A Survey on Web Usage Mining Techniques

`1R.Umamaheswari, 2K.Saraswathi
1M.Phil Research Scholar, 2Assistant Professor
PG and Research Department of Computer Science
Government Arts College, Coimbatore

Abstract: **Web usage mining is the part of web mining techniques. This web usage mining(WUM) can be used to identifies the data from the web log servers.The WUM involves three types, Data Preprocessing , Pattern Discovery & Pattern Analysis. world wide web it provides the lot of information . The web mining is used to discover and extract useful information from the web sites.web mining is used to gather the important information from customers visiting the site.Web mining can be classified into three types Web Content Mining, Web Structure mining, Web Usage Mining. Web content mining is used to search information of resource available in online. Web structure mining is used to structure of hyperlink with in the web itself. Web usage mining is used to used to log data stored in the web server.**

Keywords:
Data preprocessing, Pattern analysis, Pattern discovery, Web mining,

I. INTRODUCTION

World Wide Web is the biggest and most popular way of communication. Today, User wants search anything on the internet, within a fraction of a secondby just a single click. It serves as a platform for exchanging information.

The volume of information available on the internet is increasing now days. So analyzing user's behavior (recorded in web server log files) is an important part of web page design. Web users usually suffer from the information overloaded. Because, significant increase and the rapidly expanding growth in the amount of information available on the web.

Web mining strategies:

- Web content mining
- Web structure mining
- Web usage mining

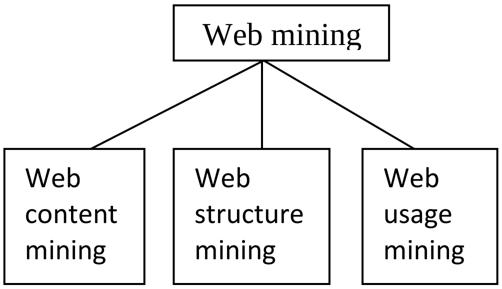


Figure 1: Kinds of web mining

The three important kinds of web mining are explained here.

Web content mining

Web content mining is used to search information of resource available in online.

It deals with discovery the useful information from the web content (data/document/services). It consist several types of data such as textual, image, audio, video, etc. Two different points of view are used to the content mining.

- Information Retrieval View
  It can be used to only retrieval the information.
- Database View
  Database view can be accessed the web site.

Web structure mining

Web Structure Mining is the process. Hyperlinks are used to access the structure of a web site within the web itself. Structure of the hyperlinks with connected to the web pages. Structure mining consists of unstructured and semi structure. Structure represents the graph of the link in sites. Structure mining consist graph theory used to analysis node of a web pages.

Web usage mining

Web Usage Mining mines the log data stored in the web server.web usage mining consists of textual logs collected by web servers all around the world. There are fourstages in web usage mining.

Data Collection:

Data Collection is the process. It used to collect the relevant data from web pages. Data are collected from server side, client side, proxy side and so on.

Preprocessing :

Preprocessing is one of the important steps in Data mining. Web log used to remove the inconsistent data from log files. Preprocessing steps include data cleaning, user identification, session identification, path completion and transaction identification.

Pattern discovery :

This is one of the important step in data mining techniques. Data mining techniques like that

1. Association rule analysis,
2. Clustering,
3. Classification
4. Sequential pattern.

Pattern analysis:

Once patterns were discovered from web logs, uninteresting rules are filtered out. Analysis is done using information uncertainty mechanism such as SQL or OLAP operations.

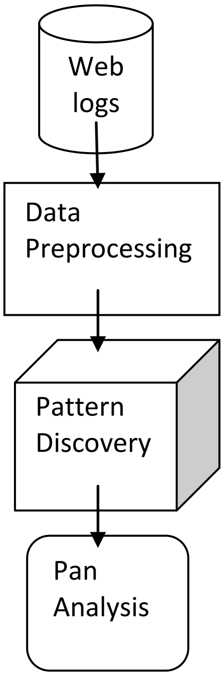


Figure 2: Steps involved in W**eb usage mining**

The four important steps in web usage mining are explained here.

The objective of this paper is to provide a evaluation of web usage mining and a assessment of preprocessing phase. Data can be collected from the various data sources and preprocessing segment review. The dissimilar works had done in session identification, path completion process.Outstanding sections briefs regarding pattern discovery, analysis and the different areas of applications where web usage mining is used.

II. DATA COLLECTION

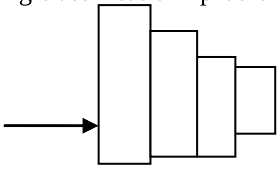
Data Collection is the first step in web usage mining process. The relevant data on web are gathered. Data's can be collected from the server-side, client-side, proxy side. Database contains business data or consolidated data on Web itself.

Server side:

Client requests are collected and stored in the server as web logs. Web server logs that convert the plain text that is independent from server platform. Most of the web servers go after common log layout as "IP address of username password.

Client side:

Client sends the request to server using JavaScript and applets. The client side waits for a few seconds for the server reply. It overcomes both the caching and gathering classification problems.



Proxy side:

Intermediate between Client and Server is called proxy. Data collected from intermediate server between browsers and web servers. Proxy caching is used to reduce the loading time of a Web pages [11]. Proxy access the record from proxy servers. It consist same format as web server log. The web page request and response for the server. Proxy Server may traces the actual HTTP requests from multiple clients to multiple Web servers.

III. DATA PREPROCESSING

Pre-processing of web log data is most important step to analysis the Web Log.

Pre-processing is a method to applying steps on web logs. Pre-processing consists of Data Cleaning, User Identification, Session Identification, and Path Completion. Data Pre-processing is a essential step before analyzing and improve the qualityof results can be displayed.[8]

Data preprocessing transforms data into a format that will be added simply, and powerfully process of the function to user. The main task of data preprocessing is to chooseconsistent data from the unique log files, prepared for user navigation pattern discovery algorithm.

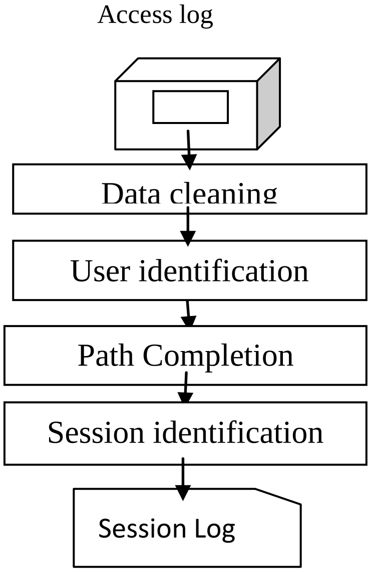


Figure 3: phases of Data Preprocessing

The four important phases of Data Preprocessing method are explained here.

Data cleaning

Data Cleaning is a first important step in web log files. Data cleaning is used to remove out the unrelated information from the web log records. This stride helps to decrease the size of the data by removing out unwanted log data and also improve the quality of data. The records referring images, graphics or video etc. are removed. And also the records with failed HTTP status codes are eliminated. Removing irrelevant items such as jpeg, gif files or sound files from web pages. Accessorial resources set in file of the HTML, robots error request [8].

Robots (spiders) are instrument of software with the purpose of search a website online to extract the comfortable material. Spiders automatically conform to the complete hyperlinks from web log page. To explore search engine is use spider for grape in

Each page from online web site and customized their search index [6].

User Identification

The next step after the data cleaning is identification of users. Here users know how to be recognized by exclusive IPaddress or the unique ID assigned by the server (cookies). Cookies help the website developer to easily identifying individual visitors.User Identification is required to categorize User accesses of websites or web pages. Each user will be identified by a unique IP address and different users will have same IP address. Then the different browser and different operating system represent different user [8]. If the requested page is not directly reachable from any of the pages visited by the user, then the user is recognized as a new customer in the same address [5]. If the IP address and user agent are same then only they display the correct web pages. Caching problem can be rectified by assigning a short expiration time to HTML pages enforcing the browser to recover every page from the server [2].

Session Identification

Session identification means set of user used to one or more web site. To group the activities of a single user from the web log files is called a session. Session Identification step used to web log pre-processing phase is identifying each individual user session [5]. The goal of this step is to group page accesses or activities of each user into individual session. One of the general methods used to classify the session and time out mechanism as long as user is connected to the web site [7]. Two methods are used to session identification.

1. *Time Oriented Heuristics*
2. *Navigation-Oriented Heuristics*

These methods are used by a lot of applications. To develop the performance in dissimilar methods were devised on the basis of Time and Navigation leaning heuristics by special researchers. Different works were done by researchers for effective reconstruction of sessions[10].

Path Completion

The path completion step is used to fill in missing page references. This techniques used to similar for identification of user can be used for completion path. Path completion which is based on the knowledge configuration and sequence of the reference from the server logs. In several times user like to click on browser's back button. While browsing the result to incomplete user access path which will remain not entered in web logs, so to work with this kind of situations the Path Completion step is necessary.

IV. PATTERN DISCOVERY

Pattern discovery is the next step after completing the data preprocessing . This is one of the important steps in data mining techniques. Data mining techniques like that

1. Association rule analysis,

2. clustering,
3. classification
4. Sequential pattern.

Association rule: Association rule analysis is used to find the relationship between pages and server sessions. The association rules may contains set of pages that are accessed together with a maintain value above some specified entry. These pages may not be directly associated to one another via links [4].

Clustering:In this technique is used to combine the group together a set of items having similar characteristics. There are two types of clusters to be discovered:
1. Usage clusters
2. Page clusters.
Usage Clustering used to establish the groups of users that represent similar browsing patterns. In page clusters, clustering of pages will discover and groups of pages having correlated content [4].

Classification: Classification means the data item can be splattered in one or more number of predefined classes. Classification uses supervised learning algorithms. Naive Bayesian classifiers, decision tree classifiers, k-nearest neighbor classifiers etc.

Sequential Pattern: This method of sequential pattern used to discovery attempts to find out some inter-session patterns.

V. PATTERN ANALYSIS

Pattern analysis is the final step on the whole Web Usage mining process. The inspiration behind pattern analysis is to clean out unexciting rules or pattern from the set of pattern found in the pattern discovery phase. The most familiar type of pattern analysis consists of a knowledge query mechanism like SQL [3]. Another way is to enter usage information into a data cube in order to execute various OLAP operations like roll-up, drill-down etc. Some Visualization techniques, graphing patterns or assignment of colors to various values, can often highlight general patterns or trends in the data [4].

VI. CONCLUSION

This paper has presented the details of web usage mining tasks that are performing the web sites. We give some rules in every phases of the web usage mining in order to develop the design and implemented them easily.The results of mining can be used to improve the website design and increase the applications on the web logs. Web usage mining is good and quality research in data mining field. We can also increase the volume of request resources for future.

REFERENCES

1.Dr. Antony SelvdossDavamani, Reader in Computer Science , NGM College (AUTONOMOUS) Pollachi, Coimbatore,Tamilnadu, India, V.Chitraa , Lecturer ,CMS College of Science and Commerce ,Coimbatore, Tamilnadu, India" A Survey on Preprocessing Methods for Web Usage Data" *(IJCSIS),Vol. 7, No. 3, 2010*

2.Chitraa.V, Dr. Davamani A,"A Survey on Preprocessing Methods for Web Usage Data" (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 3, 2010.

3.Cyrus Shahabi, Amir M.Zarkessh, JafarAbidi and Vishal Shah "Knowledge discovery from users Web page navigation, ", In. Workshop on Research Issues in Data Engineering, Birmingham, England,1997.

4. JaideepSrivastava, Robert Cooley, MukundDeshpande, Pang-Ning Tan, ―Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data‖. 1999.

5.MichalMunk, JozefKapusta, Peter Švec, Constantine the Philosopher

6. Nirali H.Panchal1, Ompriya Kale2, M.E1, 2, Assistant Professor2, computer Engineering Department1,2,, L J Institute of Engineering and Technology1, 2, Ahmadabad, Gujarat, India" A Survey on Web Usage Mining" , **(IJCTT) – volume 17 Number 4 Nov 2014**, ISSN: 2231-5381

7. RajniPamnani, PramilaChawan 1 Qingtian Han, XiaoyanGao, "Web Usage Mining. A Research Area In Web Mining".

8.Dr. P.Sumathi, B.UmaMaheswari Asst. Professor, Govt. Arts College, Coimbatore, India, Research Scholar, Dept. of Computer Science & Applications, Bharathiyar University, Coimbatore, India"**A Survey on Web Usage Mining Preprocessing", Vol. 3, Issue 11, November 2015**

9.Trousse B. ,Tanasa D. Advanced data preprocessing for intersites Web usage mining. Intelligent Systems, IEEE,2004(19): 59 – 65.

10.VijayashriLosarwar, Dr. Madhuri Joshi" Data Preprocessing in Web Usage Mining".

11.Wenguo Wu, "Study On Web Mining Algorithm Based On Usage Mining", Computer- Aided Industrial Design And Conceptual Design, 2008. CAID/CD 2008. 9th International Conference On 22-25 Nov.2008.