Middle-East Journal of Scientific Research 23 (10): 2433-2439, 2015
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2015.23.10.22480

Effective Preprocessing and Knowledge Discovery in Web Usage Mining

[1]E. Manohar and [2]D. Shalini Punithavathani

[1]Anna University Chennai, India
[2]Government College of Engineering, Tirunelveli, India

Abstract: In this paper, we present a new methodology for the effective web usage mining. Here we introduce a new technique for the effective preprocessing and web user mining by associating the offline web usage mining information with the online web usage mining information. In preprocessing, an effective preprocessing technique is introduced; so that the size of the web log file is reduced to a great extent without affecting the consistency. In the offline web usage mining, we use two methods to discover the web users' interest. Here we use web log and web ranking for the effective knowledge discovery. In our methodology, we use web ranking as an additional technique for the knowledge discovery to improve the effectiveness in knowledge discovery. In online web usage mining technique, the web log is used to discover the current web user behavior. By comparing the offline web usage knowledge discovery with that of online web usage, we can improve the knowledge discovery about the web user.

Key words: Web Usage mining · Preprocessing · Data mining · Server logs · Web user · World Wide Web · Data Cleaning · Web logs

INTRODUCTION

A large number of new websites are born every day and a number of existing websites become deleted because of the difficulty in surveying numerous websites. To survive in such a technology era, the website should be easy to access. In web usage mining, preprocessing is very important for the accurate knowledge discovery. In knowledge discovery, most of the time is spent in preprocessing. The accuracy in preprocessing leads to accurate knowledge discovery. Preprocessing consists of data collection, data integration, data cleaning, data reduction, user identification and session identification. The web design is very important in the e-business [4], [8]. The website should be efficient and it should be satisfactory for the web users [2]. The evaluation method is used to evaluate the web users' issues and it ensures the web users' satisfaction in accessing the website [4]. Web log is used to evaluate the actual web user behavior; such an evaluation method is used for assuring the utility of the website [3], [6]; this evaluation method is used for the effective web design [5], [7]. The proposed methodology consists of two major methods: the offline web usage mining method and the online web usage mining method. The offline mining method consists of two techniques. In the first technique, web log is used for discovering knowledge about the web user. In the second technique, the ranking from the leading ranking site is considered along with the discovered knowledge to improve the effectiveness in knowledge discovery. The proposed method addresses various challenges.

Offline Web Usage Mining: The offline phase of web usage mining consists of preprocessing and web usage mining.

Preprocessing: Preprocessing consists of various steps. It consists of web log collection, integration, cleaning, reduction, user identification and session identification. In web log collection phase, the web log is collected from various servers. Then all the web logs are integrated. This technique is called data fusion. The next step is the data cleaning phase. In data cleaning, all redundant and missing value data are removed. In data reduction phase, the size of the data is reduced by removing unwanted attributes in the web log file. While removing the unwanted attributes, we should make sure that it should not affect the consistency. Then the next step is the user identification. It is very challenging because the web user may access the website through proxy server.

Corresponding Author: E. Manohar, Research Scholar Anna University Chennai, India.

So the navigation method is used to identify different users from the same I.P. address. In session identification, different session of the same user is identified.

Knowledge Discovery: In our proposed mining method, it consists of two phases. First by using the web log, we discover the knowledge about the web users. In the second phase, by using the ranking of the particular page, we compare the existing knowledge with the ranking to improve the efficiency of the knowledge. In this method, ranking technique is used because ranking is mostly based on the total number of web user access in the web page. So it is useful in the improvement of effective knowledge discovery.

Online Web Usage Mining: In the online web usage mining technique, the current user activity is navigated from the web log. If the online user activity is similar to the offline user activity, then we ensure that the knowledge discovery is accurate. If the online knowledge discovery differs from the offline knowledge discovery, then we have to rediscover the knowledge of the particular web user with the support of the online knowledge discovery.

Related Works: In our system, the server side log is used for the knowledge discovery about the web user. Server side log is automatically generated whenever the user browses a website. The web logs are analyzed to improve the performance of the website [1]. The web server log is used for the knowledge discovery about web user [3]. The web log is used to analyze the performance of web user for a long period of time.

Web log preprocessing is important for collecting information about the web users. Preprocessing is the first step in knowledge discovery [9]. In preprocessing, it consists of several steps where those steps consist of several challenges such as integration and cleaning [10]. Indr e Zliobaite and Bogdan Gabrys proposed a preprocessing technique which leads to an accurate knowledge discovery [11]. The information for the preprocessing is collected through many ways such as client, server & proxy server [12]. In data reduction phase, the image files and graphics files should be removed as it is not necessary for identifying web user access information [13]. In data cleaning the unwanted web log file should be removed [14]. The referrer based method is the solution for user identification issue in proxy server. Tasawar Hussain, Dr. Sohail Asghar proposed a clustering technique in web user session identification [15]. In proxy server user identification issue, if the tree is derived from the parent page, then it is considered as the new user with same IP address [16]. For the session identification, the page accessed by the user in single login is considered as single session [17]. In session identification, the time oriented session identification is the accurate method according to [18], [19], [20]. Mehdi Heydari introduced graph traversal technique for the session identification [21]. Unique user identification is achieved by Transaction [20].

Knowledge discovery of web log consists of various techniques. The various mining techniques such as association, classification, clustering and aggregation are used in knowledge discovery. The markov chain model along with association rule is used to discover the pattern of path by the web users [22]. The web usage mining based on ontology is a very effective technique to predict the web user's behavior [23]. A new clustering technique is adapted on page's path similarity for navigation pattern mining [24]. In web usage mining, web logs are used to discover the knowledge about the web users but Xiaohui Yu and Jimmy Xiangji Huang proposed a technique where web user behavior was predicted based on online reviews [25]. Maria J. Martin Bautista, Maria Amparo and Victor H. Escobar proposed fuzzy clustering algorithm for the effective web usage mining [26].

MATERIALS AND METHODS

The proposed methodology of knowledge discovery consists of three phases 1. Offline knowledge discovery phase 2. Online knowledge discovery phase and 3. Comparative analysis Phase.

Offline Knowledge Discovery Phase: In offline knowledge discovery phase, we discover the behavior of the web user by analyzing the web user behavior in the website. For the knowledge discovery, we use two types of information. First we use the web log file which gives the actual web user access information. Second, we use the ranking for the knowledge discovery. Most of the ranking is assigned according to the number of web user access to a webpage. Here we consider ranking as one of the attributes because ranking shows the overall interest of the web users at different sessions. The web log is mined to discover the knowledge and make it more accurate and more meaningful and we compare the knowledge with the ranking. If necessary, we can recalculate the knowledge discovery of the web user. In mining the web log, the first step is the preprocessing of the web log and the second step is the knowledge discovery.

Web Log Preprocessing: The web log preprocessing is very much essential in web user mining. In our proposed methodology, we carried out an effective preprocessing methodology. In web log preprocessing, we adapt some techniques to address various challenges such as web user identification in proxy server, data reduction, user identification and session identification. For the various phases in preprocessing, we adapt a new technique to improve the effectiveness in preprocessing. The web log preprocessing consists of data collection, data integration, data cleaning, data reduction, user identification and session identification. In data collection phase, the web logs from different servers are collected. The major challenge in data collection is the web log of proxy server. In some cases, client side web log is also difficult to collect and the user may deactivate the cookies. In such cases, the information of the web user is difficult to track. To address the major challenges in proxy server, we adapt an effective user identification technique. Session identification is very important for the knowledge discovery about the web user.

Data Collection: The collection of information about the web users is a challenging task in knowledge discovery. The main source of information for our knowledge discovery is the web log file. The web log file is not enough for the knowledge discovery. Some user may log through the proxy server. In such case, we have to use cookies or by using web browser agent, we may retrieve the user identity. But it is also challenging because the user may disable the cookie and the user agent. Here in our proposed system, we use the association rule in web log file to identify the unique user. In our proposed system, the data collection is not only related with web log file but also with cookies and web browser agent. Here we use ranking as another attribute to make the discovered knowledge effective. So we have to collect ranking from some leading ranking site (Example google.com). So data collection is the primary task in the knowledge discovery process.

Data Integration: The data may be integrated from many servers. The technique of integrating data from different servers is called data fusion. Data fusion is a technique of integrating the server log files; so that we can discover the knowledge about more number of web users.

Data Cleaning: In data cleaning phase, all the irrelevant, redundant and missing value data are cleaned in web logs which consist of many unwanted irrelevant data where all these data are obstacle for the knowledge discovery. Cleaning all such data increases the accuracy in knowledge discovery and the time taken for discovering the knowledge is very less compared with the knowledge discovery without data cleaning.

Data Reduction: In data reduction phase, the size of the data should be reduced. This phase should be carried out carefully. If the important data is removed, then it will affect the consistency in knowledge discovery. But removing unwanted attribute in the web log file leads to accurate knowledge discovery and the time reduction in knowledge discovery.

Web User Identification: Web user identification is a challenging task, mainly because of the proxy server access. Here we use association rule to identify the unique user and it is also based on outlier analysis. We identify unique users from the proxy server or from the same system.

Web Session Identification: In web session identification, the various sessions of the web user is identified. For knowledge discovery, session identification is very important because the web user behavior may change based on the time. The user activity in the morning session may vary from the user activity in the night. Session identification is very important in knowledge discovery. In the web session identification we introduce the tree based classification technique to identify different sessions.

Knowledge Discovery: There are many knowledge discovery techniques. After the completion of preprocessing, some mining techniques are used to mine the data from the database. The various mining techniques include various techniques in classification, association and clustering. In the proposed system, we use apriori itemset generation technique to identify the frequently accessed web pages of web users. The web ranking for the different web pages is evaluated by the same ranking of the web page in the particular session. Then by aggregating the apriori frequently access page itemset generation along with the web ranking, we discover the knowledge of the particular user.

Figure 1 shows the offline phase of knowledge discovery. In this phase, web logs are collected from different databases; then it is preprocessed. In preprocessing stage, the web logs from different servers are integrated by using the data fusion technique. Then frequently accessed pages are mined by using apriori frequent itemset accessed page mining algorithm.

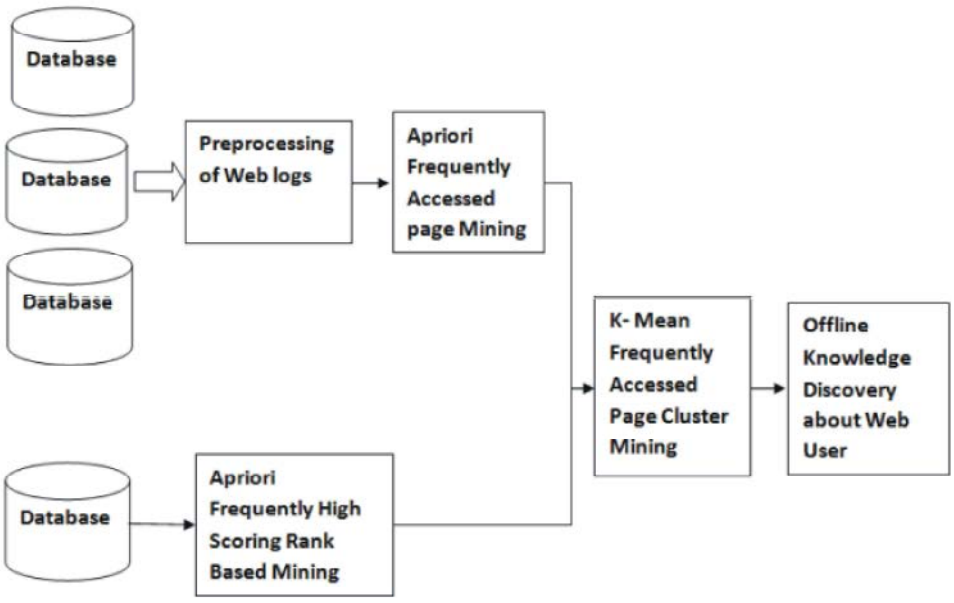

Fig. 1: Architecture of offline phase of knowledge discovery

Ranking of various web pages are collected from the ranking system and it is stored in the database. By using apriori frequently high scoring rank based mining algorithm, we discover the frequently accessed most popular web pages. Then by using k-mean frequently accessed page clustering technique, we discover the pattern of the particular web user for the particular session.

- Apriori Frequently Accessed Page Mining

Input:
Web Logs.

Output:
Frequently Accessed Pattern of Web Users.

Process:

Step 1: Web log files are classified based on different web users and their web page access session.
Step 2: Frequently accessed page of count=i to n is calculated; where initially i=1.
Step 3: While calculating frequently accessed page of count =i to n using iterative approach the web users with minsupport = 2 will be selected for the next iteration.
Step 4: Repeat the step 3 until the count=n.
Step 5: Frequently accessed page pattern is identified.

- Apriori Frequently High Score Ranking Based Mining.

Input:
Ranking of Web Pages.

Output:
Web User Access Pattern Based on Ranking.

Process:
Step 1: Web page ranking is classified based on different web pages during different sessions.
Step 2: Frequently high ranked page of count=i to n page is calculated; where initially i=1.
Step 3: While calculating frequently accessed page of count =i to n using iterative approach the web users with minsupport = 5 out of 10 point scale is considered for the next iteration.
Step 4: Repeat the step 3 until the count=n.
Step 5: Frequently accessed page pattern is identified.

- K-Mean Frequently Accessed Page Cluster Mining

Input:
Access pattern of Web Users from Web Log Based and Ranking Based.

Output:
Access Pattern of Web Users.

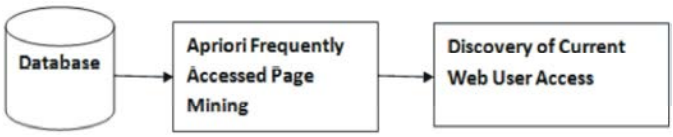

Fig. 2: Architectre of online phase of knowledge discovery

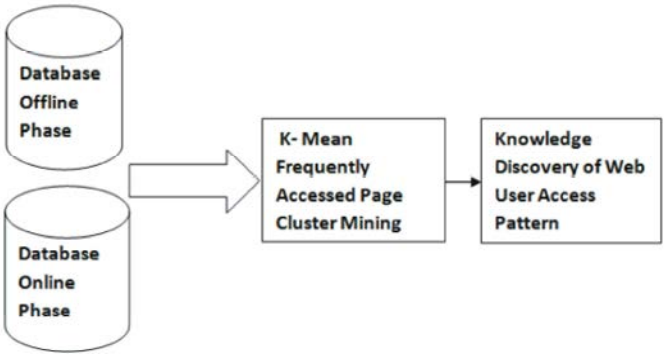

Fig. 3: Pattern Discovery

Process:

Step 1: Mean value of frequently accessed page is considered as the cluster head.

Step 2: According to the frequency there may be n number of cluster head.

Step 3: All the nearby pages to the cluster head is considered as a single cluster.

Step 4: Every cluster is considered as a discovered knowledge of the web users.

Step 5: Repeat the step 3 until the count=n.

Online Knowledge Discovery Phase: In online knowledge discovery system, the current user activity is considered for the knowledge discovery. In the online knowledge discovery also the apriori frequently itemset generation technique is used for the knowledge discovery. The online knowledge discovery phase gives the actual user activity of the particular session. So the offline knowledge discovery along with the online knowledge discovery gives the actual knowledge discovery about the web user.

In online phase, the current user activity can be navigated. The Figure 2 shows the online phase of knowledge discovery. Here the current user activities are stored in the database as web log file. By using the apriori frequently acessed page mining algorithm,we discover the frequently accessed web page pattern of the particular web user.

Comparative Anlaysis Phase: In comparative analysis phase, the result of the offline knowledge discovery phase and the online knowledge discovery phase are compared. By using the k-mean frequently accessed page cluster mining, the actual knowledge about the web user is discovered. Here the frequently accessed pages are considered as the cluster head. The nearby accessed page is considered as one particular cluster which is the pattern of the web user. Thus the pattern of access of web pages by the particular web user is identified.

The Figure 3 shows the pattern discovery of offline phase and the online phase. Here the comparative study of the offline and the online phase is carried out by k-mean frequently accessed page cluster mining. The result of the clustering technique is the discovered knowledge of the access pattern of the web user in a particular session.

RESULTS AND DISCUSSION

The effectiveness of the technique was validated with web server logs of loadvid.com website. The size of the same web logs that we take for the testing is 4.12 MB. The web logs taken for validation are 21433. For this testing, one month sample web log file is taken for the analysis. By our effective preprocessing, a higher number of unique users are identified and more number of unique sessions are also identified. The number of unique users identified is 254. The effectiveness of our methodology is verified by the number of repeated visit by the particular customer. Our experiment shows that the number of repeated visit by the customers who are provided with the information according to the knowledge discovery is higher in ratio than the web users who are not provided with the information regarding the discovered knowledge.

Table 1: Various evaluation measurements

Process	Values
Size of Web Log File	4.12 MB
Size Rduction After Preprocessing	150KB
Number of Web Log File	21433
Unique User Identified	254
Number of Web User Provided Information according to Knowledge Discovery	100
Number of Web User not provided with Information according to Knowledge Discovery	154
Number of repeated visit by the web user	120
Number of Repeated Visit by Web User who are all Provided Information according to Knowledge Discovery	86
Effectiveneness of Our Proposed System	7.1:2.2

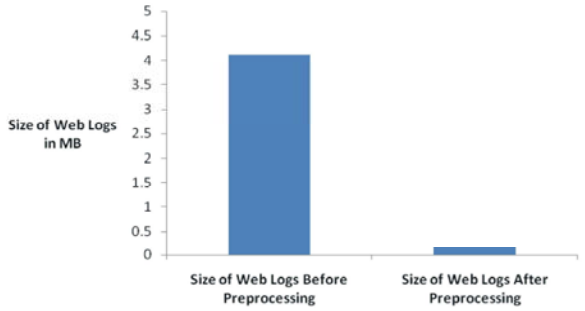

Fig. 4: Size of web log file before and after preprocessing

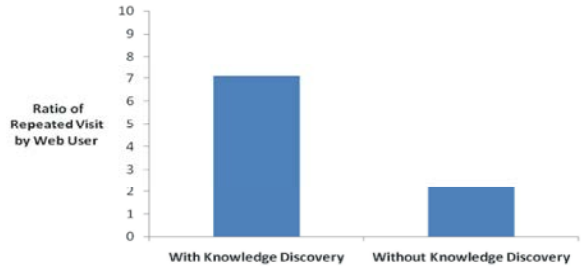

Fig. 5: Effectiveness of knowledge discovery

In Table 1 various evaluation measurements are shown with regard to preprocessing and knowledge discovery. Here the measurement value of proposed system effectiveness with respect to the ratio value is shown in the table.

The size of web log file before and after preprocessing is shown in Figure 4 where it shows that the initial data of web log file is 4.12 MB and after completing all the phases in our proposed methodology, the size is reduced to 150 KB.

All the evaluation measures show that our proposed methodology is very effective with respect to knowledge discovery. In Figure 5, the comparative study of the effectiveness of knowledge discovery is shown. Here it is shown the ratio of repeated visit by web users with knowledge discovery is higher in ratio 7.1: 2.2. Thus our knowledge discovery system is effective.

CONCLUSION

The technological development in the information technology leads to a demand for faster information access to the web users without much time on searching. Thus web usage mining is the effective technique to provide the information to the web user according to the demand of information to the web user. By effectively carrying out the web usage mining, we also curtail the dumping of information garbage to the web users. In our proposed system, we proposed an effective technique for preprocessing and knowledge discovery. In our preprocessing technique, we carried out an effective preprocessing which addresses various challenges in preprocessing. As an effective preprocessing leads to an accurate knowledge discovery about the web user, we carried out the classification technique in preprocessing which is very effective in identifying unique user and unique session of the users. The result shows the effectiveness in preprocessing by the identification of unique users and the unique sessions of the web users. In the mining phase, as we evaluate the mining with respect to past user activity along with the ranking of the particular web page and the current web user activity, it leads to an accurate knowledge about the web user access pattern. The result of the knowledge discovery shows the effectiveness of accurate web user access pattern of the web users. Thus by this effective mining technique, the information provided to the web user is faster by means of structuring the website according to the web users' need.

REFERENCES

1. Arlitt, M.F. and C.L. Williamson, 1997. Internet Web Servers: Workload Characterization and Performance Implications, IEEE/ACM Transaction on Networks, 5(5): 631-645.
2. Barnum, C.M. and S .Dragga, 2001. Usability Testing and Research, White Plains.
3. Kallepalli, C. and J. Tian, 2001. Measuring and Modeling Usage and Reliability for Statistical Web Testing, IEEE Transaction on Software Engineering, 27(11): 1023-1036.
4. Nielsen, J., 1993. Usability Engineering, Morgan Kaufmann.
5. Tauscher, L. and S. Greenberg, 1997. Revisitation Patterns in World Wide Web Navigation, ACM SIGCHI Conf., pp: 399-406.

6. Tian, J., S. Rudraraju and Z. Li, 2004. Evaluating Web Software Reliability Based on Workload and Failure Data Extracted from Server Logs, IEEE Transaction on Software Engineering, 30(11): 754-769.
7. Trinh, V.C. and A.J. Gonzalez, 2013. Discovering Contexts from Observed Human Performance, IEEE Transaction on Human Machine System, 43(4): 359-370.
8. Tullis and B. Albert, 2008. Measuring the Experience: collecting, Analyzing and Presenting Usability Metrics (Interactive Technologies), Morgan Kaufmann.
9. Dafa-Alla, Mirghani A. Eltahir and F.A. Anour, 2013. Extracting Knowledge from Web Server Logs Using Web Usage Mining, International Conference on Computing, Electrical and Electronic Engineering (ICCEEE).
10. Demin Dong, 2009. Exploration on Web Usage Mining and its Application, IEEE.
11. Indr E. Zliobaite and Bogdan Gabrys, 2014. Adaptive Preprocessing for Streaming Data, IEEE Transaction on Knowledge and Data Engineering, 26(2).
12. Jose M. Domenech and Javier Lorenzo, 2007. A Tool for Web Usage Mining, IEEE 8th International Conference on Intelligent Data Engineering and Automated Learning.
13. Doru Tanasa and Brigitte Trousse, 2004. Advanced Data Preprocessing for Intersites Web Usage Mining, Published by the IEEE Computer Society, pp: 59-65.
14. Catlegde, L. and J. Pitkow, 1995. Characterising Browsing Behaviours in the World Wide Web, Computer Networks and ISDN systems.
15. Tasawar Hussain, Sohail Asghar and Nayyer Masood, 2010. Hierarchical Sessionization at Preprocessing Level of WUM Based on Swarm Intelligence, 6th International Conference on Emerging Technologies (ICET) IEEE, pp: 21-26.
16. Sumathi, C.P., Padmaja Valli and T. Santhanam, 2011. An Overview of Preprocessing of Web Log Files for Web Usage Mining, Journal of Theoretical and Applied Information Technology, 34(2).
17. Murat Ali Bayir, Ismail Hakki Toroslu, Ahmet Cosar and Guven Fidan, 2008. Discovering More Accurate Frequent Web Usage Patterns.
18. Murat Ali Bayir, Ismail Hakki Toroslu, Ahmet Cosar and Guven Fidan, 2009. Smart Miner:A New Framework for Mining Large Scale Web Usage Data, ACM International World Wide Web Conference Committee.
19. Raju, G.T. and P. Sathyanarayana, 2008. Knowledge discovery from Web Usage Data: Complete Preprocessing Methodology, IJCSNS 2008.
20. Robert F. Dell, Pablo E. Roman and Juan D. Velasquez, 2008. Web User Session Reconstruction Using Integer Programming, IEEE/ACM International Conference on Web Intelligence and Intelligent Agent.
21. Mehdi Heydari, Raed Ali Helal and Khairil Imran Ghauth, 2009. A Graph-Based Web Usage Mining Method Considering Client Side Data, IEEE International Conference on Electrical Engineering and Informatics.
22. Mamoun A. Awad and Issa Khalil, 2012. Prediction of User's Web-Browsing Behavior: Application of Markov Model, IEEE Transaction on Systems, 42(4).
23. Mehdi Adda, Petko Valtcher, Rokia Missaoui and Chabane Djeraba, 2007. Towards Recommendation Based on Ontology –Powered Web-Usage Mining, Published by IEEE Computer Society.
24. Heidar Manosian, Amir Mosoud Rahmani and Mashalla Abbasi Dezfouli, 2010. A New Clustering Approach based on Page's Path Similarity for Navigation Patterns Mining, International Journal of Computer Science and Information Security, 7(2).
25. Xiaohui Yu, Yang Liu, Jimmy Xiangji Huang and Aijun An, 2012. Mining Online Reviews for Predicting Sales Performance: a Case Study in the Movie Domain, IEEE Transaction on Knowledge and Data Engineering, 24(4).
26. Maria J. Martin Bautista and Maria Amparo Vila, 2008. Obtaining User Profiles via Web Usage Mining, IADIA European Conference Data Mining, pp: 73-76.