Keyphrase-Based Hierarchical Clustering for Arabic Documents

Moufeda Hussein
Faculty of Engineering Shoubra
Benha University, Egypt
+201224656312
mofida.mahmoud@feng.bu.edu.eg

Abdelwahab Alsammak
Faculty of Engineering Shoubra
Benha University, Egypt
+201220588666
asammak@feng.bu.edu.eg

Tarek Elshishtawy
Faculty of Computers and Informatics
Benha University, Egypt
+20123454723
t.el-shishtawy@ictp.edu.eg

ABSTRACT
The vast amount of available Arabic web pages and text files on the internet makes it necessary to organize data in an easy way for user browsing. Document clustering is a good solution for this problem, which groups similar data into clusters with meaningful labels. In this paper, we propose a domain independent approach, which builds a hierarchical meaningful clustering tree. The proposed approach overcomes the problem of high dimensionality of feature vector by representing each document with its keyphrases. In addition, we introduced a new similarity measure by taking the common lemma form keyphrases among feature vectors of documents. This improves the accuracy of the clustering process with reduced complexity. Many experiments are carried out to evaluate the accuracy of clustering using String-based, Corpus-based, and Knowledge-based similarity measures. A dataset consists of 345 Arabic documents and covering 12 domains is used in these experiments. The results show that applying lexical similarity using keyphrase based gives more accurate clusters labels than using semantic similarity. The best purity result achieved is 0.955, which is obtained using the common lemma form keyphrases similarity method.

CCS Concepts
Information systems → Information retrieval → Retrieval tasks and goals → Clustering and classification

Keywords
Agglomerative Hierarchical document clustering; Keyphrase; Lemma; Lexical similarity; Semantic similarity

1. INTRODUCTION
Textual Data Mining is the process of discovering useful knowledge from a large collection of text documents. How to explore and navigate the large amount of text documents is a challenging task. Document clustering is one of the most important text mining methods that are developed to help users effectively navigate, and organize text documents. Document Clustering is the process of organizing documents with similar content into meaningful groups, formally known as clusters, depending on the degree of similarity between them. This can be achieved by using different similarity measure methods and clustering algorithms. The documents in the same cluster have the largest similarity between each other. In other words, the documents in one cluster share the same topic, and the documents in different clusters represent different topics. Unlike categorization, in which documents are assigned to predefined categories, clustering does not have any predefined categories. Document clustering is an unsupervised process.

The two major methods of clustering are partitioning and hierarchical clustering [6]. Partitioning clustering produces a set of clusters all belonging to the same level, while hierarchical clustering produces a tree which recursively divides the document space into related clusters. Hierarchical clustering helps user to access large dataset in the form of structured data, also it provides understandable and meaningful topics and subtopics labels for a group of documents. In addition, hierarchical clustering provides an overview of the important idea and content of large documents.

The hierarchical clustering techniques are divisive and agglomerative [18]. Divisive algorithm (Top-Down) starts with all documents in the one cluster and iteratively splits a cluster into smaller clusters until a certain termination condition is satisfied. However, the agglomerative clustering algorithm (Bottom-Up) starts with all documents belonging to their individual clusters and merge the most similar clusters until reach to one cluster contain all documents [19].

In this paper, an agglomerative hierarchical clustering system is introduced to organize unlabeled Arabic documents in meaningful topics and subtopics clusters. The proposed system applies more than one similarity techniques such as lexical and semantic ones to improve the performance of clustering of Arabic documents.

2. AGGLOMERATIVE HIERARCHICAL CLUSTERING

Agglomerative Hierarchical Clustering is a process of cluster analysis which aims to build a hierarchy of clusters, each document starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy [10]. The parent-child relationship among the nodes in the tree hierarchy can be viewed as a topic-subtopic relationship in a subject hierarchy, it starts with all documents belonging to their individual clusters and combines the most similar clusters until reach to one cluster contain all documents. This can be achieved by iterative computing the similarity between all pairs of clusters and then merging the most similar pair [9].

Linkage is the method that links documents to form clusters. It measures the inter-cluster distance and groups the documents that have maximum intra-cluster similarity and minimum inter-cluster similarity. The linkage method uses the similarity matrix as an input to determine which documents cluster together. Iteratively it merges the most similar clusters until reach to one cluster contain all documents. We used seven linkage methods, which will be illustrated, in the following subsections.

2.1 Single linkage

It is one of the most famous and oldest clustering method and also known as nearest neighbor clustering. The single linkage joins the two clusters which have the maximum similarity between them and creates a new cluster of them [4]. To calculate the similarity between the new cluster and the others, if at one-step clusters i and j have merged and create cluster m, then the similarity of cluster m and any other cluster q is determined as follows:

$$\mathrm{sim}_{\mathrm{mq}} = \max(\mathrm{sim}_{\mathrm{iq}}, \mathrm{sim}_{\mathrm{jq}})$$

2.2 Complete linkage

It is also known as farthest neighbor clustering. First, the two clusters having the highest similarity measure between them are merged together. Then, the similarity between the new cluster and the others is calculated. If at one-step clusters i and j have merged and create cluster m, then the similarity of cluster m and any other cluster q is determined as follows:

$$\mathrm{sim}_{\mathrm{mq}} = \min(\mathrm{sim}_{\mathrm{iq}}, \mathrm{sim}_{\mathrm{jq}})$$

As the document cannot join any cluster until it has a similarity level with all documents in this cluster, the probability of merge new document to a cluster become smaller as the number of documents in the cluster increases [20].

2.3 Average linkage

It is also known as the Average neighbor clustering. After merging the two clusters having the highest similarity, the new cluster similarity with others can be calculated according to the following equation:

$$sim_{mq} = \mathrm{avg}(sim_{iq}, sim_{jq})$$

2.4 Centroid

The centroid of a cluster is a point whose parameter values are the mean of the parameter values of all the points in the clusters. Join the two clusters with the closest centroid.

2.5 Ward

It uses an analysis of variance approach to evaluate the distances between clusters. Join the two clusters that will produce the smallest increase within-cluster variation. The new cluster similarity with the others can be calculated by the sum of squares of distance within the cluster.

2.6 Median

The same as the centroid, except that equal weighting is used to construct the centroid of the joined clusters. The distance between the new cluster and the others equals to the weighted distance between their centroids.

2.7 Weighted pair group method with averaging (WPGMA)

In Weighted pair group method with averaging, the distance between clusters is calculated as a simple average distance between each of the cluster members.

3. PROPOSED SYSTEM

The main objective of this paper is to improve the performance of clustering of Arabic documents. The improvement is aimed to reduce the processing time and increase the results accuracy. To achieve this objective, we adopt a new vector space model to represent the text document. Each text document is represented by its keyphrases. In addition, the common lemma forms of keyphrases for documents are used to measure the similarity between them. This reduces significantly the complexity of the similarity process by reducing the size of feature vector of each document and simplifying the similarity algorithm. Then a linkage algorithm is applied to the similarity matrix to group similar documents together and builds a hierarchy tree called dendrogram tree. The obtained tree is cut off into clusters by making use of the shared words of the feature vectors. Each cluster is labeled by the common keyphrase between feature vectors within the cluster. Finally, the hierarchy tree for each cluster is built starting from the shared keyphrase as a root and up to three levels only. Figure 1 shows the framework of the proposed system.

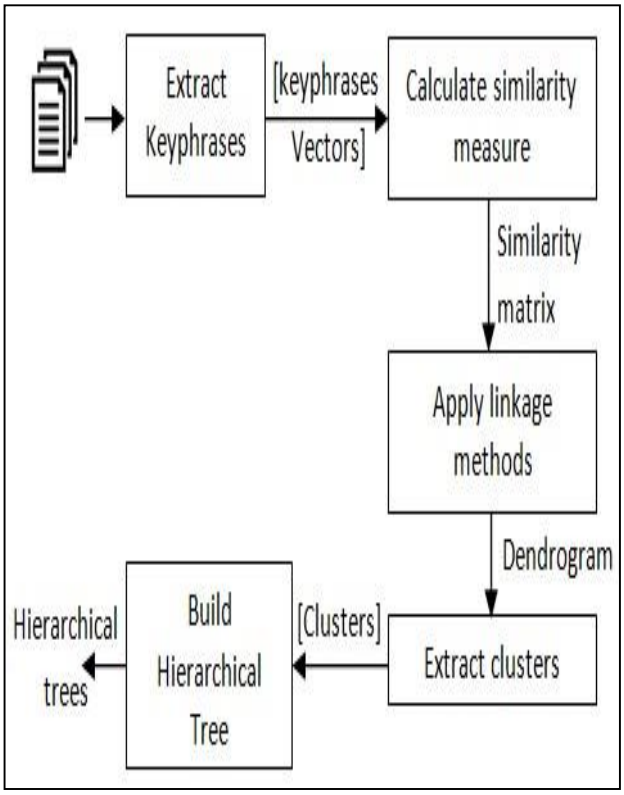

Figure 1. The framework of the proposed system

3.1 Keyphrase Extraction

Each document is being converted from the full text version to a weighted feature vector, which makes the documents easier to handle and decreases their complexity. Each text document is represented by its keyphrases. A keyphrase is defined as a meaningful and significant expression consisting of one or more words in a document. Keyphrases offer a brief summary of document content. Using the keyphrase-based model for representing document overcomes the drawbacks of using word-based model such as, high dimensionality and ignoring the syntactic structure of the word. In addition, it overcomes the drawback of the term-based model, as not all phrases have a distinct meaning. The keyphrases are extracted using AKE [2].

The similarity between documents could be measured using the shared keyphrases represented in their lemma forms. The lemma is the abstract form, which describes the canonical form of a set of words. It refers to the set of all word forms that have the same meaning. Table 1 describes the elements of the weighted feature vector that will be used to represent each document.

Table 1. Features representing a document

Feature	Description
kp_j	Set of Keyphrases in document j
lk_j	Lemma form of keyphrases in document j
w_{kj}	Weight of Keyphrase k in document j
t_{kj}	POS tag of the keyphrase k in document j

The weights of the extracted features depend on the document size. This will be misleading for the process of measuring similarity between documents by comparing their feature vectors. Thus, the weights of the extracted features of each vector will be normalized to be in the range between [0, 1] using the following equation:

$$normailzed\ weight = \frac{current\ keyphrase\ weight}{\sqrt{\sum(keyphrase\ weight)^2}}$$

3.2 Similarity Measure

Measuring the similarity between documents is an important role in document clustering, which helping clusters closed documents together. Its operation is to compare two weighted feature vectors and compute a single number, which indicates their similarity to each other. In this process, a similarity matrix of similarity between each pair of the weighted feature vectors is being built.

There are two categories to measure the similarity between text: lexically and semantically [3]. The lexical similarity depends on the word's character sequence without any lexicon dictionary. Lexical similarity is shown through string based measures which operates on the string surface's form. However, semantic similarity is a context dependent. It is shown through knowledge based and corpus based. The knowledge based similarity uses semantic networks such as WordNet lexical database to determine the degree of semantic similarity between words. The semantically similar words may be substituted for each other in context. The corpus based determines the semantic similarity between words using large corpora and compute the frequencies of co-occurrences of the term in the feature vector. The ways of similarity measure used are shown in figure 2.

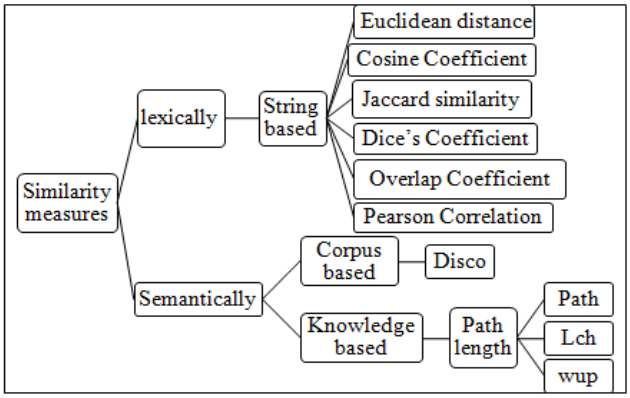

Figure 2. The ways of similarity measures

3.2.1 String Based Similarity

It operates on string sequences. It is a statistical calculation depending on comparing each two feature vectors and computing a single number which indicates their similarity to each other [5]. Many algorithms are used to measure the string similarity such as Euclidean distance, Cosine coefficient, Jaccard, Dice's coefficient, Overlap coefficient, and Pearson correlation.

In addition to the existing similarity measures, we will measure the similarity between vectors by counting the shared keyphrases represented in their lemma forms between vectors. This method decreases the processing time to measure similarity.

3.2.2 Corpus based similarity

Corpus based similarity is a semantic similarity. It measures the similarity between the weighted feature vectors using the information gained from a corpus. We apply Extracting DIStributionally related words using CO-occurrences (DISCO) as the corpus similarity. DISCO is a method that computes distributional similarity between words using Arabic Wikipedia and Ajder Corpora [11; 12].

We applied Disco similarity between all keyphrases represented in their lemma form in each two weighted feature vectors. We compared the occurrence between each two phrases using the Arabic Disco corpus. However, it takes a long time to compute the similarity. To overcome this problem, we measured the similarity between uni-gram lemma keyphrases, between uni-gram lemma and bi-gram lemma contained this uni-gram lemma, and finally between bi-gram lemma and tri-gram lemma contained this bi-gram lemma. The similarity obtained is faster and enhance the clustering results than using disco similarity only.

3.2.3 Knowledge based similarity

Knowledge based similarity is a semantic similarity. It is based on semantic network to determine the degree of similarity between words. One of the most popular semantic networks is WordNet (WN) for English language [14] and Arabic WordNet (AWN) for Arabic language [1; 17].

3.2.3.1 Using Arabic WordNet

The knowledge based similarity measure does not depend on lexical features only, but also using semantic features. Arabic WordNet (AWN) database is used to add semantic features to the weighted feature vector of each document. We add the synonyms of each uni-gram lemma to the weighted feature vector such as the synonyms of "علاج" are "معالجة", "دواء", and "مداواة ". However, as the database of the AWN is small and not all uni-gram lemma has available synonyms and some words have limited synonyms such as the synonyms of "سرطان" are only "سلطعون" and " ابوجلمبو "

neglecting the synonyms of cancer disease. Therefore, the current synonyms of AWN cannot be relied upon as semantic features.
Instead of using synonyms of words, we used the hyponym semantic relation between synsets (is-a relationship) as the semantic features. Actually, the hyponym relations for each uni-gram take the form of a chain with a certain root. As several noun words have the word "كينونة" as the root of the chain, which increases the similarity value between dissimilar words. To overcome this drawback, we take only the two parents of a word in the hyponym chain. However, not all uni-gram words have available hyponym chains.

3.2.3.2 Using WordNet

Rather than using AWN, we use WordNet (WN) as its database is larger than AWN. The steps for measuring similarity using WN are:

1. Arabic keyphrases of each document are translated into English excluding the stop words.
2. The lemma form of each word is extracted and added to the feature vector of the document.
3. The similarity between each two feature vectors is calculated using the path, wup, and lch methods from the WN [13; 21].

3.3 Building the Dendrogram Tree

The purpose of this stage is to group similar documents together and build a hierarchy tree called dendrogram tree. We will apply seven linkage methods during building the dendrogram tree and compare their results. The used linkage methods are single, complete, average, centroid, ward, weighted, and median.

3.4 Clusters Extraction

The tree of hierarchical clusters neither gives labels of the clusters nor number of clusters obtained. Therefore, in this stage the obtained tree is cut off into clusters by checking the tree from bottom to up. If there is a group of feature vectors having shared words, they belong to one cluster else, it is a new cluster and so on. To get a meaningful cluster label, we determine the shared keyphrases between the weighted feature vectors within the cluster. Samples of clusters obtained are shown in figure 3, which illustrates the label of each cluster and its constituent documents.

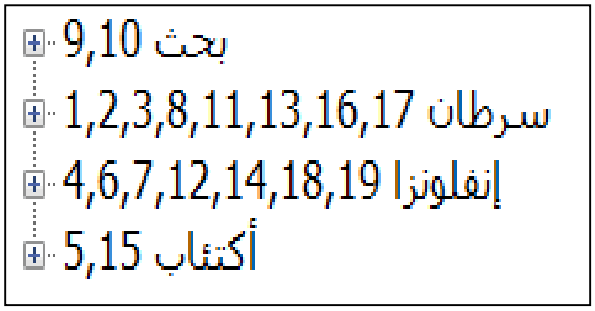

Figure 3. Sample of obtained clusters

3.5 Building Hierarchy Tree for Each Cluster

The final step is to build the hierarchy tree for each cluster using the shared keyphrase represented in its lemma form as a root of the tree. Each tree consists of three levels; the shared uni-gram keyphrase level, the bi-gram keyphrases level, and the tri-gram keyphrases level.

The bi-gram keyphrases level contains the keyphrases that consists of two words; one of them is the shared uni-gram keyphrase. For each bi-gram keyphrase in this level, there may be a tri-gram keyphrase level consists of three words, which contains this bi-gram keyphrase.

As not all bi-gram keyphrases have meaningful labels, the hierarchy tree will be improved using predefined syntactic patterns. One of the mostly used patterns is "NN DTNN" such as " انفلونزا الطيور", "سرطان الجلد", "منظومة التعليم", and "حقوق الانسان". Samples of the hierarchy tree are shown in figure 4 which describe the cluster label and the document number within cluster.

Figure 4. Sample of the hierarchy tree

4. EVALUATION MEASURES

The quality of text clustering can be measured by internal and external quality measures [9]. The internal quality measures the goodness of clustering without using any external knowledge. The external quality depends on external class labels. It measures the matching between the clusters labels obtained from the system and the externally supplied class labels. The mostly used external quality measures; purity and entropy will be used to evaluate the efficiency of the proposed system.

The externally class labels are obtained by asking three human judges to determine the suitable uni-gram topic label of each document in the dataset. Every human judge can assign up to three uni-gram topic labels for each document, i.e. a document can have the labels "سرطان","مرض", or "فيروس". For each document, we take the shared label among the judge assigns.

4.1 Purity

The purity measure represents the percentage of correctly clustered documents according to the external supplied class labels. It is calculated by counting the number of correctly assigned documents divided by the number of all documents in the document collection [16]. The purity for cluster j of size n_j is as follows:

$$P_j = \frac{1}{n_j} \max n_{ij}$$

Where: n_{ij} is the number of documents of class i in cluster j.

The overall purity for clustering is the weighted sum of individual clusters purities.

$$P = \sum_j \frac{n_j}{N} P_j$$

Where: N is the total number of documents in the collection.

The purity is bounded between [0,1]. The larger the purity, the better is the clustering efficiency.

4.2 Entropy

The entropy measures the degree to which each cluster contains documents of a single class [8]. The lower the entropy, the better is the clustering efficiency. The entropy of each cluster j is calculated as:

$$E_j = -\sum_i P_{ij} \log(P_{ij})$$

Where: P_{ij} is the probability that a document of cluster j belongs to class i.

The total entropy is calculated as the sum of the entropies of each cluster weighted by the size of each cluster n_j [7]:

$$E = \sum_{j=1}^{m} \frac{n_j * E_j}{n}$$

Where: m is the total number of clusters and n is the total number of documents.

5. EXPERIMENTAL RESULTS

The proposed clustering algorithm is applied on a dataset in different domains using string based, knowledge based, and corpus based similarity measures. The performance of the clustering algorithm is evaluated using external quality measures.

5.1 Dataset Description

The aim is to evaluate the proposed system using documents with different sizes in different domains. We collect the testing dataset from different sources such as Aljazeera online news agency, El-Ahram newspaper, and some documents of the dataset used by [15]. That collected dataset contains 345 documents and covers 12 categories (science, medicine, politic, arts, sport, economic, astronomy, technology, tourism, education, religion, and women).

5.2 Experiments

Three experiments were carried out to test the performance of the proposed system. In each experiment, the clusters of the input documents are extracted by applying the seven linkage clustering methods on the similarity matrix obtained using the different similarity measures. The purity and entropy values are computed for the clustering results of each experiment using the labels proposed by human judges. In all experiments, the weighted feature vector for each document contains the top 25 keyphrases.

5.2.1 Experiment 1: Using String-Based Similarity

This experiment aims to measure the efficiency of the proposed system when using the Arabic keyphrase using string -based similarity measure.

In this experiment, the proposed algorithm is applied on documents in a specific domain such as "سرطان", "انفلونزا", "بنوك", "تعليم", "مهرجانات", and "فنادق". Then the algorithm is applied on documents in a general domain such as "امراض", "اقتصاد", and "فنون". In addition, the algorithm is applied on documents in various domains such as "تعليم - امراض - سياسة - دينى".

Figure 5 shows the clusters labels by applying the proposed algorithm on the general domain "امراض" using average linkage clustering method. Tables 2 & 3 show the purity and entropy values using the string based similarity methods with each linkage method.

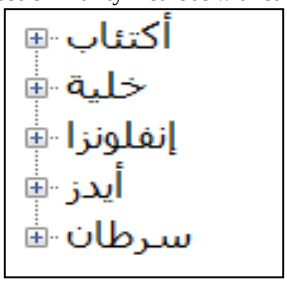

Figure 5. Clusters labels for domain "امراض" using our proposed similarity

Table 2. The purity using string-based similarity

Clustering method	Similarity method						
	Our similarity	Cosine	Jaccard	Euclidean	Dice	Overlap	Pearson
Single	0.89	0.90	0.89	0.90	0.90	0.90	0.90
Complete	0.92	0.92	0.92	0.87	0.92	0.92	0.92
Average	0.95	0.90	0.93	0.90	0.9	0.90	0.9
Weighted	0.94	0.91	0.93	0.91	0.91	0.91	0.91
Centroid	0.68	0.69	0.53	0.55	0.88	0.70	0.69
Ward	0.67	0.92	0.90	0.93	0.92	0.92	0.92
Median	0.68	0.72	0.52	0.54	0.86	0.72	0.70

Table 3. The entropy using string-based similarity

Clustering method	Similarity method						
	Our similarity	Cosine	Jaccard	Euclidean	Dice	Overlap	Pearson
Single	0.04	0.03	0.04	0.03	0.03	0.03	0.03
Complete	0.02	0.03	0.02	0.04	0.03	0.03	0.03
Average	0.01	0.03	0.02	0.03	0.03	0.03	0.03
Weighted	0.02	0.03	0.02	0.03	0.03	0.03	0.03
Centroid	0.10	0.09	0.09	0.09	0.04	0.09	0.09
Ward	0.10	0.02	0.03	0.02	0.02	0.02	0.02
Median	0.10	0.09	0.09	0.09	0.05	0.09	0.09

The results show that, our proposed similarity method with the average clustering method has the highest purity score and lowest entropy value. This comes from the fact that a document cannot join a cluster until it obtains a high average similarity value with all members of the cluster. So the probability of a cluster obtaining a new document becomes smaller as the size of the cluster increases. Our similarity method gives higher values than other traditional string based methods and gives better results with the average linkage clustering method.

5.2.2 Experiment 2: Using Corpus-Based Similarity

In this experiment, the efficiency of the proposed system using the semantic measure via corpus -based similarity is evaluated. Table 4 shows the purity and entropy values using the corpus-based similarity with the different linkage methods.

Table 4. The purity and entropy using corpus-based similarity

Clustering Method	Corpus-based similarity	
	Purity	Entropy
Single	**0.9355714**	**0.025313**
Complete	0.9112857	0.034544
Average	**0.9334286**	**0.027145**
Weighted	0.9182857	0.032017
Centroid	0.7735714	0.079441
Ward	0.8402857	0.057292
Medain	0.8027143	0.070101

The results show that both the single and average linkage methods give the best results. In addition, the string-based approach is more accurate than the corpus-based one.

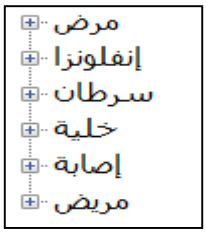

Figure 6. Clusters labels for domain "امراض" using corpus based similarity

Figure 6 represents the clusters labels for the general domain "امراض" using corpus based similarity and average linkage clustering method. Comparing the clusters labels obtained in figures 5 and figure 6 show that, the lexical similarity applied using our proposed similarity gives a more specific clusters labels than using semantic similarity via corpus-based. So, The lexical similarity is more suitable for generating detailed hierarchy tree with specific clusters labels. However, the semantic similarity using corpus-based is more suitable for generating general clusters labels.

Experiment 3: Using Knowledge-Based Similarity

This experiment aims to measure the efficiency of the proposed system using the WordNet. We translate the Arabic keyphrase into English and add the English lemma-forms of keyphrases to the weighted feature vector. Then, we use the WN to measure similarity between vectors using path, wup , and lch similarity methods. The purity and entropy values are shown in table 5 and table 6 respectively.

Table 5. The purity using knowledge-based similarity

Clustering method	Similarity method		
	Wup	Lch	Path
Single	0.4854623	0.508394	0.658151
Complete	0.8666667	0.881667	0.911
Average	0.8303333	0.841	**0.911667**
Weighted	0.866	0.853667	0.890333
Centroid	0.732	0.777	0.854
Ward	0.7583333	0.809333	0.907
Medain	0.7376667	0.788	0.875667

Table 6. The entropy using knowledge-based similarity

Clustering method	Similarity method		
	Wup	Lch	Path
Single	0.091483	0.097368	0.101382
Complete	0.049871	0.045475	0.034329
Average	0.058725	0.054951	**0.035149**
Weighted	0.050862	0.054342	0.043205
Centroid	0.069718	0.070961	0.055534
Ward	0.063632	0.060477	0.036731
Medain	0.066359	0.067557	0.047942

The results show that path similarity method with the average clustering method give the highest purity and entropy scores. In addition, the knowledge based similarity gives good specific labels as our proposed method but it takes longer.

We apply all the experiments on weighted feature vector consist of 25 keyphrases for each document. In addition, we apply the proposed system on weighted feature vector consists of 20 keyphrases for each document. The number of obtained clusters is increased as shown in figure 7. In addition, the hierarchy trees are small and some meaningful details are lost.

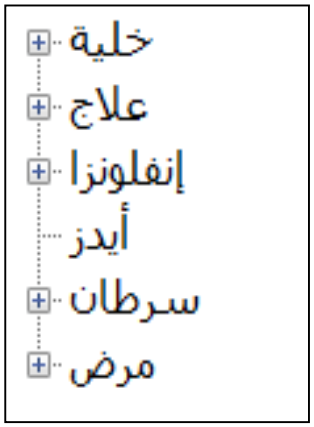

Figure 7. Clusters labels of general domain "امراض" using 20 keyphrases with our proposed similarity

6. CONCLUSION

In this paper, we present a new methodology for clustering Arabic documents based on the idea of representing each document by its keyphrases in their lemma forms. In addition, we use the common keyphrases between feature vectors of documents as a new similarity measure which gives good results and faster than the traditional similarity methods. It improves the clustering results and gives labels that are more meaningful in the hierarchy tree. We apply the proposed approach using both lexical and semantic similarity measures. We implement the semantic similarity using AWN and WN. While AWN is one of the most popular semantic networks, it gives poor results compared to lexical similarity because AWN has a limited number of word synonyms. Therefore, in this work we use the WN after adding the translated Arabic keyphrases to the feature vector. The results show that our new lexical similarity method gives more accurate similarity results than semantic similarity and is suitable for retrieving specific clustering labels. While in retrieving general cluster labels, the corpus based similarity is more beneficial.

7. REFERENCES

[1] BLACK, W., ELKATEB, S., RODRIGUEZ, H., ALKHALIFA, M., VOSSEN, P., PEASE, A., and FELLBAUM, C., 2006. Introducing the Arabic wordnet project. In *Proceedings of the Third International WordNet Conference*, 295-300.

[2] EL-SHISHTAWY, T. and AL-SAMMAK, A., 2012. Arabic keyphrase extraction using linguistic knowledge and machine learning techniques. *arXiv preprint arXiv:1203.4605*.

[3] GOMAA, W.H. and FAHMY, A.A., 2013. A survey of text similarity approaches. *International Journal of Computer Applications 68*, 13, 13-18.

[4] GRAHAM, R.L. and HELL, P., 1985. On the history of the minimum spanning tree problem. *Annals of the History of Computing 7*, 1, 43-57.

[5] HUANG, A., 2008. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, 49-56.

[6] JAIN, A.K., MURTY, M.N., and FLYNN, P.J., 1999. Data clustering: a review. *ACM computing surveys (CSUR) 31*, 3, 264-323.

[7] JAJOO, P., 2008. Document clustering Indian Institute of Technology Kharagpur.

[8] JENSI, R. and JIJI, D.G.W., 2014. A Survey on optimization approaches to text document clustering. *arXiv preprint arXiv:1401.2229*.

[9] KARYPIS, M.S.G., KUMAR, V., and STEINBACH, M., 2000. A comparison of document clustering techniques. In *KDD workshop on Text Mining*.

[10] KAUFMAN, R., 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*.

[11] KOLB, P., 2008. Disco: A multilingual database of distributionally similar words. *Proceedings of KONVENS-2008, Berlin*.

[12] KOLB, P., 2009. Experiments on the difference between semantic similarity and relatedness. In *Proceedings of the 17th Nordic Conference on Computational Linguistics-NODALIDA'09*.

[13] MENG, L., HUANG, R., and GU, J., 2013. A review of semantic similarity measures in wordnet. *International Journal of Hybrid Information Technology 6*, 1, 1-12.

[14] MILLER, G.A., BECKWITH, R., FELLBAUM, C., GROSS, D., and MILLER, K.J., 1990. Introduction to wordnet: An on-line lexical database*. *International journal of lexicography 3*, 4, 235-244.

[15] MOLIJY, A.A., HMEIDI, I., and ALSMADI, I., 2012. Indexing of Arabic documents automatically based on lexical analysis. *arXiv preprint arXiv:1205.1602*.

[16] OZGÜR, A., 2004. Supervised and unsupervised machine learning techniques for text document categorization Bogaziçi University.

[17] RODRÍGUEZ, H., FARWELL, D., FARRERES, J., BERTRAN, M., ALKHALIFA, M., MARTÍ, M.A., BLACK, W., ELKATEB, S., KIRK, J., and PEASE, A., 2008. Arabic wordnet: Current state and future extensions. In *Proceedings of The Fourth Global WordNet Conference, Szeged, Hungary*.

[18] ROSELL, M., 2006. Introduction to information retrieval and text clustering. *KTH CSC*.

[19] SAHOO, N., CALLAN, J., KRISHNAN, R., DUNCAN, G., and PADMAN, R., 2006. Incremental hierarchical clustering of text documents. In *Proceedings of the 15th ACM international conference on Information and knowledge management* ACM, 357-366.

[20] TOMBROS, A., 2002. The effectiveness of query-based hierarchic clustering of documents for information retrieval University of Glasgow.

[21] WU, Z. and PALMER, M., 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics* Association for Computational Linguistics, 133-138.