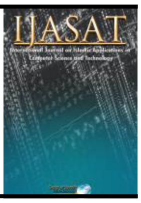

K-Means Based Algorithm For Islamic Document Clustering

Majid Hameed Ahmed[1], Sabrina Tiun[2], Mohammed Albared[3]
FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
[1]majid4000@yahoo.com, [2]sabrinatiun@ftsm.ukm.my, [3]mohammed_albared@yahoo.com

ABSTRACT

Document clustering is an unsupervised learning task. It is a form of data analysis, aims to group a set of objects into subsets or clusters. In this paper, the target domain of clustered documents is Islamic religious domain. The Islamic document clustering is considered as an important task for gaining more effective results with; the traditional information retrieval (IR) systems, organizing web text and text mining. Fast and high-quality document clustering can tremendously facilitate the user to successfully navigate, particularly on the Internet since the number of available online documents is increasing rapidly, everyday. Thus, religious domain has become an interesting and challenging area for Natural Language Processing (NLP). The aim of this paper is to evaluate the efficiency and accuracy of Arabic Islamic document clustering base on K-means algorithm with three similarity/distance measures; Cosine, Jaccard similarity and Euclidean distance. In order to implement the algorithms, we have to pre-process the data (document). The pre-processing steps are necessary in order to eliminate noise and keep only useful information so that we can boost the performance of documents clustering. Additionally, this research investigates the effect of using stemming and without stemming words on the accuracy of Arabic Islamic text clustering. Based on our experiments, we have found that the stemming process than gives better impact than without stemming process, and the K-means with Cosine similarity measure achieves the highest score of performance.

***Keywords*: Islamic document clustering, Information retrieval (IR), K-means algorithm, light stemmer and similarity/distance measures.**

1. Introduction

The world has witnessed in recent years significant and rapid growth in search for information and retrieval especially on the Internet. However, this growth has been confined only to those languages that are most prevalent, including the most dominant English language, there are some languages that share the same structure while the other languages have completely different structures. Therefore, word processing algorithms that have been developed specially to address specific language cannot be applied to other language, especially, if those languages are different in structure. One such language is Arabic language, which is the focus of this paper. Arabic is main language of Islamic document, and it is noted that nearly 1 to 30 billion of those documents are available in the internet, and of which the share of Arabic language is 1.4% (Farghaly & Shaalan 2009; Ferguson 1996). Nevertheless, there is a lack of attempts and efforts devoted to improve and develop the research and IR in the Arabic language, as against the efforts in other languages (Al-Shammari & Lin 2008). Arabic language has high derivation, where it is possible to derive a large number of words using only one root. In addition, one can derive a single word of multiple roots (Ahmed, 2000; Darwish, 2002). Unlike for the structure of forming the English word.

Clustering is a form of data analysis aimed to group a set of objects into subsets or clusters (Akbar 2008; Ding & Fu 2012; Huang 2011). The goal of clustering is to create clusters, by grouping similar data items together. Text clustering plays a vital role in many real-world applications such as, automatic IR systems, organizing web text and text mining. Fast and high-quality document clustering heavily facilitate the users to successfully navigate, summarize, and organize huge and unorganized information. It can also be used to discover the structure and content of unknown text sets. In Arabic IR, particularly with the rapid increase of the number of online documents available in Arabic language. It aims at carrying out an automatic grouping of similar documents in one cluster, by making a use of different similarity/distance measures (Froud et al. 2010). More recently, religious domain has become an interesting and challenging area for NLP and text mining (Farghaly & Shaalan 2009). In this paper, we study clustering based algorithms to cluster Islamic religious documents.

This paper is divided into four main sections: In section 2, we will explain related work on Arabic document preprocessing. Then in section 3, we will describe how we carry out our investigation. Section 4 will be on the experimental results, and lastly, section 5 will be the conclusion of our work.

2. Related Work

Arabic document clustering is the main goal in this research. In this section, we will present our review on the works related to Arabic document clustering. Most of the reviewed works are based on unsupervised machine learning approaches.

In Osama and Wesam (2012), they have evaluated stemming techniques in clustering of Arabic language documents and identified the most effective pre-processing approach for Arabic language which is more complicated than most other languages. They have used three stemming techniques: root-based stemming, light stemming, and without stemming. The data set used has been collected from BBC Arabic. The results indicate that the light stemming gets the best measurement values than without stemming and root-based stemming in Arabic document clustering.

Al-Omari (2011) has evaluated and estimated the impact of stemming in clustering algorithm. The Arabic documents pre-processing which are used in his work include: tokenization, stop word removal, and stemming function. The author used vector space model as the algorithm for clustering. The best result achieved was without stemming, and thus, it is evident that the results without stemming are better than with stemming.

Froud et al. (2010) have evaluated the impact of the stemming on the Arabic text document clustering with five similarity/distance measures: Euclidean Distance, Cosine Similarity, Jaccard Coefficient, Pearson Correlation Coefficient and Averaged Kullback-Leibler Divergence. The dataset includes Corpus of Contemporary Arabic (CCA). They have found that the Euclidean Distance, the Cosine Similarity and the Jaccard measures are effective for the partitional Arabic Documents Clustering task, and better results were achieved without performing stemming on the dataset.

3. The Study Framework

Our study framework for Islamic document clustering consists of the following stages; (i) first stage – Arabic document preprocessing, (ii) second stage – build the documents representation, (ii) third stage – apply similarity/distance measure, (iv) fourth stage – documents are clustered based on the K-means, and (v) fifth stage – the comparative analysis and evaluation of clustering is carried out by using Overall Purity and Overal F-measure. Fig.e 1 shows the framework of the Islamic document clustering.

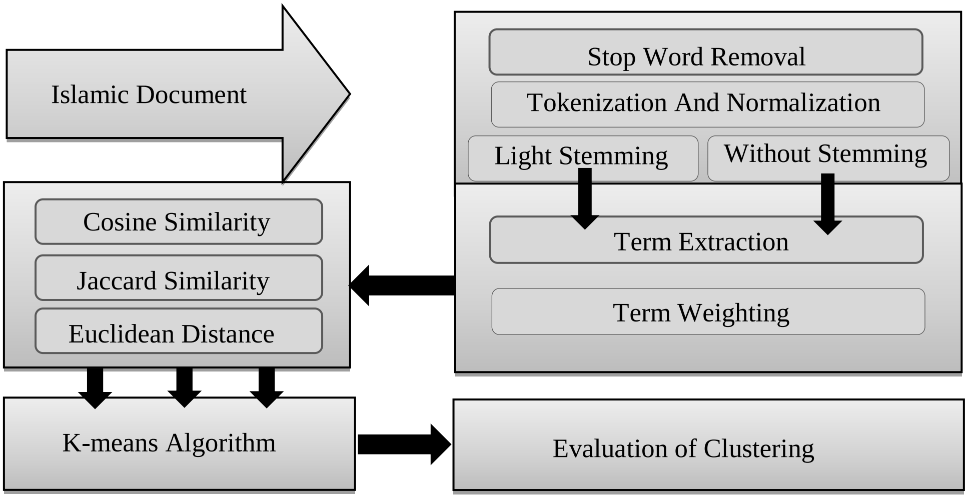

Figure 1: Framework of the Islamic document clustering.

3.1 Islamic Text Pre-processing

The Islamic document clustering consists of three stages of Arabic document preprocessing: (1) Tokenizing and Normalizing, (2) Stop Word Removal, and (3) Without Stemming or with Stemming. Detailed explanation will be given in the following subsections.

3.1.1 Arabic text tokenizing and normalizing

The first phase of the text pre-processing describes the Arabic text tokenization and normalization in details:

- Remove punctuation marks and numbers and symbols such as (1, 2, !,",:, ',?,*,[ ]).
- Remove diacritics such as (-) for example ,(بســـم) to (بسم).
- Remove non-Arab words such as (name).
- Normalize آ , أ , إ and ء with ا .
- Normalize final ىٌ , ىً , ى with ي .
- Normalize final ة with ه .
- Remove words with less than three letters.
- Split the text into set of tokens.

The reason for carrying out this normalization is that there is only one form representing all forms of hamza (ء) in dictionaries, and different forms of aleph are often misspelled by people. We have normalized the letter “ى” to “ي” and the letter “ة” to “ه”. The reason behind this normalization is that there is no single unified convention for spelling “ى” or “ي”and “ة” or “ه” when they appear at the end of a word.

3.1.2 Arabic Stop Word Removal

The Arabic stop words are present in text documents in the Arabic language, which is not beneficial to be remembered as it only linguistically enhances the structure of the text and ungrammatical, for example; prepositions, conjunction, pronouns and others. In the Arabic stop word removal process, we have removed stop words that have a negative impact on the Islamic document clustering. The removal of stop words will improve the clustering of Islamic document. The number of stop words and number ranges less than 170 words identified by Khoja and Garside (1999) and Larkey et al. (2002). More than 2200 Arabic words as stop word identified by Al-Shammari and Lin (2008). In this study, we use the stop word by Al-Shammari and Lin (2008).

3.1.3 With Stemming

This is the third phase in pre-processing. In this phase, we have used the stemming words to represent the terms. Word stemming in Arabic is the process of removing all the prefixes and suffixes and converting the word to its stem word. The importance of stemming process are for indexing and keyword filtering since it will make the clustering process more accurate and faster, and this is done by reducing the size of the vocabulary, and thus reduce the dependence on certain forms of vocabulary. There are many stemming methods available in Arabic, including light stemming (Darwish 2002) Text REtrieval Conference (TREC-2002). We use the light stemming method in our study framework (see Fig. 1). The aim of the light stemming is not to produce the linguistic root of a given Arabic the surface of form, but rather to remove the most common prefixes and suffixes of Arabic word. The advantage light stemmer is that, it does not require a dictionary like the root-based stemmer. It requires only the table of predefined affixes (prefixes and suffixes).

3.2 Text Representation

After preprocessing the text, the text should represent in a specific form, so that it will be suitable for the document clustering processing. It should be in certain text representation, because we cannot deal with raw text, due to the complex nature of text. There are several models of text representation, including the world based model or also called as the Bag-Of-Words (BOW) representation. BOW is the most common and widely used for term type of representation (Berger et al. 2006). We use BOW because it is the simplest text representation model, since it only records the frequency of the words presented in the document.

In this study, the term weighting method is based on Boolean weighting; it is a simple and easy method for term weighting. In this approach, (0) is the matrix, to which the term weight is assigned, when the term does not appear in the document, and (1) is the matrix to which the weight of a term is assigned, when the term appears in the document. Each term have its weight by using the Term Frequency – Inverse Document Frequency (TF–IDF). The TF–IDF indicates how important a term is to a document in a corpus, and this helps to control the fact that some words are important than others. Changes in the TF–IDF weighting scheme are frequently used as a central tool in the search engines to account the relevance of a document in terms of user query. TF–IDF is used in different areas including text clustering, text summarization and classification. Eq. (1) shows how the weight term, w_i is calculated:

$$w_i = tf_i.\log\left(\frac{N}{n_i}\right) \qquad (1)$$

Where N is the total number of documents in the document corpus, and n_i is the number of documents in the collection, where the term i appears.

3.3 Similarity and Distance Measure

Similarity measure is the important part and basic entity in any clustering algorithms as it facilitates to measure the percentage of similarity between entities, and put more similar elements together in group, and select smaller distance between them. In this research, we have used the following similarity/distance measures: Cosine Similarity, Jaccard Coefficient and Euclidean Distance. In the following sections, we will discuss each of the mentioned similarity/distance measures:

A. Cosine Similarity

Cosine similarity is one of the common similarity measures. In order to measure the similarity between documents using confine similarity, we have selected two documents $\vec{t}_a$ and $\vec{t}_b$, and used the equation Cosine Similarity as below (see Eq. (2)):

$$\text{Cosine similarity}\left(\vec{t}_a, \vec{t}_b\right)\frac{\vec{t}_a \;.\vec{t}_b}{\left|\vec{t}_a\right| \times \left|\vec{t}_b\right|}. \qquad (2)$$

Where $\vec{t}_a$ and $\vec{t}_b$ perceived as m-dimensional vectors models through term set T { $t_1....t_m$ }. We had represented all terms with their weight in document, by a particular dimension, and be non-negative. Thus, cosine similarity ranges between [0, 1].

B. Jaccard Coefficient

The Jaccard coefficient or Tanimoto coefficient is used to measure similarity as the intersection is divided by union of the objects. For text document, Jaccard coefficient has been used to compare the sum of shared terms weight and the sum of terms weight which are present in either of the two documents, yet they must not be shared terms. The mathematical formal definition for Jaccard coefficient is shown below, Eq. (3):

$$\text{Jaccard}\left(\vec{t}_a,\vec{t}_b\right)=\frac{\vec{t}_a.\vec{t}_b}{\left|\vec{t_a}\right|^2+\left|\vec{t_b}\right|^2-\vec{t}_a.\vec{t}_b} \qquad (3)$$

The Jaccard coefficient is a similarity measure and the measure value ranges [0,1], in which it is 0 when $\vec{t}_a$ and $\vec{t}_b$ are disjoint and 1 when the $\vec{t}_a$ equal $\vec{t}_b$.

C. Euclidean Distance

Euclidean distance, sometimes referred Euclidean metric, is widely used as a distance measure in clustering problems, and includes clustering text, and used with the K-means algorithm as the default measure distance. In order to measure the distance between the documents, given two documents d_a and d_b which are represented by their term vectors $\vec{t}_a$ and $\vec{t}_b$ consecutively; and referred to as the term set is $T=\{t_1...t_m\}$. The Euclidean distance is defined for two documents as follows, see Eq. (4):

$$\text{Euclidean}\left(\vec{t}_a,\vec{t}_b\right)=\sqrt{\sum_{t=1}^{m}\left|w_{t,a}-w_{t,b}\right|^2} \qquad (4)$$

As mentioned previously, we have used the TF-IDF value as the term weights, and $w_{t,a}$ executes from TF-IDF value, $w_{t,a}=TF-IDF(d_a,t)$.

4. Experimental results of document clustering

The goal of this research is to cluster Islamic document using K-means algorithm. We have used the dataset in three groups; (i) two categories, (ii) three categories, and (iii) four categories, with two methods and three similarity/distance measures, to evaluate the differences in the results, when the number of categories is increased.

4.1 Data Set

The dataset used in our system consists of 1600 documents and divided into four categories as shown in Table 1. Dataset is in-house collected from website http://www.islamport.com. This Islamic text consists of (العربية الحديثة) Modern Standard Arabic (MSA) and (العربية الفصحى) Classical Arabic (CA).

Classical Arabic (CA): is fully vowelized and it is the language of the holy Quran.Modern Standard Arabic (MSA): is the official language throughout the Arab world. It is used in official documents, newspapers and magazines, in educational fields, and for communication among Arabs.

Table 1: Size the Islamic dataset for each category.

Categories	size of categories set
Al Fatwa (الفتاوى)	400
Explanation Al Hadith (تفسير الحديث)	400
Al Serah (السيرة)	400
Al Aqeda (العقيدة)	400

4.2 Experiments and Result

These experiments were measured using the Overall F-measure and Overall Purity on stemmed word and without stemmed word, and with three similarity/distance measures based on the K-means. First experiment (Experiment I) had been done on the two categories, second (Experiment II) on the three categories, and third (Experiment III) experiment on the four categories as shown in Tables 2, 3, and 4 as follows:

Table 2: Experiment I: Overall Purity and Overall F-measure for two categories (Al Fatwa and Explanation Al Hadith).

Stemming	Evaluation	Cosine	Jaccard	Euclidean
With light stemming	Purity	0.85	0.81	0.83
	F-measure	0.83	0.8	0.82
Without stemming	Purity	0.81	0.77	0.79
	F-measure	0.8	0.76	0.78

Table 3 Experiment II: Overall Purity and Overall F-measure evaluation for three categories (Al Fatwa, Explanation Al Hadith and Al Serah).

Stemming	Evaluation	Cosine	Jaccard	Euclidean
With light stemming	Purity	0.83	0.8	0.81
	F-measure	0.82	0.79	0.8
Without stemming	Purity	0.78	0.76	0.78
	F-measure	0.78	0.74	0.77

Table 4 Experiment III: Overall Purity and Overall F-measure evaluation for four categories (Al Fatwa, Explanation Al Hadith, Al Serah and Al Aqeda).

Stemming	Evaluation	Cosine	Jaccard	Euclidean
With light stemming	Purity	0.79	0.77	0.78
	F-measure	0.78	0.76	0.78
Without stemming	Purity	0.77	0.75	0.75
	F-measure	0.76	0.74	0.73

Based on the results of the Experiment I, as illustrated in Tables 2, it is evident that the results with light stemming method are better than those without stemming. The highest value with light stemming is (0.85%) with Cosine similarity. Moreover, from the results of the Experiment II, as illustrated in the Tables 3, it has been proved that the light stemming method has yielded the best result with Cosine similarity of (0.83%). Furthermore, the results of the Experiment III, as illustrated in Tables 4, indicate that the highest value (0.79%) is achieved with the Cosine similarity and stemming method.

5. Conclusions

This paper aimed to investigate text clustering algorithm for Islamic Arabic texts based on the K-means algorithm, with or without stemming, with three similarity/distance measures. In this paper we had used Islamic dataset (in-house). Based on the results in section 4, the K-means algorithm has the best results with Cosine similarity compared to Jaccard similarity and Euclidean distance. We also noticed that the results with Euclidean distance are better than the results with Jaccard similarity. In addition, we also found that the results with stemming method are better than without stemming. We also found that the results depend on the number of categories and size of dataset since results from Experiment I are generally better than the results from Experiment II and III.

References

Ahmed, M. A. 2000. A Large-Scale Computational Processor of the Arabic Morphology, and Applications. A Master's Thesis, Faculty of Engineering, Cairo University, Cairo, Egypt.

Akbar, M. 2008. Fp-Growth Approach for Document Clustering. Tesis MONTANA STATE UNIVERSITY Bozeman.

Al-Omari, O. M. 2011. Evaluating the Effect of Stemming in Clustering of Arabic Documents. Academic Research.

Al-Shammari, E. & Lin, J. 2008. A Novel Arabic Lemmatization Algorithm. Proceedings of the second workshop on Analytics for noisy unstructured text data, hlm. 113-118.

Berger, H., Dittenbach, M. & Merkl, D. 2006. Analyzing the Effect of Document Representation on Machine Learning Approaches in Multi-Class E-Mail Filtering. Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on, hlm. 297-300.

Darwish, K. 2002. Building a Shallow Arabic Morphological Analyzer in One Day. Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, hlm. 1-8.

Ding, Y. & Fu, X. 2012. Topical Concept Based Text Clustering Method. Advanced Materials Research 532(939-943.

Farghaly, A. & Shaalan, K. 2009. Arabic Natural Language Processing: Challenges and Solutions. ACM Transactions on Asian Language Information Processing (TALIP) 8(4): 14.

Ferguson, C. A. 1996. Sociolinguistic Perspectives: Papers on Language in Society, 1959-1994. Oxford University Press, USA.

Froud, H., Benslimane, R., Lachkar, A. & Ouatik, S. A. 2010. Stemming and Similarity Measures for Arabic Documents Clustering. I/V Communications and Mobile Network (ISVC), 2010 5th International Symposium on, hlm. 1-4.

Huang, L. 2011. Concept-Based Text Clustering. Tesis University of Waikato.

Khoja, S. & Garside, R. 1999. Stemming Arabic Text. Lancaster, UK, Computing Department, Lancaster University.

Larkey, L. S., Ballesteros, L. & Connell, M. E. 2002. Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-Occurrence Analysis. Annual ACM Conference on Research and Development in Information Retrieval: Proceedings of the 25 th annual international ACM SIGIR conference on Research and development in information retrieval, hlm. 275-282.

Osama, A. G. & Wesam, M. A. 2012. Stemming Effectiveness in Clustering of Arabic Documents. International Journal of Computer Applications , 0975 – 8887.