Stemming Effectiveness in Clustering of Arabic Documents

Osama A. Ghanem
University Collage of Applied Science
Palestine, Gaza

Wesam M. Ashour
Islamic University of Gaza
Palestine, Gaza

ABSTRACT
Clustering is an important task gives good results with information retrieval (IR), it aims to automatically put similar documents in one cluster. Stemming is an important technique, used as feature selection to reduce many redundant features have the same root in root-based stemming and have the same syntacticalform in light stemming. Stemming has many advantages it reducesthe size of document and increases processing speed and used in many applications as information retrieval (IR). In this paper, we have evaluatedstemmingtechniques in clustering of Arabic language documents and determined the most efficient in pre-processing of Arabic language,whichis more complex than most other languages. Evaluation used three stemming techniques: root-based Stemming, light Stemming and without stemming. K-means, one of famous and widely clustering algorithm, is applied for clustering.Evaluation depends on recall, precision andF-measure methods. From experiments, results show that light stemming achieved best results in terms of recall, precision and F-measure when compared with others stemming.

KEYWORDS
Arabic text clustering, Stemming, light stemming, K-means.

1. INTRODUCTION
Most researches discussed techniques of information retrieval (IR) in English and Europe languages, but there are few researches discussed stemming 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 categorization, the process of assigning unseen documents to user-defined categories, is an important task that can help in the organization and querying of such collections [1][2].

Document clustering has become an increasingly important technique for enhancing search engine results, web crawling, unsupervised document organization, and information retrieval or filtering. Clustering involves dividing a set of documents intoa specified number of groups. The documents within each group should exhibit a large degree of similarity while the similarity among different clusters should be minimized. Some of the more familiar clustering methods are: partitioning algorithms based on dividing entire data into dissimilar groups, hierarchical methods, density and grid based clustering, some graph based methods and etc. (Jain et al. 1999; Grira et al.2005).

Partitioning methods try to partition a collection of documents into a set of groups, so as to maximize a pre-defined fitness value. The clusters can be overlapped or not. The best known partitioning algorithm is K-means (McQueen 1967) that, in a simple form, selects K documents as cluster centers and assigns each document to the nearest center. The updating and reassigning process can be kept until a convergence criterion is met. This algorithm can be performed on a large data set almost in linear time complexity. [3]

In other hand we will discuss clustering of Arabic documents using many stemming techniques. Stemming is a method, that reduces words have the same stem or root. It is essential to improve performance in information retrieval tasks especially with highly inflected language like Arabic language. The stemming process reduces the size of the documents representations by 20-50% compared to full words representations [4] which also assist to improve the retrieval for documents. Three techniques of stemming will be discussed: root-based stemming, light stemming and without stemming.

The rest of paper is organized as follows. The next section presentsrelated works; section 3describes in detailsour methodology; section 4 presents experiment design and evaluation; section 5 discusses results, finally section 6 concludes.

2. RELATED WORKS
Document Classification was discussed more widely than document clustering.In clustering there are many techniques can be used for categorize documents automatically such as partitional, density and hierarchal. Stemming has impact in clustering of Arabic language documents; it may enhance clustering process, which depends in stemming type and the language of documents.

(Yoo and Hu) [5]Performed a comprehensive comparison study of variousdocument-clustering approaches, such as K-means and Suffix Tree Clustering in terms of theefficiency, the effectiveness, and the scalability. They found that the partitional clustering algorithmsare the most widely used algorithms in document clustering.

In other hand,(Sandhya, Lalitha, V.Sowmya, Anuradha and Govardhan) [6] have studied the impact of stemming algorithm along with four similarity measures (Euclidean, cosine, Pearson correlation and extended Jaccard) in conjunction with different types of vector representation (boolean, term frequency, and term frequency and inverse document frequency) on cluster quality. They have used K Means for Clustering documents, and concluded that there are four components that affect the resultsrepresentation of the documents: applying the stemming algorithms, distance or similarity measures considered, and the clustering algorithm itself.

(Singh and Garg) [7] Have experimented English language document clustering with different representations (tf, tf.idf&

Boolean) and different feature selection schemes (with or without stop word removal & with or without stemming). The results indicate that (tf.idf) representation, and use of stemming obtains better clustering.

3. METHODOLOGY

Our interesting approach in this paper is clustering of Arabic documents. So we will apply clustering technique using K-means algorithm, which is widely used for document clustering, in Arabic documents to study impact of stemming typesin Arabic document clustering. Stemming process falls in the stage: Arabic text pre-processing (Step two in Figure 1). We will discuss the system architecture of clustering documents and concentrateon stemmingprocess, which is our problem statement, three types of stemming will be discussed: without stemming, root-based stemming and light stemming (as shown in Figure 2).

To use Arabic documents in clustering algorithm; as shown in Figure 1 there are fivetasks are necessary for clustering Arabic text:

1. Collect Arabic textdocuments
2. Arabic text pre-processing
3. Document representation
4. Apply k-means algorithm
5. Evaluation

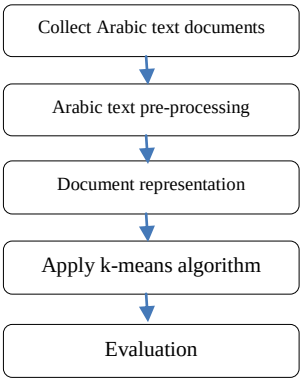

Figure 1: Arabic document clustering architecture

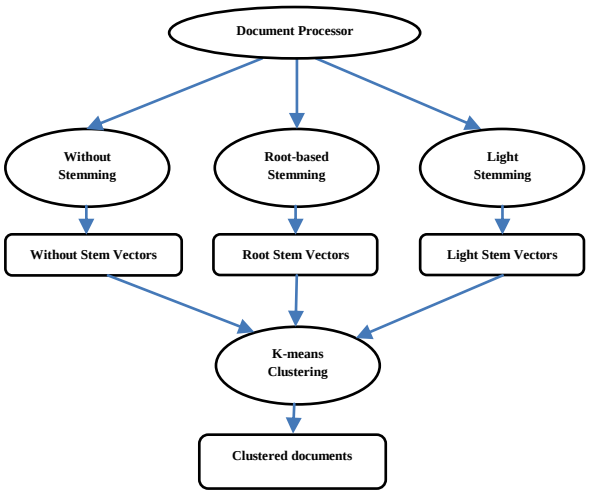

Figure 2:Stemming System Architecture

In Figure 2 is depicted stemming system of Arabic document clustering. In ourmodule three stemming techniques are applied to Arabic documents:without stemming, root-based stemming, and light stemming, and then apply K-means clustering algorithm for representing clustered documents to be evaluation.

3.1 Arabic text pre-processing

Arabic Language is one of the widely used languages in the world. 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 [8].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 discuss this technique in details.

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 [9], and the work of Elkourdi, Bensaid and Rachidi [10], and Duwairi [11]. Applying stemming algorithms as a feature selection method reduces the number of features since lexical forms (of words) are derived from basic building blocks; and hence, many features that are generated from the same stem are represented as one feature (their stem).[12]

Two types of stemming will be applied to Arabic documents in addition towithout stemming type:

a. Root-based Stemming

Stemming using root extractor which uses morphological analysis for Arabic words,Figure 2 depicts an example of using stemming for feature selection. Note that several words such as (الكتاب الكاتب المكتبة) which mean "the library", "the writer" and "the book" respectively are reduced to one stem (كتب) which means write[13] as shown in figure 3 [9] which describes preprocessing steps in root based stemming. 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 Stemming Algorithm which will be used in our experiments.

Khoja Stemming Algorithm: ShereenKhoja's developed stemmer algorithms [13]. The algorithm, developed by using both Java and C++ languages, removes the longest suffix and the longest prefix. It then matches the remaining word with verbal and noun patterns, to extract the root. The stemmer makes use of several linguistic data files such as a list of all diacritic characters, punctuation characters, definite articles, and 168 stop words. The algorithm achieves accuracy rates of up to 96%. The algorithm correctly stems most Arabic words that are derived from roots.

b. Light stemming

The main idea for using light stemming is that many word variants do not have similar meaningsor semantics. However; these word variants aregenerated from the same root. Thus, root extractionalgorithms affect the meanings of words. Lightstemming by comparison aims to enhance thecategorization performance while retaining the words'meanings. It removes some defined prefixes

andsuffixes from the word instead of extracting the originalroot [14].Light-stemming keeps the word's meanings unaffected. Figure4 demonstrates an example of using light stemming. Here we note that light stemming maintains the difference between (الكاتبون الكتاب) which means "the book" and "the writers" respectively; their light stems are (كاتب كتابي) which means book and writer. [12]

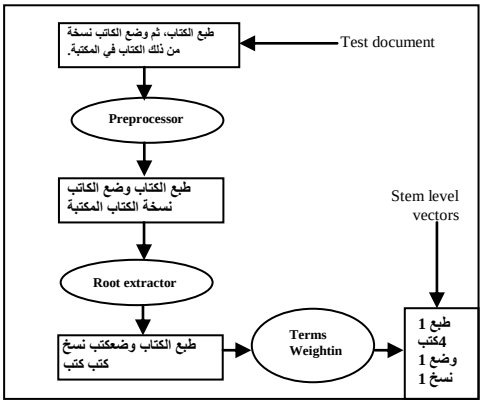

Figure 3: preprocessing withroot-based stemming

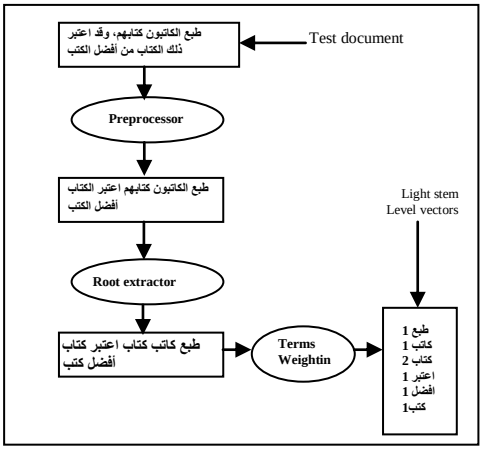

Figure 4: preprocessing with light stemming

3.2 Document representation

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

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 inorder 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 [15]. The popular schemes for term weight are Boolean model, Term Frequency (TF), Inverse Document Frequency (IDF), and Term Frequency-Inverse Document Frequency (TF-IDF).

$$TF(d,t_i) = \frac{n(d,t_i)}{\sum_i n(d,t_i)} \qquad (1)$$

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

$$IDF(t_i) = \log(\frac{D}{D_i}) \qquad (2)$$

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

$$w_{ij} = tfidf(t_i,d_j) = \frac{f_{ij}}{\sqrt{\sum_{k=1}^{M} f_{kj}{}^2}} \times \log\left(\frac{N}{n_i}\right) \qquad (3)$$

Where *N* is the number of documents in the data set, M is the number of terms used in the feature space, f_{ij} is the frequency of a term *i* in document *j*, and n_i 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 methodto enhance text document presentation as feature vector.

3.3K-means Clustering algorithm

K-means algorithm is used in our experiments to get the best clustering results.It 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-meansclustering partitions a data set by minimizing a sumof-squares cost function.

$$\boldsymbol{J} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{k}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{x}} \left\| \boldsymbol{x_i^{(j)}} - \boldsymbol{c_j} \right\|^{\boldsymbol{2}} \qquad (4)$$

Where $||x_i^{(j)} - c_j||^2$is a chosen distance measurebetween a document $x_i^{(j)}$and the cluster centerc_j, is anindicator of the distance of the *n* documents from theirrespective cluster centroids.[16]

3.4Document Clustering tool (WEKA)

WEKA is a data mining open-sourcetool in abroad, but it is rarely used at home. We provide documents preprocessing, and apply K-means algorithm in the Arabic document clustering by adjusting the parameters in WEKA.

WEKA (Waikato Environment for Knowledge Analysis) is a famous with data mining software and is well received in abroad [17]. For example, lots of document clustering and document categorization experiments have been carried out using 20 Newsgroups and Reuters-21578 corps based on

WEKA [3]. The main functions of document clustering in WEKA include 3 aspects as below: (1) Convert directory structure to arff file. (2) Convert string attributes into a set of attributes representing word occurrence. (3)11 clustering algorithm. WEKA is an open-source software, researchers can modify or add new algorithm when they needed. [18].

Text Preprocessing Tools: WEKA provides *StringToWordVector* tool. This tool converts String attributes into a set of attributes representing word occurrence (depending on the tokenizer) information from the text contained in the strings. The set of words (attributes) is determined by the first batch filtered (typically training data).There are many options for *StringToWordVector:*Boolean, TF transform,IDF transform, TFIDF transform, and minTermFreq.WEKA tools provided different algorithms that supported Arabic language show. In this paper we firstly applied Clustering algorithm without stemmer, after that applied Light stemmer and finally, Khoja stemmer. WEKA was rarely used in Arabic text processing and at home, so we used it in our research in Arabic text processing and at home and K-means algorithm is used in document clustering because of advantages we mentioned.

4. EXPERIMENT DESIGN AND EVALUATION

Experiments are applied by using Arabic BBC Arabic corpus and by applying three schemes of stemming: without stemming or row text stemming, light stemming, and root-based stemming which was Khoja algorithm based on WEKA tool whichis used at home, and the evaluation depends onprecision, recall, and F-measure. Experiment environment as follows: operating system: Windows 7, CPU: Intel Core i7 Q720 1.6 GHz, Memory:8 GB, WEKA version: 3.6.4.

4.1 Corpora

One of difficulties for Arabic language is the lack of publicly available Arabic corpus for evaluating text categorization algorithms [18]. In other side, English language has different public data set for English text clustering. Reuter's collections of news stories are popular and typical example.

In this paper freely public data set published by Saad inhttp://sourceforge.net/projects/ar-text-mining, is used for experiment. The dataset collected from BBC Arabic website for several reasons, because its free and public, contains suitable number of documents for the clustering process. BBC Arabic dataset used in experimentation and has different domains.

BBC Arabic corpus:BBC Arabic corpus is collected from BBC Arabic website bbcarabic.com, as shown in table 1; the corpus includes 4,763 text documents. Each text document belongs 1 of to 7 categories (Middle East News 2356, World News 1489, Business & Economy 296, Sports 219, International Press 49, Science & Technology 232, and Art & Culture 122). The corpus contains 1,860,786 (1.8M) words and 106,733 district keywords after stopwords removal. We converted the corpus to utf-8 encoding and stripped html tags. The corpus is available publically at. [20].

Table 1.Number of documents in each category of BBC testing data set

Text Categories	Number of documents
Middle East News	2356
World News	1489
Business & Economy	296
Sports	219
International Press	49
Science & Technology	232
Art & Culture	122
Total	**4,763**

4.2 Evaluation:

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.

F-measure: The F-measure is a harmonic combination of the precision and recall values used in information retrieval [21].

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:

$$Recall\ (i,j) = \frac{n_{ij}}{n_j} \qquad (5)$$

$$Precision\ (i,j) = \frac{n_{ij}}{n_i} \qquad (6)$$

Where$n_{ij}$is the number of documents with class label ***i***in cluster ***j*** ,$n_i$is the number of documents with class label ***i,*** and $n_j$is the number of documents in cluster***j*** , and ***n***is the total number of documents. The F-measure for class ***i***and cluster ***j***is given as:

$$F(i,j) = \frac{2 * Recall(i,j) * Precision\ (i,j)}{Recall(i,j) + Precision\ (i,j)} \qquad (7)$$

Then total F-measure of clustering process is calculated as:

$$F = \sum {}^{n_i}/_{n * maxF(i,j)} \qquad (8)$$

5. RESULTS AND DISCUSSION

In this section we will discuss the results, as mentioned above the data sets are: BBC Arabic corpus which includes 4,763 text documents, each text document belongs 1 of to 7 categories and we used WIKA as clustering tool. We compute precision, recall and F-measure for three cases the first without stemming, the second with light stemming, and finally with root-based stemming (Khoja). The results is depicted in table 2 which shows the results of average precision ,recall, and F- measure for using three types of stemming in BBC Arabic corpus dataset for seven clusters in this dataset and 4,763 documents.

Table 2. Evaluation results for three types stemming

Stemming Type	precision	recall	F-measure
Without stemming	0.6232	0.6008	0.6119
Light stemming	0.7496	0.7054	0.7269
Root-based stemming	0.5420	0.5364	0.5392

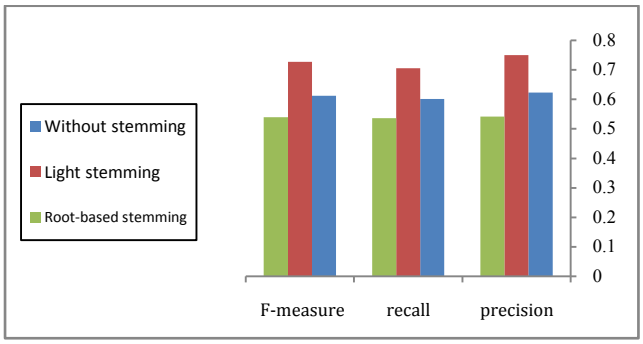

Figure 5: precision, recall and F-measure values for without stemming, light stemming and root-based stemming

Figure 5shows the precision, recall and F-measure for three stemming cases: without stemming has values 0.6, 0.6 and 0.61 for precision, recall and F-measure respectively, the second type is light stemming and has values: 0.75, 0.7 and 0.72 for precision, recall and F-measure respectively, and the last type is root-based stemming and has values: 0.54, 0.53 and 0.54 for precision, recall and F-measure respectively. From results light stemming gets the best measurement values versus without stemming and root-based stemming in Arabic document clustering, because Arabic language has a complex morphology languages, and it is a highly inflected language, so root-based stemming gives backfire in clustering documents, but light stemming gives enhancement in clustering documents.

6. 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: Middle East News, World News, Business & Economy, Sports, International Press, Science & Technology, and Art &Culture.The testing dataset consists of 4,763 documents. Three stemming techniques have been used: without stemming which remains all terms, light stemming which removes common suffixes and prefixes, 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 Light 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.

7. REFERENCES

[1] S. Ghwanmeh. "Applying Clustering of Hierarchical K-means-like Algorithm on Arabic Language", International Journal of Information and Communication Engineering 3:7 2007.

[2] Y. Fang, S. Parthasarathy, and F. Schwartz, "Using Clustering to Boost Text Classification", in Proc. of the IEEE International Conference on Data Mining, California, USA, 2001, pp. 123-127.

[3] M. Mahdavi,H. Abolhassani. "Harmony K-means algorithm for document clustering". Data Min Knowl Disc, 2008.

[4] A. A. B. Sembok T., Abu Bakar Z. "A Rule and Template Based Stemming Algorithm for Arabic Language," INTERNATIONAL JOURNAL OF MATHEMATICAL MODELS AND METHODS IN APPLIED SCIENCES,Issue 5, Volume 5, pp. 974-981, 2011.

[5]IllhoiYoo, XiaohuaH."Semantic Text Mining and its Application in Biomedical Domain." , 2006

[6] N. Sandhya1, Y. Sri Lalitha2,V.Sowmya3, Dr. K. Anuradha4 and Dr. A. Govardhan5." Analysis of Stemming Algorithm for Text Clustering", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 1, September 2011

[7] Vivek Kumar Singh, NishaTiwari, ShekharGarg."Document Clustering using K-means, Heuristic K-means and Fuzzy C-means", International Conference on Computational Intelligence and Communication Systems, 2011.

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

[9] Sawaf H, Zaplo J. and Ney H. "Statistical Classification Methods for Arabic News Articles", Presented at the Arabic Natural Language Processing Workshop; 2001 July; Toulonse, France.

[10] Elkourdi M, Bensaid A, Rachidi T. "Automatic Arabic Document Categorization Based on the Naive Bayes Algorithm", Proceedings of COLING 20th Workshop on Computational Approaches to Arabic Script-based Languages; 2004 Aug; Geneva, Switzerland.

[11] Duwairi R. "A Distance-based Classifier for Arabic Text Categorization", International Conference on Data Mining (DMIN05); 2005 Jun; Las Vegas, USA.

[12] R. Duwairil, M. Al-Refai, N. Khasawneh . "Stemming Versus Light Stemming as Feature Selection Techniques for Arabic Text Categorization", 2007.

[13] A. E. S. Sawalha M., "Comparative evaluation of Arabic language morphological analysers and stemmers".Presented at the Proceedings of COLING 2008 22nd International Conference on Comptational Linguistics, COLING 2008, 2008.

[14] Aljlayl, M. and Frieder, O. "On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach", ACM Eleventh Conference on Information and Knowledge Management; 2002 November 340-347; Mclean, VA, USA.

[15] G. C. Qiu Z., Doherty A.R., Smeaton, A.F., "Term Weighting Approaches for Mining Significant Locations from Personal Location Logs," presented at the ProceedingsinCIT(2010), 2010.

[16] M. Shameem, R. Ferdous. "An efficient K-Means Algorithm integrated with Jaccard Distance Measure for

Document Clustering", Internet, 2009. AH-ICI 2009.First Asian Himalayas International Conference on.

[17]Hall M, Frank E, Holmes B, "The WEKA data mining software: an update", ACM SIGKDD Explorations Newsletter, Vol 11, No.1, pp. 10-18, 2009.

[18] PU HAN, DONG-BO WANG, QING-GUO ZHAO. "THE RESEARCH ON CHINESE DOCUMENT CLUSTERING BASED ONWEKA ". Proceedings of the 2011 International Conference on Machine Learning and Cybernetics, Guilin, 10-13 July, 2011.

[19] K. G. Al-Shalabi R., Gharaibeh M., "Arabic Text Categorization Using kNN Algorithm," presented at the Proceedings of the Int. multi conf. on computer science and information technology, 2006.

[20] Motaz K. Saad, "Open Source Arabic Language and Text Mining Tools", (2010, August), [Online]. Available: http://sourceforge.net/projects/ar-text-mining

[21] C.J. van Rijsbergen, Information Retrieval, 2nd ed., Buttersworth, London, 1979.