Asian Journal of Information Technology 15(24): 5159-5162, 2016
ISSN: 1682-3925
© Medwell Journals, 2016

Clustering the Verses of the Holy Qur'an using K-Means Algorithm

[1]Cepy Slamet, [1]Ali Rahman, [1]Muhammad Ali Ramdhani, and [2]Wahyudin Darmalaksana
[1]Department of Informatics Engineering, UIN Sunan Gunung Djati Bandung, Indonesia
[2]Department of Hadits Science, UIN Sunan Gunung Djati Bandung, Indonesia

Abstract: The Holy Qur'an is a basic living guidance for Muslims. The depth of the sea of knowledge in the Holy Qur'an gives its own attractiveness to many researcher to conduct exploration involving automated application. This article provides a practical works of text mining applied as an initial route which begin with some Qur'anic structures phenomenon by clustering the verses. K-means algorithm has been applied to clustering experiment in a framework of text mining. This study resulted the total of 6236 verses (data corpus), using unsteamed and steamed words which then establish three clusters.

Key Words: The Holy Qur'an, data mining, text mining, clustering, k-means algorithm

INTRODUCTION

The Holy Qur'an is a basic living guidance for Muslims. It compreses all aspects of human life including Biology, Information Communication and Technology (ICT), Laws, Social, Politics, Business, Economics, Autonomy, and others (Noordin, 2013). Information technology is a system of medium and infrastructure and method to gain, transfer, access, interpret, save, organize and use the data meaningfully (Ainsisyifa, 2012 a, b). In short, al-Qur'an presents a sea of knowledge (Farooqui and Noodin, 2015).

The depth of the sea of knowledge in the Holy Qur'an gives its own attractiveness to many researcher to conduct exploration involving automated application. People in search of what Islam is and what it is not, want easier ways in which they can gain access to answers to their various queries (Yauri, et. al., 2013). In retrieving the knowledge, several research propose a lot of retrieval methods and models for extraction, build an ontology, developing new corpuses, etc. This article provides a practical works of text mining applied as an initial route which begin with some Qur'anic structures phenomenon by clustering the verses.

MATERIALS AND METHODS

Data Collection: This work uses Al-Qur'an in English translation data written by Ahmed Ali, published by http://tanzil.net as a corpus and data reference is retieved from http://corpus.quran.com which provides the data services. All the correlated data are organized in a specific directory of data storage for further processes.

Pre-Processing: Pro-processing is a second phase and it becomes main process in text mining at once. Pro-processing is conducted to the raw data to enable an excecution for the next phases. Generally, there will be including document/ words extraction (case folding, tokenizing, stopwords, and stemming) (Fig. 1).

K-Means Clustering: Integration of K-means into a system using a recommended system by (Amin & Ramdhani, 2006). K-means is one of a clustering algorithm which uses partition method. K-means is clustering algorithm that devides each data item into a cluster. The steps are as follows:

- Define a number of cluster (k) at data set;
- Define a centroid. At the first step, the centroid is defined randomly, while at the iteration uses a following formulation:

$$\nabla_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i} x_{kj} \quad (1)$$

- At each record, quantify a nearest distance to the centroid. The centroid distance used is an Euclidean Distance, by the following equation:

$$D_\theta = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} \quad (2)$$

- Group the objects based on the distance to nearest centroid; and
- Repeat the second step, and do iteration until centroid reaches optimum value.

Corresponding Author: Cepy Slamet, Department of Informatics Engineering, UIN Sunan Gunung Djati Bandung, Indonesia

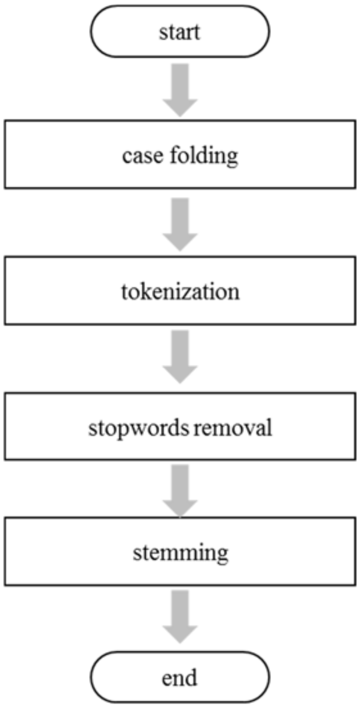


Fig. 1: Pre-processing process

Understanding the content of the Holy Qur'an requires some specific linguistic disciplines to learn. In line with the aim of a clustering which is defined as the process to find structure in data and is therefore exploratory in nature (Jain, 2010), the significance of the study is planned to deep exploring into the sea of knowledge in the Holy Qur'an through many actionable pieces of research. A knowledge portal with some determined parametrics is considered as one of the required alternatif tool to simplify the verses clustering of the Holy Qur'an. Knowledge portal is an important aspect where people enable to search knowledge (Pamoragung et al., 2006). The use of this media is a part of knowledge management concept which people are easily served to fucus on learning a required knowledge (Ainissyifa, 2012a, b).

Generally, this recent article provides a practical works applied as an initial route which begin with some Qur' anic structures phenomenon. In the structure, the topics are not gathered in a specific chapter, the similar context of topics are separated at several places (chapters or verses). Thus, in a limited perspective of data mining, the problem question is directed to, how can people easily specify the group of similar topics?

The ICT development enable to almost all data can be represented in the text form, audio, or image. One of the computer science branches is data mining. Among the wide range of data mining focuses is association rule.

One of an association rule implementation solution is aclustering data. There are a lot of research in data mining, in relation with study the content of knowledge or knowledge exploration in the Holy Qur'an, at least, recently has emerged several key features challenges for future research. These are, selection or development of correct tool with required features, extraction of knowledge, topic wise extraction of knowledge, optimal numbers of synonyms/ antonyms for Qur'an ontology matching/ development and selection of 'Tafsir' and embedding it for correct context (Noordin and Farooqui, 2015).

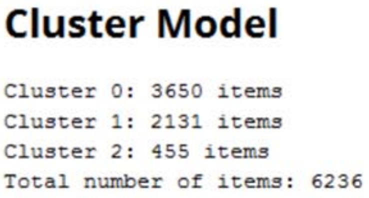
Data mining is also known as Knowledge-Discovery in Databases (KDD) is process of extracting potentially useful information from raw data. A software engine can scan large amounts of data and automatically report interesting patterns without requiring human intervention. In general, data mining has four major relationships, they are classes, clusters, associations and sequential patterns (Sankar, 2011).

Text mining is also known as Text Data Mining (TDM) and knowledge Discovery in Textual Databases (KDT). KDT plays an increasingly significant role in emerging applications, such as Text Understanding. Text mining process is same as data mining, except, the data mining tools are designed to handle structured data whereas text mining can able to handle unstructured or semi-structured data sets such as emails HTML files and full text documents, etc. (Vijayarani et al., 2015). Text Mining is used for finding the new, previously unidentified information from different written resources. There are numerous applications of text mining those can be used for knowledge extraction purpose and can be proved very useful, including text classification, opinion mining, text clustering and document summarization (Bhardwaj, 2016). Text mining entails automatically analyzing a corpus of text documents and discovering previously hidden information (Hashimi et al., 2015). In this study, Qu'anic data from http://tanzil.net is used as a corpus.

Text clustering algorithms are divided into a wide variety of different types such as: Agglomerative clustering algorithms (e.g., single linkage clustering, group-average linkage clustering and complete linkage clustering); distance-based partitioning algorithms (e.g., K-means clustering algorithm and k-mediod clustering algorithm) and standard parametric modeling based methods such as the EM-algorithm (Bhardwaj, 2016). K-means clustering algorithm is one of the popular algorithm which has gained a lot of attraction because of its simplicity and ease of implementation. K-means

Table 1: Centroid of each cluster

Attibute	cluster_0	cluster_1	cluster_2
Number of unsteamed words	12.853	33.069	68.745
Number of steamed words	5.571	14.102	29.200



Cluster Model

```
Cluster 0: 3650 items
Cluster 1: 2131 items
Cluster 2: 455 items
Total number of items: 6236
```

Fig. 2: Clustering Result

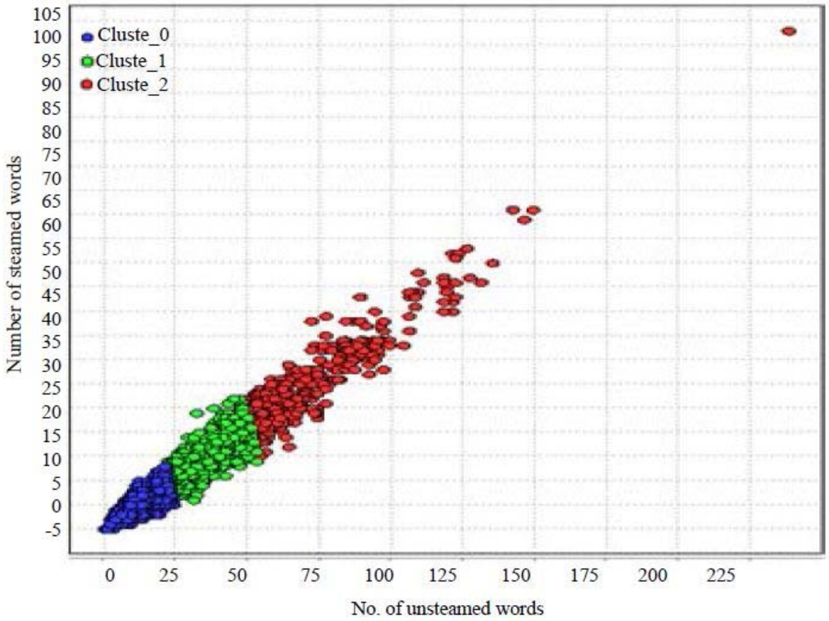


Fig. 3: Clustering with K-Means Visualization

algorithm's efficiency is limited because of random selection of k initial centers (Bhatia and Khurana, 2013). Figure 2 illustrates the result of the K-Mean algorithm where the clustering process forms three clusters.

Generally, clustering is a difference or an indifference-based sensible data grouping (unsupervised learning). Both the difference or the indifference are identified by specific type of data. A data can be categorized into a certain cluster by which it contains a similar data type.

The memberhip of cluster (Fig. 2) are grouped based on its similarities between one member and other. Analyzing a results of clustering, it is obtained that among all cluster, the highest membership is Cluster 0 (3650 items), Cluster 1 has 2131 items and the lowest is Cluster 2 has 455 items. Cluster 0 consists of verses by the amout of less then or equal to 24 unsteamed words and 13 steamed words. Meanwhile, Cluster 1 is a cluster with amount verses between 22 to 53 unsteamed words and 6 to 27 steamed words. And Cluster 2 is a cluster with lowest membership of verses which has 49 to 238 unsteamed words and 15 to 103 steamed words, these are also visualized by Fig. 3.

Table 1. illustrates the amount of centroid of every cluster categorized by numbers of unsteamed words and steamed words. Overall, the amount of two attribute within each cluster is more than two times that of the other. The lowest centroid is on cluster_0 in number of steamed words. As for these centroid, is a central value where the distance of every verse in the Holy Qur'an is calculated with each existing centroid, more closer a verse to the centroid then it includes into the cluster.

CONCLUSION

Beside requiring specific linguistic discipline, the other pathway of exploring knowledge in the Holy Qur' an can be started from comprehending its structures. This study resulted an initial practical steps in which the structures of verses in the Holy Qur' an are enable to learn. K-means algorithm has been applied to clustering experiment in a framework of text mining. The algorithm is utilized for clustering 6236 total verses, using WIBteamed and steamed words which then establish three clusters. In line with the goal of this study, the research will then continue to applying some queries regarding several thematic approaches to define more specific data/ information pattern.

REFERENCES

Ainissyifa, H., 2012a. The influence of human resources toward knowledge management implementation on secondary education institution. *Advances in Natural and Applied Sciences*, 6(6), 789-792.

Ainissyifa, H., 2012b. The influence of technology utilization toward knowledge management implementation on secondary education institution. *Journal of Applied Sciences Research*, 8(4), 2133-2136.

Amin, A. S., and Ramdhani, M. A., 2006. Konfigurasi Model Untuk Sistem Pendukung Keputusan. *Majalah Ilmiah Ekonomi Komputer*, 14(1), 11-19.

Bhardwaj, B., 2016. Text mining: Its utilities, challenges and clustering techniques. *International Journal Computer Application*, 135, 22-24.

Bhatia, M. P., and Khurana, D., 2013. Analysis of Initial Centers for k-Means Clustering Algorithm. *International Journal Computer Application*, 71, 9-12.

Farooqui, N. K., and Noordin, M. F., 2015. Knowledge exploration: Selected works on Quran ontology development. J. Theor. Appl. Inf. Technol, 72, 385-393.

Hashimi, H., Hafez, A., and Mathkour, H., 2015. Selection criteria for text mining approaches. *Computer Human Behaviour*, 51, 729-733.

Jain, A. K., 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters journal*, 31, 651-666.

Noordin, M. F., 2013. *ICT and Islam*. Kuala Lumpur, Malaysia: IIUM PRESS.

Pamoragung, A., Suryadi, K., and Ramdhani, M. A., 2006. Enhancing the Implementation of e-Government in Indonesia Through the High-Quality of Virtual Community and Knowledge Portal. 6th European Conference on e-Government (pp. 341-347). Marburg: Academic Conferences Limited.

Sankar, R., 2011. Customer Data Clustering Using Data. *International Journal of Database Management Systems*, 3, 1-11.

Vijayarani, S., Ilamathi, M. J., and Nithya, M., 2015. Preprocessing Techniques for Text Mining: An Overview. *International Journal Computer Science and Communication Network*, 5, 7-16.

Yauri, A. R., Kadir, R. A., Azman, A., Azrifah, M., and Murad, A., 2013. Quranic Verse Extraction base on Concepts using OWL-DL Ontology. *Research Journal of Applied Sciences, Engineering and Technology*, 6, 4492–4498.