Algorithm for Tracing Visitors' On-Line Behaviors for Effective Web Usage Mining

S. Umamaheswari
Research Scholar
SCSVMV University
Kanchipuram, India

S. K. Srivatsa
Senior Professor
St.Joseph College of Engineering,
Chennai, India

ABSTRACT

User behavior identification is an important task in web usage mining. Web usage mining is also called as web log mining. The web logs are mainly used to identify the user behavior. There are so many pattern mining methods which enable this user behavior identification. The preprocessing techniques will maximize the accurate and quality of pattern mining methodologies. 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. This paper brings into discussion about the basic concepts of web mining, web usage mining, general data preprocessing, how to preprocess the web data, what are the various existing preprocessing techniques and the proposed VOB algorithm.

Keywords

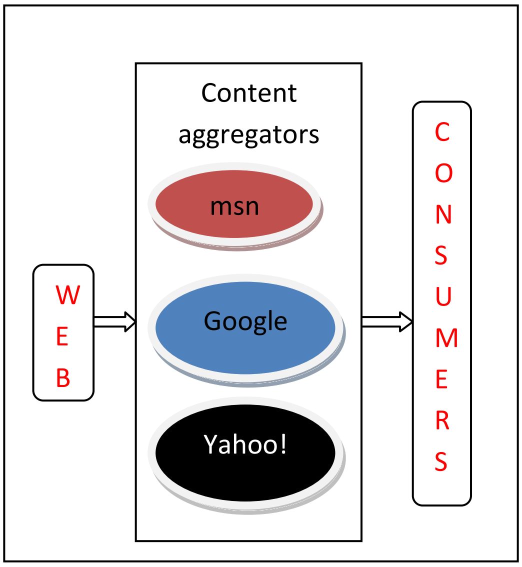
Data preprocessing methods, Web mining, Web usage mining, Web usage data, Web log.

1. INTRODUCTION

World wide web is a very large, widely distributed, global information service centre to facilitate the services such as news, advertisements, consumer information, financial management, education, government, e-commerce, etc. It consists of hyper-link information, access and usage information. World wide web gives enough number of rich sources of data for data mining .Web Mining is one of the data mining technique used to automatically discover and extract information from web documents or services. This refers a process by which we can discover useful information from the world wide web and it's usage patterns. Here the data objects are linked together for interactive access. The subtasks of web mining are resource finding, information selection and preprocessing, generalization and analysis. Resource finding refers the task of retrieving intended web documents. Information selection and preprocessing is an automatic selection and preprocessing of particular information from retrieved web resources. Also the generalization represents an automatic discovery of patterns in web sites and analysis is the validation and interpretation of mined patterns. Generally data mining techniques are used to make the web more useful and more profitable for some and to increase the efficiency of our interaction with the web.Some of the data mining techniques are association rules, sequential patterns, classification, clustering and outlier discovery. Nowadays these techniques and concepts have employed in many applications to the web like e-commerce, information retrieval and network management.

1.1 Why Mine the Web?

There are enormous wealth of information on web such as financial information like stock quotes, book/CD/video stores, restaurant information and car prices. Even though it has many sort of information, the web poses great challenges for effective resources and knowledge discovery. The web seems to be too huge for effective data warehousing and mining. Also the complexity of web pages is far greater than that of any old text documents. Only a small portion of the information on the web is truly relevant [4].It is possible to get lots of data on user access patterns and also possible to mine interesting nuggets of information. The process of searching the web is illustrated in the following "Fig. 1".



Web has recently become a powerful platform for retrieving information and discovering knowledge from web data. The idea of discovering useful patterns in data may have many names such as data mining, knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing[12].

1.2 Web Mining Applications

Web mining applications are listed such as to target potential customers for electronic commerce, to enhance the quality and delivery of internet information services to the end user, to improve the web server program's performance, to identify the potential prime advertisement locations, to facilitate adaptive sites, to improve site design, to do fraud detection and to predict the user's actions.

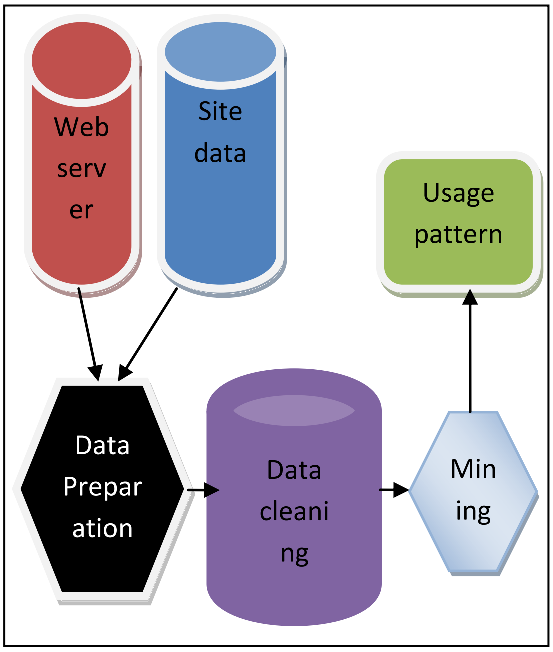
1.3 Web Mining Issues

Nowadays the web became so popular and used by many categories of people which includes school students and the business men also. The number of users who are employing the web is increasing at exponential speed [3].On the web, many different types of data such as images, text, audio/video, XML and HTML are used. Web datasets can be very large. It is in the range of tens to hundreds of tera bytes. So it cannot mine on a single server. There is a need of large forms of servers.

2. WEB USAGE MINING

Web usage mining is a category of web mining technique to discover interesting usage patterns from the secondary data derived from the interactions of the users while surfing the web.

The web pages contain information. Here actually the links are 'roads'. It tells how the people navigate the internet. The information on navigation paths is available in log files. Logs can be mined from a client or a server perspective .It is aimed to discover user 'navigation patterns' from web data and to predict the user's behavior while the user interacts with the web. Also it helps to improve large collection of resources [5].The web usage mining techniques are to construct multidimensional view on the weblog database, to perform data mining on web log records and also to conduct studies for analyzing the system performance, etc. Some of the frequently used techniques are such as data collection, data preparation and data cleaning. The web usage mining process is given in the following "Fig. 2".



2.1 What is the need for tracing visitors' on-line behaviors in web usage mining

It must to trace the visitors' on-line behaviors for website usage analysis. Actually 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. It also helps to the designers in placing the important information where the visitors look for it. It has to be done for pre-fetching and caching web pages. Also it provides adaptive website (personalization). This is represented in the figure "Fig 3." given below.

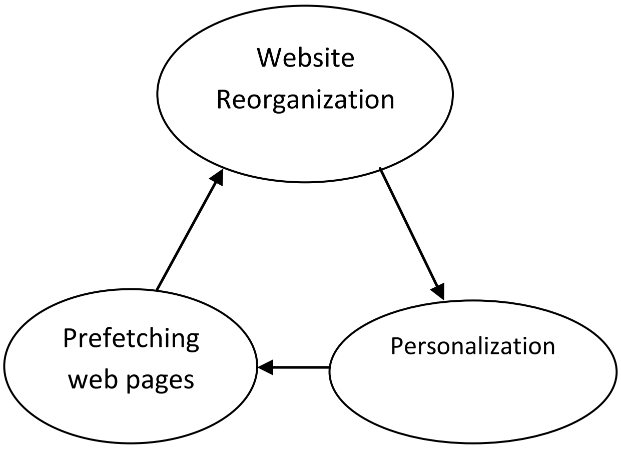


Figure 3. Website usage analysis

Many organizations have been supported by the analysis of user's browsing patterns for the purpose of giving personalized recommendations of web pages. Generally the usage-based personalized recommendation gives solution to many of the problems occurred in the web [13, 14, 15].It has created an interest between the researchers to do research. The recommendation systems listen the information overload by suggesting pages that fullfills the user's requirement. 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, etc [16]. Data sources for web usage mining are obtained in three ways [12]. 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 .It can be accomplished by using either an ad-hoc browsing application or through client side application which runs on the standard browsers. In the proxy level, the proxy side maintains user behavior information. Even though the web data is taken from many users on various web sites, only the users whose web clients pass through the proxy.

2.2 Web usage data

Generally the web pages, intra page structures, inter page structures and usage data are the input used in web usage mining. Other forms of web data resides as profiles, registration information and cookies. Web usage data is referred as the collective data about how a user utilizes a web site through his mouse and keyboard. This data can also be available in form of web server logs, referral logs, registration-files and index server logs and cookies.

2.3 Web log

The aim of web log file is to create user profile by allowing their browsing similarities with previous users. Before the data mining process, it is required to clean, condense and transform the raw data of weblog before performing data mining. Weblog information can be integrated with web content and web structure mining to help webpage ranking and web document classification. The interaction details of users with website are recorded automatically in web servers as the form of weblogs [2]. Weblogs are kept as in form of line of text in web server, proxy server and browser [8].Various forms of logs are server access logs, server referrer logs, agent logs, client-side cookies, user profiles, search engine logs and database logs. These are considered as input for knowing the end user behavior in web usage mining. Log files are those files that list the actions that have been occurred [18].Log files hold many parameters which have employed in recognizing user browsing patterns. Some of the parameters are user name, visiting path traversed, timestamp, page last visited, success rate, user agent, URL and request type [17].

2.4 Transfer / Access Log

The information on user's request from their web browsers is stored in transfer/access log.

Table 1. Transfer/Access Log

Time	Date	Host name	File requested	Amount of data Transferred	Status of report

2.5 Referrer Log

The recorded two fields of referrer log are URL and referrer URL.

Table 2. Referrer Log

URL	Referrer URL

2.6 Error Log

The list of errors and requests which have failed are collected in error log. Not only for the page which holds links to a file that does not exist, but also for the user who is not permitted to access a particular page, the user request may fail. It is depicted in the following "Fig 4."

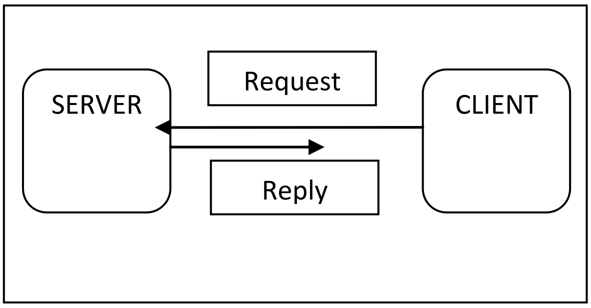


Figure 4. Error Log

When cookies are used by the websites, the information will be in the cookies field of log file. Web traffic analysis software employs the cookies to track the repeat visitors.

3.DATA PREPROCESSING METHODS

The raw data may include noise, missing values, and inconsistency mostly. The data mining results have been affected by the data quality. So it must to preprocess the data for the purpose of increasing the quality and efficiency. The process of preprocessing contains data preparation and transformation of the initial dataset. The preprocessing methods are categorized such as data cleaning, data integration, data transformation, data reduction and data discretization[4].

3.1Data cleaning

Data cleaning is an essential requirement of preprocessing methodologies. It is done for duplicate tuples. It will remove the unwanted data and shapes the required data by filling in missing values, smoothing noisy data, identifying or removing outliers and resolving inconsistencies. Always the dirty data can make confusion while processing it in the mining process.

3.2 Data integration

Many categories of databases, data cubes or files have been collected and integrated together in this step.

3.3 Data transformation

It actually pointed by the process of normalization and aggregation.

3.4. Data reduction

It leads to the reduced representation based on the volume of data collected and processed. But it gives the same or similar analytical results.

3.5. Data discretization

It is a section in the data reduction step. This is done only for the case of numerical data not for all types of data.

4. PREPROCESSING OF WEB USAGE DATA

Generally in the web usage mining, the preprocessing [9] is considered as an essential task and treated as an idea to reach the goal. As it was suggested in referred paper [1], the intelligent system web usage preprocessor splits the human and search engine accesses before using the preprocessing techniques. In the recent days, it is not possible to get good quality data. Also there is no better result for mining the quality data. But the quality decisions have been taken depending upon the quality of data. The duplicate or missing data may create incorrect or even misleading statistics. Also the data warehouse requires consistent integration of quality data. Moreover the data extraction, cleaning, and transformation take the maximum of the work in building a data warehouse. It is depicted in the following "Fig 5."

4.1 Data cleaning

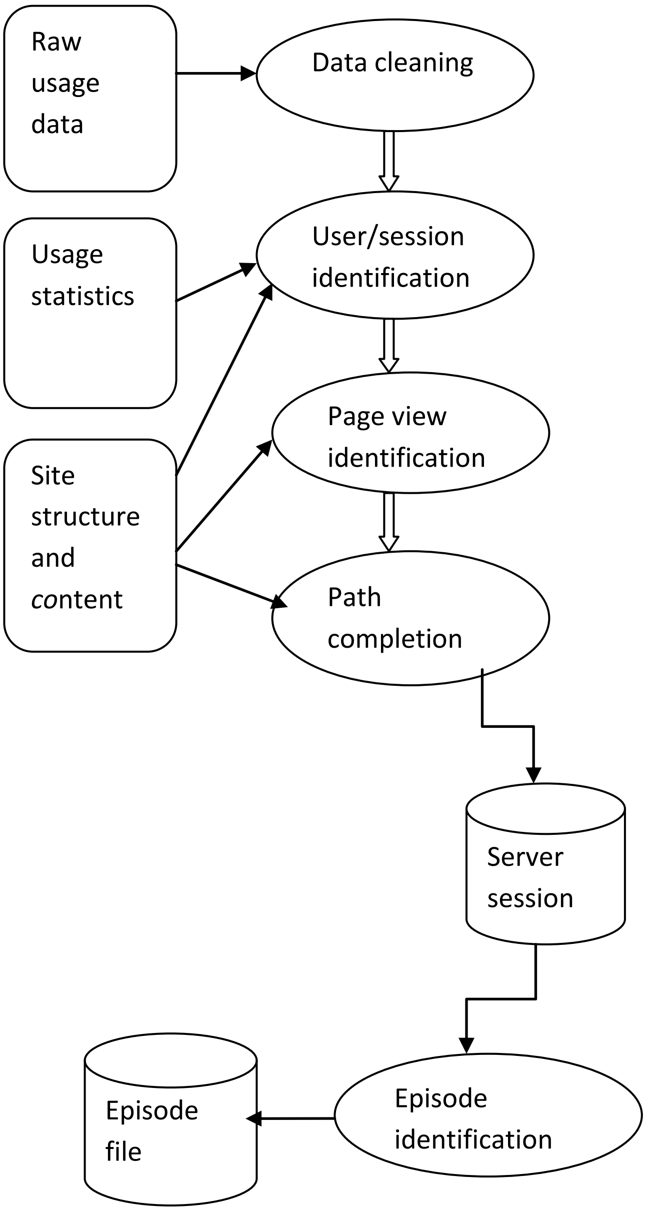
Data collection [6] is the initial step in weblog preprocessing. After collecting the data, irrelevant records are removed in the data cleaning process. Data cleaning [10] refers a process of eliminating the noisy and irrelevant data which are disturbing the process of mining the knowledge through weblogs.

4.2 User and Session Identification

From the web access log, different user sessions can be identified by user as well as session identification. Session identification [7] is the process of dividing the individual user access logs into sessions. To identify the various sessions, a referrer based method is used.

4.3 Path Completion

This is done in order to acquire the entire user access path. The incomplete access path of every user session is recognized based on user session identification. In the start of user session, the referrer has data values and delete this value of referrer by adding '_'.Also the web log preprocessing supports the unwanted click stream removal from log file and to minimize by the original file size 40-50%.



5. AN OVERVIEW OF EXISTING METHODOLOGIES

This research paper [2] studies and presents several data preparation techniques of access stream even before the mining process can be started. These are used to improve the performance of the data preprocessing, to identify the unique sessions and unique users. The methods proposed will help to discover meaningful pattern and relationships from the access stream of the user. These are proved to be valid and useful by various research tests. Yang Bin et al. in [19] used negative association rules in discovery of web visitor's patterns. Negative association rules have been deployed 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. In paper [9], a novel pre-processing technique is proposed 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 paper [1] describes the effective and complete preprocessing of access stream before actual mining process can be performed. 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 and relationships from access stream of user. Swarm based web session clustering helps in many ways to manage the web resources effectively such as web personalization, schema modification, website modification and web server performance. In this paper [2], they proposed a framework for web session clustering at preprocessing level of web usage mining. The framework will cover the data preprocessing steps to prepare the weblog data and convert the categorical weblog data into numerical data. A session vector is obtained, so that appropriate similarity and swarm optimization could be applied to cluster the weblog data. The hierarchical cluster based approach will enhance the existing web session techniques for more structured information about the user sessions. The paper [6] introduces 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. In order to create suitable target data, the further essential tasks of pre-processing like data cleaning, user identification, session identification and path completion are designed collectively. 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 are also identified using this UILP algorithm [11]. In UILP

(i) During data cleaning process, explicit image and multimedia requests from users are considered; those requests are not removed from weblogs.
(ii) Users are identified based on site topology and cookies.
(iii) Session time is calculated based on the time spent on each website by a particular user.
(iv) Frequency value is calculated based on the number of web pages visited by the user on particular website.

Here 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 website visited by the particular user. It is calculated each and every time when a user switches from one website to another and the amount of time spent in each website.

6. 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 IP address is unique then similar user cluster is created; If IP address is same and user name is not unique, agent log, operating system and browser are different then distinguish user cluster is created.

A. Steps followed

1. Create similar user cluster and distinguish user cluster based on IP address.
2. Create site clusters based on frequently accessed sites.
3. If number of sites in current site cluster is greater than previous site cluster then assign that is the most popular site.
4. Return the most popular site
5. Otherwise assume that is the least popular site
6. Return the least popular site and repeat until all user & site clusters are processed.

VOB algorithm

Input

Web log files

Output

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

Algorithm

```
If (IP address is unique) then
            Create similar_user_ cluster;
 Return similar_ user_ cluster;
If (IP address is same and user name is not unique, agent log,
operating system and browsers are different ) then
            Create distinguish_ user_ cluster.
Return distinguish_ user _cluster.
For i =sitecluster_1 to sitecluster_n do
      If (no. of. sites in current site cluster > previous site
cluster) then
          Most Popular = current_ site_ cluster
return "most popular site"
       else
Least Popular = current_ site_ cluster
         return "least popular site"
repeat until all user & site clusters are processed.
```

In the proposed method VOB, clustering plays a key role to classify web visitors on the basis of user click history and similarity measure. 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.

Here, 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.

7. EXPERIMENTAL SETUP 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.

7.1 Performance evaluation

The performance evaluation is done by analyzing the dataset taken. 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 3. Processed Dataset

Label	Processed Value
Total No. of users	5080
Similar user cluster	2279
Distinguish user cluster	2801

The following "Fig 6." shows the creation of user cluster.

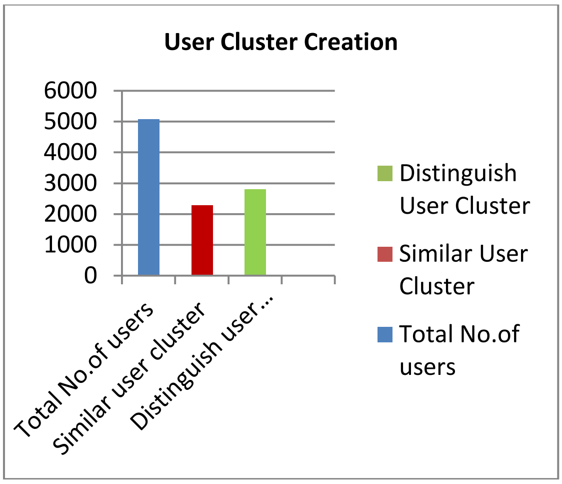


Figure 6. User cluster creation

The VOB algorithm identifies the users based on the data collected from cookies. This algorithm takes all the users in count and their request for processing. By the result, the proposed VOB algorithm outperforms to classify the similar user cluster and distinguish user cluster.

The total number of sites visited by the user is calculated as 12682. Among these sites, maximum number of visits has done for the educational websites. Totally it is of count 4700. And the users have given next preference to the social networking sites.

The number of visits made to social networking sites is 3269. Also 3031 users have referred the research sites. Only from the month of APRIL and MAY, the 1230 users have used the electronic commerce websites.

The number of visits for the case of entertainment is 452 which explicitly shows the minimum desirability of that kind of sites. The given "Fig 7." tells that site clusters are created based on frequently accessed sites.

From the following "Fig 8" it is known that the maximum weighttage has given to educational sites than other sites like entertainment, social, electronic commerce and research.

The most popular website is identified based on the condition that if no. of. sites in current site cluster is 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.

"Fig 9." shows that, the proposed algorithm proves it's efficiency for classifying the preference of users to various categories of websites.

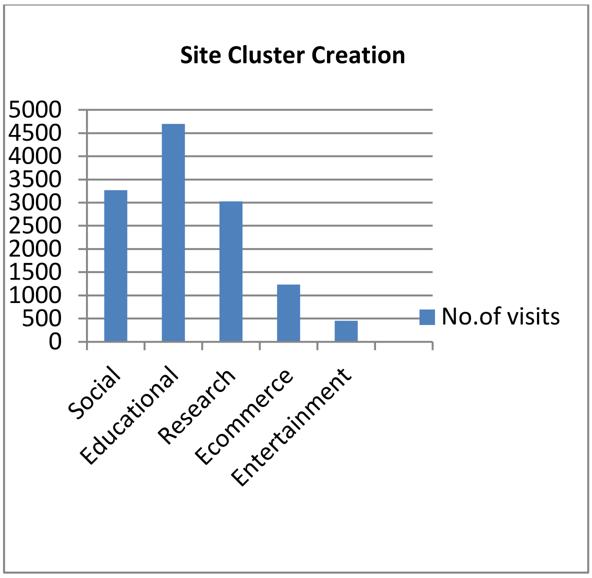


Figure 7. Site cluster creation

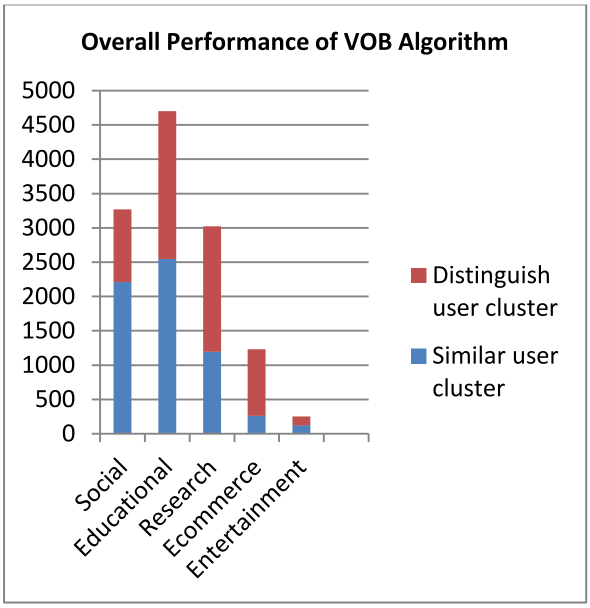


Figure 8. Overall Performance of VOB algorithm

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.

8. CONCLUSION AND SUMMARY

Web usage mining has emerged as the essential tool for realizing more personalized user-friendly and business optimal web services. The key is to use the user-click stream data for many mining purposes. Traditionally, web usage mining is used by e-commerce sites to organize their sites and to increase profits. The newly proposed algorithm is Visitors' Online Behavior (VOB) which identifies user behavior and creates user cluster, site cluster, most popular web site and the least popular web site. It must to trace the visitors' on-line behaviors for website usage analysis. Actually 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.

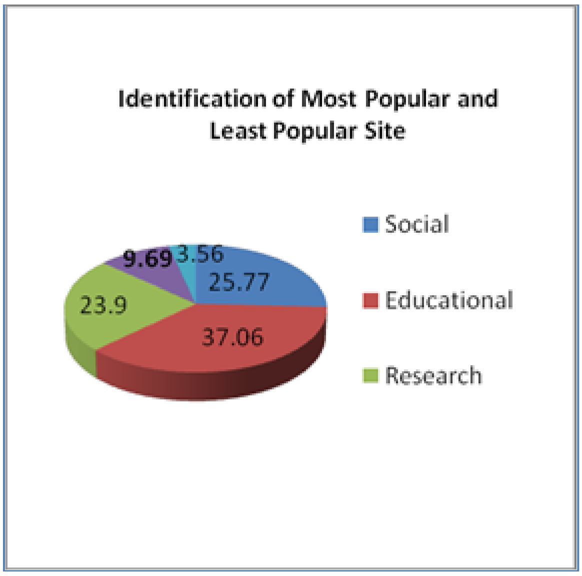


Figure 9. Identification of Most Popular and Least Popular Site

9. FUTURE ENHANCEMENT

A number of further tasks could be added by demonstrating the utility of web mining. It can be done by making exploratory changes to web sites. 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 also[1]. It can be further extended to user profiling and similar image retrieval by tracing the visitors' on line behaviors for effective web usage mining[11]. Many preprocessing techniques can be effectively applied in web log mining[7]. The preprocessing of web log data for finding frequent patterns using weighted association rule mining technique can be extended to other industrial and social organizations too[6]. In recent days, the web usage mining has great potential and frequently employed for the tasks like web personalization, web pages prefetching and website reorganization, etc[16]. So it is required to know the users' behavior when interaction is made with the web.

10. REFERENCES

[1] 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.

[2] 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, 2010, 848-851.

[3] Hussain.T, Asghar.S and Masood.N, "Web usage mining: A survey on preprocessing of web log file",Information and emerging technologies,2010, ISBN: **978-1-4244-8001-2**

[4] Jiawai Han and Micheline Kamber,"Data mining-Concepts and echniques", secondedition, Elsevier, Reprint 2010.

[5] http://www.galeas.de/webmining.html.

[6] 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.

[7] 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.

[8] J.Srivatsava, R.Cooley, M.Deshpande, and P.N. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data." ACM SIGKDD Explorat. Newsletter, 2000.

[9] 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

[10] 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.

[11] R. Suguna et.al,"User interest level based preprocessing algorithms using web usage mining", International Journal on Computer Science and Engineering.

[12] 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 July-December 2010, Volume 2, No. 2, pp. 279-283

[13] Cooley, R., B. Mobasher and J. Srivatsava, 1997. Web mining: Information and pattern discovery on the World Wide Web. Proceeding of the 9th IEEE International Conference on Tools with Artificial Intelligence, Newport Beach, CA., pp: 558-567. DOI: 10.1109/TAI.1997.632303

[14] Srivatsava, J., R. Cooley, M. Deshpande and P.N. Tan, 2000. Web usage mining: discovery and applications of usage patterns from Web data. ACM SIGKDD Explorat. Newsletter, 1: 12-23. DOI: 10.1145/846183.846188

[15] Agarwal, R. and R. Srikant, 1994. Fast algorithms for mining association rules in large database. Proceeding of the 20th Conference on Very Large Data Bases, Sept. 12-15, Morgan Kaufmann Publishers Inc., San Francisco, CA. USA., pp: 487-499. DOI: 10.1234/12345678

[16] 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, 2010, 3046-3052.

[17] Nanhay Singh1, Achin Jain1, Ram Shringar Raw, Comparison Analysis Of Web Usage Mining Using Pattern Recognition Techniques, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.4, July 2013

[18] L.K Joshila Grace, V. Maheswari, Dhinaharan Nagamalai, "Analysis of Weblogs and Web User in Web Mining," International Journal of Network Security & Its Applications (IJNSA), Vol. 3, No. 1, January 2011.

[19] 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.