Improved Technique for Path Completion in Web Usage Mining

Varun Dixit
Department of Information Technology
LNCT BHOPAL
BHOPAL, INDIA
vvarundixit@gmail.com

Abhishek Dwivedi
Department of Information Technology
LNCT BHOPAL
BHOPAL, INDIA
abhisdwivedi@gmail.com

ABSTRACT
Web Usage Mining (WUM) carries out exciting and crucial information for a group of people based on unlike work domains. We focus to generate patterns for the administrator of a website that is designed using cms. We consider a university website and try to identify the useful and meaningful information which can help the website administrator to manage the website. In the proposed they used Content management system (CMS), in which web pages are designed with same page name and it only uses the concept of unique id for retrieving the content of page. This makes the procedure of completion of path more complicated and hard. Sessions are taken into consideration for the particular user for the details. Browsers and operating systems are used for the user id which performed the tree for the visited users.

Keywords—**Web usage mining, data preprocessing, path completion, session identification.**

INTRODUCTION

Web usage mining needs to process web based data for identifying and predicting information which can be accessed. The category of web mining is the web usage mining that can be helpful in automatically discovering the patterns which are accessible to the user. Patterns and the information which are accessed indirectly from the activity which is related to that site. Here web usage mining work potentially to get the information related to the behavior of the user, which is further used for the web usage enhancements so that prediction of the upcoming next page which are likely accessed by user . it is helpful in detecting the crime on the web sites and user profiling and future prediction . it is the process where useful information is extracted from different web logs to find which all pages are accessed most of the time by people, which all web pages are being accessed together or one after the other and who are accessing what type of different web sites and for what aim. [1]

The first step of the web usage mining is data pre-processing. Data pre-processing result is directly knock the results of the next upcoming steps by including, path analysis, association rules mining, transaction identification and sequential patterns mining. Like, from the first step if we can get a better result, we could be able improve the patterns which are mined having quality and save algorithm's running time. It is important for the web log files, w.r.t the web log files structure that are not as same as the data within the database . They are complete and unstructured due to various origin . So it is necessary to web log files to be pre-processed inside the web usage mining. With the help of data pre-processing, transformed in the web log is their into another data structure, that can be easily mined.[2]

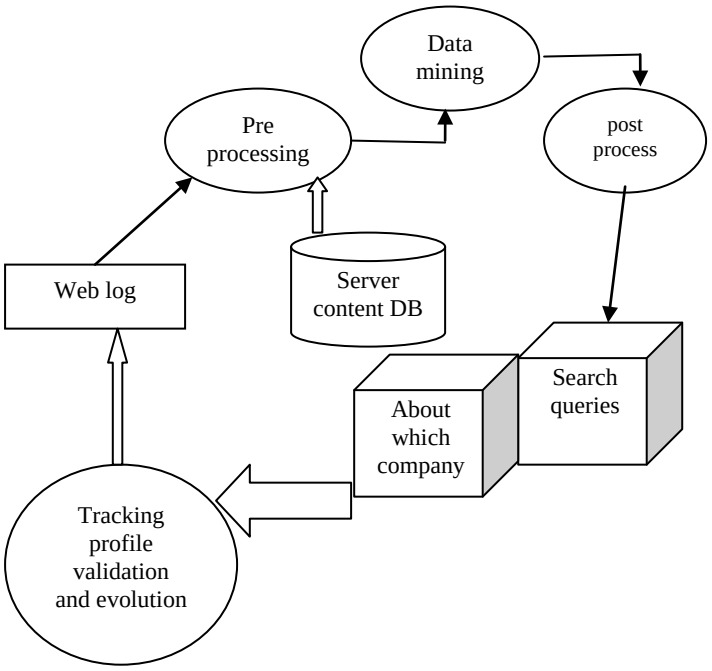


Fig 1 procedure of web usage mining.

II. DATA PREPROCESSING

The initial step within the data preparation process is the data pre processing, which aims to format the original logs that to identify the user side sessions. This type of process is much time consuming and deep step. The data pre processing is consisting of three different phases that are data cleaning, session identification, and user identification. Phase of data pre-processing is the most time consuming task because of qualitative the data will give better results. Algorithms and approaches are developed for data preprocessing. which will be discussed below:

a) Data cleaning :

In this process unnecessary and irrelevant fields from the raw data log file are removed. Extensions like gif, jpg, css in target URL are removed because this type of file actually not requested by users but automatically downloaded by HTML tags.

the records containing robots.txt inside the requested source name is removed by data cleaning. Because web robot is itself following all hyperlinks from within the web pages and like Google is periodically using web robots to collect all the pages from the particular website that to update their search indexes. These techniques are used to delete the irrelevant data from log data.[3] This is generally done for instance by mentioning to remote hostname, by mentioning to the user agent, or by examining the access to robots.txt file. However, some robots actually send a false user agent in HTTP request. In this type of cases, an heuristic based that is based on navigational behavior could be used to divide sessions of robot from actual users_ sessions.[4]

b) User Identification

A unique user can be able to identified by having the identification process of the user . By using IP address of the client , the unique user name can be identify.

1. Base on Client Information

Is the heuristic techniques used for user identification? Agent field of a log file contains browser and operating system name with version. If same IP address having different agent field it shows a different user. E.g. if user is visiting two pages of a same website by using two different browser simultaneously on a single device then this technique consider two records by the different user even they are from a single user.

2. Base on Topology:

Topology of website used to identify a user. If a user is requesting for a page that is un able to access from its previously pages that are requested is being considered as a new user, this could be done by using referrer attribute of log format and information of link from site topology .for e.g if a user is making a request by using the pages that are bookmarked which aren't concerned with via links.

3. By Using Cookies:

Cookie is a very small variable which is use to store few parameter value at the client side. A Cookie formed at server side and send to the client side. This cookie is containing some useful information regarding user so it is possible to identify a unique user. But this technique may not support, in which some browsers disable cookies.

c) Session Identification:

A user session is referring to a number of pages visited by the single user during a certain time period. We can differentiate entries into the various user sessions through a timeout.

1. Session identification by the time oriented heuristic:

Time gap is needed between the two entities, if time is exceeding to certain threshold then new session is being created.

If s.tn+1-s.tn>= the time threshold value then new session. Mainly value of threshold is half an hour or 30 minutes. This value will depend on site topology, application and on many types of parameters. Therefore, the fix threshold is not at all suitable for all applications. A Dynamic threshold is suggested according to type of application.

2. Session Identification by the time spent on page observation: Pages are classified into information pages and navigational pages that are based on time spent by the users. Information pages are a goal of user's, more time is spent by the users on information pages to study the content compared to navigational pages. This information is needed to define the session. If we are knowing the % of navigational pages inside log data, the maximum length of such page can be recognized by formula.

$$Q = -\ln(1-\gamma)/\lambda$$

Where q is a threshold value of a navigational page, γ is the % of navigational page, λ is observed duration time mean of all the pages within log data.

3. Session identification by the referrer:

W3C Extended log format having the referrer URL attribute. This attribute is occur in the same type of session. If no referrer is able to fount then it is first page of a new type of session. If two successive request a and b where p is a page and S is a session if referrer which is (r) for a page b is invoked within session S: then n is added to S, otherwise a new session is added. [5]

PATH COMPLETION

For example the design of websites with the dynamic pages inside our work. Designing of Web sites by applying the concept of CMS don't have a unique type of page name for all the page, instead of the pages which contain id through which the recovery of the pages content can be. So performance of completion of path becoming difficult and complex. After sessions identification, while pursuing for completion of path, we are building the page name by reading the page id and page name from the xml type of files like site map and RSS feed, and at the time of completion of path, we add missing type of pages inside the session, Eliminate the pages that are duplicate in the consecutive access within a session which is given and map the pages name with the page number. Adding the events permit to mine the web based logs on temporal idea, as the accessing to web through the users are not each time same . Lots of patterns can be discovered and then analyzed which are based on event

A. Path definition When we access the website, by the need of user, different path are built . It is essential to

understand the type of path before we use for path completion. Below are the definitions for many amount of path.

Path definition: A path denoted as p= {p1, p2, --, pn} (n is the different pages traversed in a session).

B. Construction of path

The path is built based on:-

Read the data first and according to the algorithm add it.

Add the url for a given type of session into single path.

Repeat all the above steps for each sessions for given file. After the algorithm, is applied to resultant path that will carry every missing pages that are added, every ID which is replaced through page names, so they are intelligible while pages are used for patterns discovery.

C. Event data generation

Event is based on the type of website . In events of online shopping case can be, brand wise sale, festival sale, season sale end. This is to highlights that web users may access and access the website in a several way at the time of different periods. We try to capture the events and the web logs mining specific to the events in accordance with the regular type of access in the website, in our work.[5]

III . LITERATURE SURVEY

Doddegowda B J(2016)et al presents about extraction of the behavioral patterns from the web usage data which is preprocessed for the web personalization . There is a use of different FSP mining algorithms, that are , WAP-tree, SPADE and Prefix Span for the Mining of FSPs from WUD of the academic web site for a period varying from weekly to quarterly. performance analysis of such FSP algorithms made against the #FSPs which they generate with a given minimum number of support . Here experimental outcomes is indicating that algorithm like PrefixSpan FSP mining do better than SPADE algorithms and WAP-tree for minimum support .[6]

Jorge Esparteiro Garcia (2016)et al presents about maintaining the requirements using the web usage data. paper shows REQAnalytics, which is a recommender system that is used to gather the information or data about the website usage , then processes it and then generates the recommendations to the website requirements specification this paper uses the case study on website of online newspaper. REQ Analytics is also helping in requirements management, which is contributing to the quality of the service of web itself.[7]

Satyaveer Singh(2016)et al shows web page Recommendation System which is based on Semantic here discussion is about web usage mining taxonomy of techniques of recommendation system and some open challenges and problems in the recommendation systems development. The paper is performing the case studies of some of web page which are main here recommendation frameworks is based on semantic web usage mining.[8]

Dr. Daya Gupta(2016)et al presents about various ranking algorithm which is the user preference based page these algorithm are page rank algorithm hits algorithm and weighted page rank algorithm discussion on the comparison of proposed algorithm by using various parameters with standard page ranking algos like PageRank, HITS and Weighted PageRank is also done. These algorithms tries to overcome limitations by giving priority to user trends and also to pages content both it also decreases the iterations number to reach the normalized page ranks . The algorithm of user preference which is based on page ranking algorithm that avoids analogy in ranking and is more powerful in nature by providing users an effective method to measure the quality of the page.[9]

Suharjito(2016)et al presents about the implementation of classification approaches of banking company with algorithm of k-nearest neighbor which is implemented with the standardized Euclidean distance which is to classify pattern which are frequently accessed . Here the result showing that the algorithm of0 k-nearest neighbor is implemented in web usage mining and also Help Company to search the interesting data inside web server log. Good result is shown by K-nearest neighbor in comparison of accuracy with the classification of naïve Bayesian so that it can be used inside web usage mining.[10]

IV. PROBLEM STATEMENT

The structure of website are measured in our method which should be deployed on a website outlined using Content Management System, which divides two features, 1) unique identifiers are considered for the generation of pages and 2) logical names are used in terms of actual names for the content fetching of the pages. These descriptions are significant to recognize before carrying out path completion. In the existing work, mainly authors focussed on the static web pages which are not suitable for today's world but we will work on dynamic web pages. In Content management system (CMS), web pages are designed with same page name and it only uses the concept of unique id for retrieving the content of page. This makes the procedure of completion of path more complicated and hard.

V. PROPOSED WORK

We proposed an algorithm in which we analyzed many patterns by discovering it from the event logs. The proposed work focussed on the full web log and gets the result on the basis of users. User based computation improves the user interaction and make it more easy to use according to the requirements. In the initial level, we fetch the data then perform identification of session for every user and construct the path. This flowchart clarifies the entire procedure of the proposed algorithm. The pseudocode for this algorithm has been simplified further and the whole process of proposed algorithm has been clarified below.

Proposed Algorithm:

Step 1:- Fetching the data stored in .xlxs file.

Step 2:- Session Identification is linked to find out the user's session.

Step 3:- For the Path Construction, following steps are established,

Read user session Si, Si= {S1, S2, - , SN} where n is the maximum number of sessions.

First url is splited into number of pages from the single session, pi= {p1, p2, - pn}, where n is the amount of pages.
Evaluate the length of the page after examine pi then checking it with the link name appeared in the link features. Record the link name, uiname and level in the path.
Determine the category of path in the presented links if the way then other than basic and path replaces it with the authentic name of the link.

Now examine the second url in the session
If url is identical as first, don't include it in the path, examine next url.
If url is dissimilar then record the link name, uiname, level and distance.

If uiname is same then insert the page name in the path, or else evaluate the nodes to be negotiated and record the new pages in the center of first and the second url.
Above steps are repeated for all urls in a specified session.
The url appended for a specified session into a single path.
Repeat the above procedures for every session in a specified file.

Step 4:- Path Completion
After the algorithm is executed to the resultant path will bring every the lost pages included, every one of the IDs are spread, so they are more readable while pages are utilized for finding patterns.

Step 5:- Pattern Discovery and Pattern Analysis
Patterns are created based on 4 aspects which are user ID based. The decisions are selected either date wise or overall. The decisions are:-

1. Page visited by the user.
2. Time spent on the web pages.
3. Browser used.
4. Operating system used.

Step 6:- Finally tree generated which show that the number of visited user and is known as visited page count tree.

VI. Result Analysis

We have done experiments on base and proposed algorithms.
Simulation tool :- MATLAB R2013.
The dataset contains synthetic data of above 1000 records. The database with all the values is kept in Microsoft Office Excel 2013. All experiments were executed well and entirely on a Dell workstation with 4 GB RAM and 32-bit OS, running windows 7.

Steps in the development of the proposed work:-
Step 1: When we run the program, a GUI is presented which is shown below:-

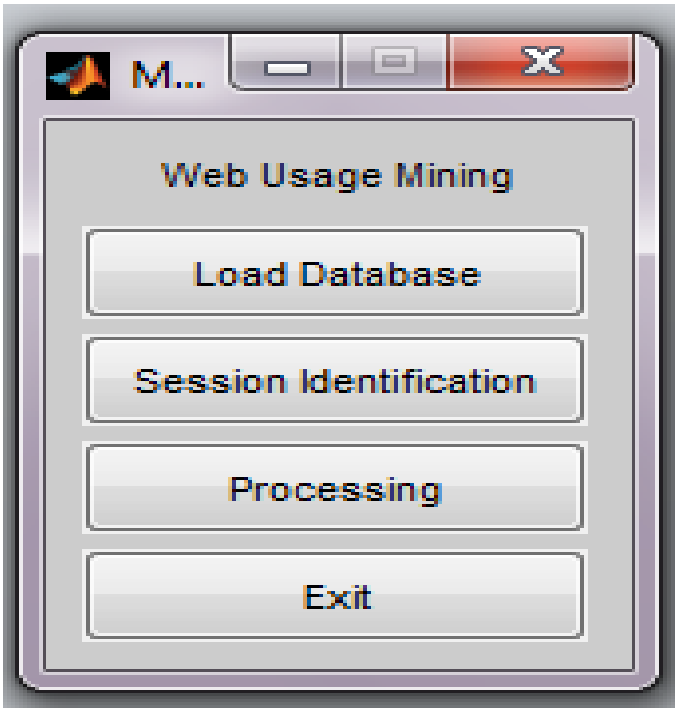
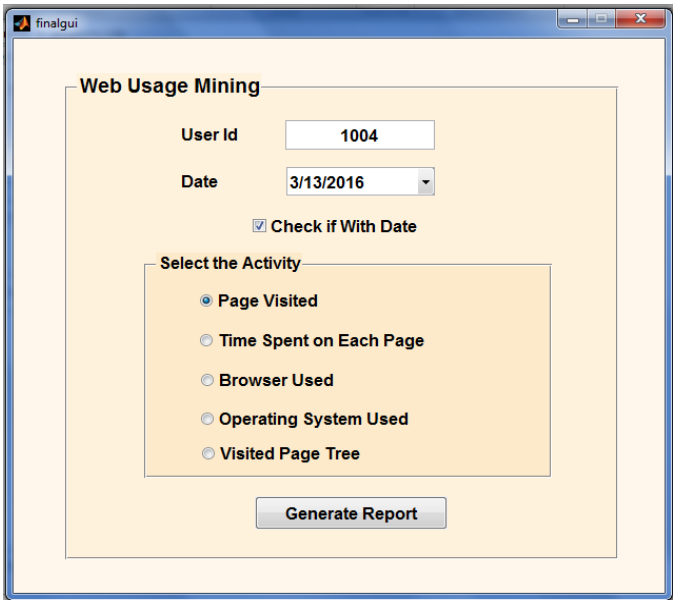


Fig 2. GUI.

Step 2: On clicking first button “Load Database”, a pop-up is created when data is loaded showing “Database loaded!”

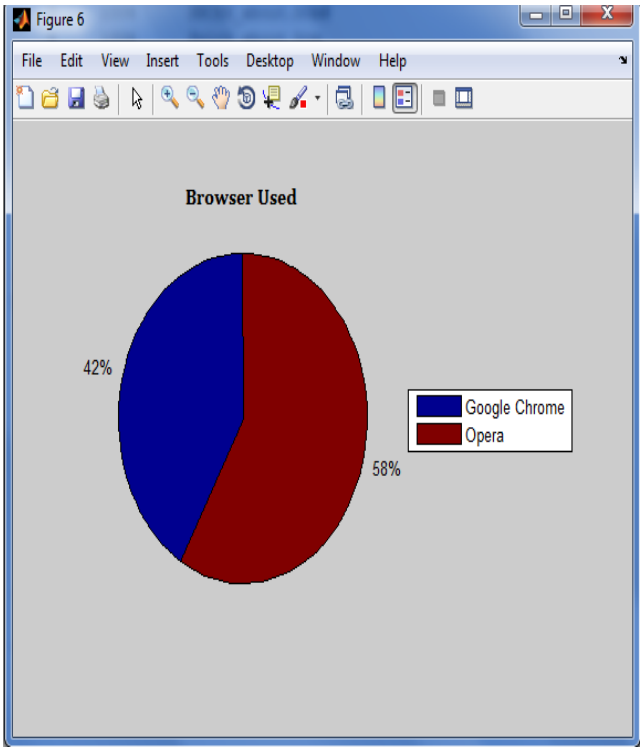
Step 3: Now on clicking button “Session Identification”, a pop-up is created when showing “Session Identified!”

Step 4: On clicking “Processing” button, a new window opens. Entering user ID and the date



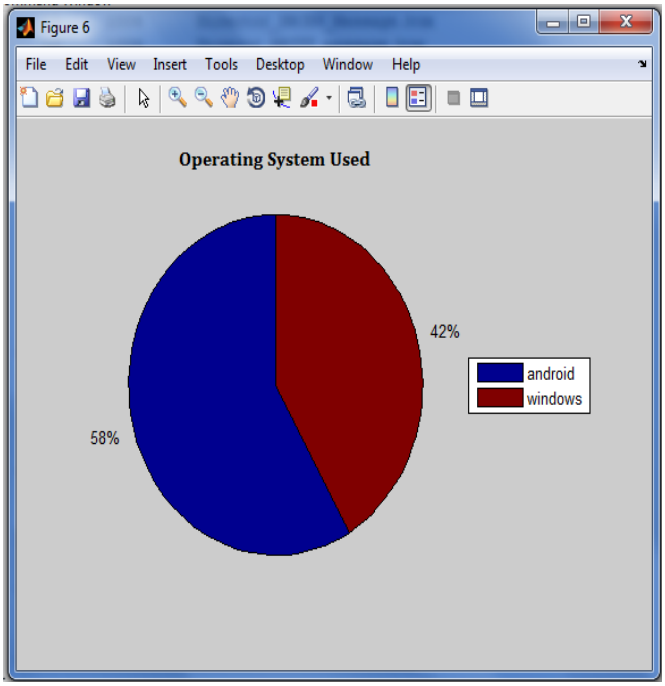
Browser Used

UserId	Browser Used
1004	Google Chrome
1004	Opera

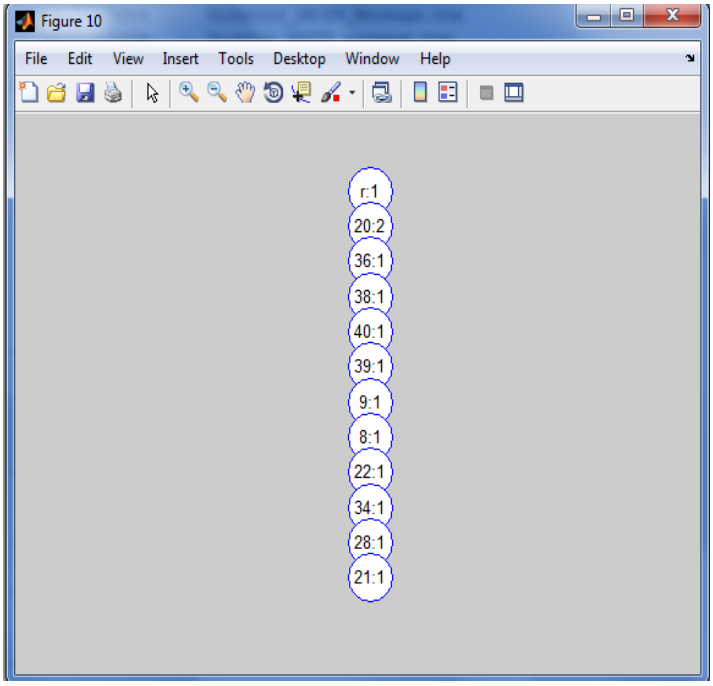


Operating System Used

UserId	Operating System Used
1004	android
1004	windows



Step 5:- now we generate the visited page tree for the particular user.



VII. CONCLUSION

With the growth of the Internet, the difficulty of information overload is increasing seriously. People have experienced the feeling of being overwhelmed by a number of new books, articles, and proceedings coming out every year. Path completion is a critical and hard task in the preprocessing phase of web usage mining. We have molded the pattern discovery and analysis phase to complete our goal to mine the data in better and appropriate way which is user based. So in our proposed work, we perform the path completion task in more accurate form to generate better results.

REFERENCES:-

Monika dhand i, rajesh kumar chakrawarti," A Comprehensive Study of Web Usage Mining, 978-1-5090-0669-4/16/$31.00 © 2016 IEEE.

[2] K. Sudheer Reddy, M. Kantha Reddy, V. Sitaramulu," An effective Data Preprocessing method for Web Usage Mining.

[3] Payal Sagar, Prof.A.V.Nimavat," Web Usage Mining: Survey on Process and Methods, Volume-2, Issue-5, May-2015 ISSN: 2349-7637

[4] ANITHA TALAKOKKULA," A Survey on Web Usage Mining, Applications and Tools, ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online) Vol.6, No.2, 2015

[5] Payal Sagar, Prof.A.V.Nimavat," Web Usage Mining: Survey on Process and Methods, Volume-2, Issue-5, May-2015 ISSN: 2349-7637

[6] Nandita Agrawal, Prof.Anand Jawdekar," Based Approach For Finding Various Results In Web Usage Mining, 978-1-5090-0669-4/16/$31.00 © 2016 IEEE.

[7] Doddegowda B J, G T Raju, Sunil Kumar S Manvi," Extraction of Behavioral Patterns from Preprocessed Web Usage Data for Web Personalization, 978-1-5090-0774-5/16/$31.00 © 2016 IEEE.

[8] Jorge Esparteiro Garcia , Ana C. R. Paiva ," Maintaining Requirements using Web Usage Data, 1877-0509 © 2016.

[9] Satyaveer Singh ,Mahendra Singh Aswal," Towards a Framework for Web Page recommendation System based on Semantic Web Usage Mining: A Case Study, 978-1-5090-3257-0/16/$31.00 ©2016 IEEE

[10] Dr. Daya Gupta, Devika Singh," User Preference Based Page Ranking Algorithm, ISBN: 978-1-5090-1666-2/16/$31.00 ©2016 IEEE

[11] Suharjito, Diana, Herianto," Implementation of Classification Technique in Web Usage Mining of Banking Company, 978-1-5090-1709-6/16/$31.00 ©2016 IEEE