International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Tracing Visitors Online Behaviors by Using VOB Algorithm for Effective Web Usage Mining

Jyoti Ashokkumar Aidale[1], Sonali Rangdale[2]

[1]Department of Information Technology, Siddhant College of Engineering, Sudumbare, SCOE Maharashtra, Pune-33, India

[2]Assistant Professor Dept. of Information Technology, Siddhant College of Engineering, Sudumbare, SCOE Maharashtra, Pune-33, India

Abstract: ***Web mining can be broadly defined as discovery and analysis of useful information from the World Wide Web. It can also be defined as automatic discovery of patterns in clickstreams and associated data collected or generated as a result of user interactions with one or more Web sites. The main important goal of Web Usage mining is to analyze the behavioral patterns and profiles of users interacting with a Web site. Web usage mining also called web log mining. Web mining is one of the techniques of Data mining. The main purpose of this paper is to identify the user behavior and pattern by using basic concept of Web mining, Web Usage mining ,general data preprocessing, what are the various existing preprocessing techniques and the proposed Visitors' Online Behavior VOB algorithm.. There are several pattern mining methods which enable user behavior identification. In existing algorithms, the preprocessing concepts are applied to calculate the unique user's count, to minimize the log file size and to identify the sessions. The newly proposed algorithm is Visitors' Online Behavior (VOB) which identifies user behavior, creates user cluster and page cluster, and tells the most popular web page and least popular web page.***

Keywords: Data Cleaning**,** Data preprocessing, Web mining, Web usage mining, Web log.

1. Introduction

Data mining is the process of finding useful information and knowledge from large databases. Web mining is the application of data mining technology, which is to extract interesting and potentially useful patterns and hidden information from web documents and web activities [2],[3]. Web Mining is broadly categorized into Web content mining (WCM), Web structure mining (WSM), and Web usage mining (WUM) [3]. The first is web content mining. The knowledge is taken from Web page contents, i.e. from the topics of different sites the useful data can be extracted. It is the automatic search and retrieval of information and resources available from millions of sites. The second sub-category is web structure mining. Here the knowledge is taken from hyperlinks and it shows how pages are connected one with another. The third sub-category is web usage mining. It helps to define the behavior of visitors and classify them into groups.

Web usage mining focuses on analyzing search logs or other activity logs to find interesting patterns. One of the main applications of web usage mining is to learn user profiles. While web structure and content mining utilize primary data on the web, web usage mining works on the secondary data such as web server access logs, proxy server logs, referrer logs, browser logs, error logs, user profiles, registration data, user sessions or transactions, cookies, user queries, and bookmark data. The Various Business Areas Where Web Mining has helped in Improving the Business Decision Making.

1.1 Web Mining Applications

Web mining can provide companies managerial insight into visitor profiles, which help top management take strategic actions accordingly. Also, the company can obtain some subjective measurements through Web Mining on the effectiveness of their marketing campaign or marketing research, which will help the business to improve and align their marketing strategies timely.

For example, the company may have a list of goals as follow:

- Increase average page views per session.
- Increase average profit per checkout.
- Decrease products returned.
- Increase number of referred customers.
- Increase brand awareness.
- Increase retention rate (such as number of visitors that have returned within 30 days).
- Reduce clicks-to-close (average page views to accomplish a purchase or obtain desired information).

The company can identify the strength and weakness of its web marketing campaign through Web Mining, and then make strategic adjustments, obtain the feedback from Web Mining again to see the improvement.

2. Web Usage Mining

Web Usage mining is one of the category of Web Mining. One of the most important tasks of Web Usage mining is capture the identity or origin of Web users along with their browsing behavior on a web site. Capturing, Modeling and analyzing of behavioral patterns of users are the goal of Web Usage Mining.

Data in Web Usage Mining as follows:

- Web server logs
- Site contents
- Data about the visitors, gathered from external channels
- Further application data

Data in Web Usage Mining collected from Web servers, proxy servers, and Web clients.

2.1 Web Log

The aim of Web Log is to create a User profile. The interaction details of users with the website are recorded automatically in web servers as the form of weblogs [6]. Several forms of logs are server access logs, server referrer logs, agent logs, client-side cookies, user profiles, search engine logs and database logs.

2.2 Why tracing visitors' on-line behaviors in web usage mining is important

It is an analysis to get knowledge about how visitors use website which could provide guidelines to web site reorganization and helps to prevent disorientation. By using Pre-fetching and Catching web pages, important information captured which helps designers to satisfy visitor's needs. Web Usage Mining provides adaptive Website or Personalization. Web personalization can be defined as any action that tailors the Web experience to a particular user, or set of users. The main advantage of using Web personalization based on Web usage mining is that it can automate the adaptation of Web-based services to their users. In recent days, the web usage mining has great potential and frequently employed for the tasks like web personalization, web pages pre-fetching and website reorganization [1].

Data sources for web usage mining are obtained in three ways [7]. In server level, the server keeps the client request details. At the client level, the client itself forwards data about user's behavior to a database.

3. Preprocessing method in web usage mining

3.1 Web Usage Data

In Web Usage mining, Data Preprocessing is an important and essential task. Data Collected from web pages, intra page structures, inter page structures and usage data are the input used in web usage mining. Other forms of web data reside as profiles, registration information and cookies. It can also be available in the form of web server logs, referral logs, registration-files and index server logs and cookies. Data preprocessing is difficult when data is unavailable, inconsistent, and noisy. The duplicate or missing data may create incorrect or even misleading statistics. The required high-level tasks are data cleaning, user identification, session identification, page view identification, and path completion as shown in fig. 1.

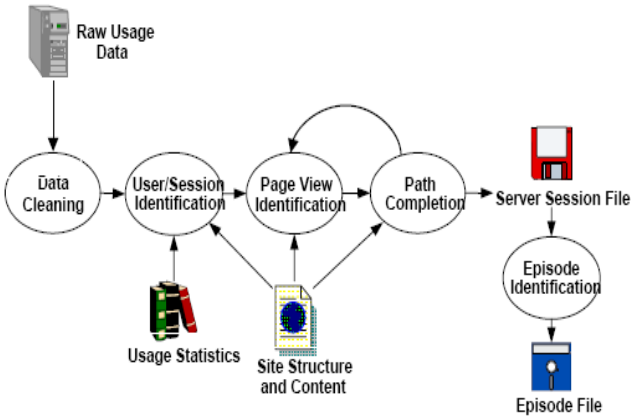

Figure 1: Preprocessing of Web Usage Mining

3.2 Data Cleaning

In Data Preprocessing First step after collecting data is data cleaning. Data cleaning [11] refers a process of eliminating the noisy and irrelevant data which are disturbing the process of mining the knowledge through weblogs.

3.3 User and Session Identification

From the web access log, different user sessions can be identified by the user as well as session identification. Session identification [9] is the process of dividing the individual user access logs into sessions.

3.3.1 Why Sessionize Require.

i. In Web usage analysis, these data are the sessions of the site visitors: the activities performed by a user from the moment she enters the site until the moment she leaves it.
ii. Difficult to obtain reliable usage data due to proxy servers and anonymizers, dynamic IP addresses, missing references due to caching, and the inability of servers to distinguish between different visits.
iii. Cookies and embedded session IDs produce the most faithful approximation of users and their visits, but are not used in every site, and not accepted by every user.
iv. Therefore, *heuristics* are needed that can sessionize the available access data.

A mechanism for User Identification is shown in below Fig.2. It is an Example: page tags (use javascript), some browser plugins.

Method	Description	Privacy Concerns	Advantages	Disadvantages
IP Address + Agent	Assume each unique IP address/Agent pair is a unique user	Low	Always available. No additional technology required.	Not guaranteed to be unique. Defeated by rotating IPs.
Embedded Session Ids	Use dynamically generated pages to associate ID with every hyperlink	Low to medium	Always available. Independent of IP addresses.	Cannot capture repeat visitors. Additional overhead for dynamic pages.
Registration	User explicitly logs in to the site.	Medium	Can track individuals not just browsers	Many users won't register. Not available before registration.
Cookie	Save ID on the client machine.	Medium to high	Can track repeat visits from same browser.	Can be turned off by users.
Software Agents	Program loaded into browser and sends back usage data.	High	Accurate usage data for a single site.	Likely to be rejected by users.

Figure 2: Page Tags and some browser plugins

3.4 Pageview identification

A Pageview is an aggregate representation of a collection of Web objects contributing to the display on a user's browser, resulting from a single user action (such as a click-through). Conceptually, each pageview can be viewed as a collection of Web objects or resources representing a specific "user event," e.g., reading an article, viewing a product page, or adding a product to the shopping cart.

3.5 Path Completion

This method is done in order to acquire the entire user access path**.** Client- or proxy-side caching can often result in missing access references to those pages or objects that have been cached. For instance, if a user returns to a page A during the same session, the second access to A will likely result in viewing the previously downloaded version of A that was cached on the client-side, and therefore no request is made to the server. This results in the second reference to A not being recorded in the server logs.

Finally, the session file can be filtered by removing very small server sessions or episodes, and low-support URI references—references to those URIs that do not appear in a sufficient number of sessions. This type of support filtering can be useful in eliminating noise from the data, such as that generated by shallow navigational patterns of "non-active" users, and URI references with minimal knowledge value for the purpose of personalization.

4. An overview of existing methods

There are several existing methods are proposed which are used to improve the performance of the data preprocessing, to identify the unique sessions and unique users. This will help to discover useful patterns and relationship from the access stream of the user. Yang Bin et al. in [13] used negative association rules in discovery of web visitor's patterns. Negative association rules have been used to solve the deficiencies in which positive rules are referred to. It is known that the data preprocessing is an essential process for effective mining process. By using a cluster with K-means algorithm analyzing customer behavior from e-commerce data [4].

A novel pre-processing technique is proposed [10] by removing local and global noise and web robots. Anonymous Microsoft web dataset and MSNBC.com anonymous web dataset are used for estimating this preprocessing technique. The log file collected from different sources undergoes different preprocessing phases to make actionable data source. It will help to automatic discovery of meaningful pattern.

The paper [5] proposed an extensive research framework which is capable of preprocessing web log data completely and efficiently. The learning algorithm of proposed research framework separates human user and search engine accesses intelligently, with less time. The framework reduces the error rate and improves significant learning performance of the algorithm. The work ensures the goodness of split by using popular measures like entropy and Gini index.

In UILP, data cleaning method is used to remove the noisy and irrelevant information from the weblog. This is one of the features in identifying the user level of interest. The second feature used is based on site topology and cookies. Frequency value, session identification, path completion is also identified using this UILP algorithm [8]. In UILP the site topology is used to identify the user and for completing the missing path. To label the session, the time duration is calculated between two nearby websites visited by the particular user.

5. Proposed Methodology

The proposed method tells user behavior and it creates user cluster and site cluster. Also, it gives the information like what sites are the most and least popular, which website is most commonly used by visitors and from what search engine are visitors coming frequently. In this method, if an IP address is unique then similar user cluster is created; If an IP address is same and user name is not unique, agent log,

operating system and browser are different then distinguish user cluster is created.

VOB algorithm

Input
Web log files

Output
User cluster, site cluster, most popular site, least popular site.

Algorithm

1) If (IP address is unique)
2) Then Create similar_user_ cluster
3) Return similar_ user_ cluster
4) If (IP address is same and user name is not unique, agent log, operating system and browsers are different)
5) then Create distinguish_ user_ cluster.
6) Return distinguish_ user _cluster.
7) For i =sitecluster_1 to sitecluster_n do
8) If (no. of. sites in current site cluster > previous site cluster)
9) Then Popular = current_ site_ cluster
10) return "most popular site"
11) else Least Popular = current_ site_ cluster
12) return "least popular site"

5.1 VOB Proposed Algorithm

This algorithm considers four entities namely IP address, user name, website name, and frequency of accessed sites. Cookies based weblogs are taken as the input which mainly classify the unique users and helps to create user clusters.

The website and webpage navigation behavior are considered as the basic source for tracing the visitors' online behavior and also to identify the interest of the user in accessing the various web sites. Based on the number of sites in the site clusters, it is concluded that it is the most popular website or the least one. Also the frequency is calculated by taking the time difference and the total number of clicks on a particular website given in a log file. Hence the VOB algorithm effectively traces the behavior of online users which supports the website usage analysis.

Clustering plays important role in VOB algorithm. Classify web visitors on the basis of user click history and similarity measure. Cookies based weblog is taken as input which classify user the unique users and helps to create user clusters. This algorithm considers four entities namely IP address, user name, website name, and frequency of accessed sites.

The website and webpage navigation behavior are considered as the basic source for tracing the visitors' online behavior and also to identify the interest of the user in accessing the various web sites. Depend on the number of sites in the site clusters, it is concluded that it is the most popular website or the least one. Also the frequency is calculated by taking the time difference and the total number of clicks on a particular website given in a log file. Hence the VOB algorithm effectively traces the behavior of online users which supports the website usage analysis.

6. Experimental Results and Performance Analysis

The weblog files are collected from college web server and browser machine for the period of 6 months from January 2013 to June 2013. For implementation, Java (jdk 1.6) is used in the system which posses Intel core i3 processor with 4GB RAM.

6.1 Performance evaluation

For performance evaluation we processed dataset for obtaining weblog. In the period of 6 months, the total no. of users is 5080. By the proposed algorithm, web visitors are classified on the basis of user click history and similarity measure. The processed dataset is given below.

Table 1: Processed Dataset

Users	No. of Users
Similar user cluster	2279
Distinguish user cluster	2801
Total No. of users	**5080**

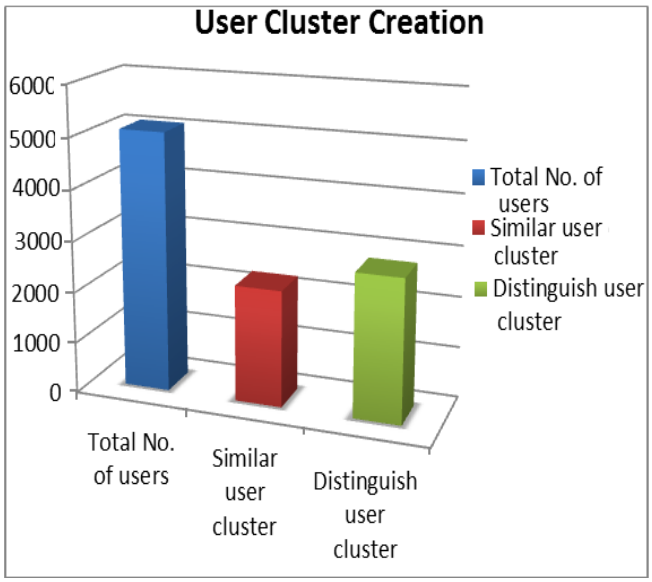

Figure 3: User cluster creation

VOB algorithm takes all the users in count and their request for processing. The proposed VOB algorithm outperforms to classify the similar user cluster and distinguish user cluster. Total no. of websites visited by users 12682. Maximum number of visits has done for the educational websites total no. of users 4700 and the users have given next preference to the social networking sites these are 3269. For Research these are 3031. Users for electronic commerce website are 1230. The number of visits for the case of entertainment is 452. Total no of visits for each site as shown in fig 4.

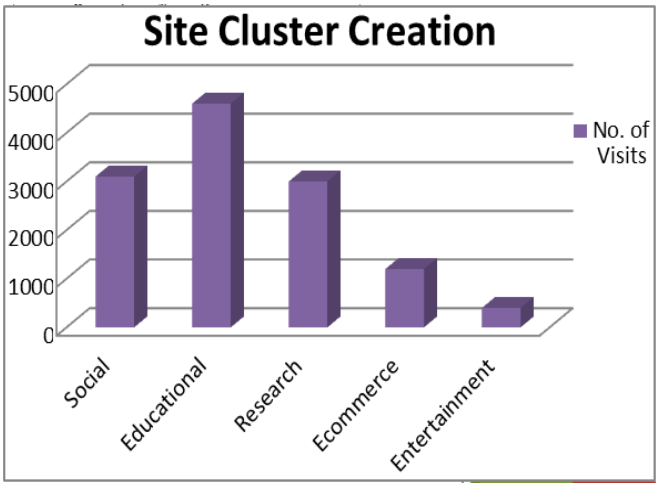

Figure 4: Site cluster creation

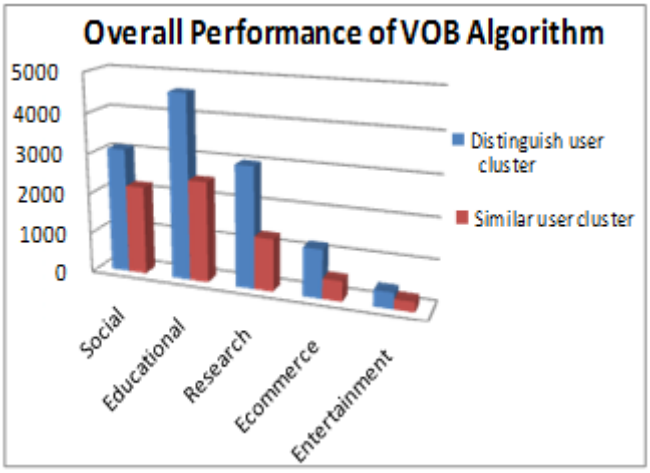

Figure 5: Overall Performance of VOB algorithm

The given fig. 4 tells that site clusters are created based on frequently accessed sites. Figure 5. Shows that the most popular website is identified based on the condition that if no. of. Sites in current site cluster are greater than previous site. Otherwise it was assumed that is the least popular site this same procedure is repeated until all user and site clusters have processed.

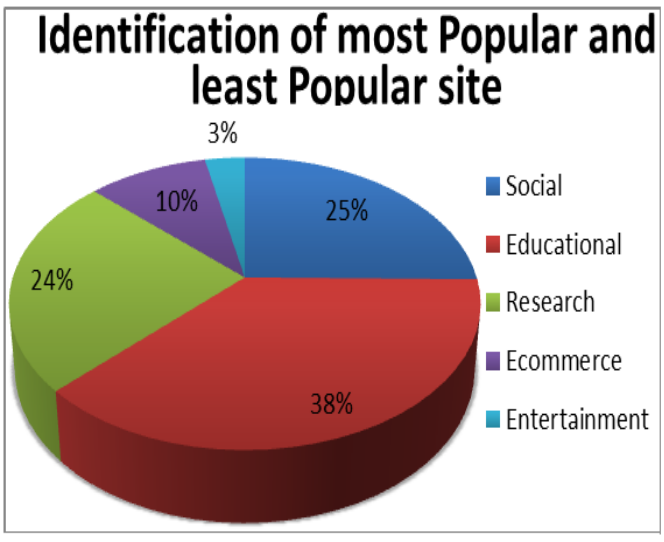

Figure 6: Identification of Most Popular and Least Popular Site

Figure 6. Shows that, the proposed algorithm proves its efficiency for classifying the preference of users to various categories of websites. In this algorithm, user cluster and site cluster creation is mainly considered as an important work and it helps to do website usage analysis based on their website surfing behavior.

7. Conclusion

Web usage mining is an essential tool for realizing more personalized user-friendly and business optimal web services. Web usage mining is used by e-commerce sites to organize their sites and to increase profits. The Proposed VOB algorithm identifies user behavior and creates user cluster, site cluster, most popular web site and the least popular web site. It traces the user behavior and gives knowledge about how visitors use website which can provide guidelines to web site reorganization and helps to prevent disorientation. Web Usage mining useful in Personalization, System Improvement, and Modification of Website and Efficient Business intelligence applications

8. Future Enhancements

Web usage mining has great potential and frequently employed for the tasks like web personalization, web pages prefetching and website reorganization, etc[1]. The intelligent system web usage preprocessor splits the human and search engine accesses before using the preprocessing techniques. This can be extended by using some other learning algorithms [12]. It can be further extended to user profiling and similar image retrieval by tracing the visitors' on line behaviors for effective web usage mining. Many different preprocessing techniques can be effectively applied in web log mining [8]. Many clustering algorithms can be applied on Weblog [IJEIT]. The preprocessing of web log data for finding frequent patterns using weighted association rule mining technique can also be extended for Discovering Pattern and User behavior [5].

References

[1] C.P. Sumathi et. al., Automatic Recommendation of Web Pages in Web Usage Mining, (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, No. 09, 3046-3052,2010.
[2] Dunham, Margaret H., Data Mining Introductory and Advanced Topics. Beijing: Tsinghua University Press, pI95-220, 2003.
[3] Han Jiawei and Kamber Micheline Data Mining Concepts and Techniques [M].Beijing: China Machine Press, p290-297, 2001.
[4] Mahendra Pratap Yadav , Mhd Feeroz , Vinod Kumar Yadav ,"Mining the customer behavior using web usage mining In e-commerce" ICCCNT'12 ,July 2012
[5] M. Malarvizhi S. A. Sahaaya Arul Mary, "Preprocessing of Educational Institution Web Log Data for Finding Frequent Patterns using Weighted Association Rule Mining Technique", European Journal of Scientific Research ISSN 1450-216X Vol.74 No.4 617-633,2012.

[6] Mr. Sanjay Bapu Thakare and Prof. Sangram. Z. Gawali, "A Effective and Complete Preprocessing for Web Usage Mining", (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 03, 848-851, 2010.
[7] Navin Kumar Tyagi1, A.K. Solanki2& Sanjay Tyagi3, An algorithmic approach to data preprocessing in web usage mining, International Journal of Information Technology and Knowledge Management, Volume 2, No. 2, pp. 279-283, 2010.
[8] R. Suguna et.al,"User interest level based preprocessing algorithms using web usage mining", International Journal on Computer Science and Engineering. Vol. 5 No. 09,Sep 2013.
[9] Sheetal A. Raiyani and, Shailendra Jain, "Efficient Preprocessing technique using Web log mining, International Journal of Advancements in Research & Technology", 1(6) ISSN 2278-7763, 2012.
[10] V.Chitraa, Dr.Antony Selvadoss Devamani, "A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing", International Journal of Computer Applications, Volume 34– No.9, 2012 .
[11] Vijayashri Losarwar and Dr. Madhuri Joshi, Data Preprocessing in Web Usage Mining, International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, Singapore, 2012.
[12] V.V.R. Maheswara Rao and Dr. V. Valli Kumari, "An Enhanced Pre-Processing Research Framework For Web Log Data Using A Learning Algorithm", Netcom 2010, Cscp 01, Pp. 01–15, 2011.
[13] Yang Bin, Dong Xianguin, Shi Fufu, "Research of Web Usage Mining based on Negative Association Rules" International Forum on Computer Science-Technology and Applications, 2009.

Author Profile

Ms. A. Jyoti completed B.E in Computer science and Engineering from T.P.C.T COE, Osmanabad in year 2009, Maharashtra, India. Later she is pursuing M.E in information Technology from Siddhant college of Engineering, Pune-33 Maharashtra, India.