Application of web usage mining for web page clustering and recommendation

Mohammad Mohammadi

mohammadimohammadi@chmail.ir

Abstract: In recent years, the explosion of items in the internet has caused difficulty of locating appropriate item to users. A personalized recommendation is an enabling mechanism to overcome information overload occurred in internet environment and delivering suitable user resources to users. This paper proposes a novel recommendation mechanism which is used for personalized recommender system (PRS). In this paper, in order to clustering in the way of organizational increase of web pages a new indexing method of web pages has been presented. 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. The experimental results show that our proposed method outperforms collaborative filtering algorithm and can perform superiorly and alleviates problems such as cold-start and sparsity.

[Mohammad Mohammadi. **Application of web usage mining for web page clustering and recommendation.** *Rep Opinion* 2016;8(1):31-38]. ISSN 1553-9873 (print); ISSN 2375-7205 (online). http://www.sciencepub.net/report. 5. doi:10.7537/marsroj08011605.

Key Words: Clustering, Web Pages, Data Mining, Indexing, K-Means

1. Introduction

The information available on the internet is increasing exponentially and it is necessary to create the technologies that can assist users to discover the most valuable information to them from all the available information. To help user deal with information overload and provide personalized recommendations, recommender systems have become an important research area since the first paper on collaborative filtering in the mid-1990s [1]. The task of delivering personalized item is often framed in terms of a recommendation task in which a system recommends items to an active user [2]. Several educational recommender systems have been proposed in the literature that the most of them focus on recommending suitable materials or learning activities [3].

In the recent years, recommender system is being deployed in more and more e-commerce entities to best express and accommodate customer's interests. According to their strategies, recommender systems can be divided into three major categories: content-based, collaborative, and hybrid recommendation [4].

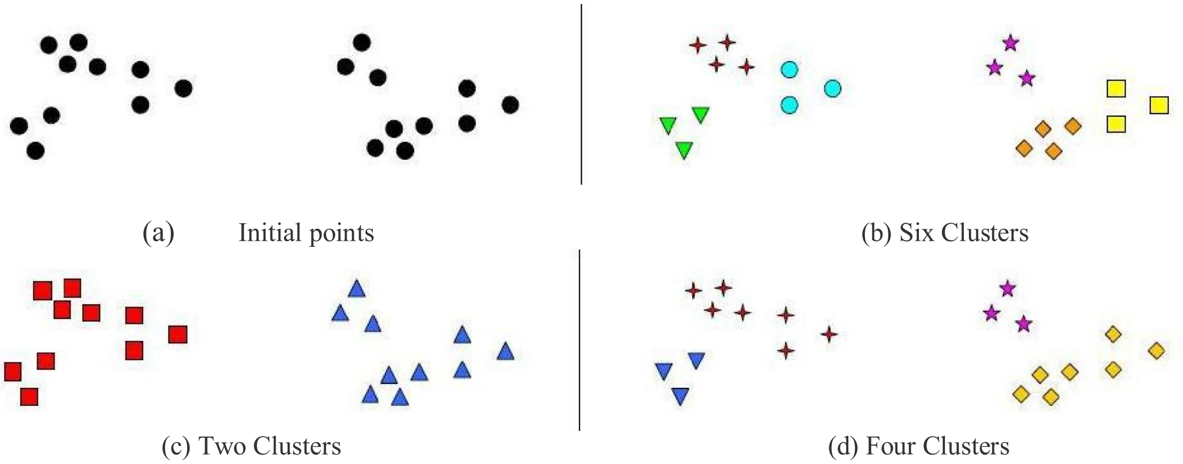
In the recent years web personalization has undergone through tremendous changes. The content, collaborative and hybrid based filtering are three basic approaches used to design recommendation systems. The content based filtering relies on the content of an item that user has experienced before. The content based information filtering has proven to be effective in locating text, items that are relevant to the topic using techniques such as Boolean queries, vector space queries etc. However, content based filtering has some limitations. It is difficult to provide appropriate recommendation because all the information is selected and recommended based on the content. Moreover, the content based filtering leads to overspecialization i.e. it recommends all the related items instead of the particular item liked by the user. The collaborative- filtering aims to identify users who have relevant interests and preferences by calculating similarities and dissimilarities between their profiles. The idea behind this method is that to one's search the information collected by consulting the behavior of other users who shares similar interests and whose opinions can be trusted may be beneficial. The different techniques have been proposed for collaborative recommendation; such as correlation based method, semantic indexing etc. The collaborative filtering overcomes some of the limitations of the content based filtering. The system can suggest items to the user, based on the rating of items, instead of the content of the items which can improve the quality of recommendations.

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 [5].

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.

To better understanding the problem of deciding what constitutes a cluster, consider figure 1(a) to figure 1(d), which show twenty points and three different ways that they can be divided into clusters. The most reasonable interpretation of the structure of these points is that there are two clusters, that each of which has three subclusters. By the way, the apparent division of the two larger clusters into three subclusters may simply be an artifact of the human visual system. Eventually, it may not be unreasonable to say that the points form four clusters. Therefore, emphasize once again that the definition of what constitutes a cluster is ambiguous, and the best definition depends on the type of data and the desired results [24].


(a) Initial points

(b) Six Clusters

(c) Two Clusters

(d) Four Clusters

Figures 1: Dividing points into clusters [6]

2. Literature review

Recommender systems have already implemented in real e-commerce applications such as Amazon [7] and CDNow [8] where they are used to recommend to online shoppers, products and services that they might otherwise never discover on their own.

Most of recommendation systems are designed either based on content-based filtering or collaborative filtering. Both types of systems have inherent strengths and weaknesses, where content-based approaches directly exploit the product information, and the collaboration filtering approaches utilize specific user rating information. In addition, to produce the accurate and effective recommendations and ensure the real-time requirement of the system, researchers proposed several different algorithms, some of which derives from the achievements of data mining. Some of recommending algorithms are user-based collaborative filtering [9], Item-based collaborative filtering [10], Cluster-based collaborative filtering [11], Dimension reduction based collaborative filtering [12], Horting Graph-theoretic collaborative filtering [13], Bayesian network based recommendation (Herlocker, 2000). In the following of this section, we explain some researches in four categories.

Collaborative filtering: Majority of researchers used collaborative filtering based recommendation system [14]. Based on the assumption that users with similar past behaviors have similar interests, a collaborative filtering system recommends items that are liked by other users with similar interests [15]. Collaborative filtering methods are completely independent of the intrinsic properties of the items being rated or recommended.

Content based filtering: the recommendations are done based only in the profile made taking into consideration the object content analysis the user has evaluated in the past [16]. The Recommendation Systems based in the content are mainly used to recommend documents, Web pages, publications, jokes or news. For example [17] used users' recent navigation histories and similarities and dissimilarities among user preferences and also among the contents of the learning resources for online automatic recommendations.

Data mining: The data mining techniques use the gathered information about the user behavior, such as navigation history, to produce recommendations. These techniques are suitable to recommend the sequence of learning resources (i.e., learning path) rather than the learning resources itself. Clustering was proposed to group learning documents based on their topics and similarities. Data mining techniques such as Association Rule mining, and inter-session and intra-session frequent pattern mining, were applied.

Hybrids: Each recommendation strategy has its own strengths and weaknesses. Hence, combining several recommendation strategies can be expected to provide better results than either strategy alone [18-20]. Most hybrids work by combining several input data sources or several recommendation strategies.

In summary, in order to improve the learning resource recommendation efficiency, developing a framework for integrating contextual information including multi-dimensional attributes of resources and user's rating information is necessary. Most of researches only use some of this information in resource recommendation process.

3. K-means algorithm

K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

The problem is computationally difficult (NP-hard); however, there are efficient heuristic algorithms that are commonly employed and converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm formixtures of Gaussian distributions via an iterative refinement approach employed by both algorithms. Additionally, they both use cluster centers to model the data; however, k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

4. Material and methods

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).

Table 1: classification of the documents' words in seven classes

Documents No.	1	2	3	4	5	6	….
Number of 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 miscellaneous words	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 value 0.45, 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 also distance covered in intended cluster. By introducing the proposed weighting approach would overcome the problem.

5. Evaluation

We have conducted a set of experiments to examine the effectiveness of our proposed recommender system in terms of accuracy of neighbour-selection, cold start and recommendation quality. In this work, the results of the proposed algorithm are compared with the algorithm presented in [21], which works based on TF-IDF method and Amalgamation K–Means Algorithm [22].

Table 2 shows 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 3 (a and b) shows scatter diagram of algorithms for accuracy and execution time in different executions.

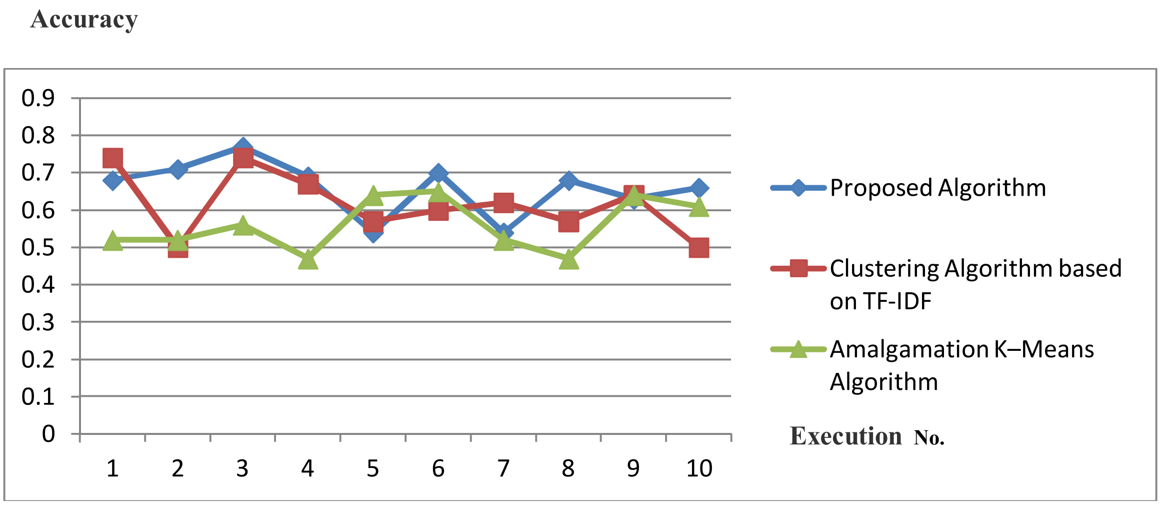


Figure 3 (a): Scatter diagram of algorithms accuracy

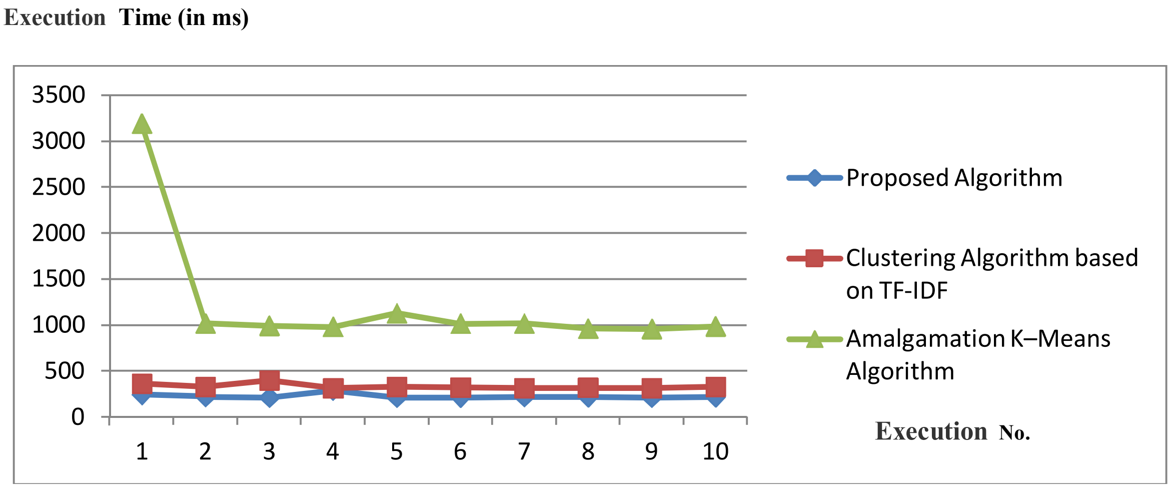


Figure 3 (b): Scatter diagram of algorithms execution time

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.

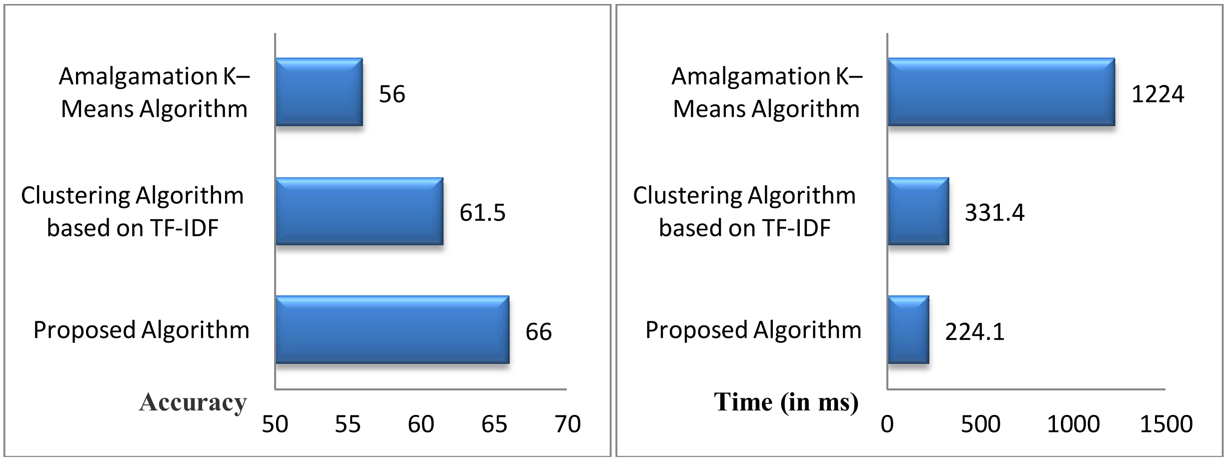
Table 3: Descriptive statistics for accuracy and execution time

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.

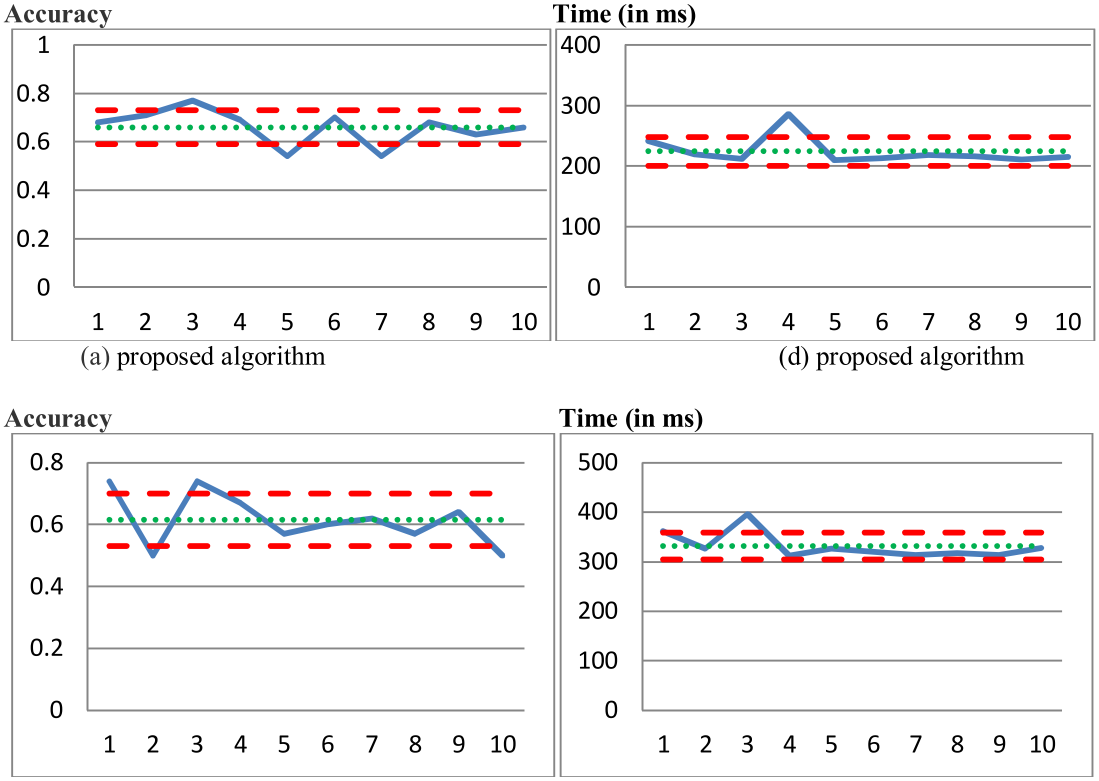
Figure 4 (a and b) shows bar diagrams of algorithms for accuracy and execution time in different executions.

Figures 5 (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.



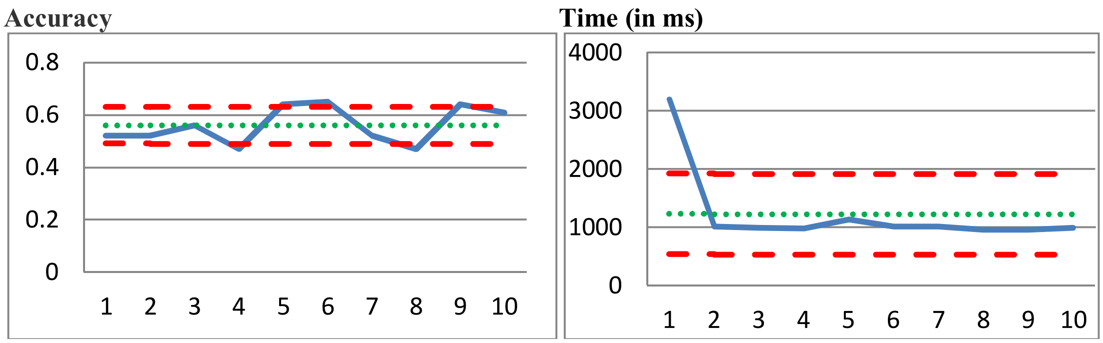
(a) Accuracy (b) Execution Time

Figure 4: bar diagram for means algorithms



(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 5 (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 [23-25]. The following table shows the results.

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 between 207 and 241 MS.

Conclusion

This paper describes a novel personalized recommender system that utilizes clustering of approach and provides the recommendations for the active user with good quality rating using similarity measures. 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.

Table 4: 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

References:

1. Ahlquist, J., Breunig, C., (2012) "Model-based clustering and typologies in the social sciences", 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 Research Institute, 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. Drineas, P., Frieze, A., Kannan, R., Vempala, S., & Vinay, V., (1999) "Clustering large graphs via the singular value decomposition", Machine Learning.
7. Duda, R., Hart, P., & Stork, D., (2001) "Pattern classification, 2 edn". New York: John Wiley & Sons.
8. Ehab A., Samhaa R., Salwa E., (2006) "A Feature Reduction Technique for Improved Web Page Clustering", IEEE.
9. Fersini, E., Messina, E., & Archetti, F., (2010) "A probabilistic relational approach for web document clustering", Information Processing and Management.
10. Han, J., & Kamber, M., (2011) “Data Mining: Concepts and Techniques, 3rd Edition”, Morgan Kaufman.
11. Hartigan, J. A., (1972) "Direct clustering of a data matrix", Journal of the American Statistical Association, pp: 123–132.
12. Jain, Anil K., Dubes, Richard C., (1988)," Algorithms for clustering data", Prentice Hall.
13. Jain, Anil K., (2009) "Data Clustering: 50 Years Beyond K-Means", Pattern Recognition Letters.
14. JSTOR, (2009), http://www.jstor.org.
15. Lioyd, S., (1982) "Least squares quantization in PCM", IEEE Transactions on Information Theory, Originally as an unpublished Bell laboratories Technical Note (1957), pp: 129–137.
16. 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.
17. Mahdavi, M., Haghir Chehreghani, 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.
18. Meila, M., (2006) "The uniqueness of a good optimum for k-means", Proceedings of the 23rd International Conference on Machine, pp: 625–632.
19. Moore, D., & McCabe, G. (2006) "Introduction to the practice of statistics, 4th ed.", New York: Freeman.
20. 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.
21. Pasha, E., & Fatemi, A., (2006) "Intuitionistic fuzzy sets clustering (IFSC) with an application in psychology", Journal of Mathematics and Applications.
22. Sokal, Robert R., Sneath, Peter H. A., (1963), "Principles of numerical taxonomy", W. H. Freeman, San Francisco.
23. Steinhaus, H., (1956), "Sur la division des corpmateriels en parties", Bulletin of Acad. Polon, pp: 801–804.
24. Tan, P., Steinbach, M., & Kumar, V., (2006), "Introduction to Data Mining", University of Minnesota.
25. Tukey, J., (1977) "Exploratory data analysis", Addison-Wesley.

1/24/2016