Volume 2, Issue 12, December 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Fuzzy Equivalent Matrix for Discovering Patterns of Web Users Navigation

V.CHITRAA
Assistant Professor
CMS College of Science and Commerce
Coimbatore, Tamilnadu, India

Dr. ANTONY SELVADOSS THANAMANI
Reader in Computer Science,
Pollachi, NGM College
Coimbatore,Tamilnadu, India

Abstract --World Wide Web provides abundance of information for the Internet users and is a huge repository of web pages and links. The growth of web is tremendous as approximately one million pages are added daily. Web logs record users' accesses. Because of the tremendous usage of web , the web log files are growing at a faster rate and the size is becoming huge. Web data mining is the application of data mining techniques in web data. Web Usage Mining applies mining techniques in log data to extract the behaviour of users which is used in various applications like personalized services, adaptive web sites, customer profiling, prefetching, creating attractive web sites etc., and consists of three phases preprocessing, pattern discovery and pattern analysis. Web log data is usually noisy and ambiguous and preprocessing is an important process before mining. To characterize users access pattern, the navigation patterns identified are expected to capture the user's interests and also be used as a prediction system. The experimental results performed on real usage data from a commercial web site show a significant improvement in the pattern identification.

Keywords - Classification, Clustering, Prediction, Recommendations, Web Usage Mining

I. INTRODUCTION

The substantial increase in the number of websites presents a challenging task for webmasters to organize the contents of the websites to cater to the needs of users. Modeling and analyzing web navigation behavior is helpful in understanding what information of online users demand. The analyzed results can be seen as knowledge to be used in intelligent online applications, refining web site maps,web based personalization system and improving searching accuracy when seeking information. In spite of that an online navigation behavior grows and extracts information intelligently from it is a difficult issue.

Web Usage Mining is the process of extracting knowledge from Web user's access data by exploiting Data Mining technologies [14]. It can be used for different purposes such as personalization, system improvement and site modification and it is a process done in both online and offline. The off-line component is aimed at building the knowledge base by analyzing historical data, such as server access log files, that is then used in the online component. The main functions carried out by this component are Preprocessing, i.e. data cleaning and session identification, and Pattern Discovery. Log Data preprocessing takes the log data in servers, cleans by removing irrelevant entries like images, unsuccessful entries and so on. Users are identified and sessions are constructed for each user by using various techniques. Pattern Discovery adopts one of several techniques like statistical analysis, classification, dependency modeling and clustering. The online component is devoted to the generation of personalized content. On the basis of the knowledge extracted in the off-line component, it processes a request to the Web server by adding personalized content which can be expressed in several forms, such as links to pages, advertisements, and information relating to products or service estimated to be of interest for the current user.

The rest of the paper is organized as follows: Section 2 describes various works done in this area. Section 3 describes the steps of log data preprocessing aimed to identify user sessions, namely data cleaning and user identification. Section 4 deals with the employment of fuzzy clustering to derive user profiles by categorizing user sessions previously determined. Finally, Section 5 presents experimental results and draws final conclusions. The obtained clusters represent user profiles which can be exploited to implement different personalization functions, such as dynamic suggestion of links to Web pages retained interesting for the user.

II. RELATED WORK

The basics of evaluating web usage mining approaches and for predicting the user's next request is discussed in [7]. In [3], it is argued that web user session identification itself is a non-trivial issue and clustering techniques have been used to characterise a user session. A survey of unsupervised and semi-supervised clustering methods was presented by Grira, Crucianu and Boujemaa in [8]. Nasraoui et al [17] have used the Competitive Agglomeration algorithm for Relational Data which yielded optimal number of clusters with non-Euclidean measures Squared error algorithms rely on the possibility of representing each cluster by a prototype.An agglomerative hierarchical clustering algorithm is defined by

Mayil and Duraiswamy [13]. Banerjee and Ghosh [2] calculated the similarity by using longest common subsequence (LCS) applied the clustering algorithm to cluster the sessions. Morzy et al [16], a bottom-up approach of clustering based on Web Access Sequences is given, where frequent sequence patterns among web user sessions are identified. The users are then clustered based on their access sequence similarity. Shi [22] has used the approach of fuzzy modelling taking into account the time duration that a user spends at a URL.. A survey of classification in data mining is given in [21]. Perkowitz and Etzioni [20] discuss adaptive Web sites that learn from user access patterns.

The Page Gather algorithm uses the page co-occurrence frequencies to find clusters of related but unlinked pages. A technique for capturing common user profiles based on association-rule discovery and usage-based clustering is discussed in [15]. Users are classified using a hypergraph partitioning technique by Cooley[6]. Cooley's method is used to identify particularly interesting and similar path histories, but it cannot be used to gain an overall picture of all usage of a Web site. Unsupervised robust multiresolution clustering techniques to discover Web user groups[18].Xie and Phoha [24] use belief functions to cluster Web site users. They separate users into different groups and find a common access pattern for each group of users. Unfortunately the approach still needs to identify sessions.

Fuzzy versions of methods based on the squared error are also discussed, such as Fuzzy C-Means in many works. When compared to their 'crisp' counterparts, fuzzy methods are more successful in avoiding local minima of the cost function and can model situations where clusters actually overlap. In A sequence based clustering for web usage mining using K-means algorithm with artificial neural networks and Markov models is given in [8]. It also demonstrates how a fuzzy approach yields superior accuracy. It is concluded from the literature review that for session clustering, we so far lack of a complete preprocessing methodology. Hence our proposed methodology will not only improve the quality and efficiency of data for later steps but also to enhance the log file visibility and to structure the informationin hierarchical clustering.

III. PROPOSED METHOD

A.*Web Usage Data*

Log data is the primary source of data which is collected automatically in Web servers and it represents the accurate navigational behavior of visitors. There are different forms of log files like Apache, IIS etc., Each log entry may contain fields such as date time s-ipcs-method cs-uri-stem cs-uri-query s-port cs-username c-ipcs(User-Agent) sc-status sc-substatus sc-win32-status sc-bytes cs-bytes. A sample log is given below

```
2007-12-06 05:22:16 ::1 GET /iisstart.htm - 80 - ::1
Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+N
T+6.0;+SLCC1;+.NET+CLR+2.0.50727;+Media+
Center+PC+5.0;+InfoPath.1;+.NET+CLR+1.1.4322
```

B. *Data Cleaning*

The task of data cleaning is to remove the irrelevant and redundant log entries for the mining process. There are three kinds of irrelevant or redundant data to be removed. They are.

a. Additional Requests: A user‟s request to view a particular page often results in several log entries. Graphics and scripts are downloaded in addition to the HTML file, because of the connectionless nature of the HTTP protocol. Since the main intention of Web Usage Mining is to get a picture of the user‟s behavior, it does not make sense to include file requests that the user did not explicitly request. Suffix part of an URL is checked and eliminates suffixes like gif, jpg, GIF, JPEG, css, map etc.

b. Robots‟ requests: Web robots are software tools that scan a Web site to extract its content. Spiders automatically follow all the hyperlinks from a Web page. To remove robots‟ request, we can look for all hosts that have requested the page "robots.txt", which is checked by robot while browsing.

c. Entries with error: Status code shows the success or failure of a request. Entries with status code less then 200 and greater than 299 are failure entries which are to be removed.

Only necessary fields like date, time, IPaddress, User Agent, URL requested, URL referred, time taken are considered for further experiments to reduce the processing time So attribute subset selection is done.

C. *User Identification*

The strategy of User Identification based on the log entries without considering the topology structure of site. The description of concrete strategy algorithm is as follows.

Input: N records of web log file

Output: User sets identified

Algorithm:

Repeat steps

1: Compare ipaddress of first log entry with ipaddress of second log entry.

2: If both are same compare the user agent of both entries else assume as different users.

3:If both user agents are same identify both entries are fromsame user.

until last entry

User‟s IP addresses of two consecutive entries are compared. If the IP address is the same, user's browser and operating system is verified and if both are same, both the records are considered from the same user. These experiments prove that the algorithm significantly improves the efficiency and the accuracy of user identification without usage of site topology.

D.*Session Identification*

User‟s traversal is the input to thismethod[5]. Matrix consists of rows and columns in which columns are the web pages and rows are users and their sessions identified from the above algorithm. Browsing time for a particular page BT is determined by finding the differences between the time fields of two consecutive entries of a same user. In IIS 7.0 time-taken is another field which is the processing time of the server. So this time is also deducted from the Browsing time

Approximate Browsing time :Website Administrators fix the minimum time and maximum time for all web pages as per the contents. For example home page will take less time to browse. This arbitrary time fixed by web site designer is designated as BTmin and BTmax [11]. Compare the browsing time BT of a user with BTmin and BTmax. Weights are calculated which are numeric values to store in the matrix cells. The advantage of codification of weights is that the behavior of users such as navigation pages , interested pages, longer duration pages is also known accurately for grouping them while discovering patterns. The codes for weights are fixed from 0,1 to 9, 10,100.

If BT <BTminwt = 0
Else If BT >BTmin and BT <BTmax then wt = 1 to 9
Else if BT >BTmax then wt =10
Else if referURL == null then wt = 100.

1 to 9 is the valid Browsing time, 10 is the longer time taken by user and if referrerURL field is null a weight 100 is assumed. Valid Browsing Time is considered as a range because of the Cache memory problem. Whenever a user uses „BACK‟ key to view visited page, and if the page is not having an „expires‟ property within a short duration, a copy of webpage which is available in client side is displayed and so it‟s entry does not reach the log file. So it is expected that the same page may be traversed and an entry is stored in log file in „BACK‟. In this regard the weight is incremented in Matrix. So a range is fixed and with an assumption that not more than nine times the same page is revisited in same session.

E. *Session Construction*

Sessions are constructed as follows :

Input: User sets with N records, BTmin, BTmax, 2D matrix
Output: Constructed Sessions
Algorithm:
Repeat steps
1: Calculate the browsing time of a web page by a user by finding the difference between two consecutive entries and subtract the time taken value
2: Compare the browsing time with minimum and maximum time of each web page
3: If the browsing time is less than minimum time fix the weight as „0‟ else if it is between minimum and maximum, then weight is fixed as „1‟ , if the weight exceeds maximum fix as 10 and if referrer URL is null weight is fixed as 100.
4: If the same page is visited by the user again in user‟s set increment the corresponding entry.
5: Weights are stored in the matrix in the corresponding cells. The value aij represents a weight based on users browsing time in pagej. until last row in users set.

F.*Pattern Discovery*

Clustering is a widely used technique in data mining application for discovering patterns in underlying data clustering algorithms can be divided into two types partitioned and hierarchical. Partitioning algorithms construct a partition of a database D of n objects into a set of clusters where k is a input parameter. Hierarchical algorithms create decomposition of the database D. There are two types Agglomerative and divisive. Hierarchical clustering builds a tree of clusters, also known as a dendrogram. Every cluster node contains child cluster. An agglomerative clustering starts with one-point Clusters and recursively merges two or more most appropriate clusters. A divisive clustering starts with one cluster of all data points and recursively splits into the most appropriate clusters. The process continues until a stopping criterion is achieved. One important criteria to be considered in the choice of the clustering method is the possibility of creating overlapping clusters. This is a fundamental facet in Web personalization, where the ambiguity of the navigational data requires that a user may belong to more than one category or profile. Fuzzy clustering turns out to be a good candidate method to handle ambiguity in the data, since it enables the creation of overlapping clusters and introduces a degree of item-membership in each cluster[4].

1)Transformation: A web source matrix of the form R = (rij) n x m can be converted into a web fuzzy matrix R' = (rij) n x n where every rij∈ [0,1]. This conversion is done by applying the transformation step[23]. That is by applying the below formula on the web source matrix the web fuzzy matrix is obtained.

$$\text{rij} = 1 - c \sum_{k=1}^{m} |\text{xik} - \text{xjk}| \quad (1)$$

The c in the formula is the factor that is selected to make the rij in the range of [0,1] and it can be adjusted according to the practical situation.

2)*Clustering:* The next step is performing the clustering of the web users. The Clusteringis based on the Relative Active Degree [1] of the Web Users. For this clustering, the elements rij (i = 1,…,n; j = 1,…,m) that are different from each other in the web fuzzy matrix R. Using Euclidean distance similarity fuzzy matrix is created. Of these elements of the when threshold is 1. When evaluating Clusters, put xi and xj into one cluster if rijis between 0 and 1.we can acquire the final matrix based on fuzzy equivalent relation and the final classification under different λ –threshold. If the fuzzy relation matrix is fuzzy equivalencerelation, for any level of λ∈[0,1] , the matrix isalso equivalent relations.

IV. EXPERIMENTAL RESULTS

The web log data considered for evaluation is collected from reputed college web server during the period of May to August, 2012. Initially the log file consists of 9464 raw log entries with noisy entries like gif, jpeg etc which are not necessary for web log mining. So data cleaning is performed to remove the unnecessary log which will reduce the processing in determining the web usage pattern. This cleaning phase involves the removal of records with graphics and videos format such as gif, JPEG, etc., and records with robots traversal is also removed. The number of records resulted after cleaning phase is 1476 as shown in the figure.

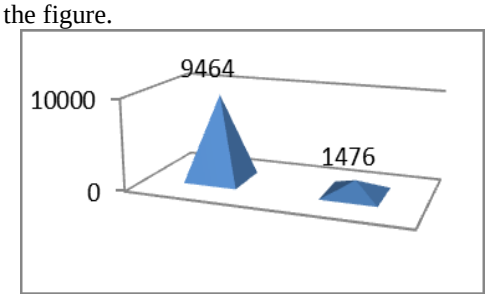

Fig 1:Comparison of initial and Cleaned log

Session identification process is carried out. A part of the result obtained by using the session identification process is represented in matrix format is presented in figure 2. The values in the matrix represent the appropriate weight which is determined by the proposed approach. These sessions will helps in determining the significant URL in the web site.

TABLE 1. SAMPLE MATRIX FOR USER-SESSION IDENTIFICATION

	A	B	C	D	E	F
U1	2	2	0	2	4	1
U2	1	2	1	2	4	2
U3	3	3	1	1	1	0
U4	2	2	1	3	3	1
U5	3	2	1	1	0	0

TABLE 2.STANDARDISED MATRIX

	A	**B**	**C**	**D**	**E**	**F**
U1	0.500	0.000	0.000	0.333	1.000	0.500
U2	0.000	0.000	0.333	0.333	1.000	1.000
U3	1.000	0.500	0.333	0.000	0.250	0.000
U4	0.500	0.000	0.333	0.667	0.750	0.500
U5	1.000	0.000	0.333	0.000	0.000	0.000

TABLE 3. TRANSFORMED FUZZY SIMILAR MATRIX

	U1	**U2**	**U3**	**U4**	**U5**	**U6**
U1	1	0.6542	0.2051	0.7639	0.1823	0.1434
U2	0.6542	1	0.0494	0.6369	0.0303	0.1823
U3	0.2051	0.0494	1	0.2163	0.7527	0.0963
U4	0.7639	0.6369	0.1823	1	0.2085	0.2472
U5	0.1823	0.0303	0.7527	0.2085	1	0
U6	0.1434	0.1823	0.7527	0.2472	0	1

When threshold is fixed as 0.6, the users can be divided into five categories: {U1, U4}, {U2}, {U3, U5}, {U6}, {U7}. Whenλ =0.6, the users can be divided into four categories :{U1, U2, U4}, {U3, U5}, {U6}, {U7} and this is an optimum solution for clustering.

TABLE 4. SIMILARITY MATRIX WITH THRESHOLD 0.6

	U1	**U2**	**U3**	**U4**	**U5**	**U6**
U1	1	1	0	1	0	0
U2	1	1	0	1	0	0
U3	0	0	1	0	1	0
U4	1	1	0	1	0	0
U5	0	0	1	0	1	0
U6	0	0	1	0	0	1

V. CONCLUSION

The traditional clustering is a hard clustering with some limitations in web log user applications. The algorithm introduced in this paper based on the matrix based fuzzy equivalent approach. This approach greatly reduced the computing space and time. The calculation results show that the algorithm is effective and it produces effective web clusters of web users. In future the matrix can be extended for page clusters

REFERENCES

1. Arotaritei and S. Mitra, "Web Mining: a survey in the fuzzy framework," in *Fuzzy Sets and System, 2000, vol. 148, pp. 5-19*
2. Banerjee, A. and J. Ghosh (2001). "Clickstream Clustering using Weighted Longest Common Subsequences", in *Proceedings of the Web Mining Workshop at the 1st SIAM Conference on Data Mining, Chicago (2001)*
3. Bianco.A, G. Mardente, M. Mellia, M. Munafò, and L. Muscariello, "Web user session characterization via clustering techniques," in *Proc. IEEE GLOBECOM 2005, St. Louis, MO, Nov. 2005, vol. 2, pp. 1102-1107.*
4. Castellano G.et al. "Mining user profiles from access data using fuzzy clustering" *Proceedings of the 6th WSEAS International Conference on Simulation, Modelling and Optimization, Lisbon, Portugal, 2006*
5. Chitraa. V, Antony SelvadossThanamani,"A Novel Technique for Sessions Identification in Web Usage Mining Processing*", IJCA(0975-8887) Volume 34-No.9, November 2011.*
6. Cooley .R, "Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data", *Ph.D. Thesis, University of Minnesota, May 2000*
7. Gery M. and H. Haddad, "Evaluation of Web Usage Mining Approaches for User's Next Request Prediction",*Proceedings of the fifth ACM international workshop on Web information and data management, 2003, pp.74–81.*
8. N. Grira, M. Crucianu, and N. Boujemaa,"Unsupervised and semi-supervised clustering: a brief survey." In A Review of Machine Learning Techniques for Processing Multimedia Content. *Report of the MUSCLE European Network of Excellence, July 2004.*
9. Houqun Yang, Jingsheng Lei, Fa Fu, "An approach of Multi-path Segmentation Clustering Based on Web Usage Mining", *Proceedings ofFourth International Conference on Fuzzy Systems and Knowledge discovery, IEEE,2007.*
10. [5].Ji He, ManLan, Chew-Lim Tan,Sam-Yuan Sung, Hwee-Boon Low, "Initialization of Cluster refinement algorithms: a review and comparative study", *Proceeding of International Joint Conference on Neural Networks, Budapest,2004*
11. KoichiroMihara, Masahiro Terabe and Kazuo Hashimoto," A Novel web usage mining method Mining and Clustering of DAG Access Patterns Considering Page Browsing Time*", Proceedings of Conf. Web Information Systems and Technologies, 2008*
12. Li Chaofeng, " Research and Development of Data Preprocessing in Web Usage Mining*", International Conference on Management Science and Engineering , 2006.*
13. Mayil, V. V. and Dr.K.Duraiswamy (2008). "Similarity Matrix Based Session Clustering by Sequence Alignment Using Dynamic Programming." *Computer and Information Science, Vol. 1, No. 3, August 2008*
14. Mobasher.B, R.Cooley, J.Srivastava,"Automaticpersonalization based on web usage mining", *Communications of the ACM*,ACM, 2000, pp. 142–151.
15. Mobasher, Cooley R., and Srivastava, J. "Creating Adaptive Web Sites Through Usage-based Clustering of URLs", *Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange, 1999.*
16. Morzy.T, M. Wojciechowski, and M. Zakrzewicz."Web users clustering." *International Symposium on Computer and Information Sciences, 2000.*
17. Nasraoui O., Frigui H., Joshi A., and Krishnapuram R., "Mining Web Access Logs Using Relational Competitive Fuzzy Clustering", *Proceedings of the Eighth International Fuzzy Systems Association Congress, Hsinchu, Taiwan, August 1999*.
18. Nasraoui .O, R. Krishnapuram, "A new evolutionary approach to Web Usage and Context Sensitive Associations Mining", *International Journal on Computational Intelligence and Applications-Special Issue on Internet Intelligent Systems, September 2002.*
19. Park, N. Suresh, Jeong, "Sequence-based clustering for Web usage mining: A new experimental framework and ANN-enhanced K-means algorithm", *Data &Knowledge Engineering 65 (2008) 512–543.*
20. Perkowitz, M., Etzioni, O. "Adaptive Web sites: automatically synthesizing Web pages", *Proceedings of Fifteenth National Conference on Artificial Intelligence, Madison, WI, 1998*
21. Phyu.T, "Survey of Classification Techniques inData Mining", *Proceedings of the International Multi Conference of Engineers and Computer Scientists 2009 Vol I, IMECS 2009, Hong Kong, March 18 - 20,2009.*
22. Shi, P., "An Efficient Approach for Clustering Web Access Patterns from Web Logs", *International Journal ofAdvanced Science and Technology*, Vol. 5, April 2009.
23. Sudhamathy .G., JothiVenkateswaran ," Matrix Based Fuzzy Clustering for Categorization of Web Users and Web Pages" *International Journal of Computer Applications (0975 – 8887) Volume 43– No.14, April 2012*
24. Xie.Y, Phoha. V, "Web user clustering from access log using belief function", *Proceedings of the ACM K-CAP'OI, First International Conference on Knowledge Capture, Victoria, British Columbia, Canada, (2001).*

AUTHORS PROFILE

Mrs.V. Chitraa is a doctoral student in ManonmaniamSundaranar University, Tirunelveli, Tamilnadu. She is working as an Assistant Professor in CMS college of Science and Commerce, Coimbatore. Her research interest lies in Database, Web Mining, Knowledge mining. She has presented many papers in conferences and published many papers in reputed international journals. She is an IEEE student member

Dr. Antony SelvadossThanamani is working as Reader in NGM college, Pollachi with a teaching experience of about 25 years. His research interests includes knowledge management, web mining, networks, mobile computing, telecommunication. He has guided 41 M.Phil scholars, attended 15 conferences, presented 35 papers, published about 8 books and many papers.

