# Arabic Text Mining Based On Clustering And Coreference Resolution

Salma Mahmood
University of Basrah- Iraq
salma_abdulbaki@yahoo.com

Faiez Musa Lahmood Al-Rufaye
Middle Technical University- Wasit, Iraq
faizmossa2000@gmail.com;

***Abstract*-Text mining discover and extract useful information from documents, whenever increase the size and number documents leads to redouble features. The huge features for the documents adds challenge to text mining called high dimension. The aim of this proposed study is minimize the high dimension of the documents, and improve Arabic text mining using clustering. In order to achieve this goal, we propose to applied coreference resolution technique using the clustering algorithms k-mediods and k-means. This study uses 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.**

***Keywords*-**Arabic language processing , Text mining, Clustering**.**

## I. Introduction

Text Mining finds a new knowledge in the corpus using appropriate tools, which exceeds human's ability to detect knowledge patterns and extract meaning from corpus due incompatibility with human ideas and expectations (Kaur & Garg, 2015) (Aggarwal & Zhai, 2012). This technique uses multi-disciplinary fields, such as information retrieval, text analysis, social networks, information classification, and database technology (Irfan et al., 2015). Clustering one of the unsupervised learning methods. The documents clustering is used in Texts Mining applications, which collects documents in the similar groups (Rogério dos Santos Alves; Alex Soares de Souza, 2014). This study is interested in Arabic Text Mining using Clustering Technique and coreference resolution as a tool to reduce the features involved in the clustering process. This study is organized as follows, The next section Clustering System requirements, section 3 explain the Documents Clustering and using coreference resolution technique, section 4 contain evaluation and discusses of

results, finally section 5 contain conclusions and future works.

## II. Clustering system requirements

The proposed system requires building linguistic lexicon and corpus of the Arabic language, Arabic lexicon consists of several tables, Arab roots verbs table contains approximately 3366 root , Nouns table contains approximately 1749 noun and Pronouns and nouns linked, signal and question table. Another requirement is the Arabic Corpus, which is compiled from several web sites (alsumaria news, kooora, Arabia, bein sport) and these texts is a sports news of different sizes include football ( 70 documents), swimming (70 documents) and formula1(60 documents). It is stored in txt type file, and saved encodes UTF-8.

## III. Documents clustering based on coreference resolution

The objective of the system is to implement the k-medoids and k-means on the Corpus using measurements of similarity (Euclidean, Cosine). Figure (1) shows the operations of the proposed system which contain the following steps:

### A. Preprocessing

The aim of preprocessing process is splitting the document into sentences and words to use in subsequent processors. The process include Tokenization to divide sentences relies on punctuation tools (point, comma). Also, include determining word sequence in the text. As well as deletes numbers and special symbols to yield words only.

### B. Natural Language Processing and Features Collection

Aim of this phase extraction characteristics of Arabic documents based on three consecutive Analyzers , the first one is **Lexical Analyzer,** it determine the types (such as noun , verb, preposition) and features of words, such as (feminization and masculinization) , numbers (single , dual, plural), it searches and matching the words with the lexicon. Next, it is **Morphology Analyze** works with words are not recognized in the lexical analyzer. It removes precedents and suffixes of



978-1-5386-2955-0/17/$31.00 ©2017 IEEE

the word based on morphology of the Arabic language, then re-implement the lexical analyzer. Finally , it is **Syntactic Analyzer,** this Analyzer removes the confusion resulting from the multiple classification of the words which obtained from the lexical and morphology analyzer. For example, the word (كتب – write ) can be classified a verb or noun.

### C. Coreference Resolution

This phase solve the huge extracted features by applying three algorithms, each one has its function to collect words together produce terms that represent concepts in text.

**a. Nouns Identification algorithm :** This algorithm adds new features to the nouns, it has a significant impact on the following create terms algorithm, the features includes (defined, indefinite, proper nouns, and others) which obtained by applying number of rules that described in table (1). At the end of this stage the collected features includes (location, type, class, sub-class, number, gender) in addition to the feature nouns Identification.

**b. Create terms algorithm :** This algorithm discriminates and highlights the important entities in the document, these entities often consist of successive nouns. For example " وزارة الشباب والرياضة - Ministry of Youth and Sports" , which successive nouns indicate to a single entity, as well as the nouns " هاشم خميس – Hashem Khemees" . Many nouns loses significance when individually treated , such as the "وزارة" alone does not indicate a word meaning in the document. As well as "الشباب" and so on. This algorithm Benefits from a proposed set of axiomatic rules that are described in the table (2).

**c. Corefrence Resolution Using Clustering :** The goal of this process of extracting important topics contained in the document. Clustering collects all matched nouns whole or part with clusters centers (terms, proper nouns single). Next implements clustering process by using the k-means algorithm and using a similarity measure described in equation (1), which calculates the value of the similarity between the clusters centers and others nouns, and pronouns, demonstrative pronoun, and relative pronoun (Angheluta, Jeuniaux, Mitra, & Moens, 2004)(Wang & Ngai, 2006). Other nouns ignore if are not existing in clusters. The algorithm (1) describes the corefrence resolution using clustering

$$Didt(NP_i, NP_j) = \sum_{f \in F} W_f * \text{incompatibility}_f(NP_i, NP_j) \quad \ldots.(1)$$

$NP_i$ : The first word (the term, proper noun single)

$NP_j$ : The second word (pronouns, demonstrative pronoun, and relative pronoun). $F$: Features used to determine the distance. $W_f$ : Property weight.

After finishing of the previous operations yields number of clusters. The centers of the clusters includes terms or proper nouns single are chosen only.

Table 1: describes Nouns Identification feature

| Precedent | Feature | Feature Symbol | Examples |
|---|---|---|---|
| ال | Defined | Def | اللاعب |
| والـ | Defined | WDef | والرياضة |
| و | Indefinite | xInd | وملعب |
| ل، لل، بـ | Indefinite | xInd | لنادي |
| —— | proper noun | Per | احمد |
| —— | Others words | Other | هذا، هو |

### D. Features Representation and Features Selection

The results stored VSM matrix, where terms represent as rows and documents as columns. Features Selection determines documents properties using the following methods , **Term Frequency (TF) ,** it calculates the Frequency of terms in documents, **And, Inverse Document Frequency (IDF)** , that gives high values of the rare terms frequency and small values high-frequency of the terms, **And, Term Frequency-Inverse Document Frequency (TF-IDF),** that 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:

| Table (2) :proposed set of axiomatic rules. | | |
|---|---|---|
| Seq. | Symbol | Example |
| 1 | Ind Ind Def Per Per | مدرب حراس المرمى هاشم خميس |
| 2 | Per Per Per Per | هاشم خميس احمد محمد |
| 3 | Ind Def Def Def | لاعبي الفريق الأربعة المبعدين |
| 4 | Ind Def Def Per | حكومة الرئيس فرانسوا هولاند |
| 5 | Per Per Per | زين الدين زيدان |
| 6 | Def Per Per | اللاعب علي حصني |
| 7 | Ind Def WDef | وزارة الشباب والرياضة |
| 8 | Ind Def Def | نادي الميناء الرياضي |
| 9 | Ind Ind Def | مدرب حراس المرمى |
| 10 | Per Per | كريستيانو رونالدو |
| 11 | Def Def | اللاعبين المبعدين |
| 12 | Ind Ind | أتلتيكو مدريد |

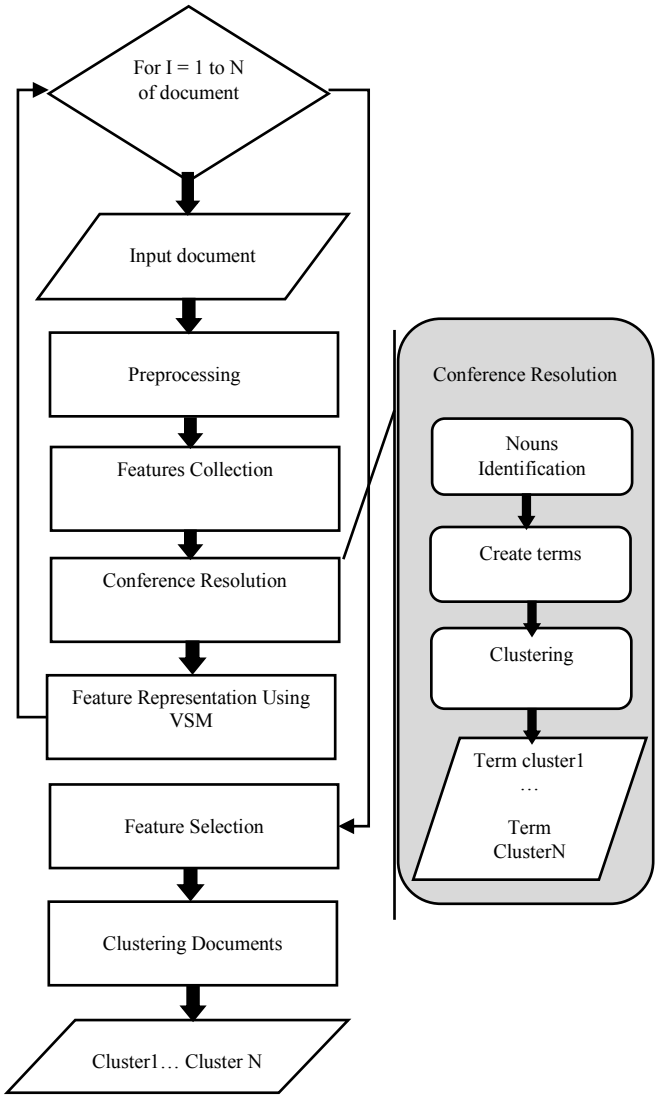



Fig 1. The documents clustering system with using coreference resolution

### E. Documents Clustering

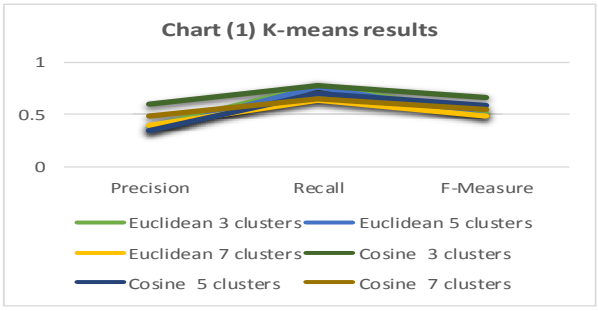
This phase builds the clusters based on two algorithms k-means, k-mediods, **k-means** method based on the concept of finding the center point of the documents, where this point is calculated by finding the mean distance between points within the cluster. The algorithm starts random to choose of the 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-mediods** 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). The proposed method minimize the high dimension of the documents, and improve Arabic text mining using clustering as described in following table:

Table (2a) Arabic text mining using clustering as described

| No. of documents | With corefrence resoltion | Without corefrence resoltion | Minimize dimention percentage |
|---|---|---|---|
| 100 | 1212 words × 100 documents | 1806 words × 100 documents | 23% |
| 200 | 1970 words × 200 documents | 2652words × 200 documents | 35% |

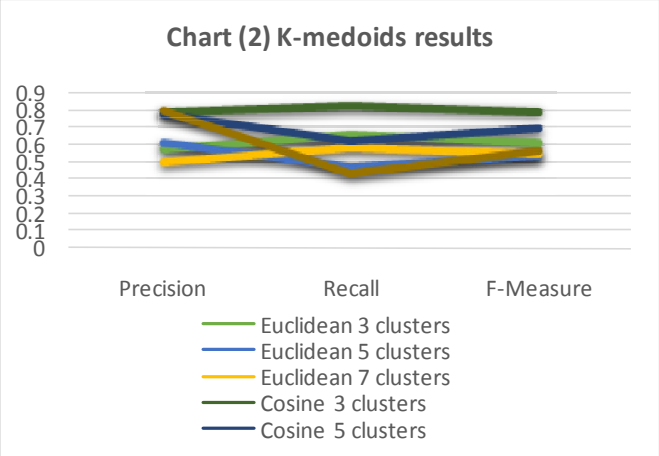
## IV. Evaluation and Results

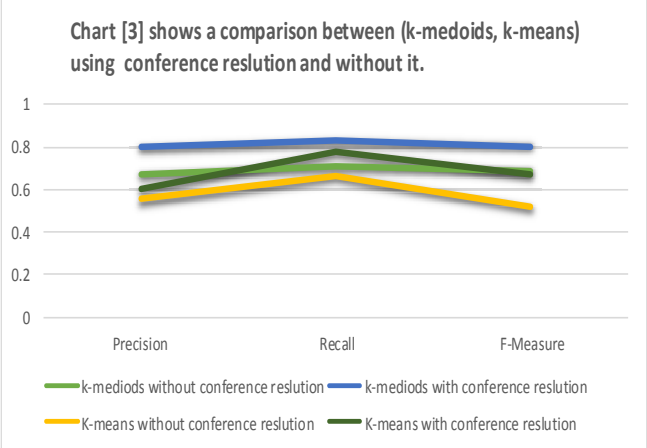
This study use three metrics, **Precision, Recall and F-measure.** In experiments used K-means with similarity measurements (Euclidean, Cosine), the best results in three metrics when we uses Cosine measure ، as shown in Table 3, and chart 1. While we uses k-mediods with similarity measurements (Euclidean, Cosine), the good results when we uses Cosine function, as shown in Table 4, as chart 2. On the other hand, the result of the K- mediods is better than K-means due K-mediods chooses centers randomly in each iteration instead of the mean, as shown in chart 3.



| Table [3]: K-means results | | | | |
|---|---|---|---|---|
| Number of clusters | Similarity | Precision | Recall | F-Measure |
| 3 | Euclidean | 0.38 | 0.78 | 0.51 |
| | Cosine | 0.60 | 0.78 | 0. 67 |
| 5 | Euclidean | 0.35 | 0.76 | 0.48 |
| | Cosine | 0.50 | 0.70 | 0.58 |
| 7 | Euclidean | 0.40 | 0.64 | 0.49 |
| | Cosine | 0.49 | 0.65 | 0.55 |
| | | | | |

| Table [4]: K- mediods results | | | | |
|---|---|---|---|---|
| Number of clusters | Similarity | Precision | Recall | F-Measure |
| **3** | **Euclidean** | **0.58** | **0.67** | **0.62** |
| | **Cosine** | **0.80** | **0.83** | **0.81** |
| **5** | **Euclidean** | **0.62** | **0.48** | **0.54** |
| | **Cosine** | **0.79** | **0.63** | **0.7** |
| **7** | **Euclidean** | **0.51** | **0.59** | **0.55** |
| | **Cosine** | **0.81** | **0.44** | **0.57** |




## V. Conclusions and future works

Many different factors affect the precision scale, such as , the nature and size of Corpus, variety of subjects that it contained. also, the powerful and effectiveness of the Analyzers , lexical Analyzer, morphological Analyzer, syntactical Analyzer that effect on overall to processing of Arabic language, the results may be more improve if we use semantic and pragmatic analyzers. The challenges include determination initial points and outliers of clustering algorithm in order to yield more clusters cohesion. We can propose using conference resolution with hierarchical algorithms or with fuzzy clustering. As well as can be used the techniques to reduce features such as particles swarm optimization (PSO) and topic model as a future solution to high dimension challenge.

## References


1. Abdel, O., & Ghanem, F. (2014). *Evaluating the Effect of Preprocessing in Arabic Documents Clustering*. Islamic University, Gaza, Palestine, Gaza,.
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.
4. Alkoffash, M. S. (2012). Comparing between Arabic Text Clustering using K Means and K Mediods, *51*(2), 5–8.
5. Angheluta, R., Jeuniaux, P., Mitra, R., & Moens, M.-F. (2004). Clustering algorithms for noun phrase coreference resolution. *Proceedings of the 7es Journes Internationales d'Analyse Statistique Des Donnes Textuelles*, 60–70.
6. Huang, A. (2008). Similarity measures for text document clustering. *Proceedings of the Sixth New Zealand*, 49–56.
7. Irfan, R., King, C. K., Grages, D., Ewen, S., Khan, S. U., Madani, S. A., … others. (2015). A survey on text mining in social networks. *The Knowledge Engineering Review*, *30*(2), 157–170.
8. Kaur, M., & Garg, S. K. (2015). Survey on Clustering Techniques in Data Mining for Software Engineering, (MAY 2014).
9. Rai, P., & Singh, S. (2010). A Survey of Clustering Techniques. *International Journal of Computer Applications*, *7*(12), 1–5.
10. Rogério dos Santos Alves; Alex Soares de Souza, et all. (2014). *Data Mining. Igarss 2014*. Elsevier.
11. Wang, C., & Ngai, G. (2006). A clustering approach

for unsupervised Chinese coreference resolution. In *Proc. of the Fifth SIGHAN Workshop on Chinese Language Processing. Sydney: Association for Computational Linguistics* (pp. 40–47).

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.