An Approach in Web Content Mining for Clustering Web Pages

R. Etemadi[1] , N. Moghaddam[2]
[1]Department of Electrical and computer engineering, Islamic Azad University branch of bonab,Tabriz, Iran ,
[2]Department of Electrical engineering and computer engineering,Tarbiat Modarres university,Tehran, Iran,
[1]Roohalahetemadi@gmail.com, [2]Charkari@modarres.ac.ir

Abstract

Nowadays, using web and Internet as a world wide information system faces us with so many data. In this direction, the necessity of accessing some tools for data processing in web level which helps the man intelligently to transform these data into useful knowledge seems so important. Clustering the web pages is one of these techniques. In this paper, a new algorithm has been represented to cluster web pages based on data content. The new algorithm has been suggested based on the expressions and key words existed in web pages, and their bit display a vector and using a new similarity criterion obtained from Cosine and Jaccard similarity criterion. To evaluate the efficacy of suggested algorithm, some pages with five subjects of software engineering, computerized networks, architecture of computer, parallel processing and operating system have been investigated and after preparing a suitable data bed the represented algorithm has been simulated separately through two similarity criteria of Cosine and that of represented in this pager and has been evaluated using Dunn index. The results obtained from simulation show high efficiency of the algorithm proposed in separating web pages and their clustering. The represented algorithm can be used in most of the problems related to clustering web pages.

Key words: *Clustering algorithm, validation of clusters, web mining, Cosine similarity criterion, jaccard coefficient.*

1. Introduction

Nowadays World Wide Web is a popular and interactive medium to distribute the information which develops increasingly. The World Wide Web includes the various amounts of dynamic data and documents, so various methods are required so that these much information and data resources existed in World Wide Web are used efficiently based on these materials and techniques. Web data have different formats; therefore about 90% of data remain without use and are not represented in user mining [1]. Confusion among these information with so many stored data as well as data manipulation for a simple search requires suitable and sufficient tools to extract the information involved. In simple language web mining is using the data mining techniques to retrieve, extract and evaluation (diagnose and analysis) of information automatically from web data, documentaries and its services [3].

Considering the studies performed in this field, web mining techniques are classified into three groups:

Web content mining: web content mining refers to description and detection of useful information from the web contents/data /documents. There are two views on web content mining: view of information retrieval and data base. The aim of web content mining according to data retrieval based on content is to help the process of data filtering or finding data for the user which is usually performed based on extraction or demand of users; while according to the view of data bases it means attempt for modeling the data on web and its combination such that most of the expert query required for searching the information can be executed on this kind of data mode [1] , [2] , [4].

Web structure mining tries to discover the model underlying the link structures of the web. This model can be used to categorize web pages and is useful to generate information such as the similarity and relationship between different web sites [2], [4].

Web usage mining: web usage mining using the data derived from using effects on the web detects the behavioral models of the users to access the web services automatically [1] , [2], [4].

Detection of models is a key component in web mining which includes algorithms and different techniques in different investigational fields such as data analysis, machine learning, statistics and modeling. [4].One of the important processes of model detection in web is clustering. Clustering indicates the process of sample grouping in which similar samples lie in one group [Gose 96], the individual groups are called cluster. Cluster analysis is a technique for grouping the users or data items (web pages) based on a similar property.

The content of web pages documentaries includes three main parts which include key words, key terms, and ordinary words and terms which are important. In this paper an algorithm to cluster web pages will be represented based on data content after studying clustering methods in second section and the related works in third section based on the variety of words events and key terms.

2. Clustering

2.1. Definition of clustering

The process of grouping a collection of physical or abstract objects into similar categories groups is called clustering. The objects in a cluster are more different and are different from the objects of other group [5], [7].

Definition: Consider the set $x=\{x_1,x_2,\ldots,x_n\}$ including n object, the aim of clustering is grouping the objects in k cluster as $c=\{c_1,c_2,\ldots\ c_k\}$ such that every cluster is as following:

$1)C_1 \cup C_2 \cup \ldots \cup C_k = X$

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

Regarding the above definition, the various states for clustering n object to k cluster equals:[25]

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

In most of methods, the rate of clusters that is K is determined by the user. From equation (1) it is understood that even if K is obvious, finding the best state of clustering is not easy. In addition , the methods of clustering n object to k cluster increases as $K^n / k!$,so finding the best state for clustering n object to k cluster is considered a NP-Complete and complicate problem and should be solved optimally by some techniques [6].

2.2. Algorithms and clustering methods

In general, clustering technique and algorithms have been represented based on different methods which are classified into five categories: Partition methods, hierarchy methods, Density methods , Grid methods and model based methods.

In partition method, k-means and k-medoids are two famous hard-type detectional algorithms. The equal of these two algorithms in fuzzy partitioning are fuzzy k-means and fuzzy k-medoids [5],[9],[10],[11].

The hierarchical clustering allows us to review the data and samples in different levels of magnification. Algorithms of single – link, complete link, Average link, Group Average link, Median distance and word are among this method. [5],[7],[8],[11],[12],[13] and [14].

Algorithms DBSCAN and OPTICS are samples of the Density based methods.

Algorithm STING and CLIQUE are examples of Grid Based methods. The main advantage of this method is its higher rate which is independent on data examples [5].

In the methods based on model for every cluster, a model is regarded and it is tried that the data are concordant with those models. The main strategies for this work are the statistical methods like COBWEB, CLASSIT and neural networks like SOFM [5].

2.3. Validation of cluster

The results obtained from exerting clustering algorithm on data considering the selection of algorithm parameters are different.

The aim of validating algorithm is to find the cluster which has the best concordance with the considered data. Two basic criteria of measurement suggested for evaluation and selection are clusters of compactness and separation.

Also of the main methods of evaluating the clustering clusters are external, internal and relative criteria. Both of the external and internal criteria are based on statistical methods having more calculational complexity. The evaluation of clusters is accomplished by external criteria using the special view of the users. Internal criteria perform the assessment of clusters using the rates of clusters and their display. The basis of relative criteria is different as compared with the rates of clustering (algorithm plus its parameters).

In this method, the basis of comparison is validity indexes. Various validation indices have been suggested the important of which are Dunn Index, Davies Bouldin Index, the mean square root of standard deviation (RMSSDT) and R root (RS), validity of SD and S-Dbw [15],[16].

3. Related Works

The concept of web mining was raised by Etzioni for the first time in 1996. Based on its definition, web mining was using the data mining techniques for extracting data from [3].

According to data mining various algorithms and techniques have been represented to data mining, the most important of which are single-link clustering technique[23], complete-link, Average – link, Group Average link[24], medium Distance , ward's, k-means (C-means or c-centered)[22], clustering algorithm of LBC which can be as basic methods for clustering web pages. Various algorithms and techniques have been represented for clustering web pages based on data content of which are Hard and fuzzy algorithm of clustering web pages document tarries based on key words inside web pages and Cosine similarity measure[17] and clustering web pages based on the behavioral models of the users [18], clustering of web pages based on link structure between them which is based on the textual information in links [19], clustering web documentaries using neural networks based on key words inside the documentaries [20].

4. Algorithm of clustering web pages based on data content

Before representing the suggested algorithm, consider the following definitions: suppose p is a set of m web page which is clustered based on data contents of internal documentaries. In this algorithm , the subject of web pages is not limited ; however to cluster, an n term or a key word is hypothesized in operational environment; These terms in content of web pages documentaries have been distributed based on their background and subject.

$P=\{x_1,x_2,\ldots,x_m\}$

Now the following definitions are represented for the suggested algorithm for clustering web pages.

A) Vector correspondent with key words and terms: suppose that the number of key terms and words in operational environment equally n, then vector correspondent with key words is defied as following:

$$Kp=\{kp_1,kp_2,\ldots kp_n\}$$

B) Vector correspondent with web pages: Vector correspondent with web page xi is defined as follows:

$$X_i=\{x_{i,1},x_{i,2},\ldots,x_{i,j},x_{i,n}\}$$

$$1\le i\le m\ ,\ 1\le j\le n$$

Where m is the web pages of clustering, n key terms in operational environment and j and xi are the events of key term of jth in web page xi.

C) Vector correspondent with cluster: vector correspondent with cluster c_i is defined based on key terms and words in web pages in the considered cluster.

$$C_i=\{c_{i,1}\ ,\ c_i,2,\ldots\ c_{i,j},\ldots c_{i.n}\}\qquad 1\le j\le n$$

If the number of web pages in c_i cluster equals L, then:

$$c_{i,j}=\sum_{k=1}^{L}x_{k,j}\qquad ,1\le j\le n\quad ,1\le k\le L\qquad (2)$$

Where n is the number of key terms and words and $x_{k,j}$ is the number of jth key word in k page.

D) Similarity between web pages and clusters: In order to obtain the similarity of a web page with the considered cluster, Cosine similarity criterion is used. The similarity of two vectors in Cosine criterion is determined based on the angle between two vectors (θ). The most similarity is obtained when two vectors are parallel [17].

Using the similarity criterion of Cosine between two vectors of web page and clusters vectors is defined as following:

$$Sim_C(X_i,C_j)=\frac{\sum_{k=1}^{n}x_{i,k}*c_{j,k}}{|X_i|*|C_j|}\qquad (3)$$

Considering this issue which has been created inside the cluster, based the web pages, the value of all key terms and words is not the same; that cluster in more pages will have more value in the cluster. Considering Criterion of relation (3) is changed as follows:

$$Sim_C(X_i,C_j)=\frac{\sum_{k=1}^{n}\frac{Nx_{i,k}}{N_j}x_{i,k}*c_{j,k}}{|X_i|*|C_j|}\qquad (4)$$

Where $Nx_{i,k}$ are the number of web pages in c_i cluster and Nj is the number of web pages in cluster c_j .

If the relation (3) is reevaluated it is determined that in similarity of one web page with one cluster, occur of one key term or word is a dominant parameter in our similarity criterion, the variety of key terms should be exerted under a separate parameter in clustering.

To calculate the effect of variety in key term under a separate parameter, a similarity criterion of Jaccard similarity criterion between two vectors of a and b with 0 or 1 entries is defined as follows:[21]

$$Sim(a,b)=\frac{\#(a_i=b_i=1)}{\#(a_i=1)+\#(b_i=1)-\#(a_i=b_i=1)}\qquad (5)$$

Now the similarity criterion between the vector of one web page and the vector of considered cluster based on the variety of key terms using similarity criterion of Jaccard key terms using similarity criterion of Jaccard is stated as following:

$$Sim_J(X_i,C_j)=$$

$$\frac{\#(x_{i,k}\ne 0\quad if\,C_{j,h}\ne 0\quad \&\quad k=h)}{\#(x_{i,k}\ne 0)+\#(C_{j,h}\ne 0)-\#(x_{i,k}\ne 0\quad if\,C_{j,h}\ne 0\,\&\quad k=h)}\qquad (6)$$

Now using the relations (4) and (6) , the similarity matrix for clustering is corrected as follows.

$$Sim_{JC}(X_i,C_j)=Sim_J(X_i,C_j)*Sim_C(X_i,C_j)\qquad (7)$$

E) Similarity between clusters: considering the similarity between web page and cluster, the similarity between two clusters is defined as follows:

$$Sim_C(C_i,C_j)=\frac{\sum_{k=1}^{n}(\frac{Nc_{i,k}}{N_i}c_{i,k}*\frac{Nc_{j,k}}{N_j}c_{j,k})}{|C_i|*|C_j|}\qquad (8)$$

Where Ni and Nj are the number of i and jth cluster members, and k, Ncj,k, Nci,k are the number of web pages in jth and ith cluster which have kth key word.

The similarity criterion of Jaccard is calculated as follows:

$$Sim_J(C_i,C_j)=\frac{\alpha}{\#(C_{i,k}\neq 0)+\#(C_{j,h}\neq 0)-\alpha} \qquad (9)$$

Where α is obtained from the following relation:

$$\alpha=\#(C_{i,k}\neq 0 \, if \, C_{j,k}\neq 0 \, \& \, \frac{N_{C_{I\ K}}}{N_i}>0\ 5 \, \& \, \frac{N_{C_{J\ K}}}{N_j}>0\ 5) \qquad (10)$$

Now considering the relations (8) and (9) the similarity between two clusters is defined as follows:

$$Sim_{JC}(C_i,C_j)=Sim_J(C_i,C_j)*Sim_C(C_i,C_j) \qquad (11)$$

F) Determining the rate of threshold for the similarity of web page with clusters:

Considering the way of calculating similarity criterion, the rate of threshold for the similarity between one page with the clusters is obtained.

Indeed, the rate of threshold is a parameter which determines if the considered page can add to a cluster based on the rate obtained for similarity matrix?

4.1. Representing suggested algorithm

To represent the suggested algorithm, it is hypothesized that the key terms in the considered operational environment have been extracted, and the key words vector have been extracted for the individual web pages considering the mentioned hypothesize, the algorithm represented to clustering web pages in this stage includes the following phases:

1. Firstly every page is considered as a separate cluster.

2. For all clusters, a cluster pair with the most similarity is found and is combined in the case of necessary conditions.

3. The second phase is repeated if the number of clusters is more than threshold rate.

Supposing I index for selected cluster and J index for found cluster with the most similarity with I cluster, the states occurred in the second phase of algorithm are as follow:

Combination: If the cluster J is not selected for combination in current phase, it will combined with I cluster (figure 1).

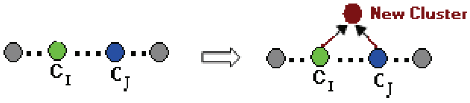

Figure 1. Combination of clusters

Breaking: In breaking operation, the combined clusters are separated. The states in which breaking occur are:

1. Cluster J has been already selected for combination but at the same time STM (CJ-m,CJ) < SIM (CI,CJ) (figure2.a)

2. Cluster I has been already selected for combination but the similarity matrix is SIM (CI-n,CI) < SIM (CI,CJ) (figure 2.b)

3. Clusters I and J have been selected for the combination previously but similarity matrix is as SIM (CI-n,CI) < SIM (CI,CJ) and SIM (CJ-m,CJ) < SIM (CI,CJ) (figure 2.c).

Non-combination:

If two clusters I and J have the following states, cluster I resides without combination in one cluster.

1. Cluster J is selected for combination, and relation SIM (CJ-m,CJ) > SIM (CI,CJ) is established for clusters similarity (figure 3.a).

2. Clusters I and J have been selected for the combination previously and their similarity is SIM (CJ-m,CJ) > SIM (CI,CJ) (figure 3.b).

3. Cluster J has been before cluster I and has been combined with different cluster (figure 3.c).

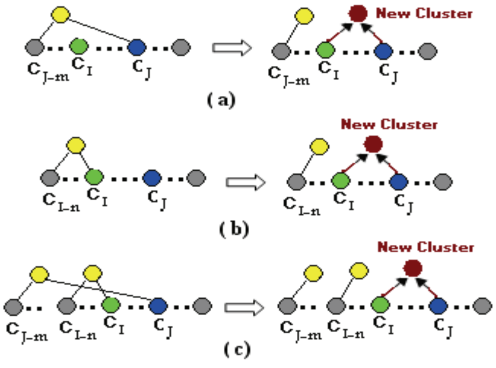

Figure 2. Breaking of Clusters.

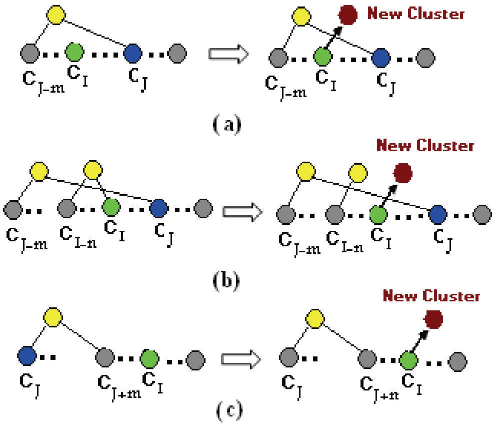

Figure 3. Non-Combination State

In figure (4) the cluster algorithm has been shown.

4.2. Main features of algorithm

The represented algorithm uses the hierarchical method. In general, the main idea of suggested algorithm has been extracted from word's algorithm, with this difference that in algorithm word's, in every stage, similar pair clusters are found and combined [11].

But in our suggested algorithm, to create the best state of clustering, only cluster pairs with the most similarity are combined; and probably in one specific stage, some clusters are not combined or the combined clusters are separated. Of the main advantages of this algorithm in proportion to the previous algorithm is using the combination of several similar criteria with less calculations to create clusters and lack of algorithm dependence on the combination of clustering data and primary selection of data.

The time complexity of suggested algorithm in best case is an O (n2) and in worse case is an O(n3).

4.3. Algorithm simulation

To simulate algorithm, at first the operational environment of some pages with five subjects of software engineering, computerized networks, computer architecture, parallel processing and operating system have been considered. For the mentioned subjects, the corresponding web pages have been studied.

Algorithm1. pseudo-code of the Proposed algorithm

```
Input: Set of web page $P = \{X_1, X_2, ..., X_n\}$
Output: Web page Clusters
Begin
  For i=1 to m do
    $X_i \rightarrow Cluster_i$;
  numcluster=m;
  while numcluster > $\tau$ do
    Begin
      For i=1 to numcluster do
        Begin
          FindTwoClusterWith max similarity();
          if Condition( ) then
            Begin
              CombineTwoCluster( );
              numcluster --;
            End
        End
    End
End
```

Figure 4. cluster algorithm of web pages

The considered web pages contain vectoral structure with 40 entries in which every entry corresponds with one key expression and has been distributed unequally between subjects and have been stored with 10000 records in 4 files. The represented algorithm has been simulated in two separated cases with similar criteria, and the results obtained for simulation have been evaluated by Dunn index which is in diagram (1). In every state of clustering, the proportion of minimum distance between clusters on maximum distance between two samples is obtained inside one cluster. The distance between clusters is calculated using relation (12) as following:

$$DISTANCE = 1 - Similarity \qquad (12)$$

The way of calculating Dunn index is observed in figure (5).As it is observed in figure (5), Dunn index is calculated as follows:

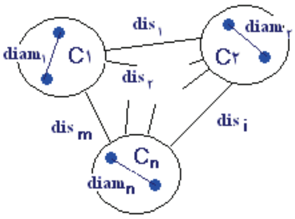

Figure 5. The way of calculating Dunn Index.

$$DUNN\ Index = \frac{\mathbf{min}\{dis_1, dis_2, .., dis_i, .., dis_m\}}{\mathbf{max}\{diam_1, diam_2, ..., diam_n\}} \qquad (13)$$

Regarding the operational environment and selecting five different subjects for web pages in simulating suggested algorithm, the clustering operation has been continued less than or equal to five. Considering the different parameters for assessing the suggested algorithm, the results of simulation have been represented in table (1),(2).

5. CONCLUSION

Considering the results obtained in table (1),(2) for clustering the pages based on similar criterion of cosine, Dunn index has been influenced on the distance between separated clusters which results from inappropriate allocation of clusters during cluster process. Now if the results represented in table (1),(2) is considered in table (2) for clustering pages with similar criterion Jaccard – cosine, it is observed that in the number of minimum distance between clusters in every state is a fixed number, and Dunn index changes through changing maximum diameter inside the clusters, which results from suitable separation of clusters and combination of clusters in every stage. Considering the results obtained in table (2), in average Dunn index 0.4019 units in size in suggested algorithm with similar criterion Jaccard- Cosine has been more than cosine criterion, and this is a reason for better clustering of the suggested algorithm with Jaccard- cosine similar criterion.

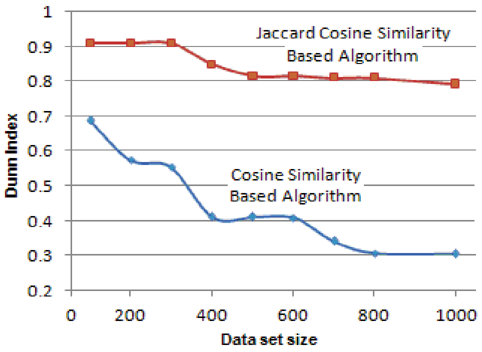

Diagram 1: the results obtained for simulation of suggest algorithm have been evaluated by Dunn index.

Table1: the results obtained for simulation of suggest algorithm.

	Jaccard-Cosine Similarity Based Algorithm(JCSBA)Cosine Similarity Based Algorithm(CSBA)Number of cluster(N.C)											
Data Size	**Data set1**			**Data set2**			**Data set3**			**Data set4**		
	DUNN Index		**N. C**	**DUNN Index**		**N. C**	**DUNN Index**		**N. C**	**DUNN Index**		**N. C**
	CSBA	**JCSBA**		**CSBA**	**JCSBA**		**CSBA**	**JCSBA**		**CSBA**	**JCSBA**	
500	0 2622	0.5340	5	0.3188	0.4951	5	0 3731	0.5498	8	0.1820	0.6135	9
1000	0 2707	0.5190	5	0.3071	0.4408	5	0 2380	0.5274	7	0.2455	0.5193	7
2000	0 2270	0.451	5	0.2508	0.4493	5	0 2076	0.4144	8	0.1552	0.4502	7
3000	0 2368	0.3705	5	0.1639	0.3129	5	0.1792	0.3849	9	0.2702	0.3460	6
4000	0 2363	0.3388	5	0.1240	0.3030	5	0.1570	0.2512	7	0.1640	0.2892	8
5000	0.1552	0.2942	7	0.1834	0.3115	5	0.1558	0.2512	7	0.1486	0.2743	7
6000	0.1552	0.1928	5	0.2151	0.3537	5	0.1454	0.2512	7	0.1793	0.2800	8
7000	0.1390	0.2814	6	0.1695	0.2583	5	0.1138	0.2188	8	0.1319	0.2800	9
8000	0.1661	0.2981	6	0.1644	0.2436	5	0.1702	0.2020	6	0.1450	0.2207	7
9000	0 2020	0.3069	4	0.1035	0.2800	5	0.1556	0.1918	7	0.0906	0.1525	7
10000	0.1274	0.2650	5	0.0812	0.2296	5	0 2036	0.2507	5	0.0956	0.2547	5

Table2 the results obtained for simulation of suggest algorithm

	Cosine Similarity Based Algorithm			**Jaccard-Cosine Similarity Based Algorithm**		
Data Size	**Min Distance**	**Max Diameter**	**DUNN Index**	**Min Distance**	**Max Diameter**	**DUNN Index**
100	0 668	0 972	0 687	0 764	0 841	0 908
200	0 565	0 984	0 574	0 764	0 841	0 908
400	0 406	0 989	0 410	0 764	0 900	0 848
600	0 406	0 993	0 408	0 764	0 938	0 814
800	0 304	0 993	0 306	0 764	0 943	0 809
1000	0 304	1	0 304	0 764	0 964	0 792

References:

[1] Dragos Arotaritei,Sushmita,Web mining: a survey in the fuzzy framework, Fuzzy Sets and Systems 148,p.5-19,2004.

[2] Raymond Kosala, Hendrik Blockeel,Web Mining Research: A Survey, ACM SIGKDD, 2000.

[3] O.Etzioni,The World Wide Web: Quagmire or gold mine, Communications of the ACM, ,p.65-66,1996

[4] WangBin,LiuZhijing,Web Mining Research, IEEE, 2003.

[5] J. Han, and M. Kamber, Data Mining: Concepts and Techniques, San Francisco: Morgan Kaufmann, 2001.

[6] E.R. Hruschka, N.F.F. Ebecken,A genetic algorithm for cluster analysis, Intelligent Data Analysis 7(1) 15–25, 2003.

[7] F. Keller, "Clustering", Computer University Saarlandes, Tutorial Slides.

[8] Maria Irene Miranda, “Clustering methods and algorithms” http://www.cse.iitb.ac.in/dbms/Data/Courses/CS632/1999/clustering/dbms html

[9] Brian T. Luke: “K-Means Clustering”, Tutorial Slides, http://fconyx ncifcrf.gov/~lukeb/kmeans.html

[10] Andrew Moore: “K-means and Hierarchical Clustering”, Tutorial Slides, http://www-2.cs.cmu.edu/~awm/tutorials /kmeans html

[11] Hee-su Kim , Sung-bae Cho,An Efficient Genetic Algorithm Whith Less Fitness Evaluation By Clustering,IEEE, p.887-889,2001.

[12] A. R. Web, Statistical Pattern Recognition, John Wiley & Sons, 2002.

[13] Q. He, A Review of Clustering Algorithms as Applied in IR , Graduate School of Library and Information Science University of Illinois at Urbana-Champaign, 1999.

[14] X. Huang, A. Acero, H. W. Hon, Spoken Language Processing, Printice Hall, 2000.

[15] F. Kov, C. Leg, A. Babos, Cluster Validity Measurement Techniques, Department of Automation and Applied Informatics, Budapest University of Technology and Economics, 2003.

[16] Frigui Hichem: Similarity Measures and Criterion Functions for clustering, http://prlab ee memphis edu /frigui/ELEC7901/UNSUP2/SimObj html

[17] Menahem Friedman, Mark Last,Yaniv Makover,Abraham Kandel, Anomaly Detection in web documents using Crisp and fuzzy-based cosine clustering methodology, Information sciences 177(2007) 467-475.

[18] Qinbao Song, Martin Shepperd, Mining Web browsing Patterns for E-commerce, Computers in Induststry 57 (2006) p.622-630.

[19] Xiaofeng He, Hongyuan Zha, Chris H.Q. Ding, Horst D. Simon, Web Document Clustering Using Hyperlink Structures, Computational Statistics & Data Analysis 41(2002) 19-45.

[20] M.Shamim Khan, Sebastian W.Khor, Web Document Clustering Using a hybrid neural network, Applied Soft Computing 4 (2004) 423-432.

[21] Xiaodi Huang, Wei Lai, Clustering Graphs for visualization via node similarities, Journal of Visual Languages and Computing 17(2006) 225-253.

[22] D.R. Cutting, D.R. Karger, J.O. Pederson, J.W. Tukey, Scatter/gather: cluster-based approach tobrowsing large document collections, SIGIR’92, ACM, New York, 1992, pp. 318–329.

[23] C.J. Van Rijsbergen, W.B. Croft, Document clustering: an evaluation of some experiments with the Cranfield 1400 collection, Inf. Process. Manage. 11 (1975) 171–182.

[24] M. Sahami, S. Yusufali, M.Q.W. Baldonado, SONIA: a service for organizing networked information autonomously, Digital Libraries, ACM, New York, 1998, pp. 200–209.

[25] G.L. Liu, Introduction to Combinatorial Mathematics, McGraw-Hill, 1968.