A Novel Web-Page Clustering Method Using K-Means Improved with Cellular Learning Automata And Genetic Algorithm

Peyman Almasinejad† and Mohammad Javad Kargar††,

†Department of computer Engineering and information technology, Payame Noor University, I.R. of Iran
†† Department of Computer Engineering, University of Science and Culture, Tehran, Iran

Summary
Improvement of accuracy and optimal function of search engines has always been an area of concern for designers and researchers. Although these search engines work relatively well in simple searches, because the most existing search algorithms are based on search keywords, it can be expected for search engines to face trouble and confusion in some states of advanced searching. A possible solution is implementing Web Resources Categorization before performing the search. This study examines the basic web page clustering algorithms with the help of a k-means algorithm and optimizes its performance by solving its problems. The main issue is in the initial selection of clusters which can have a significant impact on the final clustering. Therefore, this research study proposes a new method for optimizing the core algorithm using cellular learning automata algorithm based on the Genetic Algorithm.

Key words:
k-means Algorithm, Evolutionary Computation Algorithm, Cellular Learning Automata, Web Page Clustering, Genetic Algorithm

1. Introduction

Data clustering is an important subject in data mining which acts as one of the pre-phases of data processing and provides valuable results [1]. Web clustering is significant because it can be used to improve search engine results or the web creep operations, in addition to preprocessing [2]. This not only resolves the need for manual organization of information but also it can improve the recovery efficiency by restricting the search to a limited number of clusters. Ultimately, this allows users to have easier access to the collection of documents [3-6].
Clustering is defined as positioning the data in different groups so that the intra-group similarities are minimum and inter-group similarities are maximum [4]. Regarding application of clustering on the web, there are two categories of works: clustering of domain-specific web documents and clustering of search results [30]. Clustering of web documents involves several challenges including defining the characteristic of the documents and determining an appropriate weight for each of them, choosing a clustering approach and establishing a suitable criterion for similarity, as well as the limitations of computational resources and memory [21, 30].
We continue to section 2 where procedures and criteria of web page clustering are provided. The clustering is then defined completely in section 3. Section 4 introduces the k-means clustering algorithm which is applicable to web pages. Section 5 addresses the improvement of the given algorithm using cellular learning automata algorithm based on revolutionary computation and evaluates the clustering results of some common methods compared to proposed method.

2. Procedures and Criteria for Web Page Clustering

In this section, we will discuss the conventional procedure of clustering operations as well as the criteria to be considered for clustering.

2.1 Clustering Procedure

Clustering of the patterns usually includes the following steps [8]:

1) Pattern Recognition (this can also include extraction or selection of criteria and features),
2) Defining a measure for pattern Proximity (similarity) for the data range,
3) Clustering or Grouping,
4) Data Abstraction (if needed),
5) Evaluation or Validation of outputs (if needed).

Pattern Recognition involves with determining the number of classes and the number of available patterns for the existing clustering algorithm. Feature Selection is the process of determining the most effective subset of the main features to be used in clustering. Feature Extraction uses one or more conversion of input parameters to generate other new and significant features. Pattern Proximity (similarity) is usually measured by defining a distance function on each pair of patterns. Different criteria are used for measuring the distance between patterns, the most famous of which is the Euclidean Distance [10]. The output of clustering can be hard or

Manuscript received June 5, 2017
Manuscript revised June 20, 2017

fuzzy groups (each pattern may have a different membership level in each group [26]). The feedback path indicates that cluster analysis is not a sudden process and in many cases, one can sense the need for iteration and rotation between stages. Data Abstraction is the process of extracting a (simple and compact display of) dataset [11]. Data Abstraction in content clustering is the description of a summary of each cluster such as cluster centroid [12].

2.2. Clustering Criteria

One of the features that the web provides for information exploration is the application of new criteria for clustering. Generally, the criteria for web clustering can be categorized as Link-based criteria, content- (and structure-) based criteria and criteria which use a combination of the two [9].

2.2.1. Link-Based Criteria

Three criteria are considered in link analysis of web structures: co-link, coupling, and co-linkage. The co-linkage of a pair of web domains is the number of domains that they are both linked to. Coupling of two domains is also the number of domains that they are both linked to. Together, these two criteria are known as Cohesion. The reason that we use coupling and co-linkage along with co-link is that those sites which do not belong to the same community usually do not link out to each other due to competitive reasons. However, these pages can be linked through coupling or co-linkage. The results show that the combination of the three criteria results in the highest probability for identifying similar sites. However, the main improvement is achieved through using the links [13].

2.2.2. Content and Structure-Based Criteria

Conventional methods of documents clustering valued all parts of the text, including the title, keywords, etc. the same. Snippets and anchor texts are two significant sources of web page clustering beside the content of the document. Linking terms or phrases written around hyperlinks are other groups that can be used for clustering [13].

2.2.3. Combined Criteria

After the development of link-based criteria, it appears that these criteria alone cannot be sufficient for clustering and stating similarity since these criteria have very low remembrance with a high possibility of noise (false and biased links). Thus combining the link-based criteria with content-based criteria to create more accurate and comprehensive criteria can be useful [24].

2.3. Exploring the Web

There are three types of web exploring methods: exploring the web content, exploring the web structure, and exploring the web applications [19].

Exploring Web content is concerned with describing and discovering useful information from the web content, data, and documents. There are two approaches to exploring the web content: information retrieval approach and database approach. Exploring data content in data retrieval based on the content aims to help the data filtering process or to find the data for users which is usually done based on data extraction or user demand whereas database approach refers to web data modeling and combining, e.g. the majority of specific queries required for searching information.

Exploring the web structure attempts to discover the linking structures of the web. This model can be used to classify web pages and to generate information such as similarities and connections between different web sites.

Exploring web applications uses the data obtained from the results used to identify users' behavior models to automatically achieve web services. Identifying the models is the key component of web search that includes different algorithms and techniques in various fields of research, such as data analysis, machine learning, statistics, and modeling. One of the important model identification processes in the web is clustering. Clustering concerns with the process of grouping objects, so that similar objects are in the same group. Each of these groups is called a cluster. Cluster analysis is a technique for grouping data users or items (web pages) based on similar features.

3. Problem Definition

The process of grouping a set of physical or abstract objects in similar groups is called clustering. Objects of a cluster are different from one another and objects in other groups.

Consider a set of n objects as x= $\{x_1, x_2, \ldots, x_n\}$. Clustering attempts to group these objects in k clusters such as c= $\{c_1, c_2, \ldots, c_k\}$. Each cluster would be as follows:

$$1)\ C_1 \cup C_2 \cup \cdots \cup = X$$

$$2)\ C_i \neq \phi \quad i=1,2,3,\ldots,k$$

In the above definition, difference scenarios of clustering n objects into k cluster would be as follows:

$$NW(n,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^i \binom{k}{i}(k-i)^n \qquad (1)$$

In most methods, the use determines the value of k. The above equation shows that even when k is a known value,

finding the best clustering scenario is not easy. Moreover, the clustering methods of n objects into k clusters increase the value of $k^n/k!$. Therefore, the clustering problem for clustering n objects into k clusters is an NP problem in the best case scenario.

For clustering, each web page is considered to be a point in the space that we will draw later on. To do so, we would need information about each page. This information is obtained by web creep which gathers the required information about the pages and puts them in a database. Here, we will study the algorithm with a standard Dataset obtained from the UCI Repository in order to achieve a standard and reliable conclusion. These data possess nine features and 650 records, the information of which is shown in Table 1.

Table 1: Features of the Web Layout Data

Feature	Minimum	First Quartile	Median	Average	Third Quartile	Maximum
1	2	55	108	108	161	214
2	1.511	1.517	1.518	1.518	1.519	1.534
3	10.73	12.9	13.3	13.41	13.83	17.38
4	0	2.09	3.48	2.676	3.6	3.98
5	0.29	1.19	1.36	1.447	1.63	3.5
6	69.81	72.28	72.79	72.66	73.09	75.41
7	0	0.13	0.56	0.4991	0.61	6.21
8	5.43	8.24	8.6	8.958	9.18	16.19
9	0	0	0	0.1759	0	3.15

Therefore, each web page is a point positioned in a nine-dimensional space in which the clustering is going to be conducted.

Examination of the clusters with criteria is one of the important parts of clustering. We will use Davies-Bouldin Index for this purpose [28]. Davies-Bouldin Index is a function of the ratio of total inter-cluster diffraction to distance between clusters. Davies-Bouldin validation index is shown by equation (2). This method works based on minimization.

$$DB = \frac{1}{n}\sum_{\substack{i=1 \\ i\neq j}}^{n} max\left\{\frac{S_n(Q_i)+S_n(Q_j)}{S(Q_i,Q_j)}\right\} \qquad (2)$$

Where n is the number of clusters, S_n is the average distance of the cluster data from the cluster center, and $S(Q_i, Q_j)$ is the distance between cluster centers. Therefore, when the inter-cluster objects are close, and the clusters are far from each other, we will have a smaller ratio. The smaller the value of Davies-Bouldin index, the higher the validity of clustering.

4. K-means Algorithm for Web Page Clustering

K-means algorithm is one of the common and widely used algorithms in web page clustering that is less dependent on the problem and application type compared to other algorithms. In its basic state, this algorithm finds a representative for each cluster and assigns the web page to these representatives. Then, each cluster center is updated considering to its members. The assign-update process repeats until a specific ending criterion is met [2]. Despite its simplicity, this algorithm is considered to be a basic method for many other clustering methods. There are different types of this algorithm. However, all of them have the same routine that attempts to estimate the following for a fixed number of clusters:

- Finding points for cluster centers. These points are in fact the mean of the points that belong to each cluster.
- Assigning each sample data to a cluster that has the shortest distance from the center of that cluster.

In the basic form of this method, first, a number of points equal to required clusters are selected randomly. Then the data is assigned to one of these clusters according to their proximity (similarity), and thus, new clusters are created. By repeating this routine, each time we can calculate new centers by finding the means of the data and then re-assign them to new clusters. This will continue until the data would not change any more. K-mean algorithm is shown in Figure 1.

```
01  Select an Initial Partition (k centers)
    Repeat
02  Data Assignment: Re-compute Membership
03  Relocation of "means": Update Centers
04  Until (Stop Criterion)
05  Return Solution
```

Fig. 1 k-means Algorithm

As is shown in the Figure, the k-mean algorithm consists of two measurable parameters. The first being the number of clusters (k) which should be determined from the beginning of the process and second being the starting points of the algorithm. These starting points are selected randomly in the standard algorithm.

The algorithm should be run as follows for web page clustering in a nine-dimensional space:

1- First, k points are selected in the nine-dimensional space as the cluster centers.

2- Each web page (which is shown as points in this space) will be assigned to the cluster that its center is closest to that data. The distance between each web page and the

center of each cluster is determined using Euclidean Distance method which is shown in equation 3.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad , p = (p_1, p_2, \dots, p_n), q = (q_1, q_2, \dots, q_n) \qquad (3)$$

3- After all the web pages are assigned to a cluster, a new point is calculated for each cluster as the center (the mean of the points of that cluster).

4- Steps 2 and three are repeated until the center of clusters no longer change.

All the running steps of the algorithm in a two-dimensional space is shown in Figure 2.

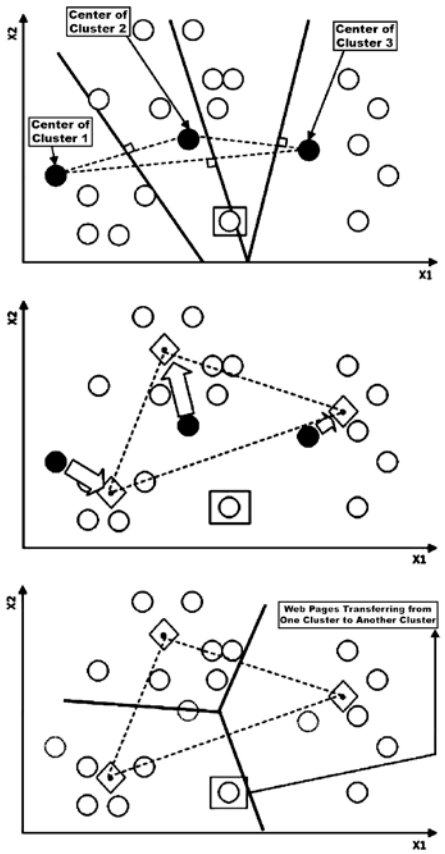

Fig. 2 Performance of the k-means Algorithm in the Two-Dimensional Web Page Environment

Knowing the number of clusters (k) from the beginning is important. Now we want to find the appropriate k for the solution by running the above-mentioned algorithm in a nine-dimensional space. We will start by setting k=2. Then we increase its value one unit at a time while running the algorithm each time. This should be repeated until the optimal number of clusters with the lowest value of Davies-Bouldin validation index is determined.

To find the optimal k, its value increases while the clustering validation index decreases. This, however, won't be the case forever since after some time, the index value starts to increase. Figure 3 shows how k is selected. The index decreases for k=2 to 5 and then it starts to increase again. Thus, the optimal value of k in Figure 3 is equal to 5.

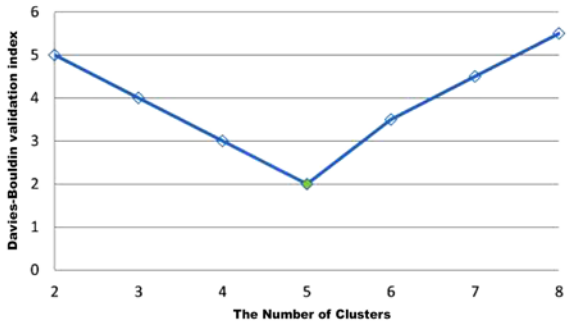

Fig. 3 Finding the Optimal Number of Clusters

This algorithm has problems, some of which are:

- The final result depends on the selection of initial clusters.
- There is no specific routine for initial calculation of cluster centers.
- If in one running iteration, the number of data assigned to a cluster turns out to be zero, there is no way for changing and improving the rest of the process.

To solve these problems as well as improving the consistency of the algorithm with web features, we have developed and improved the algorithm in the next section.

5. Improving the k-means Algorithm for Web Page Clustering

This section intends to optimize the k-means algorithm for web page clustering using cellular learning automata based on revolutionary computation [29] which is tasked to determine the starting points of the algorithm.

5.1. The Structure of the Cellular Learning Automata

First, we need to specify the cellular learning automata. For this problem, we are going to use a one-dimensional CLA (Cellular Learning Automata). The number of cellular learning automata elements is equal to records of the standard data of web pages. We assign a number to each record. Figure 4 shows an overall view of 8 records where records are numbered. In fact, figure 4 has eight records, each of which has two features. That is why this figure is shown in a two-dimensional space whereas if three features were used, space would have been three-

dimensional, and if n features were used, space would have been n-dimensional.

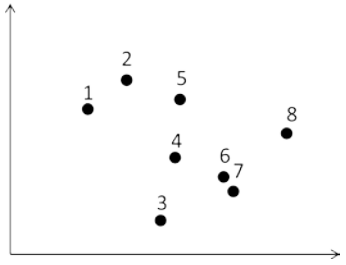

Fig. 4 Overall View of Numbering Records with 8 Records in a Two-Dimensional Space

Next, we will determine the number of clusters and randomly select the same number of points. For example, in figure 4, the number of clusters is two and two points of 2 and 7 are selected as initial cluster centers. These two points are colored blue in the overall view of Figure 5.

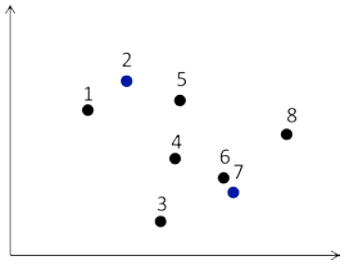

Fig. 5 Two Points Are Randomly Selected

The appropriate automata structure for Figure 5 is one-dimensional, and its values are in bits. If the value is equal to zero, that means no point should be selected. If it is equal to 1, then one point should be selected as the cluster center. Figure 6 shows the appropriate automata for Figure 5.

1	2	3	4	5	6	7	8
0	1	0	0	0	0	1	0

Fig. 6 The Appropriate Cellular Learning Automata Structure for Figure 5

The overall structure of proposed cellular learning automata is one-dimensional, and it has n cells. The values of these cells are in bits. This structure is shown in Figure 7 where n is the number of records.

1	2	3	…	N
0~1	0~1	0~1	…	0~1

Fig. 7 Structure of Proposed Cellular Learning Automata

Therefore, the number of elements equals the number of clusters, and the cellular learning automata are calculated by equation 4 when the number of clusters is k. In this equation, n is the number of records.

$$\sum_{i=1}^{n} Cellular\ Learning\ Automata_i = k \qquad (4)$$

The following algorithm shown in Figure 8 is used to determine the number of initial points of cellular learning automata.

```
Initialization Cellular Learning Automata
1.  For i=1 to k
2.    Do
3.      Rand=Random Between 1 and n
4.    While (Cellular Learning Automata Rand< >1)
5.    Cellular Learning Automata Rand=1
6.  End for
End
```

Fig. 8 Initialization of Cellular Learning Automata

Another important point regarding cellular learning automata is defining its Neighborhood. The proposed neighborhood definition is that each point has a neighborhood with all the other points. That is so that we could exchange the values of each cell in a CLA with zero points. Therefore, the transfer function is defined so that only one cell with the value of 1 is exchanged with a cell with the value of zero.

The cellular learning automata should be evaluated after each exchange of the cluster centers. This evaluation is done with the help of determining the distance of each page with its cluster center using Euclidean Distance method shown in equation 3.

The quality of clustering will then be evaluated based on equations (5).

$$\begin{aligned}
&O = \{c^n \mid n = 1,\ldots,k\} \\
&O^n = \{C_i \mid i = 1,\ldots, \|T^c - O\|\} \\
&\rho(k) = \frac{1}{k}\sum_{n=1}^{k}\left(\min\{\frac{\eta_n + \eta_m}{\delta_{nm}}\}\right) \\
&\eta_n = \frac{1}{\|O^n\|}\sum_{c_i \in O^n} Sim(c_i, c^n) \\
&\eta_m = \frac{1}{\|O^m\|}\sum_{c_j \in O^m} Sim(c_j, c^m) \\
&\delta_{nm} = Sim(c^n, c^m)
\end{aligned} \qquad (5)$$

As an ending criterion for cellular learning automata, after each CLA exchange operation, the best clustering is saved. If no better cluster was created after some times (e.g. 100 times calculated by trial and error), the cellular learning automata would stop.

5.2. Using Genetic Algorithm as Part of the Solution

The genetic algorithm receives the cellular learning automata as an input to determine the starting points of the k-mean algorithm.
To determine the initial population of the algorithm we need to turn the final output of each cellular learning automata into a chromosome. Figure 9 shows an example of turning the cellular learning automata to clustering.

1	2	3	4	5	6	7	8
1	0	0	0	0	0	1	0

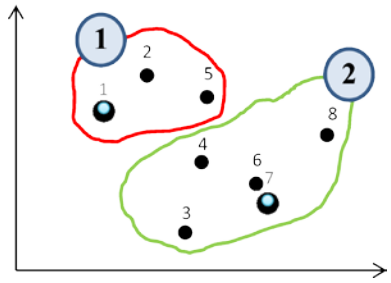

Fig. 9 Example of Turning the Cellular Learning Automata to Clustering

Then we will define the chromosome based on the created clusters. The number of chromosomes is equal to defined records in the standard data and gens contain the number of clusters. For example, in Figure 9, records 1, 2 and five are related to the first cluster and the rest of the records are related to the second cluster. The equivalent chromosome of Figure 9 is shown in Figure 10.

1	2	3	4	5	6	7	8
1	1	2	2	1	2	2	2

Fig. 10 Chromosome Created Based on Cellular Learning Automata in Figure 9

When the formation method for the initial population was determined, then it is time to determine the size of the population. The size of the population is an important factor in the efficiency of the algorithm. If the population is too small, a small section of the solution space would be searched and the solution would tend towards a local optimum quickly and with high probability. If the population is too large, reaching the solution would need numerous calculations and, as a result, running would take a very long time. The best scenario is an initial population of 200, which has been calculated with trial and error.
Considering that the chromosome is the same as cellular learning automata, the fitness function would be equal to fitness function section of learning automata. Now, we will use the Ranked-Based Selection method to select the parents. The reason is preventing the premature convergence of genetic algorithm and creating divergence capability in the algorithm.
In the next step, we will use the uniform crossover operator. This operator chooses the value of the child gene according to the values of the corresponding genes of both parents. In this method, the values of each parent's genes have equal chance to participate in corresponding child genes. Based on a random distribution this operator determines that the value of each child gene will be selected from which corresponding parent gene. An example of this operation is shown in Figure 11.

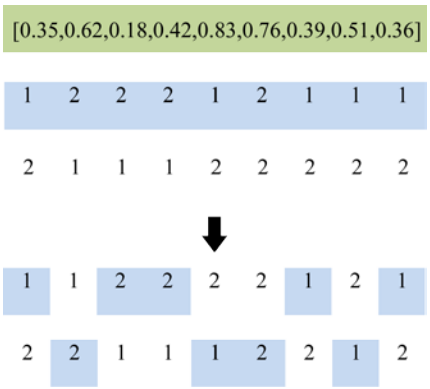

Fig. 11 Structure of Uniform Crossover

In the second part of the composition, in the roulette wheel selection scenario, we will use a one-point crossover operator, the structure of which can be seen in Figure 12.

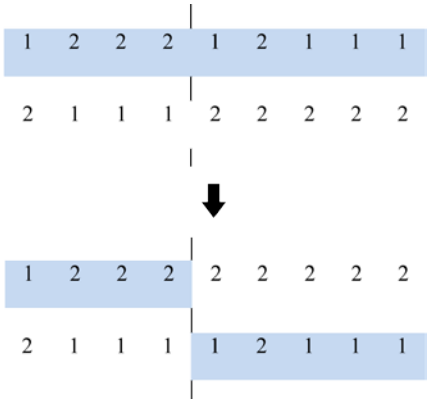

Fig. 12 Structure of One-Point Crossover

Mutation Operator is the other genetic algorithm operator that will be discussed in this section. The mutation operator selects a gene randomly and re-initializes it. Obviously, due to lack of exploitation of the population information, this is completely compatible with the definition of mutation operator, and it attempts to make the algorithm divergent. The mutation structure can be seen in Figure 13

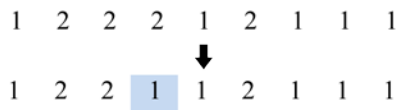

Fig. 13 Mutation Structure

We will now use the generational Replacement Method because this method can both be divergent and convergent by changing its variables. Thus, 50% of the parents and 50% of the children are replaced in the next generation which makes the problem dynamic.

In the end, the ending criterion of the genetic algorithm is set to be the production of 100 unchanged generations in the fitness function and the best chromosome with the best exit criteria will be selected. Then, clusters are drawn, and the means of the clusters are selected as the starting point of the k-means algorithm. Now we un the k-means algorithm. Figure 14 illustrates the formation of the starting points of k-means algorithm.

1	2	3	4	5	6	7	8
1	1	1	1	1	2	2	2

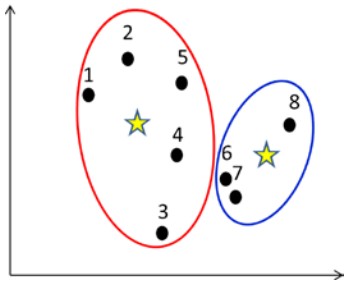

Fig. 14 Formation of the Starting Points of k-means Algorithm

The proposed algorithm flowchart can be seen in Figure 15.

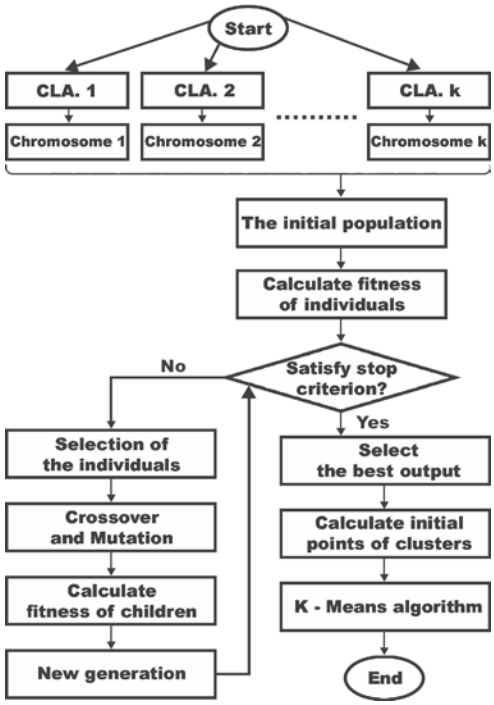

Fig. 15 Flowchart of the Proposed Algorithm

6. Evaluation of the Proposed Algorithm

This section compares the results of the proposed method on the UCI standard dataset which contains the web page clustering information with the results of several other evolutionary clustering offered in recent years. This comparison is conducted by the help of Davies-Bouldin index which works on the basis of minimization. Algorithms that this study has selected for comparison include Basic Algorithm, k-means, Fuzzy c-means, ACO, PSO, GA, CLA, K-Medoids, DBSCAN and the proposed algorithm. The results of evaluation of these algorithms are shown in Figure 16.

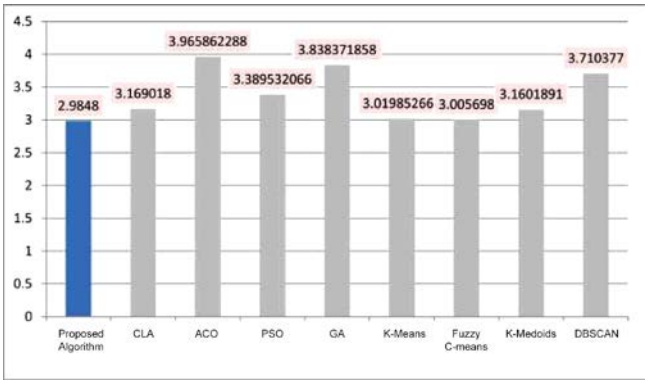

Fig. 16 Results of Evaluation of the Proposed Algorithm with other Algorithms

As is shown in Figure 15, the proposed algorithm performs better than the other eight methods. In order to find out how better does the proposed method work compared to other methods, we will use a statistical parameter called relabel which can be seen in (6).

$$\text{relabel} = \frac{|new\ value - old\ Value|}{old\ Value} \times 100 \qquad (6)$$

The proposed algorithm is evaluated compared to other methods based on the statistical parameter of relabeling. The result of this evaluation can be seen in Figure 17.

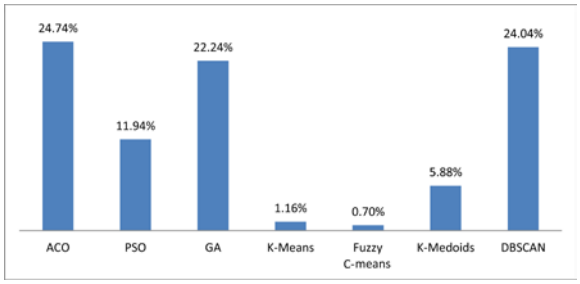

Fig. 17 Evaluation of the Result of the Proposed Algorithm Compared to Other Methods Based on Statistical Parameter of relabel

As is shown in Figure 17, Fuzzy c-means clustering method works better for the standard web page data; the proposed method is better than that method by 0.69%. The proposed method works 24.7% better than ACO algorithm which is the worst method for these data.

7. Conclusion

Web pages are rich sources of information, and they can be used as new features to improve clustering algorithms. Part of this information and features come through the link structure of the web, and the other part comes from the web content. However, since there is no specific discipline in web development, these features can also create profound challenges. As a result, examination of various aspects of these criteria and the associated algorithms is essential. On the other hand, there is a need for clustering the set of web pages so that the need for manual organization of information as well as increasing the efficiency of information retrieval operation by restricting the search to a limited number of clusters can be met. K-means algorithm is known as one of the most famous methods for web page clustering. However, this method has some shortcomings despite its high efficiency. This study attempts to examine these shortcomings while providing a method for resolving them and improving its performance. Since this algorithm begins its work by randomly selecting initial points for determination of clusters and this choice will definitely affect the final result of clustering, the initial points of clusters can be considered as an improvement in the performance of this algorithm. Therefore, the proposed method uses the cellular learning automata to determine the initial points of the cluster centers. Then we exploit the genetic algorithm to receive our input from the cellular learning automata and determine the initial points of clusters for the beginning of the k-means algorithm with the help of that algorithm.

References

[1] B. Arzanian, F. Akhlaghian, P. Moradi, A multi-agent based personalized meta-search engine using automatic fuzzy concept networks, Third International Conference on Knowledge Discovery and Data Mining (2010) 208–211.

[2] F. Akhlaghian, B. Arzanian, Moradi, P, A personalized search engine using ontology-based fuzzy concept networks, International Conference on Data Storage and Data Engineering (2010) 137–141.

[3] A. Asllani, A. Lari, Using genetic algorithm for dynamic and multiple criteria website optimizations, Eur. J. Oper. Res. 176 (3) (2007) 1767–1777.

[4] S. Bandyopadhyay, S.K. Pal, Classification and Learning Using Genetic Algorithms: Applications in Bioinformatics and Web Intelligence, Springer-Verlag, Hiedelberg, Germany, 2007.

[5] M. Boughanem, C. Chrisment, J. Mothe, C.S. Dupuy, L. Tamine, Connectionist and genetic approaches for information retrieval, Soft Comput. Inf. Retr. Stud. Fuzziness Soft Comput. 50 (2000) 173–198.

[6] M. Boughanem, C. Chrisment, L. Tamine, On using genetic algorithms for multimodal relevance optimization in information retrieval, J. Am. Soc. Inf. Sci. Technol. 53 (11) (2002) 934–942.

[7] H.J. Bremermann, The Evolution of Intelligence. The Nervous System as a Model of Its Environment, Technical Report No. 1, Department of Mathematics, University of Washington, Seattle, WA, 1958.

[8] D. Cai, S. Yu, J.-R. Wen, W.-Y. Ma, VIPS: a vision-based page segmentation algorithm, Microsoft Technical Report, MSR-TR-2003-79, 2003.

[9] J.P. Callan, Passage-level evidence in document retrieval, in: Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, 1994, pp. 302–310.

[10] S. Chawla, P. Bedi, Personalized web search using information scent, International Joint Conferences on Computer, Information and Systems Sciences, and Engineering, Technically Co-Sponsored by: Institute of Electrical& Electronics Engineers (IEEE), University of Bridgeport, published in LNCS (Springer) (2007) 483–488.

[11] C.C. Cheng, P.L. Chen, F.R. Chiu, Y.K. Chen, Application of neural networks and Kano's method to content recommendation in web personalization, Expert Syst. Appl.: Int. J. 36 (3) (2009) 5310–5316.

[12] E.H. Chi, P. Pirolli, K. Chen, J. Pitkow, Using information scent to model user information needs and actions on the web, in: International Conference onHuman Factors in Computing Systems, New York, NY, USA, 2001, pp.490–497.
[13] F. Crestani, G. Pasi, Soft Computing in Information Retrieval: Techniques and Application, 50, Physica-Verlag, Heidelberg, Germany, 2000.
[14] C. Ding, X. He, P. Husbands, H. Zha, H. Simon, Link Analysis: Hubs and Authorities on the World. Technical Report: 47847, 2001.
[15] W. Fan, M.D. Gordon, P. Pathak, Personalization of search engine services for effective retrieval and knowledge management, in: International Conference on Information Systems, Brisbane, Australia, 2000, pp. 20–34.
[16] W. Fan, M.D. Gordon, P. Pathak, Discovery of context-specific ranking functions for effective information retrieval using genetic programming, IEEE Trans. Knowl. Data Eng. 16 (4) (2004) 523–527.
[17] D.E. Goldberg, Genetic Algorithms in Search, Optimization and MachineLearning, Addison-Wesley Longman Publishing Co., Boston, MA, USA,1989.
[18] M. Gordon, Probabilistic and genetic algorithms in document retrieval, Commun. ACM 31 (10) (1988) 1208–1218.
[19] J. Heer, E.H. Chi, Separating the swarm: categorization method for user sessions on the web, International Conference on Human Factor in Computing Systems (2002) 243–250.
[20] J.T. Horng, C.C. Yeh, Applying genetic algorithms to query optimization in document retrieval, Inf. Process. Manag. 36 (5) (2000) 737–759.
[21] J. Jayanthi, K.S. Jayakumar, An integrated page ranking algorithm for personalized web search, Int. J. Comput. Appl. 12 (11) (2011) 1–5.
[22] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, Angela Y. Wu, An efficient k-means clustering algorithm: analysis and implementation, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002).
[23] S. Kim, B.T. Zhang, Web document retrieval by genetic learning of importance factors for html tags, in: International Workshop on Text and Web Mining, Melbourne, Australia, 2000, pp. 13–23.
[24] H. Kim, S. Lee, B. Lee, S. Kang, Building concept network-based user profile for personalized web search, 9th International Conference on Computer and Information Science (2010) 567–572.
[25] K. Kise, A. Sato, M. Iwata, Segmentation of page images using the area voronoi diagram, Comput. Vision Image Underst. 70 (3) (1998) 370–382.
[26] G.J. Klir, B. Yuan, Fuzzy Sets and Fuzzy Logic: Theory and Applications, Prentice Hall, 1995.
[27] K.W.-T. Leung, W. Ng, D.L. Lee, Personalized concept-based clustering of search engine queries, J. IEEE Trans. Knowl. Data Eng. 20 (11) (2008)1505–1518.
[28] F. Liu, C. Yu, W. Meng, Personalized web search for improving retrieval effectiveness, J. IEEE Trans. Knowl. Data Eng. 16 (1) (2004) 28–40.
[29] V. Loia, P. Luongo, An Evolutionary Approach to Automatic Web Page Categorization and Updating, Conference on Web Intelligence: Research and Development, Springer-Verlag, 2001, pp. 292–302.
[30] M.P. Selvan, A. Sekar Chandra, A. Dharshin Priya, Survey on web page ranking algorithms, Int. J. Comput. Appl. 41 (19) (2012) 1–7.
[31] G. Nagy, S. Seth, M. Viswanathan, A prototype document image analysis system for technical journals, Computer 7 (25) (1992) 10–22.
[32] O. Nasraoui, C. Petenes, Combining web usage mining and fuzzy inference for website personalization, International Conference on Knowledge Discovery and Data Mining (2003) 37–46.
[33] P. Navrat, M. Kovacik, A.B. Ezzeddine, V. Rozinajova, Web search engine working as a bee hive, J. Web Intell. Agent Syst. 6 (4) (2008) 441–452.

Peyman Almasinejad received his Bachelor of Software Engineering from the Islamic Azad University, Iran in 2006 and Master of Software Engineering from Payame Noor University (PNU), Tehran, Iran in 2009. Then, He worked in a software company. Currently, he is working as an instructor at Department of Computer Engineering and Information Technology, Payame Noor University (PNU), Iran. Email: p.almasinejad@pnu.ac.ir

Mohammad Javad Kargar is an Assistant Professor at the Department of Computer Engineering at University of Science and Culture, Tehran, Iran. He received his Bachelor in Software Engineering, M.Sc. in Computer Architecture from University of Sciences and Researches, and Ph.D. in Information Technology and Multimedia System from University Putra Malaysia (UPM). He has published dozen of articles in the science – research journals, and , IEEE and ACM conferences. Dr. Kargar has also been serving on the Editorial Review Board for the International Journal of Advancements in Computing Technology and International Journal of Science and Advanced Technology. He is also reviewer for a number of ISI and Scopus indexed journal. He founded International Conferences on Web Research in Iran which is unique Web event in the country and the region. His research interest is Web and data mining, distributed systems and quality assessment.