ORIGINAL RESEARCH PAPER | Computer Science | Volume : 6 | Issue : 11 | November 2016 | ISSN - 2249-555X | IF : 3.919 | IC Value : 74.50

Arabic Text Clustering Using Different Similarity Measures

KEYWORDS: Arabic language processing , Text mining, Clustering

Dr. Salma A. Mahmood
Department of computer science, collage of Science, Basra University, Basra, IRAQ

Firas H. Neama
Department of computer science, collage of Science, Basra University, Basra, IRAQ

ABSTRACT *Text clustering considers an efficient way to the text mining, and the text mining is important in many applications such as retrieval of texts on the Web, browsing, and indexing texts. The aim of this proposed study is improved the technique for Arabic text mining using clustering as a learning unsupervised method. In order to achieve the goal of our study the clustering algorithms k-mediods and k-means is used, with using the similarity metrics Euclidean and Cosine. The system implements using a corpus contains on 200 sport news Arabic. Finally, evaluation measures are used including (Precision, Recall and F-measure) to evaluate our system, and we get satisfy results.*

1. Introduction

Text mining is a branch of data mining, defines as the ability to explore or find a new knowledge in the documents collection, which enables the user to interact with the corpus using appropriate tools, and identifies and detects important patterns to process various documents. In addition, text mining can execute processes enormous on the documents that exceed the capacity of the human to read these documents at an appropriate time. Moreover, it exceeds human's ability to detect knowledge patterns and extract meaning from document because incompatibility with human ideas and expectations (Kameshwaran & Malarvizhi, 2014) (Fejer & Omar, 2015) (Kaur & Garg, 2015). Text mining applies in various fields such as medicine, science, agriculture, political science, marketing, economics and computer applications (translation mechanism).

Documents clustering one of the texts mining techniques is a natural activity in every organization, and used in technologies such as document retrieval, classification patterns, and decision-making, furthermore, clustering is effective technique to organize document efficient(Rogério dos Santos Alves; Alex Soares de Souza, 2014)toxic chemical products formed as secondary metabolites by a few fungal species that readily colonise crops and contaminate them with toxins in the field or after harvest. Ochratoxins and Aflatoxins are mycotoxins of major significance and hence there has been significant research on broad range of analytical and detection techniques that could be useful and practical. Due to the variety of structures of these toxins, it is impossible to use one standard technique for analysis and/or detection. Practical requirements for high-sensitivity analysis and the need for a specialist laboratory setting create challenges for routine analysis. Several existing analytical techniques, which offer flexible and broad-based methods of analysis and in some cases detection, have been discussed in this manuscript. There are a number of methods used, of which many are lab-based, but to our knowledge there seems to be no single technique that stands out above the rest, although analytical liquid chromatography, commonly linked with mass spectroscopy is likely to be popular. This review manuscript discusses (a.

This study is organized as follows: The next section include Clustering system requirements, section 3 explain Arabic Documents clustering, section 4 evaluation and discusses of results, finally section 5 concludes and future works.

2. Clustering System requirements

The proposed system requires building linguistic lexicon and corpus of the Arabic language illustrated as bellow :

2.1 Arabic lexicon

The Proposed system needs to build Arabic lexicon, which is uses to extract the features during execution, lexicon consists of several tables, as follows:

- Arab roots verbs table, it contains approximately 3366 roots.
- Nouns table, it contains approximately 1749 noun.
- Pronouns and The Relative Pronouns, The Demonstrative Pronouns.

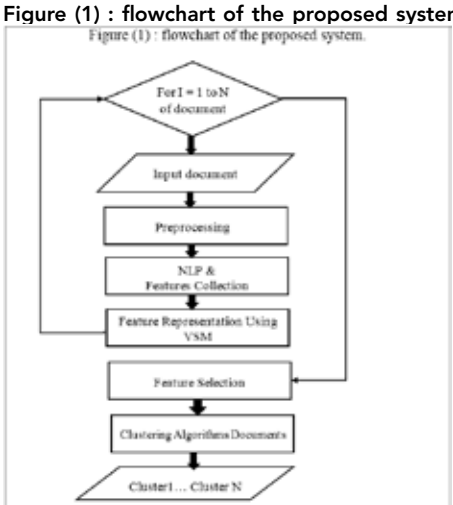
2.2 Arabic News Corpus

Corpus is compiled from several web sites (alsumaria news, kooora, Arabia, bein sport) and included in 200 text files of sports news in different sizes, they include football , swimming and formula1.

3. Arabic Clustering proposal system

The aim of this system is to implement k-medoids and k-means on the Corpus using measurements of similarity (Euclidean, Cosine). The figure (1) shows the clustering system diagram.

Figure (1) : flowchart of the proposed system.



Volume : 6 | Issue : 11 | November 2016 | ISSN - 2249-555X | IF : 3.919 | IC Value : 74.50

The following steps explain system processes:

3.1 Preprocessing

The aim of preprocessing process is perform the Tokenization to divide sentences relies on punctuation marks (point, comma). Then, separate the Arabic words depend on spaces. Also, include determining word sequence in the text. As well as deletes numbers and special symbols.

3.2 NLP and Features Collection

Aim of this phase is to extract the documents features based on three consecutive stages **Lexical Analyzer,** its objective is determining the types and features of words. The input of this stage is the words obtained from Tokenization phase and output list of words features, this analyzer searches for words in the lexicon to reach their features. And **Morphological Analyzer,** which works on words that are not recognized in lexicon. it removes precedents and suffixes of the word based on morphology of the Arabic language, then re-implement the lexical analyzer. Finally, **Syntactical Analyzer,** that removes the ambiguity resulting from multiple uses of Arabic words , this ambiguity cannot solved in lexical and morphological analyzer. For example, the word (كتاب) can be classified a verb or noun.

At end of above phases, the results is words and their features in a single list. The resulting list contained nouns only, because they distinguishing the subjects of the document.

3.3 Features Representation Using Vector Space Model

The results stores matrix VSM, where terms represent as rows and documents as columns. As following algorithm:

Vector Space Model algorithm

Input: Number of the document, list names with frequency in document

Output: Vector Space Model

N: count list names.

For i = 0 **To** N – 1

Number the row= Select Vector Space Model like ' name[i] '

If Number the row=0 **then**

Update filed Number of the document

Else

insert name[i] & Number of the document

End if

End for

3.4 Features Selection

It determines documents properties using the following methods(Alelyani, Tang, & Liu, 2013):

Term Frequency (TF)

This method calculates the Frequency of terms in documents, using the following equation:

$$TF(fi) = \sum_{j \in Df_i} tf_{ij} \qquad \ldots\ldots (1)$$

Inverse Document Frequency (IDF)

This method gives high values of the rare terms frequency and small values high- frequency of the terms, using the following equation:

$$idf(fi) = \log \frac{|D|}{|Dfi|} \qquad \ldots\ldots (2)$$

Term Frequency-Inverse Document Frequency (TF-IDF)

The TF-IDF gives small values for high-frequency terms (or words) in documents, the value of the largest small-frequency , and thus give a greater ability to distinguish the features of documents and get the best clustering(Zhao, Zhang, & Wan, 2013) , using the following equation:

$$tf - idf(fi, dj) = tfij * idf(fi) \qquad \ldots\ldots (3)$$

3.5 Clustering Algorithms

We use two clustering algorithms k-means and k-mediods to clustering Arabic documents, as follows:

k-means

This method based on the concept of finding the center point of the documents, where this point is calculated by finding the mean distance between points within the cluster. The algorithm starts random to choose of the centers K, documents are added to the cluster depending a function similarity with the Centre. Cluster centers are modified in each iteration until, threshold is achieved or Clusters stay fixed(Alkoffash, 2012) (Rai & Singh, 2010).

k-means Algorithm

Input: Array (Vector Space Model), K (the number of suggested clusters),

Output: Clusters Group

1. Select k centers randomly

2. Assign each document to the closest cluster based on the distance between centers and the document (Euclidean or Cosine).

3. compute means for each cluster

4. Select k centers based on step 3.

5. If current clusters < > previous clusters then Repeat steps 2-4. Else Complete Clusters End If

K-mediods

This method is similar to K-means But, it differs to choose new centers random rather than relies on the mean elements of the cluster (Rai & Singh, 2010) (Alkoffash, 2012).

Input: Array (Vector Space Model), K (the number of suggested clusters),

Output: Clusters Group

1. Select k centers randomly

2. Assign each document to the closest cluster based on the distance between centers and the document (Euclidean or Cosine).

3. compute k centers randomly for each cluster

4. Select k centers based on step 3.

5. If current clusters < > previous clusters then Repeat steps 2-4. Else Complete Clusters End If

Similarity Measurements

We use two different Similarity Measurements, Euclidean Distance :

Volume : 6 | Issue : 11 | November 2016 | ISSN - 2249-555X | IF : 3.919 | IC Value : 74.50

$$S_{x,y} = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \quad \ldots\ldots\ (4)$$

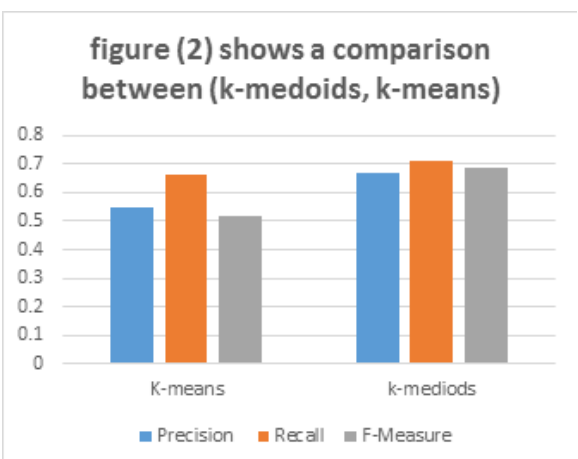
And Cosine Similarity witch define as : (Huang, 2008) (Aggarwal & Zhai, 2012).

$$cos(d_1, d_2) = \frac{(d_1.d_2)}{\|d_1\|\|d_2\|} \quad \ldots\ldots \quad (5)$$

4. Evaluation and Discusses

In this paper we uses three metrics, **Precision, Recall, and F-measure** as(Abdel & Ghanem, 2014) to evaluate the results . In experiments used k-means with functions similarity (Euclidean, Cosine), the best results when it uses Cosine the three metrics, as shown in Table 2. While we uses k-mediods with the same similarity measurements , the good results when it uses Cosine function, as shown in Table 2. On the other hand, the result of the k- mediods better than k-means due k- mediods chooses randomly to the center in each iteration instead of the mean, as shown in chart in figure 2.

Number of clusters	Similarity	Precision	Recall	F-Measure	Precision	Recall	F-Measure
3	Euclidean	0.36	0.41	0.38	0.49	0.51	0.50
	Cosine	0.33	0.59	0.42	0.67	0.71	0.69
5	Euclidean	0.41	0.31	0.35	0.55	0.44	0.48
	Cosine	0.44	0.69	0.54	0.73	0.68	0.70
7	Euclidean	0.55	0.26	0.35	0.59	0.44	0.50
	Cosine	0.55	0.66	0.52	0.78	0.60	0.67
Table [2]:		K-means results			K- mediods results		



5. Concludes and future works

The results of such system influenced mainly by the nature of corpus and NLP processing, the lexical, morphological and syntactical analyzers must be comprehensive to increase efficiency of system. The comparison show that K-mediods is better than k-means results, We suggest that combined with other algorithms such as hierarchical algorithms and Particles Swarm Optimization (PSO). As well as can be used the methods to reduce features such as conference resolution, topic model which future solution to solve high dimension.

References

1. Abdel, O., & Ghanem, F. (2014). Evaluating the Effect of Preprocessing in Arabic Documents Clustering. Gaza,: Islamic University, Gaza, Palestine.
2. Aggarwal, C. C., & Zhai, C. (2012). *Mining text data*. book, Springer Science & Business Media.
3. Alelyani, S., Tang, J., & Liu, H. (2013). Feature Selection for Clustering : A Review. *Data Clustering: Algorithms and Applications*, 1–37. http://doi.org/10.1.1.409.5195
4. Alkoffash, M. S. (2012). Comparing between Arabic Text Clustering using K Means and K Mediods, *51*(2), 5–8.
5. Bholat, D., Hansen, S., Santos, P., & Schonhardt-Bailey, C. (2015). Text Mining for Central Banks. *Centre for Central Banking Studies, Handbook*, *33*, 1–19.
6. Fejer, H. N., & Omar, N. (2015). Automatic Multi-Document Arabic Text Summarization Using Clustering and Keyphrase Extraction, 1–9. http://doi.org/10.3923/jai.2015.
7. Huang, A. (2008). Similarity measures for text document clustering. *Proceedings of the Sixth New Zealand*, (April), 49–56. Retrieved from http://nzcsrsc08.canterbury.ac.nz/site/proceedings/Individual_Papers/pg049_Similarity_Measures_for_Text_Document_Clustering.pdf
8. Kameshwaran, K., & Malarvizhi, K. (2014). Survey on Clustering Techniques in Data Mining, *5*(2), 2272–2276.
9. Kaur, M., & Garg, S. K. (2015). Survey on Clustering Techniques in Data Mining for Software Engineering Survey on Clustering Techniques in Data Mining for Software Engineering, (MAY 2014).
10. Rai, P., & Singh, S. (2010). A Survey of Clustering Techniques. *International Journal of Computer Applications*, *7*(12), 1–5. http://doi.org/10.5120/1326-1808
11. Rogério dos Santos Alves; Alex Soares de Souza, et all. (2014). *Data Mining. Igarss 2014*. http://doi.org/10.1007/s13398-014-0173-7.2
12. Zhao, J., Zhang, K., & Wan, J. (2013). Research of Feature Selection for Text Clustering Based on Cloud Model. *Journal of Software*, *8*(12), 3246–3252. http://doi.org/10.4304/jsw.8.12.3246-3252