American Journal of Information Systems, 2013, Vol. 1, No. 1, 26-30
Available online at http://pubs.sciepub.com/ajis/1/1/4
© Science and Education Publishing
DOI:10.12691/ajis-1-1-4

Adapt Clustering Methods for Arabic Documents

Boumedyen Shannaq*

Computer science and Information Technology Department, Mazoon College, “University College”, Muscat, Sultanate of Oman
*Corresponding author: aboumedyen@gmail.com

Received August 04, 2013; Revised November 21, 2013; Accepted November 26, 2013

Abstract This research paper develops new clustering method (FWC) and further proposes a new approach to filtering data collected from internet resources. The focus of this research paper is clustering groups’ data instances into subsets in such a manner that similar instances are grouped together, while different instances belong to different groups. The instances are thereby organized into an efficient representation that characterizes the population being sampled thereby reducing the gigantic size of retrieved data. This has been done by removing dissimilar text files, and grouping similar documents into homogeneous clusters. Arabic text files of 974 MB has been collected, processed, analyzed and filtered by using common clustering methods. This new clustering methods are presented, divided into: hierarchical, partitioning, density-based, model-based and soft-computing methods. Following the methods, the challenges of performing clustering in large data sets are discussed and tested by the proposed new clustering method. Two experiments were conducted to establish the effectiveness of FWC methods and the obtained results show that the new FCW method suggested in this paper produced better results and outperformed existing clustering methods.

Keywords: *clustering, knowledge management, information retrieval system*

Cite This Article: Boumedyen Shannaq, “Adapt Clustering Methods for Arabic Documents.” *American Journal of Information Systems*, no. 1 (2013): 26-30. doi: 10.12691/ajis-1-1-4.

1. Introduction

Nowadays Search and Navigation activities become one of the common and needed services on the Internet Technology. How to be fast and smart in your search? Is now a top priority for most individuals and Organization? The internet technology has flooded the world with online information [1,2]. Organizing the abundant information available on the internet has become a major challenge to the researchers [3]. In today’s world, children, students, schools, universities, colleges, companies, government etc… used internet as a main source for collecting their essential information. The dependence on the internet increases the traffic on the net, accordingly there is a challenge faces by search engines to find new helpful techniques to deal with enormous volume of information available on the internet. Furthermore serving the great number of internet users [4,5]. Since we live in an information age, it is thus necessary to design or make a search engine which can successfully index or classify the web pages in a manner that help its users to derive the exact information required by them. However despite of companies claiming their success in producing a search engine which will satisfy the internet users, still the user complain of the lack of accuracy and relevance of retrieved information [6] particularly the Arabic user. The attitudinal change in the user behavior, has forced many IT companies to seriously think and develop advance search technologies that may enable the user to retrieve the desired information. Google, Yahoo, Microsoft etc. companies are well aware of this fact [7]. This work put into practice a strategy for filtering webpage's retrieved from search engines, as a result reduced the gigantic size of retrieved text collection, as well as improved the presentation and performance of knowledge base.

2. Literature Review

A growing amount of research has studied the organization of web data. [8] discuss standards and evaluations in test collections Using different clustering models and various text transformation approaches were proposed, to arranged text collection for searching and building knowledge based systems. Most of these techniques were concerned with text operation i.e. lexical analysis, elimination of stop words, stemming etc…, moreover such text operation are useful for selection of index terms and building thesauri. However, there is no evidence that such text operations improved or removed unnecessary text from text collection [7]. Other researchers used the operation of compressing text aims at reducing space and communication cost, there is no doubt such compressing operations requires less storage space and takes less time, to be good compression ratio, fast coding, fast decoding, and this basically not easy task. [3] includes an analysis of the inverted index, inverted lists, suffix arrays, Pat arrays and Huffman coding, however such techniques are dealing with text retrieval to implement query operation, but not for removing unnecessary text files from text collection. [9] Explained three classic models in Web Information Retrieval: the

Boolean, the vector and Probabilistic models, those models were used as a ranking algorithms. [9] Presented new clustering techniques to discover when two documents are similar. The proposed techniques aimed to resolve the difficulties of articles repetition; the new clustering technique was based on LIPNS, SAMA1, Text Normalization, DNSA and, NADST techniques respectively. [10] Describes the process of analyzing text considers various measurement to evaluate corpus and collections by zip's and Mandelbrot distribution law [11]. [12] works on analyzing corpus data in order to generate useful text analysis and categorization over different file formats, such application may help and support researchers in selecting the best text collection for example, building a knowledge base; an ontology; Thesaurus, and glossary [13]. Currently numerous questions are raised on how the internet provides the electronic document? How to structure this electronic document? Which document to be retrieved? Can we trust this information retrieved by search engines? The challenge is how to describe what document is about? One of the common approaches is to select terms to represent what the document is about. Today, the current trends looking to find a group of activities to facilitate the access to a specific information and knowledge which often can be seen implicitly, and most of discussion methods above have not clearly introduced a helpful techniques for answering the raised questions above.

3. Experiments

3.1. Text Collection

For the implementation of this work, data text was collected and downloaded from wekipedia http://wikipedia.org/wiki/Text_corpus, size of 974 MB (1,021,566,976 bytes), text collection was in Arabic, XML format, and stored in one file only. The text collection has been partitioned to group of files and transformed to. Txt format using Python software. The splitting process consumed 2hrs 8 min. There after statistical information from the new text collection was obtained, and the processed data was 1.10 GB (1,187,983,360 bytes) containing 228,308 Files and 65,704 Folders. HTML tags, Stop words like

"لن له من هو هي قوة كما لها منذ وقد ولا لم كل
هناك وقال وكان وقالت وكانت فيه كلم لكن وفي وقف
ولم ومن وهو وهي يوم فيها منها"

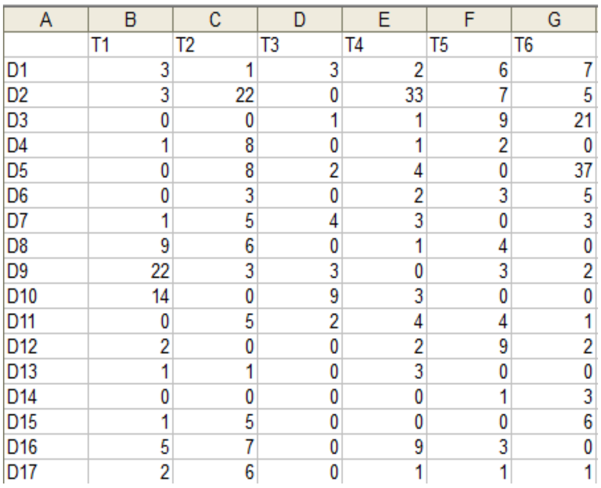
, numbers, punctuations and spaces between adjacent words have been removed, Super Arabic morphological analyzer (SAMA1) [14] has been used for stemming purposes, the stemming process was done for handling many issues raised from Plurals, gerund forms and past tenses. Simple stemming algorithm can be described as follow: If word starts in "ال “then replace with " ",

"الانترنت" → " انترنت"

C# program has been developed to remove the Arabic stop words.

3.2. Experiment # 1

The text collection were transformed to document/Term matrix, first column represents Documents, first row represents terms in text collection, and elements(values) represent the frequencies of terms in a specific document. Figure 1 shows fragments of document/Term matrix, we provide this fragment as an example to illustrate our contribution to the existing methods.



A	B	C	D	E	F	G
	T1	T2	T3	T4	T5	T6
D1	3	1	3	2	6	7
D2	3	22	0	33	7	5
D3	0	0	1	1	9	21
D4	1	8	0	1	2	0
D5	0	8	2	4	0	37
D6	0	3	0	2	3	5
D7	1	5	4	3	0	3
D8	9	6	0	1	4	0
D9	22	3	3	0	3	2
D10	14	0	9	3	0	0
D11	0	5	2	4	4	1
D12	2	0	0	2	9	2
D13	1	1	0	3	0	0
D14	0	0	0	0	1	3
D15	1	5	0	0	0	6
D16	5	7	0	9	3	0
D17	2	6	0	1	1	1

Figure 1. Fragments of document/Term matrix

Different cluster methods like, ‘groups linkage’, ‘centroid’ and ‘ward's’ method respectively have been used and tested, in addition to available alternatives like ‘euclidean distance’ and ‘cosine’, after many iterative operations, ‘Ward's’ method and Euclidean distance measure have been selected, for the reason that they provide best clustering results against others methods. Figure 2 shows the proximity matrix of Document/Document (distance matrix D)after applying ‘Ward's’ method and ‘Euclidean distance’ measure. The next illustrate the implemented formulas used to obtained Distance matrix D. all formulas were obtained from [15,16].

‘Centroid method’

$$\frac{\overline{x}_1 n_1 + \overline{x}_2 n_2}{n_1 + n_2}$$

‘Ward’s method’

The total deviance (T) of the p variables, corresponding to n times the trace of the variance–covariance matrix, can be divided in two parts: the deviance.

Within the groups (W) and the deviance between the groups (B), so T = W + B.

The total deviance (T) can be denoted by:

$$T = \sum_{s=1}^{p} \sum_{i=1}^{n} (x_{is} - \overline{x}_s)^2$$

Groups (W) are given by the sum of the deviances of each group and can be denoted by:

$$W = \sum_{k=1}^{g} W_k$$

W_k represents the deviance of the p variables in the k_{ith} group and can be denoted by:

$$W_k = \sum_{s=1}^{p} \sum_{i=1}^{n_k} (x_{is} - \overline{x}_{sk})^2$$

The deviance between the groups, (B) is given by the calculated sum on all the variables and can be denoted by:

$$B = \sum_{s=1}^{p}\sum_{k=1}^{g} n_k\left(\bar{x}_{sk} - \bar{x}_s\right)^2$$

'Cosine measure'

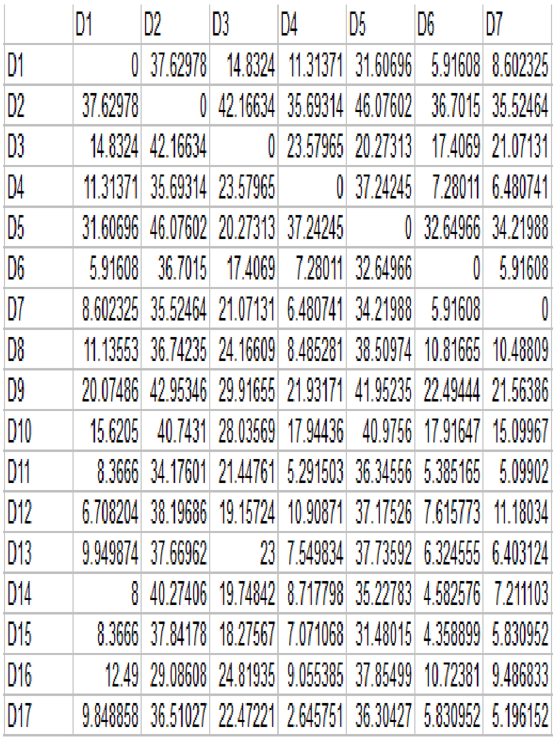
$$sim\left(d_j, q\right) = \frac{\bar{dj} \bullet \bar{q}}{\left|\bar{dj}\right| \times \left|\bar{q}\right|}$$

$$= \frac{\sum_{i-1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i-1}^{t} w_{i,j}^2} \times \sqrt{\sum_{j-1}^{t} w_{i,q}^2}}$$

'Euclidean distances'

$$d_{ij} = \left(\sum_{K=1}^{N}\left(x_{ik} - x_{jk}\right)^2\right)^{\frac{1}{2}}$$

The introduced methods above amid to put the similar documents in different groups by calculating the similarities between the documents. Let us consider the following input matrix obtained from matrix, assumed all the way that the input data are in the form of a matrix, illustrated in Figure 1, the obtained Proximity matrix from matrix shown in Figure 1 is described in Figure 2.



	D1	D2	D3	D4	D5	D6	D7
D1	0	37.62978	14.8324	11.31371	31.60696	5.91608	8.602325
D2	37.62978	0	42.16634	35.69314	46.07602	36.7015	35.52464
D3	14.8324	42.16634	0	23.57965	20.27313	17.4069	21.07131
D4	11.31371	35.69314	23.57965	0	37.24245	7.28011	6.480741
D5	31.60696	46.07602	20.27313	37.24245	0	32.64966	34.21988
D6	5.91608	36.7015	17.4069	7.28011	32.64966	0	5.91608
D7	8.602325	35.52464	21.07131	6.480741	34.21988	5.91608	0
D8	11.13553	36.74235	24.16609	8.485281	38.50974	10.81665	10.48809
D9	20.07486	42.95346	29.91655	21.93171	41.95235	22.49444	21.56386
D10	15.6205	40.7431	28.03569	17.94436	40.9756	17.91647	15.09967
D11	8.3666	34.17601	21.44761	5.291503	36.34556	5.385165	5.09902
D12	6.708204	38.19686	19.15724	10.90871	37.17526	7.615773	11.18034
D13	9.949874	37.66962	23	7.549834	37.73592	6.324555	6.403124
D14	8	40.27406	19.74842	8.717798	35.22783	4.582576	7.211103
D15	8.3666	37.84178	18.27567	7.071068	31.48015	4.358899	5.830952
D16	12.49	29.08608	24.81935	9.055385	37.85499	10.72381	9.486833
D17	9.848858	36.51027	22.47221	2.645751	36.30427	5.830952	5.196152

Figure 2. fragment of obtained Proximity matrix

To extend the procedure of clustering described in [15], we perform the following:

Consider every row vector as a point and rearrange the row vectors according to their similarity, in Figure 2 the row vectors ordersare d1d2d3d4d5d6d7d8d9d10d11d12 d13d14d15d16d17, next if we rearrange the row vectors orders to d4d17d7d11d6d15d13d14d1d12d8d16d9d10d3d5.

Table 1 illustrates the steps of creating clusters from Figure 2.

Table 1. The steps of creating clusters

Number of Cluster	Clusters
1	4,7
2	7,11
3	6,15
4	13,14
5	1,12
6	8,16
7	9,10
8	3,5
9	4,17,7,11
10	6,15,13,14
11	3,5,2
12	4,17,7,11,6,15,13,14,1,12,8,16
13	4,17,7,11,6,15,13,14,1,12,8,16,9,10
14	4,17,7,11,6,15,13,14,1,12,8,16,9,10,3,5,2

3.3. Experiment # 2

The proposed idea illustrates that, to cluster the collected text into groups of similar documents based on the idea of analyzing and extracting the first word only from each document then group all documents which have the same first word together.

The next procedure illustrates the proposed approach:

1-Read First Document Di from Corpus

2-Extract only first keyword (fk) (eliminate any stop words or signs)

3-Find the frequency (fr) of extracted first keyword over the document Di

4-Add Address of the document, first keyword (fk) and frequency (fr)

5-Repeat all steps for all documents in the corpus

Table 2 shows a sample of the output after performing the above procedure:

Table 2. Extracted sample from the corpus

Document number(Di)	Keyword (fk)	Frequency(fr)
Document 1	معالج	23
Document 2	انترنت	13
Document 3	بحث	17
Document 4	انترنت	17
Document 5	شبكة	9
Document 6	بحث	7
Document 7 -	انترنت	5

Reorganize the table horizontally until formulating groups of similar documents based on matching first keywords. (Table 3 shows this operation after reorganization the rows in the above table).

Table 3. Reorganized rows

Document number(Di)	Keyword (fk)	Frequency(fr)
Document 2 (Collection # 1)	انترنت	13
Document 4 (Collection # 1)	انترنت	17
Document 7 (Collection # 1)	انترنت	5
Document 3 (Collection #2)	بحث	17
Document 6 (Collection #2)	بحث	7
Document 1	معالج	23
Document 5	شبكة	9

4. Evaluation and Results

Evaluating the results of the obtained grouping means verifying that the groups are consistent with the primary objective of the cluster analysis, to satisfy the conditions of internal cohesion and external separation. Choosing the right number of groups is fundamentally important. Here we evaluate the results obtained from experiment one and experiment two, considering the factor of how many new words appear each time versus new extracted words from new documents. It was hard to compare all the obtained clusters from experiment one and Two, furthermore we select only the largest clusters considering size factor(number of documents).

Figure 3 shows the results derived from largest clusters obtained from experiment.

One using basic clustering methods.

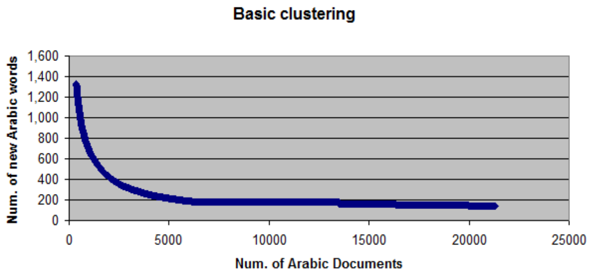


Figure 3. Number of new words against number of documents

Figure 3 illustrate that out of 200 to 170 new words continue to appear in this cluster. This means that there are some files which is not related to the cluster, and this in general will affect the final results of any research experiment oriented to a specific domain.

Figure 4 shows the obtained results from experiments two.

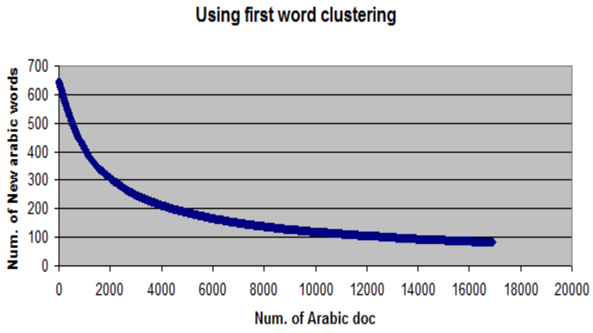


Figure 4. Using first word clustering

Figure 4 illustrate that out of 100 to 75 new words continue to appear in this cluster. This means that there are some files which are not related to the cluster, but the later is better than the former. Our consideration is, as the numbers of extracting new words each time reduces as we keep on testing other documents. This factor can be considered as a fact that the cluster contains similar and related documents, which are used to describe a specific domain. Regarding the obtained results from both experiments, the proposed approach was able to remove/filter about 100 dissimilar documents from the text collection.

Table 4 shows the finding and comparison between the basic clustering methods and the new developed method.

Table 4. Finding and results

Method	Number of total words	Number of new words
Basic method	>20000	200 to 170
First word method	>20000	100 to 75

5. Discussion

Most of the experimenters and researchers depend on concrete dataset, to test their hypothesis and algorithms. However finding and collection the related data set become a challenge, when the question is how to obtain the related data set for a specific topic i.e. data set must describes a specific domain and contains only the related terms of this domain [17]. Organizing the data considering as the hottest topic today in research and development area. Aims to build ontology of a specific domain. The advantage of building such ontology is to unified ideas and create standards over the web, more ever to enable self communication between machines. As a matter of fact, to build the ontology, first you need to prepare a glossary for your ontology, the glossary must contains all terms, their synonyms and other attributes [18]. Such terms must be used only to describe a specific domain. Extracting those terms manually is time and efforts consuming, thus automation this process is highly appreciated, however you must be sure that your text collection contains only related terms of your interested topic. How to choose which method to apply, In practice there is not a method that can give the most qualified result with every type of data. Experiment with the different alternatives and compare them in terms of the chosen criteria. This work proposes a novel approach to filter the text collection from dissimilar documents, by developing new clustering method. The developed clustering method, depends on the factor of considering the first word only from documents, the developed technique was tested, and the obtained result outperform the other existing clustering methods. We believe this developed techniques will be as a new development for the available clustering methods, there were some limitation of documents processing, such as selecting stop words and stemming issues, but we believe that these errors and limitation was not affect the obtained results since they are applied to the both experiments. This work can be applied to English and other languages to test the performance of the proposed approach. Other text collection can be used also to test the developed clustering method.

6. Conclusion

This work introduces one of the most interesting fields in computer-oriented data analysis. Developing new clustering technique FCW, aims to cluster the collected text into groups of similar documents, it is based on the idea of analyzing and extracting the first word only from each documents then grouping documents which have the same first word together. We were able to solve the problem of cluster analysis, and to group similar documents into homogeneous clusters. It is still too early to reach a consensus on the advantages of using this developed technique for the web, but we believe that this work and obtained results will prove to be effective in web application. Therefore this developed strategy will have wide application in the domain of E-Learning, Knowledge management and web management.

References

[1] Allen J., Aslam J., Belkin N., Buckley C., CallanJ.,"Challenges in information retrieval and language modeling", *Special Interest Group on Information Retrieval(SIGIR)*, Vol 37,No. 1, pp. 31-47, 2003.

[2] Shannaq B., Aleksandrov V.," Clustering the Arabic Documents(CAD) ", *Universal Journal of Applied computer Science and Technology (UNIASCIT)*, Vol. 1 No. 3, pp. 90-94, 2011.

[3] Araujo M.,Navarro G., Zivani N.," Large text searching allowing errors", *4th South American Workshop on String Processing (*WSP *'97)*, pp. 2-24, 1997.

[4] Allan J.,Carterette B., LewisJ., "When will information retrieval be "good enough?", *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 443-440, 2005.

[5] Shannaq B., " Using Russian and English Ontology In Expanding The Arabic Query", *Universal Journal of Applied computer Science and Technology (UNIASCIT)*, Vol. 1 No. 3,pp. 95-100, 2011.

[6] Shannaq B., Aleksandrov V.,"Using Product Similarity for Adding BusinessValue and Returning Customers ", *Global Journal of Computer Science and Technology*, Vol. 10, No. 12. pp. 2-8, 2010.

[7] MorleyD., Parker C., " *Understanding Computers Today and Tomorrow Comprehensive* " 13th edition, 2010.

[8] Shaw W., Burgin R., Howell P., " Performance standards and evaluations in IR test collections: Cluster-based retrieval models". *Information Processing and Management*, Vol. 33, No. 1, pp. 1-14, 1997.

[9] Baeza R., Ribeiro B., "*Modren Information Retrieval* ", ACM Press, New York, 1999.

[10] Mason, Oliver, Berglund, Ylva, "Low-level parameters reflecting the naturalness of texts". *Proceedings of JADT2002, 6th International Conference on Textual Data Statistical Analysis*, Saint Malo, March 13-15. Vol.2, pp. 507-516, 2002.

[11] Giinther R., Levitin L., Chapiro B., Wagner P.," Zipf's law and the act of ranking on probability distributions", International Journal of Theoretical Physics, Vol. 35, pp. 395-417, 1996.

[12] Shannaq, Boumedyen. "Investigating the Distribution of Arabic and English Keywords and Their Progress Over Different Text File Formats." American Journal of Computing Research Repository 1.1 (2013): 1-5.

[13] Kokorin P.,Shannaq B., "Algorithm of Normalization and Ontological Clusters Texts", Information-measuring and operating systems *Journal*, Vol. 7, No. 8,pp 60-64, 2010.

[14] Shannaq B., Aleksandrov V., " Super Arabic Morphological Analyzer (SAMA1)", *information-measuring and operating systems Journal*, Vol.11, No. 7,pp. 60-63, 2009.

[15] Witten H., Frank E., " *Data Mining & Practical Machine Learning Tools and Techniques*",Elsevier,2005.

[16] Giudici P., "*Applied Data Mining, Staistical Methods for Business and Industry* ", Wiley,England, 2003.

[17] Boumedyen Shannaq,” Methods and Algorithms for Searching Arabic Name Entity”, *International Journal of Computer Applications*, Vol.82 - Number 8, 2013.

[18] Boumedyen Shannaq, Kaneez Fatima, " Hierarchy Concept Analysis in Accounting Ontology ", *Asian Journal Of Computer Science And Information Technology*, Vol.2: 2, 2012 13-20.