J. Basic. Appl. Sci. Res., 3(3)1195-1201, 2013

ISSN 2090-4304
Journal of Basic and Applied Scientific Research
www.textroad.com

A New Approach for Persian Web Page Clustering Using K-Means Algorithm

OmidZiyarati[1], HabibRostami[2], MashaallaAbbasi-Dezfuli[3]

[1,3]Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Khouzestan, Iran
[2]Computer Engineering Department, School of Engineering, Persian Gulf University of Bushehr, Bushehr 75168, Iran

ABSTRACT

Today, the Internet is considered as one of the most important information sources that has allocated a lot of internet users to it. These users include the researchers, scholars and even the general public. The volume of web, based on the carried out researches is beyond billions of the pages, and millions of the pages are added to it in every second and moment. The heterogeneity of web documents is to the extent that the resulted disturbance or confusion from it is uncontrollable. The web environment researchers have felt that if the retrieval and the organization of web pages do not take place, the existing data in web would operationally/practically be unusable. The search engines were come into existence therefore, by conducting a lot of research works. However, in spite of several advantages of research engines, they do not cover the embedded web layers that retain a lot of information in them. In addition, the resulted output from search engines is always associated with irrelevant pages, redundancy and multiplicity that, quest is very costly in it. Hence, the researchers put forth the idea of automated classification and clustering of web pages, based on it the web pages will be organized in a systematic structure. In this paper, a new indexing method of web pages has been presented based on the content, in order to clustering in the way of organizational increase of web pages. This algorithm, at first by selecting the desired parameters of web documents, each document is given weight considering the presented technique and finally by using K-Means, we will cluster the documents. After the simulation of suggested algorithm and its comparison with the other algorithms, it was observed that its accuracy and performance is better than that of the previous algorithms.

KEYWORDS: Clustering, Web Pages, Data Mining, Indexing, K-Means

1.INTRODUCTION

Recent advances in storage technology and dramatic growth in applications such as the Internet search and digital imaging have created many high-dimensional data sets. To provide huge potential for the development of automatic data analysis and retrieval techniques, often the data stored digitally on electronic media. In addition to the growing amount of data, the variety of data has also increased. Blogs, e-mails, and billions of Web pages create terabytes of new data every moment.

The increase in the variety of data requires advances in methodology to automatically understand, process, and summarize the data. Overall, the data analysis techniques are divided into two main types [19]: (i) descriptive, meaning that the investigator does not have pre specified models or hypotheses but wants to understand the general characteristics or structure of the high-dimensional data, and (ii)inferential, meaning that the investigator wants to confirm the validity of a hypothesis/model or a set of assumptions given the available data. There are many statistical techniques for data analysis, such as analysis of variance, principal component analysis, linear regression, cluster analysis, and factor analysis (a useful review is given in [3]).

The operation of grouping a set of physical or abstract objects into classes of similar objects is referred to as clustering. A cluster is a set of data objects that are like to one another within the same cluster and are unlike to the objects in other clusters. By automated clustering, dense and sparse regions in object space are identified and, therefore, discover overall distribution patterns and interesting correlations among data attributes[9].

Clustering is also referred as data segmentation in some applications because clustering partitions large data sets into groups according to their similarity. Cluster analysis has been used to group related for browsing and to find similar web pages. However, in other cases, cluster analysis is only a useful starting point for other functions, e.g., data compression or efficiently finding the nearest neighbors of points.

Generally speaking clustering methods are divided into two main groups known as hard and fuzzy. Hard methods are divided into two kinds known as partitioned and hierarchical. The hierarchical method is also divided into two kinds known as divisive and Agglomerative. These methods are shown in figure 1.

***Corresponding Author:** Omid Ziyarati, Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Khouzestan, Iran. Email : omidziyarati@yahoo.com

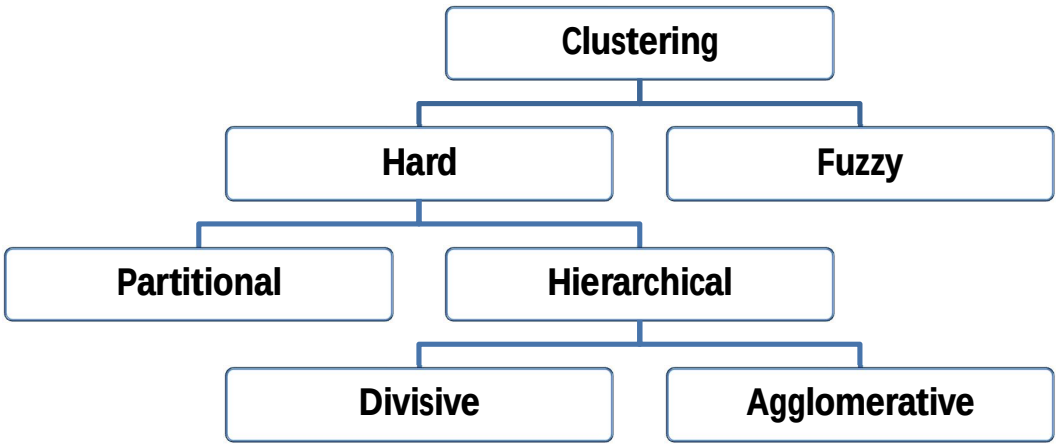


Figure 1: Clustering Methods

In hard clustering each object is just placed in one cluster; while in fuzzy clustering each object is placed in a cluster with a degree of membership. The higher this degree of membership, the more the dependency of that object to that cluster. Hard methods can be developed and changed into fuzzy ones.

2.Related works

Clustering methods in different scientific fields have improved. Psychologists[16], taxonomists [5], social sciences experts [1], biologists[20], statisticians, mathematicians, engineers, computer experts, medical researchers[21], and those who collect and process real data cooperate in presenting clustering methods. Data clustering was presented in 1954 articles with anthropological data. Also clustering as analyst, typology and cumulative mass was known as taxonomy which is dependent on the field in which it is employed. Several books are published on data clustering as following: Principals of numerical taxonomy[17], Cluster analysis for applications[2], Direct clustering of a data matrix[10], Algorithms for data clustering[11], Pattern classification[6]. Also clustering algorithms are vastly studied in data analysis. Some of them are as following; Data Mining: Concepts and Techniques[9], Introduction to data mining[18], Pattern recognition and machine learning [4].

Clustering the web pages is one of the most important approaches for extracting knowledge from the web. One of the most agreeable trends in clustering the high dimensional web pages has been tilt toward the learning and optimization approaches. Mahdavi et al. proposed novel hybrid harmony search (HS) based algorithms for clustering the web documents that finds a globally optimal partition of them into a specified number of clusters[13]. Also, Fersiniet al. presented a document clustering approach, which takes into account both hyperlink structure and contents information of web page collection, where a document is displayed as a set of semantic units[8].In the following, we will first describe the K-means algorithm, and then discuss the major approaches that have been developed for Web page clustering.

3.K-means algorithm

K-means method is the most applicable method for data clustering. This method was presented by Macqueen for the first time[12]. In this method the number of clusters is fix and predetermined. In this method objects are randomly divided into a number of k clusters. In the next step the distance between the object and the center of its cluster is calculated. If the distance between the object and the mean of its cluster is much and it is closer to another cluster, this object is assigned to that cluster which is closer to it. This is done repeatedly until the error function reaches to the least amount or the members of the clusters do not change. If D is the data set with n objects and C_1, C_2, … ,C_k represent k distinct clusters of D, the error function (EF) defines the total distance of each object to its cluster center:

$$EF = \sum_{i=1}^{k} \sum_{X \in C_i} d(X, \mu(C_i))$$

μ represents the (mean) cluster center, and $d(X, \mu(C_i))$ is the distance of each object from its cluster center. The distance of each object from its cluster can be calculated based on Euclidean or other methods. Since there is an error function in the centripetal clustering which we want to make it the least, one can regard clustering problems as optimization problems. There is an objective function in this kind of clustering which is an error

function and we want to make it the least; there are some limitations such as:a) the number of clusters are predetermined and we cannot make them more or less, b) The number of none of the clusters can be zero.
There are some steps in k- means clustering as following:

Initial step: separation of preliminary data into k clusters optionally.

Repetitive step:

a) Calculation of the distance of each object from the center of its cluster.
b) Calculation of error function.

Improve step: Relocation of a member with the most distance from the center of its cluster with the cluster that has the least distance from it.

Stop order: a situation in which the number of cluster members does not change or the error function does not decrease.

4.Proposedapproach

Most of researches on clustering the web pages have been implemented on English language. The proposed approach would cluster the Persian web pages. This approach choose the desirable parameters of web pages, then using our technique weighs pages is calculated, and the weights are given as input to the k-means algorithm.

The web pages are basically semi-structured. In the preprocessing step, documents' text is tokenized, html tags and stop words(such as or, and …) are removed, and remaining words of are classified using a defined directory in this research (see table 1). Considering various Persian language fields, we put the words in seven classes include: social, economic, political, cultural, sport, scientific and miscellaneous(Table 1).

Table1: classification of the documents' words in seven classes

DocumentsNo.	1	2	3	4	5	6	….
Numberof the social words	4	3	5	1	10	5	….
Number of the economic words	2	18	12	4	3	4	….
Number of the political words	0	4	4	2	8	0	….
Number of the cultural words	2	2	10	7	9	15	….
Number of the sport words	20	0	5	17	2	6	….
Number of the scientific words	1	10	10	1	4	7	….
Number of the miscellaneouswords	7	12	18	15	7	10	….

After preprocessing step, the following steps are applied:
1. We calculate weight for each document using the following equation:

$$w_i = \frac{\varepsilon}{n} + (1-\varepsilon) \times \frac{p \times k}{q \times m}$$

Where ϵ is constant value0.45, n is the total number of words in the current document, p is the largest number of the words in the seven classes, k is the number of classes which in this research is equal to 7,$m$is constant value 2 and q is obtained using the following equation:

$$q = n - p$$

As an exception case, when q is zero we calculate weight with the following equation:

$$w_i = \frac{\varepsilon}{n} + (1-\varepsilon) \times (p \times k)$$

2.Weights calculated in the previous step, is sent as input to the clustering algorithm, which in this research applied the K-means algorithm.

The main reason for applying this method is to determine the quality of the document context through describing the context. The best method for classifying a document is contextual analysis; because proposed approach causes to separate topical document and indexation based on document's context to clustering. In the field of clustering, the main problem of more algorithms is related to hide the efficient framework for appearing distance criterion to documents and distance covered in intended cluster. By introducing the proposed weighting approach would overcome the problem.

5. Evaluation

In consider of clustering algorithms, two factors are examined, accuracy and execution time. So, we gave amounts of these factors, then analyses the statistical indexes and in the end will present the results.

In this work, the results of the proposed algorithm are compared with the algorithm presented in [7], which works based on TF-IDF method and Amalgamation K–Means Algorithm [15].
Table2shows the accuracy and Execution time of the algorithms in different executions.

Table 2: Accuracy and execution time in different execution.

The accuracy of the algorithms										
Execution No.	1	2	3	4	5	6	7	8	9	10
Proposed Algorithm	0.68	0.71	0.77	0.69	0.54	0.70	0.54	0.68	0.63	0.66
Clustering Algorithm based on TF-IDF	0.74	0.50	0.74	0.68	0.57	0.60	0.62	0.57	0.64	0.50
Amalgamation K–Means Algorithm	0.52	0.52	0.56	0.47	0.64	0.65	0.52	0.47	0.64	0.61
The Execution Time(in ms) of the algorithms										
Execution No.	1	2	3	4	5	6	7	8	9	10
Proposed Algorithm	241	219	212	286	210	213	218	216	211	215
Clustering Algorithm based on TF-IDF	361	326	396	312	327	320	313	318	313	328
Amalgamation K–Means Algorithm	3191	1017	992	979	1128	1014	1017	960	957	985

As in table 2 has showed, the accuracy and execution time of algorithms in several executions are different. So, the results have uncertainty. This uncertainty resulted to choose the k-means algorithm of first clusters data at randomly. To contrast with uncertainty in measuring, there are many statistical analyses that will present in the rest of the paper.

Figure 2 (a and b) shows scatter diagram of algorithms for accuracy and execution time in different executions.

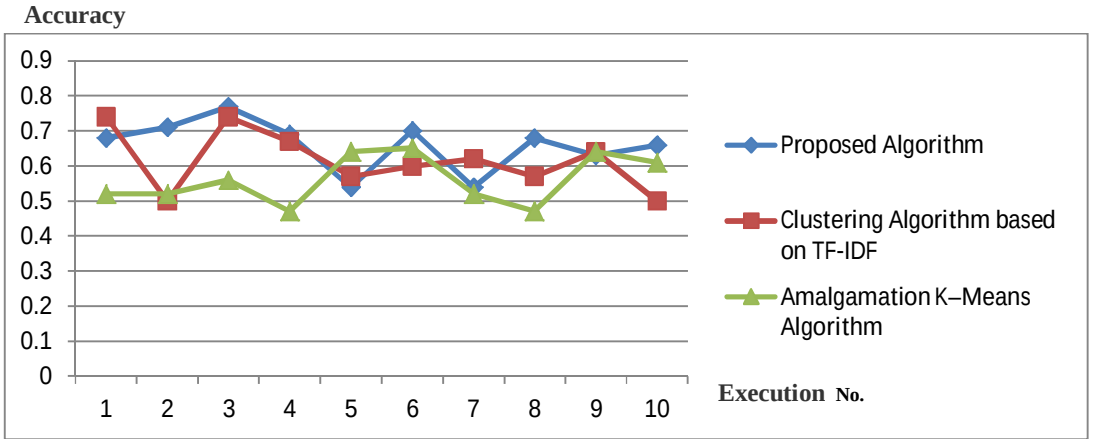


Figure 2(a): Scatter diagram of algorithms accuracy

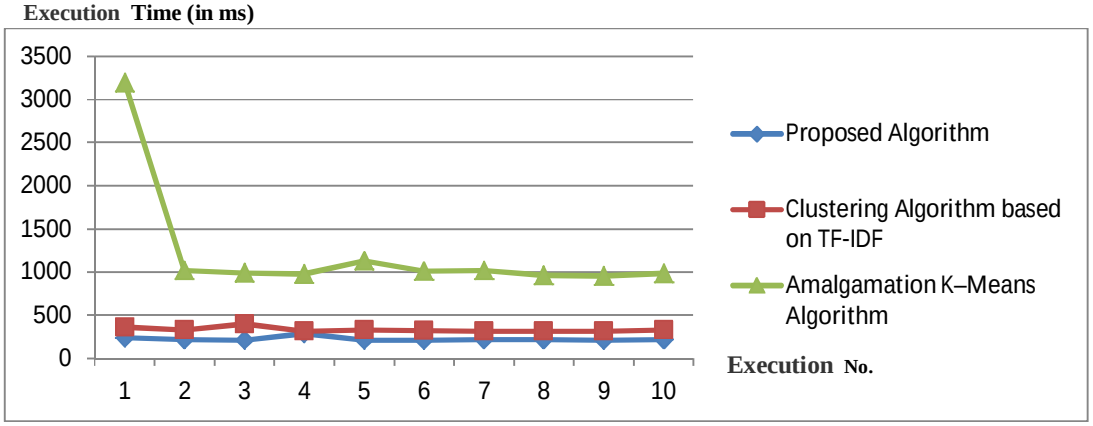


Figure 2(b): Scatter diagram of algorithms execution time

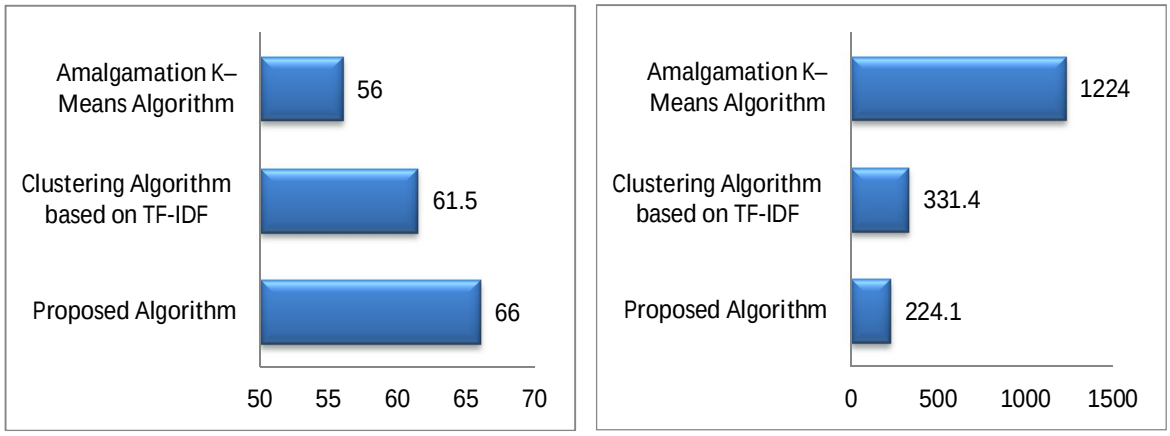
One method to calculate of unreliability in statistic is mean of data. Whatever the amount of data, that we want to test, be more, to the same ratio, the value of mean amount will be better. Table 3 shows the descriptive statistics numbers to evaluate accuracy and execution time of algorithms, which we want to test them.

Table3: Descriptive statistics for accuracy and execution time

Descriptive Statistics Accuracy						
	N	Min	Max	Mean	Std.Deviation	Variance
Proposed Algorithm	10	54	77	66	52.889	7.27247
Clustering Algorithm based on TF-IDF	10	50	74	61.5	72.944	8.54075
Amalgamation K–Means Algorithm	10	47	65	56	49.333	7.02377
Descriptive statistics Execution Time						
	N	Min	Max	Mean	Std. Deviation	Variance
Proposed Algorithm	10	210	286	224.1	23.28209	542.056
Clustering Algorithm based on TF-IDF	10	312	396	331.4	26.81708	719.156
Amalgamation K–Means Algorithm	10	957	3191	1224	692.82497	480006.444

According to table 3, we concluded that the accuracy of proposed algorithm is better than other algorithms. Also, the execution time of proposed algorithm is less than other algorithms, so this algorithm has higher performance.

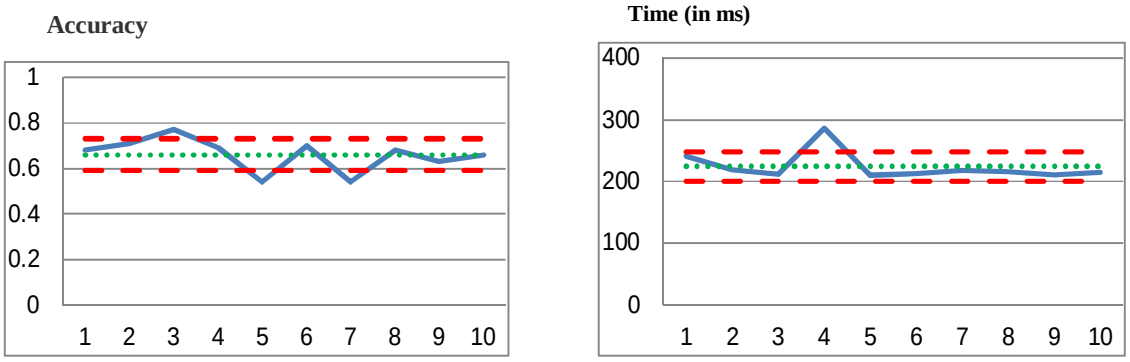
Figure 3 (a and b) shows bar diagrams of algorithms for accuracy and execution time in different executions.



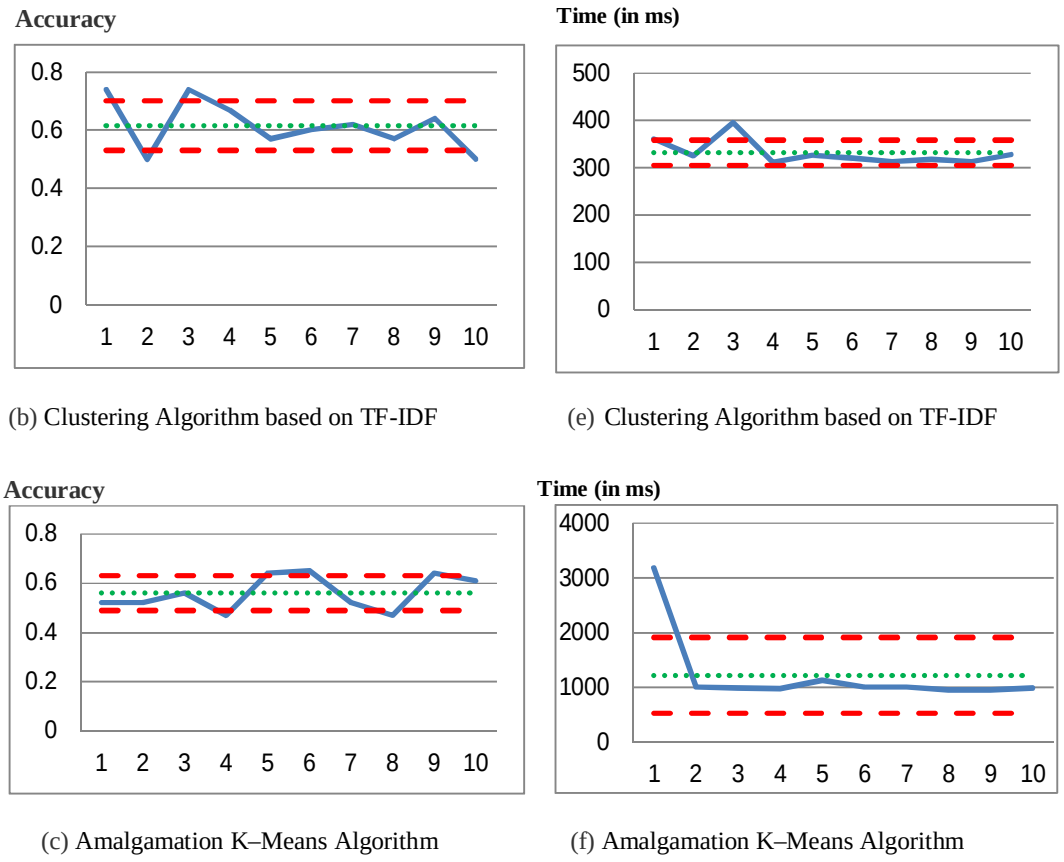
(a) Accuracy (b) Execution Time

Figure 3: bar diagram for means algorithms

Figures 4 (a through f) show the accuracy and execution time of scatter diagrams around the mean axes, which the scope of the standard deviation marked with red lines and means with green dots.



(a) proposed algorithm (d) proposed algorithm



(b) Clustering Algorithm based on TF-IDF

(e) Clustering Algorithm based on TF-IDF

(c) Amalgamation K–Means Algorithm

(f) Amalgamation K–Means Algorithm

Figure 4 (a) & (b) & (c): Scatter diagram of algorithms accuracy; (d) & (e) & (f): Scatter diagram of algorithms execution time.

The confidence interval generated an approximation range of values which is likely to include an unknown population parameter, the approximation range being calculated from a given set of sample data. To calculate confidence intervals, we used the one sample t-test[14]. The following table shows the results.

Table4: One Sample T-Test result

One Sample T-Test For Accuracy						
	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence interval of the Difference	
					Lower	Upper
Proposed Algorithm	28.699	9	0.000	66	60.7976	71.2024
Clustering Algorithm based on TF-IDF	22.771	9	0.000	61.5	55.3903	67.6097
Amalgamation K–Means Algorithm	25.213	9	0.000	56	50.9755	61.0245
One Sample T-Test For Execution Time						
	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence interval of the Difference	
					Lower	Upper
Proposed Algorithm	30.493	9	0.000	224.1	207.85	241.16
Clustering Algorithm based on TF-IDF	39.079	9	0.000	331.4	312.22	350.58
Amalgamation K–Means Algorithm	5.587	9	0.000	1224	728.38	1719.62

As in table 4 has showed, with 95% confidence, the accuracy of the proposed algorithm is between 60.80 and 71.20%.Also, execution time of the proposed algorithm is between207 and 241 MS.

6.Conclusion

Clustering algorithms are used extensively in various applications. The methods of web page clustering is considering in less level. Thus, the use of clustering methods will be suitable to dynamic environments, such as web in which are thousand pages add to this area every day. In this paper, a new approach present to indexing web pages based on content, in order to increases the organizing the web pages. Finally, proposed method was compared with some previous works on Persian web pages and the results showed that proposed method outperforms the previous ones in terms of accuracy and performance.

REFERENCES

[1] Ahlquist, J., Breunig, C., (2012) "Model-based clustering and typologies in the socialsciences", Political Analysis.

[2] Anderberg, M. R, (1973) "Cluster analysis for applications", Academic Press.

[3] Ball, G., Hall, D., (1965) "ISODATA, A novel method of data anlysis and pattern classification", Stanford ResearchInstitute, Stanford, CA.

[4] Bishop, Christopher M., (2006) "Pattern recognition and machine learning", Springer.

[5] Cailloux, O.,Lamboray, C., &Nemery, P., (2007) "A taxonomy of clustering procedures", Proceedings of the 66th Meeting of the European Working Group on MCDA, Marrakech, Maroc.

[6] Duda, R., Hart, P., & Stork, D., (2001) "Pattern classification, 2 edn". New York: John Wiley & Sons.

[7] Ehab A., Samhaa R., Salwa E., (2006) "A Feature Reduction Technique for Improved Web Page Clustering", IEEE.

[8] Fersini, E., Messina, E., &Archetti, F., (2010) "A probabilistic relational approach for web document clustering",Information Processing and Management.

[9] Han, J., &Kamber, M., (2011) “Data Mining: Concepts and Techniques, 3rd Edition”, Morgan Kaufman.

[10] Hartigan, J. A., (1972) "Direct clustering of a data matrix", Journal of the American Statistical Association, pp: 123–132.

[11] Jain, Anil K., Dubes, Richard C. , (1988)," Algorithms for clustering data", Prentice Hall.

[12] Macqueen, J., (1967) "Some methods for classification and analysis of multivariate observations", Fifth Berkeley Symposium on Mathematics, Statistics and Probability, University of California Press, pp: 281–297.

[13] Mahdavi, M., HaghirChehreghani, M., Abolhassani, H., &Forsati, R., (2008) "Novel meta-heuristic algorithms for clustering web documents ", Applied Mathematics and Computation 201,Elsevier Science Inc, Pages 441–451.

[14] Moore, D., & McCabe, G. (2006) "Introduction to the practice of statistics, 4th ed.", New York: Freeman.

[15] Napoleon,D., &Pavalakodi, S. , (2011) "A New Method for Dimensionality Reduction using K-Means Clustering Algorithm for High Dimensional Data Set " ,International Journal of Computer Applications.

[16] Pasha, E., &Fatemi, A., (2006) "Intuitionistic fuzzy sets clustering (IFSC)with an application in psychology", Journal ofMathematicsand Applications.

[17] Sokal, Robert R., Sneath, Peter H. A., (1963), "Principles of numerical taxonomy", W. H. Freeman, San Francisco.

[18] Tan, P., Steinbach, M., & Kumar, V., (2006), "Introduction to Data Mining", University of Minnesota.

[19] Tukey, J., (1977) "Exploratory data analysis", Addison-Wesley.

[20] Voevodski, K.,Balcan, M.,Roglin, H.,Teng, S., & Xia,Y., (2012) "Active Clustering of Biological Sequences", Journal of Machine Learning Research.

[21] Zhong,T., Gefeng,Y.,Xu, O., &Zhisheng,L., (2011), "Application research of fuzzy clustering approach in new rural cooperative medical insurance system module", IEEE Computer Society.