A Survey on Various Approaches to Find Frequent Item-sets from web logs

[1] Asim Munshi, [2] Dr. Shyamal Tanna
[1]M.E. Student, [2]Faculty
[1]Department of Information Technology
[1]L.J.I.E.T, Ahmedabad, India.

***Abstract* - Web usage mining refers to programmed revelation of examples and related information, gathered or created as an after effect of client connections with one or more Web destinations. Principle objective is to investigate the behavioral examples and profiles of clients interfacing with a Web page. The found examples are represented as collections of pages, objects, or resources that are frequently accessed by groups of users with common interests. Web usage mining consists of three phases, namely pre-processing, pattern discovery, and pattern analysis. . In the pattern discovery phase, frequent pattern discovery algorithms applied on raw data. In the pattern analysis phase interesting knowledge is extracted from frequent patterns and these results can be used further.**

***Keywords* - Web Mining, Web usage mining, Web log mining, Pattern discovery**

I. INTRODUCTION

The internet features have influenced virtually every part of our universe. Since the number of web sites along with website pages are increasing and their features are improved rapidly, discovering and understanding web users surfing behavior are essential for the development of successful web monitoring and recommendation systems. One of the ways to achieve the above is through mining of Web Logs Web Usage mining [2] is the process of applying data mining techniques to the discovery of usage patterns from Web data, targeted towards various practical applications such as personalized web search and surfing, web recommendation systems. Data mining efforts associated with the Web, called Web mining, can be broadly divided into three classes, i.e. web content mining, web structure mining, and web usage mining. It attempts to discover useful knowledge from the secondary data, especially those contained in Web log files. Other sources can be browser logs, user profiles, user sessions, bookmarks, folders and scrolls. These data are obtained from the interactions of the users with the Web.

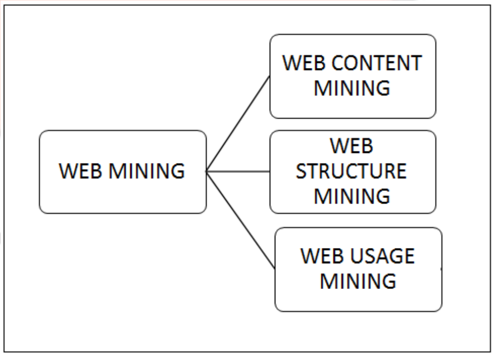


Figure 1 Taxonomy of Web Mining

II. TAXONOMY OF WEB MINING

Web mining can be broadly divided into three distinct categories, according to the kinds of data to be mined. Figure 1 shows the taxonomy

Web content mining

Web content mining is the process of extracting useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain. It may consist of text, images, audio, video, or structured records such as lists and tables. Application of text mining to web content has been the most widely researched. Issues addressed in text mining include topic discovery and tracking, extracting association patterns, clustering of web documents and classification of web pages.

Web Structure Mining

The structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Web structure mining is the process of discovering structure information from the web. Both hyperlinks and document structure can be mined

Web Usage Mining

The Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications. Usage data captures the identity or origin of web users along with their browsing behavior at a web site. web usage mining itself can be classified further depending on the kind of usage data.

III. OVERVIEW OF WEB USAGE MINING

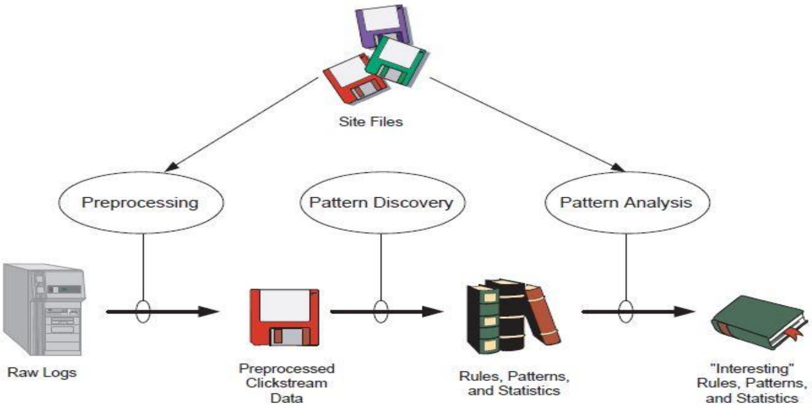
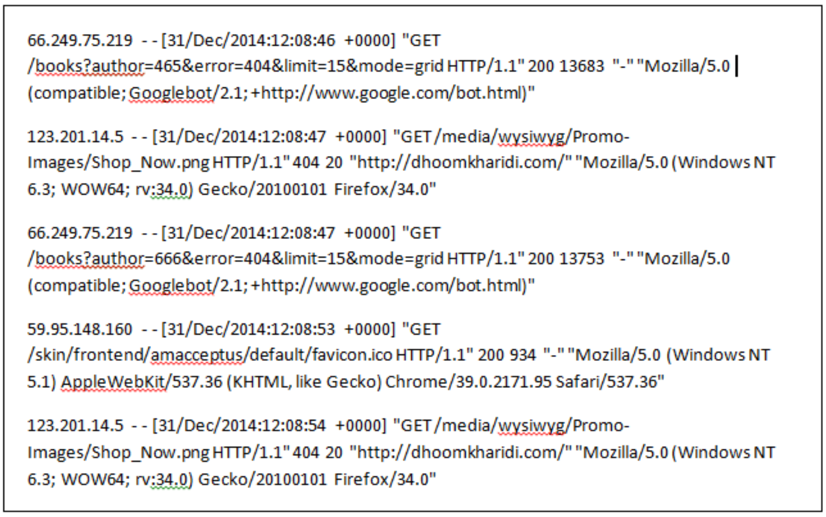

Site Files
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Logs
Preprocessed Clickstream Data
Rules, Patterns, and Statistics
"Interesting" Rules, Patterns, and Statistics

Figure – 2 Web Usage Mining Process

The web logs available over the internet can be termed as raw data and are not available in a required format hence for pattern discovery phase the data logs available over the internet need to be pre-processed hence pre-processing is the first and the most important step in the web usage mining process. There are three main sources to get the raw web log file [5] such as Client Log File, Proxy Log File and Server Log File. Uses of these sources have their own pros and cons but their importance to collect the data for web usage mining is invaluable.



```
66.249.75.219 - - [31/Dec/2014:12:08:46 +0000] "GET
/books?author=465&error=404&limit=15&mode=grid HTTP/1.1" 200 13683 "-" "Mozilla/5.0
(compatible; Googlebot/2.1; +http://www.google.com/bot.html)"

123.201.14.5 - - [31/Dec/2014:12:08:47 +0000] "GET /media/wysiwyg/Promo-
Images/Shop_Now.png HTTP/1.1" 404 20 "http://dhoomkharidi.com/" "Mozilla/5.0 (Windows NT
6.3; WOW64; rv:34.0) Gecko/20100101 Firefox/34.0"

66.249.75.219 - - [31/Dec/2014:12:08:47 +0000] "GET
/books?author=666&error=404&limit=15&mode=grid HTTP/1.1" 200 13753 "-" "Mozilla/5.0
(compatible; Googlebot/2.1; +http://www.google.com/bot.html)"

59.95.148.160 - - [31/Dec/2014:12:08:53 +0000] "GET
/skin/frontend/amacceptus/default/favicon.ico HTTP/1.1" 200 934 "-" "Mozilla/5.0 (Windows NT
5.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/39.0.2171.95 Safari/537.36"

123.201.14.5 - - [31/Dec/2014:12:08:54 +0000] "GET /media/wysiwyg/Promo-
Images/Shop_Now.png HTTP/1.1" 404 20 "http://dhoomkharidi.com/" "Mozilla/5.0 (Windows NT
6.3; WOW64; rv:34.0) Gecko/20100101 Firefox/34.0"
```

Figure 3 - Example of web logs

The true user behavior can be portrayed from a log file. The next step in web usage mining process is Pattern Discovery. According to the data pre-processing, the raw data is used to discover the knowledge and to implement the techniques which will be used for machine learning. This makes use of data mining procedures. The last process of the web usage mining is Pattern Analysis. It is the process after pattern discovery. It checks whether the pattern is correct on the web and guides the process of extraction of the information/ knowledge from the web

Pre-Processing

Various research works are carried in this area for grouping sessions and transactions, which is used to discover user's navigation patterns. In brief, the whole process deals with the conversion of raw Web server logs into a formatted user session file in order to perform effective pattern discovery and analysis phases. Generally, data pre-processing has four main tasks that are called data cleaning, user identification, session identification and path completion, as shown below.

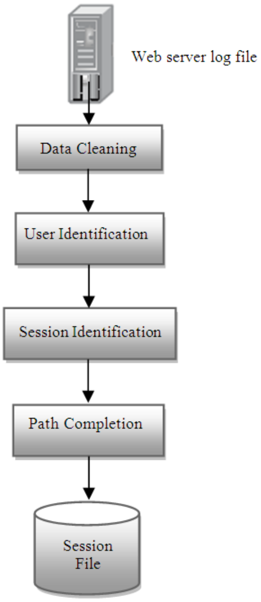


Figure 4 – Various steps involved in the pre-processing phase

Pattern Analysis

Pattern analysis is the last step of the Web Usage Mining process. In this phase we look to extract the interesting rules, patterns or statistics from the output of the pattern discovery process by filtering the rules or statistics which are not required. Navigational behavior and useful information about users can be obtained. For example, given a group of pages in a website we can know which page is visited more compared to other pages. Visualization techniques, such as graphing patterns or assigning colors to different values, can often highlight overall patterns or trends in the data.

Association Rule Mining

In the context of Web usage mining, once patterns have been identified association rules can be used to relate patterns or items that are most often referenced together in a single session. Such rules indicate the possible relationship between items that are often viewed together even if they are not directly connected, and can reveal associations between groups of users with specific interests. Since usually such transaction databases contain extremely large amounts of data, current association rule discovery techniques try to prune the search space according to support for items under consideration.

Support is a measure based on the number of occurrences of user transactions within transaction logs. The typical rule mined from database is formatted as:

$X \rightarrow Y$ [Support, Confidence]

It means the presence of item (page) X leads to the presence of item (page) Y, with [Support]% occurrence of [X,Y] in the whole database, and [Confidence]% occurrence of [Y] in set of records where [X] occurred.

Support = P (A /\ B)

= Number of sessions that contain A and B / Total number of Sessions

Confidence ($X\rightarrow Y$) = Support (X /\ Y) / Support(X)

Many algorithms can be used to mine association rules from the data available one of the most used and famous is the Apriori algorithm. This algorithm, given the minimum support and confidence levels, is able to quickly give back rules from a set of data through the discovery of the so-called large item set.

Apriori Algorithm

Apriori is a typical algorithm for frequent item set mining and association rule learning over transactional databases. It is proceed by recognize the frequent individual items in the database and extend them to big and big item sets as long as those item sets appear sufficiently often in the database. The frequent item sets find out by Apriori can be used to find out association rules which highlight general trends in the database: this has applications in domains such as market basket analysis. . It is is a typical algorithm for frequent item set mining and association rule learning over transactional databases. It is proceed by recognize the frequent individual items in the database and extend them to big and big item sets as long as those item sets appear sufficiently often in the database. The frequent item sets find out by Apriori can be used to find out association rules which highlight general trends in the database: this has applications in domains such as market basket analysis. It uses important property called Apriori property is used to reduce the search. All the non-empty subsets of frequent item sets must also be frequent this property belongs to a special category of properties called Anti-monotone. If a set can't pass a certain test than all of its supersets will fail the same test this property is called anti-monotone.

Apriori follows two steps approach:

- In the first step it joins two item sets which contain k-1 common items in kth pass. The first pass starts from the single item the resulting set is called the candidate set Ck.

- In the second step, the algorithm counts the occurrence of each candidate set and prunes all infrequent item sets. The algorithm ends when no further extension found

FP-Growth Algorithm

FP-growth is a well-known algorithm that uses the FP-tree data structure to achieve a representation of the database transactions and employs a divide-and-conquer approach to decompose the mining problem into a set of smaller problems. In essence, it mines all the frequent item sets by recursively finding all frequent item sets in the conditional pattern base which is efficiently constructed with the help of a node link structure.

The algorithm consists of two steps:

- Compress a large database into a compact, Frequent Pattern tree (FP-tree) structure – highly condensed, but complete for frequent pattern mining and avoid costly database scans
- Develop an efficient, FP-tree-based frequent pattern mining method (FP-growth) – A divide-and-conquer methodology: decompose mining tasks into smaller ones and avoid candidate generation: sub-database test only

FP growth is a noble approach that allows frequent patterns to be identified without generating candidate. But for large database and frequently changing or real time database, creating this tree can be a time consuming process.

IV. RELATED WORK

In [1] an algorithm to predict user's behavior is proposed. The algorithm is named as single scan pattern recognition algorithm as it scans the database only once. The connectivity of the web data is taken into consideration. In usual approach to find frequent pattern a pattern tree is created and then analysis is done but in the approach proposed by this paper there is no need for tree creation and analysis is done on website architecture. In this algorithm first transaction is taken and then path is taken and followed, each node has 0 as account initially the value of count is increased by one whenever a path is traversed. When the structure of website structure is very complex the graph creation can become very complex and can result in poor outcome. Also in today's world the structure of website keeps on changing and every time the structure changes anew scan will be required. Hence a system for extracting user's navigational behavior is presented. An undirected graph based on connectivity between each pair of the Web pages was considered and also proposed a new formula for allocating weights to each edge of the graph.

In [2] a graph based approach is suggested to mine frequent sequential access patterns for users. It uses frequent sequential pattern algorithm. It compromises of three basic steps which are - Construct a graph, Prune a graph, Mine a frequent sequential pattern from web usage graph. It emphasizes on to showing how frequent pattern discovery tasks can be accomplished by capturing complex user's browsing behavior in to a graph data structure in order to obtain hidden useful user's access patterns. In [7] the same above work is extended to generate useful recommendations Recommender System is one kind of filtering system which is used for ranking or priority of the objects. The recommendations are retrieved for a given user's web access sequence. Length of the user web access sequence must satisfy the thresholds. If its length is greater than max length then we have to remove first element. If it contains next item then the recommendation rule order by the support is returned

In [3] a method to predict the user's navigation patterns is proposed using clustering and classification from Web log data. First phase of this method focuses on separating users in Web log data, and in the second phase clustering process is used to group the users with similar preferences. Finally in the third phase the results of classification and clustering are used to predict the users' next requests.

Theint Theint, Aye in [4] stressed the importance of pre-processing step in the mining of web logs. Data pre-processing is an important task of Web usage mining. Therefore, data must be processed before applying data mining techniques to discover user access patterns from web log. The data preparation process is often the most time consuming. This paper presents two algorithms for field extraction and data cleaning. Not every access to the content should be taken into consideration. So this system removes accesses to irrelevant items and failed requests in data cleaning. After that necessary items remain for purpose of analysis. Speed up extraction time when user's interested information is retrieved and users' accessed pages is discovered from log data. The information in these records is sufficient to obtain session information.

In [6], a model for association rules to mine the generated frequent k-itemset is proposed. This process is taken as extraction of rules which expressed most useful information. Therefore, transactional knowledge of using websites is considered to solve the purpose. In this paper interestingness measure that plays an important role in removing invalid rules thereby reducing the size of rule data sets is used. The performance analysis attempted with Apriori, most frequent rule mining algorithm and interestingness measure to compare the efficiency of websites. The proposed work reduces large number of immaterial rules and produces new set of rules with interesting measure. The current algorithm is not capable of handle very large number of log entries and thus suffers from scalability point of view.

V. CONCLUSION

In Discovering hidden information from large amount of Web log data collected by Web servers is very difficult, pattern discovery has become one of the most important phases in Web usage mining. This paper presented a brief introduction to Web usage mining and focused on methods that can be used for the task of pattern extraction from Web log files. After discovering patterns, the result will be used for pattern analysis phase. Analyzing of the Web user's navigational patterns can help understand the user behavior's and Web structure; therefore the design of Web components and Web applications will be improved. Various methods to mine web data suggests that the methods shown above work fine with smaller amount of data but real time databases are very large and hence they suffer from scalability point of view. To solve this issue we propose the use of data mining technique which is more scalable. More scalability can be achieved by improving the time and memory parameters of the existing algorithm.

VI. REFERENCES

[1] Murli Manohar Sharma, Anju Bala, “An approach for frequent access pattern identification in web usage mining” IEEE Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference in September 2014, Pages 730 – 735, ISBN 978-1-4799-3078-4.

[2] Dheeraj Kumar Singh, Varsha Sharma, Sanjeev Sharma “Graph based Approach for Mining Frequent Sequential Access Patterns of Web pages”, International Journal of Computer Applications 2012 Applications (0975 – 8887) Volume 40– No.10, February 2012

[3] V. Sujatha, Punithavalli “Improved user Navigation pattern prediction technique from web log data”, Science Direct, Procedia engineering, Volume 30, 2012, Pages 92 – 99.

[4] Theint Theint, Aye “Web Log Cleaning for Mining of Web Usage Patterns” IEEE Computer Research and Development (ICCRD), 2011 3rd International Conference (Volume:2), March 2011, Pages 490 – 494, ISBN 978-1-61284-839-6.

[5] Hussain S, Asgar S, Masood N” Web usage mining: A survey on pre-processing of web log file” IEEE Information and Emerging Technologies (ICIET), 2010 International Conference, June 2010, Page 1 – 5, ISBN 978-1-4244-8001-2.

[6] Avadh Kishor Singh, Ajeet Kumar, Ashish K. Maurya, “Association Rule Mining for Web Usage Data to Improve Website”, IEEE Advances in Engineering and Technology Research (ICAETR), 2014 International Conference, Aug 2014, Page 1 – 6, ISSN 2347-9337.

[7] Valera M, Chauhan U, “An efficient web recommender system based on approach of mining frequent sequential pattern from customized web log pre -processing”, IEEE fourth conference on computing, communications & networking technologies (ICCCNT), July 2013, Page 1 – 6, ISBN 978-1-4799-3925-1.