Arabic Text Clustering

Basel Altoom

High Institute for Applied Sience and Technologies
Damascus, Syria
basel.altoom@hiast.com

***Abstract*— the main focus of this paper list studies about Arabic text clustering and show different methodologies that have been used, comparing its stages and evolution metrics by applying and implementing the most common ones**

Keywords— Arabic text processing, clustering, document representation, Kmeans

I. INTRODUCTION

Most researches discussed techniques of text clustering in English and Europe languages, but there are few researches discussed clustering in Arabic language. In recent years we have seen a tremendous growth in the number of text document collections available on the Internet. Automatic text clustering, the process of assigning unseen documents to clusters, is an important task that can help in the organization and querying of such collections

Document clustering is among the methods employed to group documents containing related information into clusters, which facilitates the allocation of relevant information. This technique can efficiently enhance the search process of a retrieval system [1]

II. TEXT CLUSTERING

Text clustering is the process of organizing a set of text documents to be clustered according to similarities. The aim is to discover natural document groupings, as text clustering achieves an overview of the classes or topics in a corpus.

Clustering is an unsupervised learning process because its properties or class memberships are unknown in advance .

A typical text clustering algorithm involves the following stages (Fig. 1) [2]:

Stage 1: Document representation as an option to include feature extraction or selection methods.

Stage 2: Determining and calculating the document similarity measure.

Stage 3: Applying clustering or grouping rules.

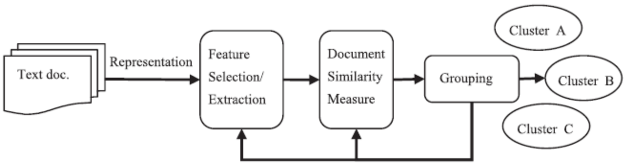


Figure 1 Text clustering methodology

III. RELATED WORKS

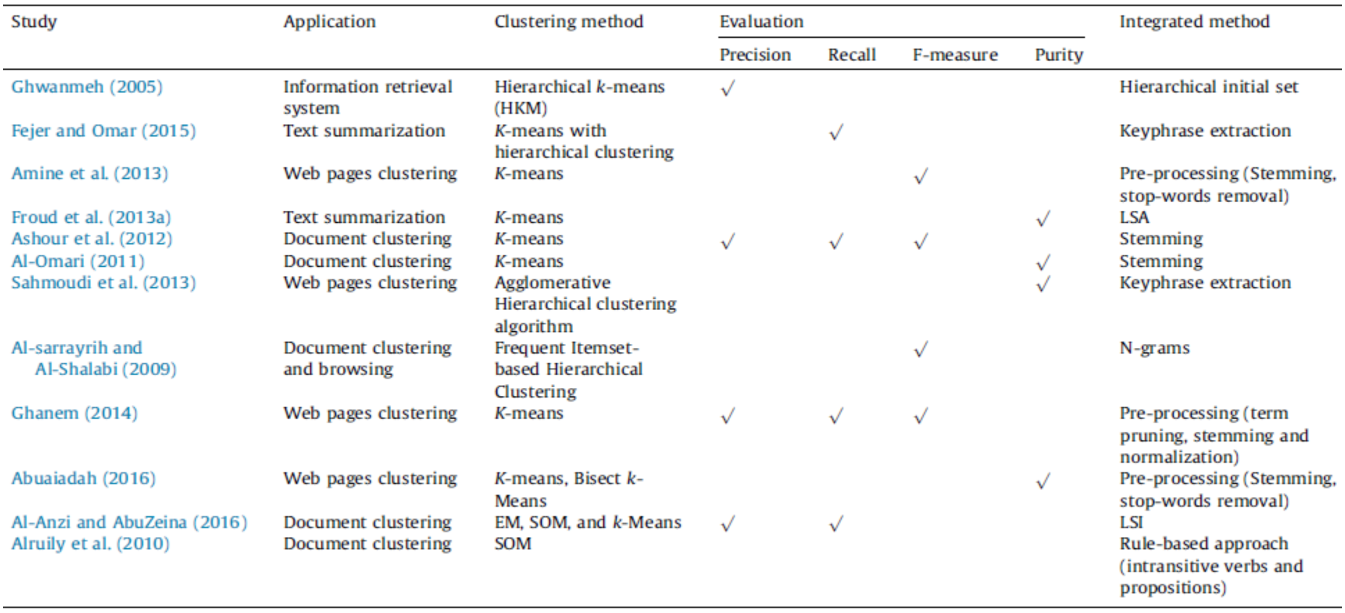
We have noted that most studies are focusing on particular step of Document Clustering above. "Preprocessing steps are necessary to eliminate noise and keep only useful information to enhance document clustering performance" [3], "Pre-processing and stemming methods such as a root-based stemmer or light stemmer can be used to obtain relevant features"[4]. carried out comparative studies on light stemming, root-based stemming and no stemming. The studies suggest that light stemming is more appropriate than root-based stemming or no stemming using precision and recall evaluation measures "light stemming is more appropriate than root-based stemming or no stemming"[5][3].

On the other hand, some studies examined the impact of using an Arabic root-based stemmer (ISRI) with different similarity measures "stemming with ISRI improves clustering quality" [6]. But generally we should consider that Arabic language may affect stemming accuracy "Arabic stemmers tend to produce high stemming error ratios" [7], "stemming is not always beneficial for Arabic text-based tasks, since many terms may be combined with the same root form" [4].

Other studies focused on Document Representation, This problem results from the large number of variables involved in text clustering methods. All terms found in the documents are included in the clustering process, which leads to a very large number of dimensions in the document vector representation. Therefore, high-dimensional data reduces clustering algorithm efficiency and maximizes execution time [8].

On Clustering Algorithm level, FIHC (Frequent Item set-based Hierarchical Clustering), was proposed to obtain the most frequently shared item sets among document sets in clusters. They used N-grams based on word level and character level Trigrams and Quad grams to extract the most frequent item sets [9]. Recommended to use probabilistic topic models for text representation to improve Arabic language clustering, EX. Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), and Latent Dirichlet Allocation (LDA) [4].

Another highlighted the influence of the morpho-syntactic characteristics of the Arabic language on document clustering performance. They compared LDA with k-means clustering by applying both techniques on a set of Arabic documents, the authors suggested that using probabilistic topic models such as LDA provides substantial performance improvement over k-means [10]. Latent Semantic Indexing (LSI) was utilized to group similar unlabeled documents into a pre-specified number of topics. They compared three different clustering methods: Expectation-Maximization (EM), Self- Organizing Map (SOM), and k-means algorithm. According to their research, LSI is recommended for labeling documents as well as improving clustering results [4]. Figure [2] below illustrate a comparison of studies with its methodologies and evolution metrics used to rate result, and "integrated method" shows the method used to improve performance.



Study	Application	Clustering method	Evaluation				Integrated method
			Precision	Recall	F-measure	Purity	
Ghwanmeh (2005)	Information retrieval system	Hierarchical *k*-means (HKM)	√				Hierarchical initial set
Fejer and Omar (2015)	Text summarization	*K*-means with hierarchical clustering		√			Keyphrase extraction
Amine et al. (2013)	Web pages clustering	*K*-means			√		Pre-processing (Stemming, stop-words removal)
Froud et al. (2013a)	Text summarization	*K*-means				√	LSA
Ashour et al. (2012)	Document clustering	*K*-means	√	√	√		Stemming
Al-Omari (2011)	Document clustering	*K*-means				√	Stemming
Sahmoudi et al. (2013)	Web pages clustering	Agglomerative Hierarchical clustering algorithm				√	Keyphrase extraction
Al-sarrayrih and Al-Shalabi (2009)	Document clustering and browsing	Frequent Itemset-based Hierarchical Clustering			√		N-grams
Ghanem (2014)	Web pages clustering	*K*-means	√	√	√		Pre-processing (term pruning, stemming and normalization)
Abuaiadah (2016)	Web pages clustering	*K*-means, Bisect *k*-Means				√	Pre-processing (Stemming, stop-words removal)
Al-Anzi and AbuZeina (2016)	Document clustering	EM, SOM, and *k*-Means	√	√			LSI
Alruily et al. (2010)	Document clustering	SOM					Rule-based approach (intransitive verbs and propositions)

Figure 2 Text Clustering Studies

IV. Methodologies

1. *Arabic text preprocessing:*

Arabic language is a Semitic language that has a complex and much morphology than English; it is a highly inflected language and that due to this complex morphology [11]. There are essential steps required to Arabic text pre-processing: tokenize string to words, word normalizing tokenized words, stop word removal, apply stemming algorithm, and term weighting for each word in document.

Our problem statement related to stemming process, so we will show next discuss of this technique in details (from A. Ghanem and M. Ashour, 2012 publication [12]).

Stemming: stemming algorithms are needed in many applications such as natural language processing, compression of data, and information retrieval systems. In Arabic, the stemming approaches are applied in information retrieval field, very few works exist in the literature that utilize stemming algorithms for Arabic text categorization such as the work of Sawaf; Zaplo, and Ney , and the work of Elkourdi, Bensaid and Rachidi , and Duwairi .

Two type of Stemming:

A. Root-based Stemming

Stemming using root extractor which uses morphological analysis for Arabic words, Figure (3) depicts an example of using stemming for feature selection. Several algorithms have been developed for this approach such as: A Rule and Template Based Stemming Algorithm, Al-Fedaghi and Al-Anzi Stemming Algorithm, and Khoja.

B. Light stemming

The main idea for using light stemming is that many word variants do not have similar meanings or semantics. However; these word variants are generated from the same root. Thus, root extraction algorithms affect the meanings of words. Light stemming by comparison aims to enhance the categorization performance while retaining the words' meanings. It removes some defined prefixes and suffixes from the word instead of extracting the original root. Light-stemming keeps the word's meanings unaffected. Figure (4) demonstrates an example of using light stemming.

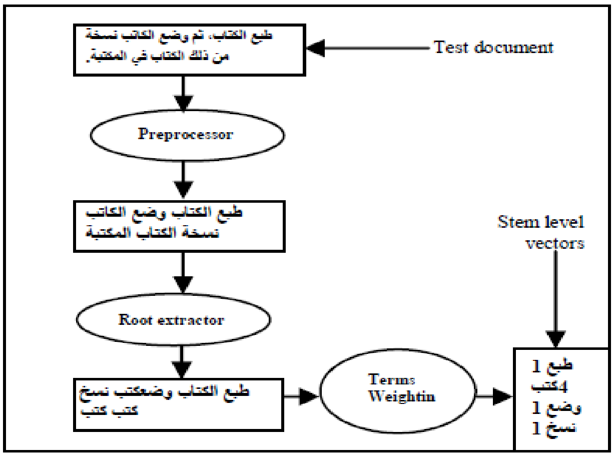


Figure 3 Root-Based Stemmer example

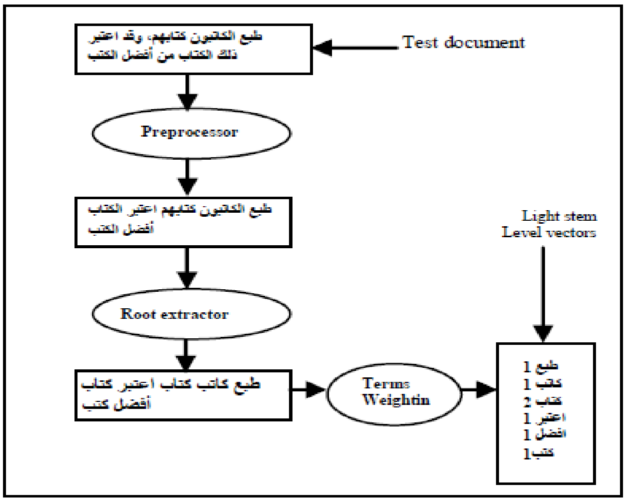


Figure 4 Light-Based Stemmer example

2.DOCUMENT REPRESENTATION

To reduce complexity, document will be converted from full text version to a document vector describes contents of the document [12]

Vector Space Model: In the Vector Space Model, the contents of a document are represented by a multidimensional space vector. Later, the proper classes of the given vector are determined by comparing the distances between vectors.

The procedure of the Vector Space Model can be divided into three stages:
- The first step is document indexing, when most relevant terms are extracted.
- The second stage is based on the introduction of weights associated to index terms in order to improve the retrieval relevant to the user.
- The last stage classifies the document with a certain measure of similarity.

Term Weight: Term weighting is one of pre-processing methods; used for enhanced text document presentation as feature vector. Term weighting helps us to locate important terms in a document collection for ranking purposes [12]. The popular schemes for term weight are Boolean model, Term Frequency (TF), Inverse Document Frequency (IDF), and Term Frequency-Inverse Document Frequency (TF-IDF).

$$TF(d,t_i) = \frac{n(d,t_i)}{\sum_i n(d,t_i)}$$

Where n(d,ti) is the number of occurrences of ti in a document and Σi n(d, ti) is the total number of tokens in document. Inverse document frequency IDF(t) is scale down the terms that occur in many documents.

$$IDF(t_i) = \log(\frac{D}{D_i})$$

Where Di is the number of documents containing ti and D is the total number of documents in the collection.

$$w_{ij} = tfidf(t_i, d_j) = \frac{f_{ij}}{\sqrt{\sum_{k=1}^{M} {f_{kj}}^2}} \times \log\left(\frac{N}{n_i}\right)$$

Where N is the number of documents in the data set, M is the number of terms used in the feature space, fij is the frequency of a term i in document j, and ni denotes the number of documents that term i occurs in at least once. We will apply Term Frequency-Inverse Document Frequency (TF-IDF) pre-processing method to enhance text document presentation as feature vector.

3.3 Clustering algorithm

K-means algorithm follows a simple and easy way to classify a given document set through a certain number of clusters (assume k clusters). The main idea is to define k centroids, one for each cluster. The simple K-means algorithm chooses the centroid randomly from the document set. The next step is to take each document belonging to a given data set and associate it to the nearest centroid. The K-means clustering partitions a data set by minimizing a sum of-squares cost function.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{x} \left\| x_i^{(j)} - c_j \right\|^2$$

Where ||xi(j) - cj||2is a chosen distance measure between a document xi(j)and the cluster center cj, is an indicator of the distance of the n documents from the irrespective cluster centroids.

V. EXPRIMENT AND EVOLUTION

We have applied our experiment by using SANAD corpus and by applying two schemes of stemming: without stemming, and root-based stemming which was Khoja algorithm based on hand coded Kmeans Algorithm to go with our data structure.
The evaluation depends on precision and recall measure.
Experiment environment as follows: operating system: Windows 10, CPU: Intel Core i5 2.0 GHz, Memory:6 GB, NetBeans IDE.

1. Corpora

In this paper freely public data set published by Omar Einea, Ashraf Elnagar and Ridhwan Al Debsi called SANAD, it is a large Single-labeled Arabic News Articles Dataset of textual data collected from three news portals. The dataset is a large one consisting of almost 200k articles distributed into seven categories that we offer to the research community on Arabic computational linguistics. We anticipate that this rich dataset would make a great aid for a variety of NLP tasks on Modern Standard Arabic (MSA) textual data, especially for single label text classification purposes. We present the data in raw form. SANAD is composed of three main datasets scraped from three news portals, which are AlKhaleej, AlArabiya, and Akhbarona. Available freely on: https://data.mendeley.com/datasets/57zpx667y9.

Table 1 Distribution of articles per category.

Label	*AlArabiya*	*Akhbarona*	*AlKhaleej*
Finance	30,076	9,280	6,500
Sports	23,058	15,377	6,500
Culture	5,619	6,746	6,500
Tech	4,411	12,199	6,500
Politics	4,368	13,979	6,500
Medical	3,715	12,947	6,500
Religion	–	7,522	6,500
Total	71,247	78,050	45,500

2. Evolution

There are many evaluation standard in information retrieval used in document clustering such as Entropy, cluster purity, and F-measure which will be used in this paper.

Precision shows how many documents are in right cluster with respect to the cluster size. Recall shows how many documents are in the right cluster with respect to total documents. Precision and recall for class ***i*** and cluster ***j*** is defined as:

$$Recall\ (i,j) = n_{ij} / n_j$$

$$Precision\ (i,j) = n_{ij} / n_i$$

Where ***nij*** is the number of documents with class label ***i*** in cluster ***j***, ***ni*** is the number of documents with class label ***i***, and ***nj*** is the number of documents in cluster ***j***, and ***n*** is the total number of documents.

3. Results

In this section we will discuss the results, as mentioned above the data sets are: SANAD and we use only part of it *Akhbarona*, each text document belongs 1 of to 7 categories and we used hand coded Kmeans Algorithm to apply clustering and we have used Vector Space Model of IR to represent documents. We compute precision, recall for two cases the first without stemming, the second with root-based stemming (Khoja). The results is depicted in table 2 which shows the results of average precision ,recall, for using two types of stemming in SANAD Arabic corpus dataset for 7 cluster in this dataset.

Table 2 Evolution Result

Stemming Type	precision	recall
Without stemming	0.50	0.522
Root-based stemming	0.494	0.480

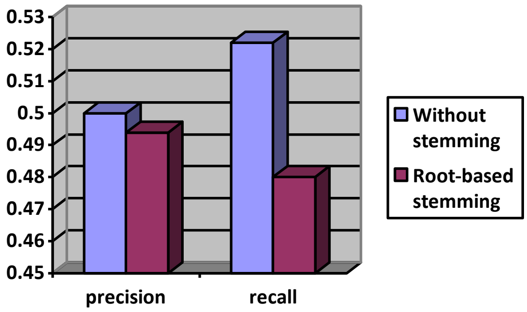


Figure 5 Evolution Results

As Figure (5) and Table 2 above we have found that using no stemmer is better than using root-based stemmer in clustering Arabic Documents process.

VI. Conclusion

In this paper we have applied feature selection methods and stemming techniques for Arabic text clustering. The data set was collected and classified manually into seven clusters: Finance

Sports, Culture, Tech, Politics, Medical, Religion. The testing dataset is *Akhbarona* consists of 7850 documents. Two stemming techniques have been used: without stemming which remains all terms and root-based (Khoja) stemming which removes words have the same root. K-means was used to cluster the test documents; it was run for each technique of stemming individually. The experiments depicted that Without Stemming is the best technique for feature selection in Arabic language document clustering, but root based stemming get deterioration results for Arabic language document clustering; because Arabic language has a complex morphology, and it is a highly inflected language.

VII. References

[1] Alsmadi, I., Alhami, I., 2015. Clustering and classification of email contents. J. King

[2] Jain, A.K., Dubes, R.C., 1988. Algorithms for Clustering Data. Prentice Hall

[3] Ahmed and Tiun, 2014; Al-Omari, 2011.

[4] Al-Anzi and AbuZeina, 2015,2016;

[5] Ghanem (2014)

[6] Bsoul and Mohd (2011)

[7] Al-Shammariand Lin, 2008

[8] Awajan, 2015a,2015b.

[9] by Al-sarrayrih and Al-Shalabi (2009)

[10] Amine et al. (2013)

[11] A. A. Al-Harbi S., Al-Thubaity A., Khorsheed M., Al-Rajeh A. "Automatic Arabic Text Classification" presented at the 9es Journéesinternationals, France, 2008.

[12] A. Ghanem and M. Ashour, Stemming Effectiveness in Clustering of Arabic Documents, Palestine, Gaza.

[13] Hanan M. Alghamdi a, Ali Selamat, Arabic Web page clustering: A review, Umm Al-Qura University, Alqunfudah, Saudi Arabia