

WEB USAGE MINING-AN OUTLINE

C.Veerakumar,

Assistant Professor,

Department of Computer Applications,
Erode Arts and Science College,
Erode, Tamilnadu, India.

Abstract: Web usage mining is the area of data mining which deals with the discovery and analysis of usage patterns from Web data, specifically web logs, in order to improve web based applications. Web usage mining consists of three phases, pre-processing, pattern discovery, and pattern analysis. The web access log file is saved to keep a record of every request made by the users. In this paper, we have considered the Data Mining and Web Usage Mining for overall study.

Keywords: Data Mining, Web Mining, Web Usage Mining, Web Log.

I. INTRODUCTION

Data is the collection of values and variables related in certain sense and differing in some other sense. Data should be processed to extract some useful knowledge from it. Mining refers to the large-scale of information extraction from your target content. If you extract information from just one or two items of content – that's 'data extraction'. But if you extract information from thousands of separate items of content – that's 'mining'. A Web resource is a resource accessible through any version of the HTTP protocol (for example, HTTP 1.1 or HTTP-NG). A Web server is the server that provides access to Web resources. A Web page is the set of data constituting one or several Web resources that can be identified by a URI. If the Web page consists of n resources, the first $n - 1$ are embedded and the n th URI identifies the Web page. A page view occurs at a specific moment in time, when a Web browser displays a Web page. A Web browser or Web client is client software that can send Web requests, handle the responses, and display requested URIs. A user is a person using a Web browser. A Web request is a request a Web client makes for a Web resource.

II. DATA MINING

Data Mining is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. It is defined as the automatic extraction of unknown, useful and understandable patterns from large database. Utilizing these massive volumes of data is becoming a major problem for all enterprises. Low quality

data will lead to Low-quality missing results. Data mining also called Knowledge Discovery in Databases (KDD) or Knowledge Discovery and Data Mining.

III. WEB MINING

Web mining is the application of data mining techniques to discover and analysis of useful information from the World Wide Web.

Types of Web Mining

The web mining techniques can be divided into a major three areas as follows;

Web content mining: The web content mining focuses on the information discovery/retrieval of the useful information from the Web documents. It may consist of audio, text, images or structured records such as tables and lists.

Web structure mining: The web structure of a typical web graph consists of web pages as nodes and hyperlinks as edges connecting between two related pages. The web structure mining as a process of discovering structure information from the Web. This type of mining can be performed either either hyperlinks nor document structures.

Web usage mining: Web usage mining integrates the technique data Mining and Internet. WWW is an immense source of data that can be collected from web content or from the web usage. Web usage mining is the type of web mining activity that involves the automatic discovery of user access patterns from one or more web servers.

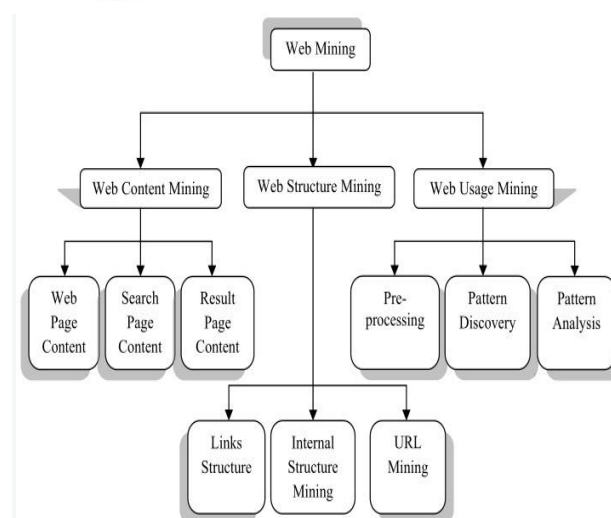


Fig 1: Types of Web Mining

IV. WEB USAGE MINING

Web pages may not only contain textual content but also other types of Web data which may include audio files, video files and images. The Web data is stored in Web servers, client machines, proxy servers or organizational databases. The primary data sources used in Web usage mining are the server log files which include Web server access logs, referrer logs and agent logs. Web Usage Mining is that part of Web Mining which deals with the extraction of knowledge from server log files; source data mainly consist of the (textual) logs that are collected when users access web servers and might be represented in standard formats. When exploiting log information from web servers, the major issue is the identification of user's sessions.

A) Data Collection

Web Usage Mining applications are based on data collected from three main sources:

- **Web servers:** The log file that resides in the web server notes the activity of the client who accesses the web server for a web site through the browser.
- **Web clients:** This kind of log files can be made to reside in the client's browser window itself. Usage data can be tracked on the client side by using JavaScript, java applets, or even modified browsers.
- **Web proxy :** A web proxy server is said to be an intermediate server that exit between the client and the web server. These web proxy servers maintain a separate log file for gathering the information of the user.

B) Web Log File

Web log files are files that contain information about website visitor activity. Log files are created by web servers automatically. Each time a visitor requests any file (page, image, etc.) from the site information on his request is appended to a current log file. Most log files have text format and each log entry (hit) is saved as a line of text.

C) Type of Web Log File

There are four types of server logs.

- **Access log file:** Data of all incoming request and information about client of server. Access log records all requests that are processed by server.
- **Error log file:** list of internal error. Whenever an error is occurred, the page is being requested by client to web server the entry is made in error log. Access and error logs are mostly used, but agent and referrer log may or may not enable at server.
- **Agent log file:** Information about user's browser, browser version.
- **Referrer log file:** This file provides information about link and redirects visitor to site.

D) Web Log File Format

Web log file is a simple plain text file which record information about each user. Display of log files data in three different format

- W3C Extended log file format
- NCSA common log file format
- IIS log file format

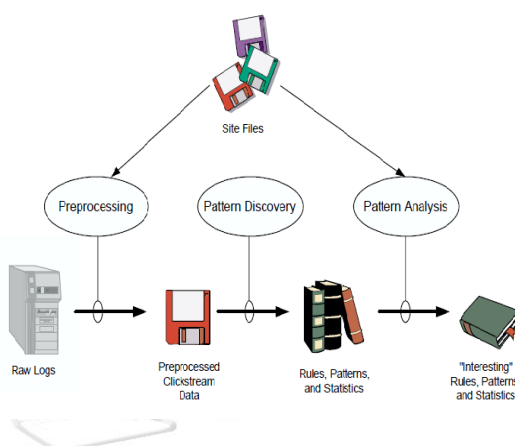
NCSA and IIS log file format the data logged for each request is fixed. W3C format allows user to choose properties, user want to log for each request.

```
<ip_addr><base_url> . <date><method><file><protocol><code><bytes><referrer><user_agent>
```

V. STEPS IN WEB USAGE MINING

The process of Web usage mining also consists of three main steps:

- Pre-processing,
- Pattern discovery
- Pattern analysis



(i) Pre-processing

The goal of pre-processing is to transform the raw click stream data into a set of user profiles. Pre-processing removes many entries from the data files that are considered uninteresting for pattern discovery. A simple Web site of simple HTML pages will generate less usage data. A graphical site or dynamic site will generate more data and noise. Data pre-processing has four main tasks that are called data cleaning, user identification, session identification and path completion.

Data Cleaning : Data Cleaning is a process of removing irrelevant items such as jpeg, gif files or sound files and references due to spider navigations. Improved data quality improves the analysis on it.

User Identification : Identification of individual users who access a web site is an important step in web usage mining. Various methods are to be followed for identification of users. The simplest method is to assign different user id to different IP address.

Session Identification : A user session can be defined as a set of pages visited by the same user within the duration of one particular visit to a web-site. A user may have a single or multiple sessions during a period. Once a user was identified, the click stream of each user is portioned into logical clusters. The method of portioning into sessions is called as Sessionization or Session Reconstruction.

Path Completion : There are chances of missing pages after constructing transactions due to proxy servers and caching

(ii) Pattern Discovery

Once the raw logs have been pre-processed, data mining techniques can be applied on the dataset to discover new patterns. Pattern discovery describes various methods and algorithms developed on statistics, data mining, and pattern recognition. Pattern discovery describes the type of mining technique that has been applied to the Web domain. In Web Usage Mining, a server session is an ordered sequence of pages requested by a user. Furthermore, due to the difficulty in identifying unique sessions, additional prior knowledge is required. Some important techniques of pattern discovery are as follows.

- Association rules mining
- Sequential pattern mining
- Clustering

(iii) Pattern Analysis

The WUM has a strange consequence and the analysis of these patterns allows distinguishing interesting results from non-interesting ones. The issue with WUM results is that it is extremely hard to capture and define the notion of interestingness [4]. In this step of the WUM process, the analyst is interested in projecting the patterns discovered on the Web site structure or on its content. For instance, the analyst may be interested to see on which sections of the Web site are situated, the pages contained in the most frequent sequential.

VI. APPLICATIONS OF WEB USAGE MINING

Personalization of Web Content: Web Usage Mining techniques can be used to provide personalized web user experience. For instance, in real time, it is possible to predict the user behaviour by comparing the current navigation pattern with typical patterns which were extracted from past web log.

Prefetching and Caching: The results produced by Web Usage Mining can be exploited to improve the performance of web servers and web-based applications. With the use of weblogs that store user's access history can be used predict future accesses. Web Usage Mining can be used to develop proper prefetching and caching strategies so as to reduce the server response time.

E-Commerce : Mining business intelligence from web usage data is dramatically important for e-commerce web-based companies. Web usage mining techniques can also be

problems. So missing pages are added as follows: The page request is checked whether it is directly linked to the last page or not. If there is no link with last page check the recent history. If the log record is available in recent history then it is clear that "back" button is used for caching until the page has been reached. If the referrer log is not clear, the site topology can be used for the same effect. If many pages are linked to the requested page, the closest page is the source.

useful in Customer Relationship Management (CRM). The issues specific to business such as customer attraction, customer retention, cross sales, and customer departure are mainly in focus.

Support to the Design: Usability is one of the major issues in the design and implementation of web sites. The results produced by Web Usage Mining techniques can provide guidelines for improving the design of web applications. Adaptive Web sites represent a further step. In this case, the content and the structure of the web site can be dynamically reorganized according to the data mined from the users' behaviour.

VII. CONCLUSION

Web sites are one of the most important tools for advertisements in international area for universities and other foundation. The quality of a website can be evaluated by analyzing user accesses of the website. To know the quality of a web site user accesses are to be evaluated by web usage mining. The results of mining can be used to improve the website design and increase satisfaction which helps in various applications. Log files are the best source to know user behaviour. But the raw log files contains unnecessary details like image access, failed entries etc., which will affect the accuracy of pattern discovery and analysis.

REFERENCES

- [1]. Aye, T.T., 2011. Web Log Cleaning for
- [2]. Mining of Web Usage Patterns. In the proceedings of 3rd International Conference on Computer Research and Developments, Pages 490-494.
- [3]. Han, J., and Kamber, M., "Data Mining: Concepts and Techniques", A. Stephan. San Francisco, Morgan Kaufmann Publishers is an imprint of Elsevier, 2006.
- [4]. Rajni Pamnani and Pramila Chawan, "Web Usage Mining: A Research Area in Web Mining", Department of computer technology, VJTI University, Mumbai, 2013.
- [5]. Srivastava, J., Cooley, R., M. Deshpande, P. Tan, Web usage mining: Discovery and applications of usage patterns from web data' SIGKDD Explorations newsletter, 1(2), 2000, Pages 12-23.
- [6]. http://en.wikipedia.org/wiki/Web_mining.