Structure-Based Web Pages Clustering

Leila Shahmoradi

North Tehran University

shahmoradileila@yahoo.com

Somaieh Goudarzvand

North Tehran University

S_goudarzvand@yahoo.com

Soheila Bagheri

Mumbai University

Soheilaaa.bagheri@gmail.com

Abstract

Recognizing similarities among the documents of a set is one of the objectives of retrieving information. The information related to the similarities of web pages can be used to present similar documents to users in order to retrieve considered information. In the present study, a new algorithm has been proposed to cluster web pages based on their structure. The proposed algorithm is based on hierarchical clustering designed based on SCAN algorithm. Cross-page link and in-page link structures have been tested to create a new similarity function and clustering algorithm based on volume criteria has been created for web pages in order to determine hierarchical relations. As the research findings reveal, the proposed method is effective to be used in uncovered structures of web pages in some organizations` websites.

Key words: Web pages, Clustering, Web mining, Web structure mining, Hyperlink

Introduction

Data has been turned into a highly important resource by developing information systems. Therefore, there is a need to methods and techniques of efficient access to data, information extraction from data and their application. Web is a wide and dynamic environment in which many users publish their various documents. At the present time, there are more than two million pages in web and this number is increasing with the rate of 7/3 millions in a day. Considering the volume and range of information in web, it is approximately impossible to manage it using conventional tools and there is a need of new tools and techniques to its management. Today, global web is a public medium in order to publish information which is increasing day by day. Web data are in various formats and accordingly, almost 90% of them is useless and often are not presented in users` searches (Arotaritei, 2004). Confusion among a huge volume of data stored in web and manipulating them for a simple search needs to appropriate processing tools to extract pertained information. Web mining can be simply defined as using data mining tools to retrieve, extract and evaluate information automatically from web data, documents and its services (Etzioni, 1996). Using web mining is the process of discovering interesting patterns of users` behavior. One of the applications of the discovered pattern is to cluster visited web pages by users in specific time intervals. Clustering helps users to classify pages and make their investigations easier (Sardasht et al, 2013). Determining the similarity between users` sessions is considered the most important case of this phase. There are different methods to determine the similarity in the method of displaying sessions and their computation. However, few methods consider the similarity among pages in similarity computation (Safar Khani and Mohsen Zade, 2008). Web mining techniques can be categorized into three classes of content-based, structure-based and usage-based. Content-based web mining describes and discovers useful information through contents, data and documents available in web (Arotaritei, 2004; Etzioni, 1996; Bin et al, 2003). Structure-based web mining is used to discover configuration model and structure of web links (Etzioni, 1996; Bin et al, 2003). Discovering patterns is a key component of web mining covering various algorithms and techniques in several research areas such as data mining, machine learning, statistics, and benchmarking. Clustering is a main process of pattern discovery in web mining. Therefore, after investigating clustering methods and referring to the works done in this regards, the structure-based algorithmic pages are presented to cluster web pages.

Web mining

Web mining is a branch of data mining dealing with useful knowledge extraction from wide web network (Blockeel and Kosala). Web mining and data mining are closely related. Data mining is the process of presenting searches as well as extracting patterns, useful and unknown information which are usually saved in databases. In fact, many data mining techniques can be used in web mining but the scope of web mining is wider than data mining so that these two research scopes are different in various aspects. Web mining does not mean to learn from web or applying machine learning techniques in web. There are some instances of machine learning in web which are not considered as the instances of web mining. One example is to apply machine learning techniques to find the best way of web search through Spiders.

On the other hand, other techniques are also used in addition to machine learning techniques in web mining. For example, there are specific algorithms to find Hubs and Authorities in web. In fact, machine learning technique support web mining and can be used in web mining. For instance, the studies report that using machine learning technique to classify documents can increase classification accuracy compared with using conventional information retrieving methods.

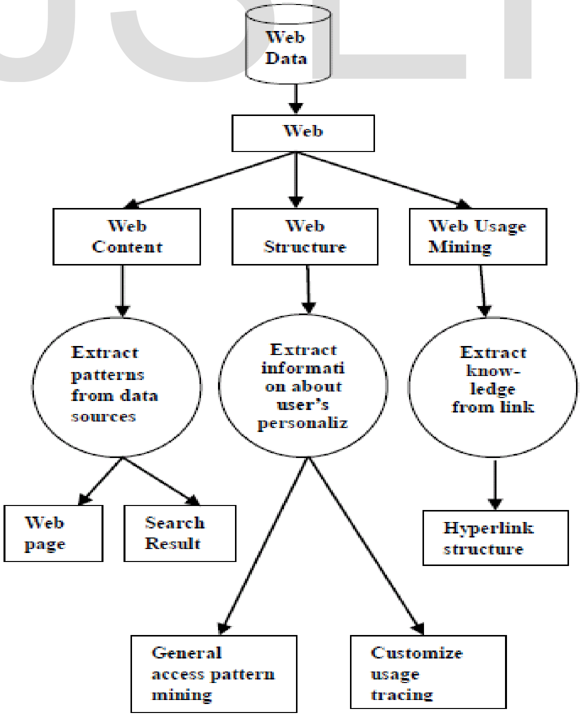
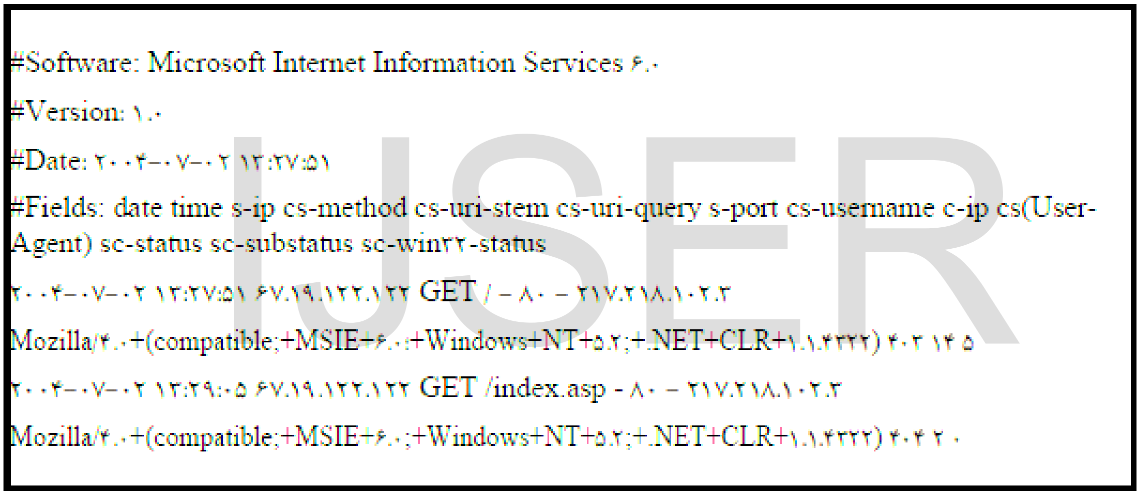

Web Data
Web
Web Content
Web Structure
Web Usage Mining
Extract patterns from data sources
Extract informati on about user's personaliz
Extract know-ledge from link
Web page
Search Result
Hyperlink structure
General access pattern mining
Customize usage tracing

Figure 1- Web mining (Sharma & Rupali Bhartiya, 2012)

Web mining can be classified into three categories (Sharma and Rupail Bhartiya, 2012). The first category is web content mining focusing on information available in web pages and most of its data are text type. Their most common application is content classification and content ranking of web pages. The second category is web structure mining focusing on the structure of web site. It mostly investigates structural information of web pages and link-based classification, a combination of content-based and structure-based ranking of web site pages and reverse engineering of web site`s models can be referred as its applications. Web can be represented both in the form of a graph in which nodes are document and files are the links among documents. The third category is web usage mining dealing with knowledge extraction from log file and its reference data includes log files stored by server in standard forms. As its applications, it can be referred to the techniques of users` web site modeling such as personalization and adaptive.

Web server is the most important and richest resource of web mining (Nnopoulos et al., 2002). Servers store a large quantity of data in log files. The log files entail databases like user name and Ip, time and date of access to web pages, list of all visited web pages, and so forth which are usually saved in standard form. The data are maintained in text files and sometimes, in databases` files. Figure 2 presents an example of log file.



```
#Software: Microsoft Internet Information Services ۶.۰
#Version: ۱.۰
#Date: ۲۰۰۴–۰۷–۰۲ ۱۳:۲۷:۵۱
#Fields: date time s-ip cs-method cs-uri-stem cs-uri-query s-port cs-username c-ip cs(User-
Agent) sc-status sc-substatus sc-win۳۲-status
۲۰۰۴–۰۷–۰۲ ۱۳:۲۷:۵۱ ۶۷.۱۹.۱۲۲.۱۲۲ GET / – ۸۰ – ۲۱۷.۲۱۸.۱۰۲.۳
Mozilla/۴.۰+(compatible;+MSIE+۶.۰;+Windows+NT+۵.۲;+.NET+CLR+۱.۱.۴۳۲۲) ۴۰۳ ۱۴ ۵
۲۰۰۴–۰۷–۰۲ ۱۳:۲۹:۰۵ ۶۷.۱۹.۱۲۲.۱۲۲ GET /index.asp - ۸۰ – ۲۱۷.۲۱۸.۱۰۲.۳
Mozilla/۴.۰+(compatible;+MSIE+۶.۰;+Windows+NT+۵.۲;+.NET+CLR+۱.۱.۴۳۲۲) ۴۰۴ ۲ ۰
```

Figure 2- An example of log file

As shown in figure 2, log file include some informational category containing a specific concept such as time and date of visit, visited page, event error and etc. the size of log files sometimes reaches to millions of lines. Therefore, extracting useful results require using specific techniques of extracting knowledge and data mining. Association rules, path analysis, sequential patterns, clustering, and classification are of the most important techniques. Among them, path analysis has been used in the present study to identify and analyze the paths and set pages visited by customers.

Clustering

Clustering data has used widely in medical engineering and industry. The importance of clustering in various sciences, type of used data, clustering speed and accuracy, time saving, and many other parameters lead to introduce various methods and algorithms of data clustering. Clustering is a supervision free technique of classification in which data (usually as multi-dimensional vectors) are divided into a specific number of clusters based on the criteria of similar or non similar. Hierarchical and separation clustering are the most commonly used techniques of data clustering. In hierarchical clustering, clustering has a structure similar to a tree in which all data belong to the first node of the tree and clustering become more accurate as proceeding in branches. In separation technique, data are divided based on cluster centers and allocated to one cluster based on the determined similarity.

The process of grouping a set of concrete or abstract objects into similar classes is called clustering and each of created group is called a cluster. The objects of a group have the most similarities to each other and the most difference with the n objects of the next groups. Suppose a set of X includes n objects. The purpose of clustering is to grouping objects in k clusters of C = $\{C_1, C_2, \dots, C_k\}$ so that each cluster has the following conditions:

1) $C_1 \cup C_2 \cup \dots \cup C_k = x$

2) $C_i \neq \emptyset$ i$= 1 - k$

3) $C_i \cap C_j = \emptyset$

With respect to the above definitions, the numbers of various modes for clustering n object into k clusters equals:

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

In most techniques, the number of clusters, i.e. k, is determined by user. Eq. (1) reveals that is it not easy to find the best mode of clustering. Also, the number of clustering ways of n objects into k clusters is approximately increased as $k^n/k!$. So, the problem of finding the best mode of clustering n objects into k clusters is considered as NP-Complete problem and should be solved optimally using techniques.

Methodology

Generally, clustering algorithms and techniques are divided into five classes including portioning, hierarchical, grid-based, model-based, and density-based.

Partitioning technique entails two known heuristic algorithm belonging to NP-hard class including k-means and k-medoids. In Fuzzy portioning, Fuzzy k-means and Fuzzy k-medoids can be mentioned as the equivalents of k-means and k-medoids (Han et al., 2001; Kim, 2001).

Hierarchical clustering allows reviewing data, templates and resources available at various zoom levels. Single-Link, Complete-Link, Average-Link, Average group-Link, Median Distance, and Word can be referred as instances of this technique (Han et al., 2001; Kim, 2001; He, 1999; Huang, 2000; Kov et al., 2003). OPTICS and DBSCAN are examples of density-based algorithm method (Han et al., 2001). CLIQUE and STING are the instances of grid-based methods; and as the main advantage, it can be referred to their high speed regardless of data templates (Han et al., 2001). In model-based clustering method, a model is considered for each cluster tending to adjust data with their models. Statistical methods such as cobweb and classit and neural networks such as SOFM are of the main strategies of model-based technique (Han et al., 2001).

Measuring validity

The purpose of measuring validity is to find the best fitted clusters of considered data. There are two basic criteria proposed to evaluate and select optimum clusters including separation and compactness. There are also three main methods of evaluating resulted clusters including internal criteria, external criteria and relative criteria. In relative criteria, the base of comparison is validity indices. There are various validity indices among which, Dum index, Davies Bouldin index, SD index, S-Dbw index, RMSSDT index, and Rs index can be mentioned (Kov, et al., 2003).

Review of the study literature

The concept of web mining was proposed by Etzioni (1996) for the first time. As Etzioni defined, web mining techniques were used to retrieve and extract information from web documents and services. Various algorithms and techniques have been presented from clustering in terms of data mining such as Single-Link (Van Rijsbergen, 1975), Complete-Link, Average-Link, Average group-Link, Median Distance, Word, and k-means (C-Centreriod or C-Means) (Cutting et al., 1992) which can be used as base methods to cluster web pages as well. Clustering we pages based on key words of web page and cosine similarity criterion (Friedman et al., 2007), clustering web pages based on users` behavior and cosin criterion (song, et al, 2006) and clustering web documents through neural networks based on key words of documents (Khan and Khor, 2004) can be also mentioned as the various techniques and algorithms of clustering.

Strehl and Ghosh (2003) proposed three consensus clustering methods including MCLA, CSPA and HGPA. Before clustering, these three algorithms change clustering aggregation into a graph. Topchy et al. (2003) formulized a target function as mutual information between final clustering and primary clustering groups. There are also other methods acting based on the corresponding similarity including FC (Gionis et al., 2005) and HAC (Fern and Bradley, 2003; Fred and Jain, 2002).

Chu (2001) analyzed the received links of 12 web sites of librarianship faculties authorized by American librarians. He examined the status of these 12 websites` links using clustering classification and multidimensional scale. One of the findings of his study revealed that putting various and wide materials in site causes to attract more visitor and links to the site. He also found that webometrics present a method of evaluation which is absent in bibliometrics. He belives that webometrics is needed to be studied carefully and accurately since both data (web-based data) and data gathering tools (web search engines) suffer from obvious defects.

The suggested method

Considering the early discussions, the present study has presented a new hierarchical web page clustering method to achieve a semantic structure regarding web pages. The main activities can be presented as follows:

- Extracting parallel links from DOM tree to create a new similar performance between pages
- Designing innovative clustering algorithm to cluster web pages linked based on compactness in order to identify hierarchical characteristics

For consistency matrix, a set of pages in an organizational web site is considered as input data and the consistency matrix between pages is presented for clustering algorithm. As mentioned, web pages contain a large amount of information about link structure helping web clusters discovery and are mostly divided into Cross-page (link graph between web pages) and In-page (organizing links within a page) web structures. If Cross-page link structure is considered as wide web structure, link structure of this page can have the first rank at low level. By combining high and low levels of web structures, webs can be clustered based on links strongly.

In the proposed design, co-citation and bibliography-coupling are two parameters and target estimation in link chart analysis. On one hand, for P_j and P_i, their co-citation C(i,j) and bibliography-coupling B(i,j) are common frequencies of in link and out link, respectively. Thus: $C_{(I,J)} = \sum_k E(i,k)E(j,k)$ and$B_{(I,J)} = \sum_k E(k,i)E(k,j)$, where$E(i,j) = 1$, if there is a higher link in the point of Pi to Pj; otherwise, $E(i+j) = 0$. The cosine function to compute the similarity Sim$_{CB}$(I,j) obtained from C(i,j) and B(i,j) for P_j and P_i is as follow:

$$(1)\quad Sim_{CB}(i,j) = \frac{C(i,j)}{\sqrt{C(i,j),C(j,i)}} + \frac{B(i,j)}{\sqrt{B(i,j),B(j,i)}}$$

On the other words, DOM (document object model) is a standard and language independent model used to present HTM or XML documents. Creating DOM trees from web pages is necessary for most of extracted algorithms. Also, the points on Dom tree have been written in short form indicating web pages` structure and organizing text contents hierarchically. Usually, DOM is used to show DOM tree extracted from resource code of a specific page (Pi) with HTML tags. For a µ point in DOM (i), sub-branches related to µ are shown with DOM (i). In this sub-branch, a parallel link is introduced as a new concept of Dom trees. Parallel links are extracted independently from each web site.

Discussion and results

The algorithm proposed in the present study is based on SCAN algorithm. SCAN is a single level network clustering algorithm. Linear time complexity can be referred as one of its advantages making it distinctive from other methods. In a SCAN test:

- Various paired-parameters are used;
- Scoring function is used to evaluate clustering various parameters;
- Pages are clustered using optimal parameters;
- The same trend is repeated for each cluster to achieve final conditions.

Since SCAN is linear for the margines number, time complexity of the proposed algorithm is also linear. The proposed algorithm was tested on real and virtual data in Microsoft visual studio (2008). To evaluate the efficiency of parallel links, figure 3 and 4 indicate the results of clustering with and without the similarities obtained from link structure of this page which are similar in high levels but in low levels of figure 4, structurally different and semantically similar pages may be combined.

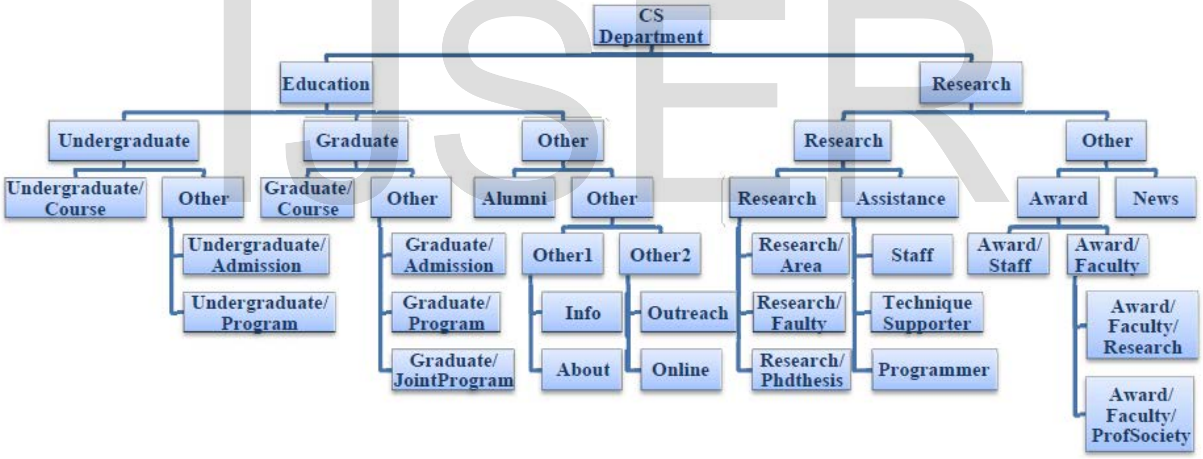


Figure 3- The results of the proposed algorithm with the similarities obtained from parallel links

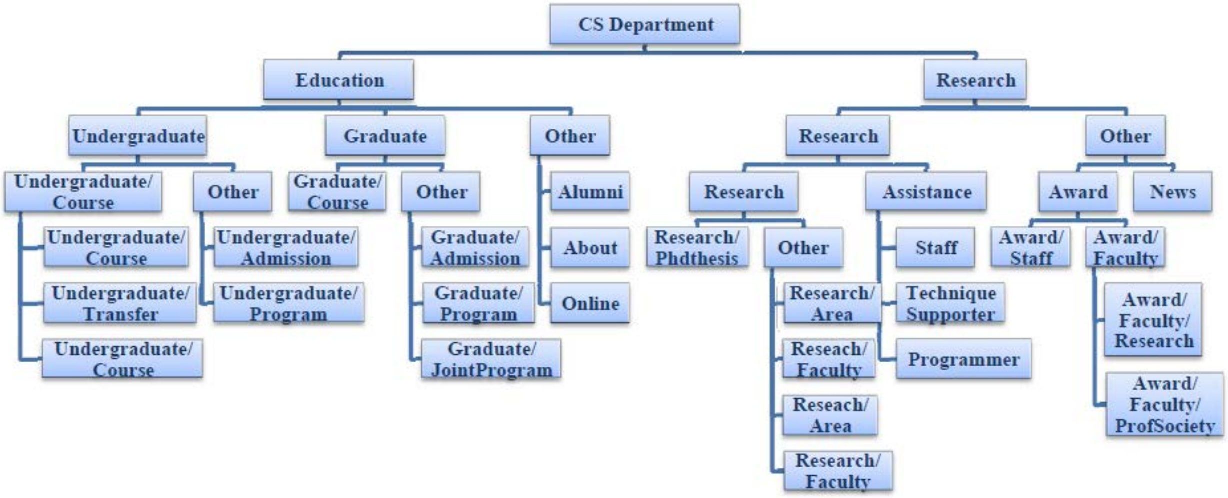


Figure 4- The results of the proposed algorithm without the similarities obtained from parallel links

To distinguish the speed of the proposed algorithm with the linear speed in comparison with networks with different sizes, 8 virtual charts are formed the number of their margins to test implementation time are from 2/414 to 61/713/102.The implementation time compared to input data size is linear. Fast modularity also is increased vey sooner than linear and k-mediods will be increased significantly.

Conclusion

In the present study, a new method of hierarchical web page clustering based on their structure was proposed. The Cross-page link and In-page link structures were tested to created a new similarity function and volume-based clustering algorithm was created for web pages group to determine hierarchical relations. The experiment results revealed that the proposed method is effective to be used in uncovered structures of web pages in some organizations` websites. This work also had a focus on investigating virtual structures of web.

Reference

1. Sardasht, P; Bali Eslami, Z; Dehghani, M. (2013). Presenting a compound method for web pages clustering through k-Means algorithm, the sixth trans-regional conference of new advancements of engineering sciences, Ayandegan high education of institution, Tonekabon
2. Arotaritei, D., & Mitra, S. (2004). Web mining: a survey in the fuzzy framework. Fuzzy Sets and Systems, 148(1), 5-19.
3. Bjorneborn, L., Ingwerson, P. (2004.)Towards a basic framework of webometrics. Journal of American Society for Information Science and Technology, 55(14), 1216-27.
4. Chu, H. (2001). A webometric analysis of ALA accredited LIS school websites. In M. Devis and C. S. Wilson (Eds ,)Proceedings of the 8th International Conference on Scientometrics & Informetrics, 20-16 July 2001. Sydney: BIRG, UNSW.
5. Cutting, D. R., Karger, D. R., Pedersen, J. O., & Tukey, J. W. (1992, June). Scatter/gather: A cluster-based approach to browsing large document collections. In Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 318-329). ACM.
6. Etzioni, O. (1996). The World-Wide Web: quagmire or gold mine? Communications of the ACM, 39(11), 65-68.
7. Fern, X. Z., & Brodley, C. E. (2003, August). Random projection for high dimensional data clustering: A cluster ensemble approach. In ICML (Vol. 3, pp. 186-193). Fred, A. L., & Jain, A. K. (2002). Data

clustering using evidence accumulation. In Pattern Recognition, 2002. Proceedings. 16th International Conference on (Vol. 4, pp. 276-280). IEEE.

8. Friedman, M., Last, M., Makover, Y., & Kandel, A. (2007). Anomaly detection in web documents using crisp and fuzzy-based cosine clustering methodology. Information sciences, 177(2), 467-475.
9. Gionis, A., Mannila, H., & Tsaparas, P. (2007). Clustering aggregation. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1), 4. Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. ACM Sigkdd Explorations Newsletter, 2(1), 1-15.
10. Huang, X., & Lai, W. (2006). Clustering graphs for visualization via node similarities. Journal of Visual Languages & Computing, 17(3), 225-253.
11. Khan, M. S., & Khor, S. W. (2004). Web document clustering using a hybrid neural network. Applied Soft Computing, 4(4), 423-432.
12. Kim, H. S., & Cho, S. B. (2001). An efficient genetic algorithm with less fitness evaluation by clustering. In Evolutionary Computation, 2001. Proceedings of the 2001 Congress on (Vol. 2, pp. 887-894). IEEE.
13. Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. ACM Sigkdd Explorations Newsletter, 2(1), 1-15.
14. Kovács, F., Legány, C., & Babos, A. (2005, November). Cluster validity measurement techniques. In 6th International Symposium of Hungarian Researchers on Computational Intelligence.
15. Nanopoulos, A., Manolopoulos, Y., Zakrzewicz, M., & Morzy, T. (2002, November). Indexing web access-logs for pattern queries. In Proceedings of the 4th international workshop on Web information and data management (pp. 63-68). ACM.
16. OD◌LIS (Online Dictionary of Library and Information Science). (2005). Retrieved December 7, 2008, from http://lu.com/odlis/.
17. Sharma, P., & Bhartiya, A. P. R. (2012). An Efficient Algorithm for Improved Web Usage Mining. International Journal of Computer Technology and Applications, 3(2).
18. Song, Q., & Shepperd, M. (2006). Mining web browsing patterns for E-commerce. Computers in Industry, 57(7), 622-630.
19. Strehl, A., & Ghosh, J. (2003). Cluster ensembles---knowledge reuse framework for combining multiple partitions. The Journal of Machine Learning Research, 3, 583-617.
20. Thelwall, M., Vaughan, L., & Björneborn, L. (2005). Webometrics. ARIST, 39(1), 81-135. Han, J., Kamber, M., & Pei, J. (2006). Data mining: concepts and techniques. Morgan kaufmann.
21. Topchy, A., Jain, A. K., & Punch, W. (2003, November). Combining multiple weak clusterings. In Data Mining, 2003. ICDM 2003. Third IEEE International Conference on (pp. 331-338). IEEE.
22. Van Rijsbergen, C. J., & Croft, W. B. (1975). Document clustering: An evaluation of some experiments with the Cranfield 1400 collection. Information Processing & Management, 11(5), 171-182.