysis modules are applied to all Hadith from pre-processing stage to keyword extraction process. A minimum threshold of 2 is specified for extracting keywords which means that words which appear more than 2 times in each hadith chapter are deemed to possible significant keywords. For this paper, only 26 keywords are selected from the set of extracted keywords (see Table 2).

Table 2: Parts of extracted keywords with column labels

Column labels	$\mathbf{k_1}$	$\mathbf{k_2}$	$\mathbf{k_3}$	$\mathbf{k_4}$	$\mathbf{k_5}$	$\mathbf{k_6}$	$\mathbf{k_7}$	$\mathbf{k_8}$	$\mathbf{k_9}$	$\mathbf{k_{10}}$	$\mathbf{k_{11}}$	$\mathbf{k_{12}}$	$\mathbf{k_{13}}$
Keywords	Allah	apostle	hajj	camel	cloth	come	command	companion	day	enter	face	father	hand
Column labels	k_{14}	k_{15}	k_{16}	k_{17}	k_{18}	k_{19}	k_{20}	k_{21}	k_{22}	k_{23}	k_{24}	k_{25}	k_{26}
Keywords	head	harm	ihram	ka'ba	marwa	messenger	mecca	perform	prophet	safa	said	tawaf	umra

Then, the chapter-keyword matrix (as shown in Table 1) becomes an input for the SOM document clustering module. The module produces two maps: (i) SOM cluster prototype map and (ii) Features cluster map. Figure 3 shows the SOM cluster map (clusters) which consists of 36 sub-nodes. Each sub-node has a bar chart that describes keyword combination for the cluster. For example node 3 is a cluster that has keywords 1, 2, 3 (high), 17, 21, 22, 23, 25 and 26 combinations. Majority of clusters have similar keywords combination however the height variations of bar charts indicate the frequency counts of keywords. Another example, node 6 has keywords 1, 2, 3, 6, 16, 17, 18, 21, 22, 23, 24, 25 and 26 and these keywords have a frequency count higher than the keyword combination in node 3.

Figure 4 illustrates a Feature Cluster Map which contains details of number of Hadith Chapters in each cluster and also frequency counts of the keywords. The figure shows that majority of sub-nodes consist of 1 Hadith Chapter with distinct keyword combination, this is due to the small number of chapters used as an input. However, there are clusters that consist of 2 or more Hadith Chapters such as node 6 which has 2 Hadith Chapters with keywords 1, 2, 3, 6, 16, 17, 18, 21, 22, 23, 24, 25 and 26 combination and node 12 contains 3 Hadith Chapters that has "high" count on keywords 1, 2, 3, 6, 8, 9, 16, 17, 18, 21, 22, 23, 24, 25 and 26 combinations.

As a result, we are able to interpret the set of extracted keywords describe the activities during Hajj. Moreover, SOM maps illustrate the numbers of Hadith chapters collected in each cluster based on the chapters' content similarity. For example, we can conclude that node 12 is a cluster with keywords describing tawaf between Marwa and Safa during umra and it contains 3 similar Hadith chapters

5. Conclusion

This paper describes a combination of text mining and SOM clustering approaches for extracting significant keywords and clustering Hadith chapters that contain similar keyword combinations. The proposed FKDC consists of 2 main modules (i) Word Pattern analysis that integrates content preprocessing and keyword extraction sub-modules and (ii) SOM document clustering that groups hadith chapters with similar keyword combinations. The proposed FKDC successfully extracts keywords from multiple hadith chapters and forms a chapter-keyword matrix. The matrix comprises the distribution of keyword co-occurrences in a set of Hadith Chapters. Then, the SOM clustering maps are presented to illustrate cluster types of keyword combinations and also detailed bar chart maps for clusters' details.

In future, the work will be extended by using bigger set of Hadith Chapters and use of keywords as phrases. Phrases are able to interpret meaning in the content of Hadith Chapters or theme of the chapters more effectively.

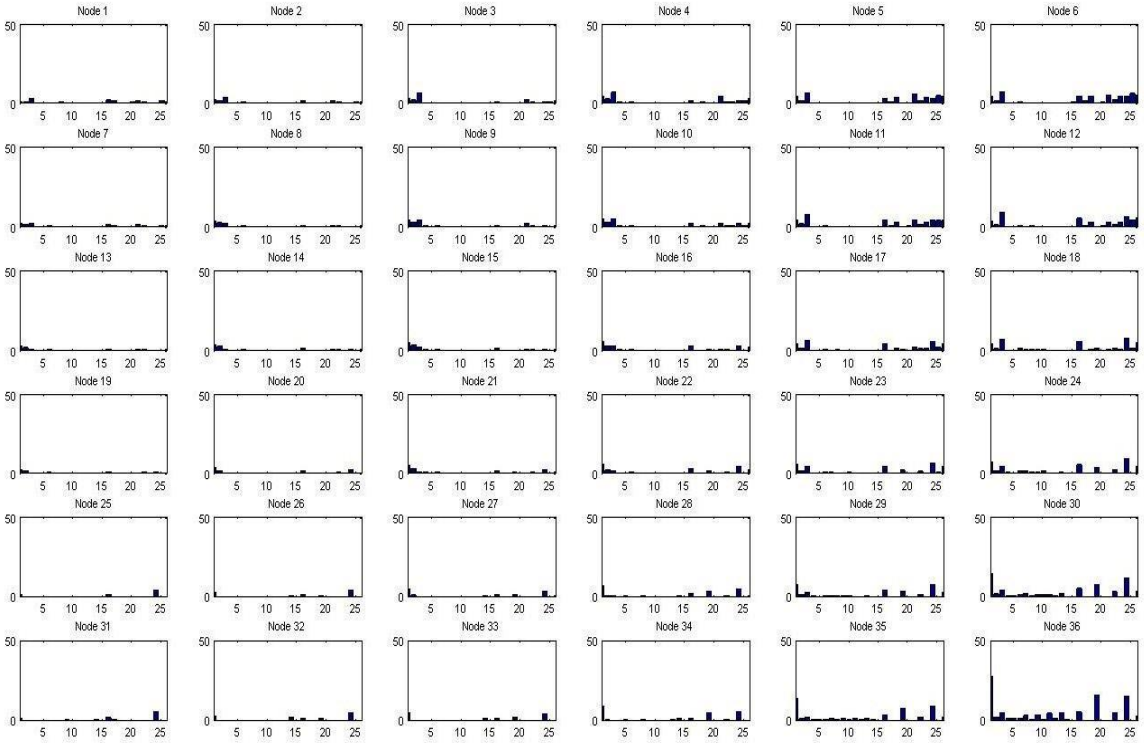

Figure. 3 SOM Prototype Map

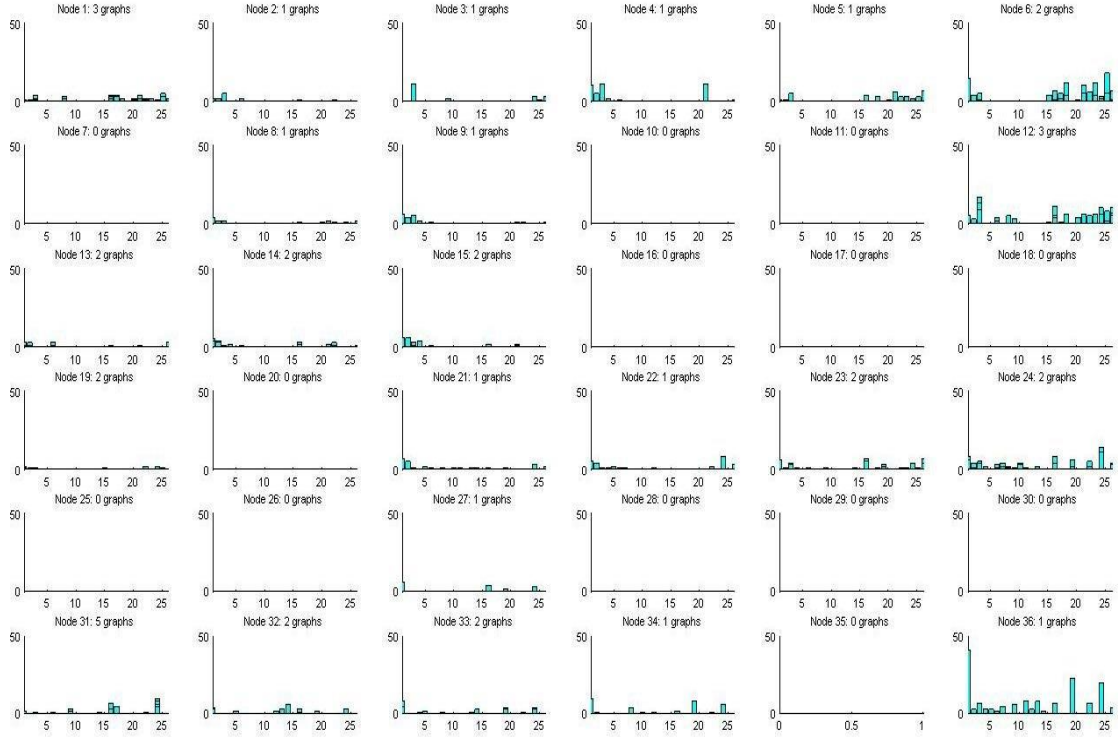

Figure. 4 Features Cluster Maps

Acknowledgment
The authors wish to thank to sponsors, Universiti Pertahanan Nasional Malaysia (UPNM) and Kementerian Pendidikan Malaysia (KPM) for providing the RACE research grant.

References

Ahmed, M. H., & Tiun, S. (2013). K-means based algorithm for islamic document clustering. Paper presented at the IMAN 2013.

Akbar, M. (2008). FP-growth approach for document clustering. Montana State University-Bozeman, College of Engineering.

Bação, F., Lobo, V., & Painho, M. (2005). Self-organizing maps as substitutes for k-means clustering Computational Science–ICCS 2005 (pp. 476-483): Springer.

Bakus, J., Hussin, M., & Kamel, M. (2002). A SOM-based document clustering using phrases. Paper presented at the Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on.

Berkhin, P. (2006). A survey of clustering data mining techniques Grouping multidimensional data (pp. 25-71): Springer.

Chifu, E. S., & Cenan, C. (2004). Discovering Web Document Clusters with Self-Organizing Maps. Sci. Ann. Cuza Univ., 15, 38-47.

Chumwatana, T., Wong, K. W., & Xie, H. (2010). A SOM-Based Document Clustering Using Frequent Max Substrings for Non-Segmented Texts. Journal of Intelligent Learning Systems and Applications, 2(03), 117.

Ding, Y., & Fu, X. (2012). Topical Concept Based Text Clustering Method. Paper presented at the Advanced Materials Research.

Gharib, T. F., Fouad, M. M., Mashat, A., & Bidawi, I. (2012). Self organizing map-based document clustering using wordnet ontologies. IJCSI International Journal of Computer Science Issues, 9(1), 1694-0814.

Gupta, H., & Srivastava, R. (2014). K-means Based Document Clustering with Automatic "k" Selection and Cluster Refinement.

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. Journal of emerging technologies in web intelligence, 1(1), 60-76.

Isa, D., Kallimani, V., & Lee, L. H. (2009). Using the self organizing map for clustering of text documents. Expert Systems with Applications, 36(5), 9584-9591.

Kohonen, T. (1998). The self-organizing map. Neurocomputing, 21(1), 1-6.

Musa, S. A., Ahmed, A. F., & Mustapha, A. R. (2012) Studies on the Hadith: National Open University of Nigeria.

Nohuddin, P.N.E., Coenen, F., Christley, R., Setzkorn, C., Patel, Y. And Williams, S. (2011). Finding "Interesting" Trends in Social Networks Using Frequent Pattern Mining and Self Organizing Maps. Knowledge-Based Systems

Rahman, N. A., Bakar, Z. A., & Sembok, T. M. T. (2010). Query expansion using thesaurus in improving Malay Hadith retrieval system. Paper presented at the Information Technology (itsim), 2010 International Symposium in.

Roy, M. S., & Bandyopadhyay, S. K. (2014). Gender recognition using Self Organizing Map (SOM)-an unsupervised ANN approach.

Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., Shen, B. (2013). Biomedical text mining and its applications in cancer research. Journal of biomedical informatics, 46(2), 200-211.