# WEB MINING TECHNIQUES & TOOLS FOR INTERNET INFORMATION RESEARCH (WWW)

**Mohammad Sahid**

[1]Department of Computer science &Engineering, GH Raisoni College Engineering
Nagpur, India.

**Abstract:** ***Because of the tremendous the number of information sources available on the World Wide Web has increased dramatically in recent years., it has become world's the most comprehensive information source. The focus of such study would be on online usage mining. Exploration of web usage is divided into 3 stages: pattern, pre-processing finding, & pattern recognition. Each component would be described in detail, but the discovery and analysis of user navigation patterns will receive special focus.***

**Keywords:** KD (Discovering new information), WUM (web traffic analysis, web analysis), and (WM) we-blog mining are some of the terms used in this paper.

## 1. Introduction

Web data mining is a method that uses data mining techniques to analyse the features of the Web and web-based data in order to uncover the inherent linkages among the info on the internet that is expressed in the form of text form, linking, or details on use. We are primarily interested in determining how people use the Internet patterns using Web use analysis, followed by leveraging learned usage information to give Users who spend more time on the internet personalised internet part, that is internet suggestions [4].

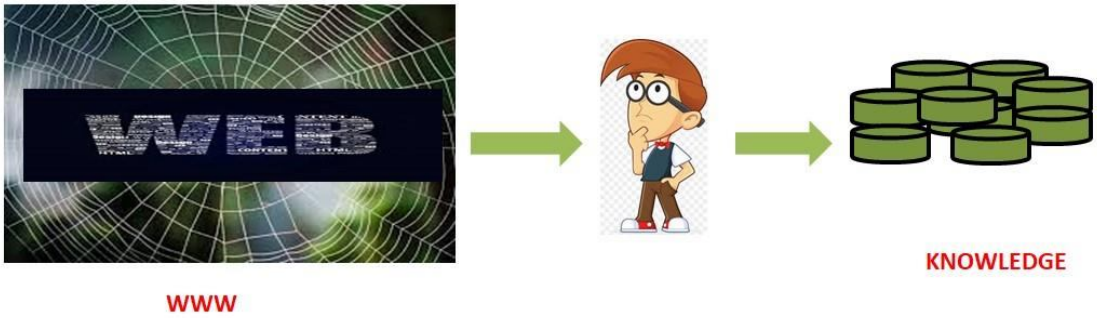

**Fig (a)** Web-based research

## 2. WWW Mining

The creation of an appropriate target data set for data mining operations is one of the most important processes in Knowledge Discovery in Databases. Data can be acquired at the server-side, client-side, proxy servers, or from an organization's database (which holds business information in Web Mining. data or Web data that has been aggregated).

Each method is kind of information collecting differs not just in terms of the location of the data source, but also in terms of the kind of data source in terms of types the amount of data accessible, the population the data was extracted from obtained, & the process used to acquire it. A wide variety of data types can be used in web mining. This study categorises the following types of data:

**2.1 Part:** The true data on Web pages, often known as web pages, are the documents that are shown on the internet information that the web page created to communicate with users. Text and visuals are commonly used, but are not restricted to.

**2.2 Way of Representation:** Data that defines how the content is organised. The placement of various HTML or XML tags within a given page is included in intra-page structure information.

**2.3 Blog on the Internet:** Server logs, error logs, and cookie logs are the three categories of log files[4]. The Extended Log File Format (ELF) is a more recent version of the Common Log File Format are used to store server logs. The following is the anatomy of a log file:

1. The IP address of your Internet service provider: This may be existing.
2. A dash, "-," is typically used to symbolise this.
3. AuthUser: An ID or password that is used to gain access to a secure place.
4. The Days are as follows: July 17, 1999, 12:38:09
5. Transaction: Typically, a "GET" filename, such as /index.html/products.htm, is used.
6. Normally, the transaction as follows
7. Transaction Extended Log Format
8. Referrer: the search engine and keyword that led people to your website, for example, http://search.yahoo.com/bin/search?p=data+mining/index.html.
9. Agent: Your visitor's browser, such as Mozilla/2.0 (Win95; I).
10. A cookie is a little text file that is stored on your computer. Snap.com is an example of a cookie. 00ed7085 946684799 youtube TRUE/FALSE 9. Agent: the browser your visitor is using,)
11. Is the tenth cookie. 946684799 u vid 0 0 00ed7085 TRUE / FALSE

Inaccuracy logs save information about unsuccessful requests including broken links, authentication difficulties, and timeout issues. Aside from detecting erroneous links or server capacity difficulties, which, when properly rectified, may be regarded an obligatory type of customer satisfaction, the utilisation of error logs can also be used to track out other issues. for the discovery of actionable marketing intelligence has thus far proven to be relatively restricted. Cookies are little text files that the web server creates and stores on the clients' computers.

Cookie log data helps to improve the transactionless state of web server interactions by allowing servers to trace client access throughout their hosted web pages. The layout and content of the marketing data, as well as the logged cookie data, are all adjustable. Query data to a web server is a fourth data stream that is commonly generated on e-commerce sites. Customers of an online store, for example, might look for products, while users of a research database might look for papers. Through cookie data and/or registration information

Although new specification ideas have reached draught status, there are presently no explicit draughts for query data handling standards. A stage of the Knowledge Discovery Process that is Internet-enabled, such as RDF is an abbreviation for Resource Description Framework. For query data to be usable, it must be organised into logical clusters, the majority of which are marketing-related[6].

**2.4 Information Resources:**

Usage data obtained from multiple sources will be used to depict the navigation patterns of various sectors of total Web traffic includes everything from single-user, single-site surfing to multi-user, multi-site access.

**2.5 Data Collection at WWW**

Because it specifically records the browsing behaviour of site users, A key source for Web Usage Mining is a Web server log. The information in server logs represents several users' (possibly simultaneous) access to a website. These log files are available in a number of different formats, including Common log and Extended log.

### 2.6 Collection at the Side of Client

Data collecting on the client side might be difficult accomplished through the use of a remote agent (for example, Javascripts or Java applets) or extending the data collection capabilities of an existing browser (such as Mosaic or Mozilla). The installation of client-side data collection techniques requires user participation, either by enabling the capability of Java applets and scripts, or by willingly using the browser that has been changed. The benefit of client-side collecting of handling cache and session identification concerns, whereas server-side collection does not. When it comes to determining the true view time of a page, Java applets are no better than server logs.

In reality, it's possible that it'll add to the overhead, especially when the Java applet is first launched. Java scripts, on the other hand, are simple to learn yet unable to catch all user clicks (such as reload or back buttons). Only single-user, single-site browsing behaviour will be collected using these methods. A customised browser is significantly more adaptable, allowing data about a single user to be collected across different Web sites. Convincing consumers to utilise the browser for their daily surfing activity is the most difficult component of employing this strategy. Which could be accomplished by compensating users who are willing to use the browser, in a manner similar to the incentive programmes offered by NetZero and All Advantage, which compensate customers using their products.

### 2.7 Proxy Level Collection

Between client browsers and Web servers, a Web proxy acts as a level of caching middleman. Proxy caching can be used to reduce the amount of time it takes for a user to load a Web page as well as the amount of network traffic on both the server and client sides. The ability of proxy caches to effectively estimate future page requests determines their performance. Proxy traces may show the true HTTP requests from several clients to multiple Web servers. This could be used to characterise the surfing habits of a group of anonymous users who utilise the same proxy server.

## 3. Research On The Use of The Internet

Web use mining is the automatic detection of user access patterns from Web servers.
Huge quantities Information are collected by organisations in their day-to-day operations[7], which are routinely created by Web servers and recorded in server access logs. Referrer logs, which contain information on the referring URLs for each page reference, and user registration or survey data gathered via CGI scripts are two more sorts of user information.

Analyzing such data can help firms determine, among other things, client lifetime value, cross-product marketing strategies, and the success of promotional initiatives.

It can also advise on how to redesign a website in order to create a more successful organisational presence, as well as throw light on how to handle workgroup communication and organisational infrastructure more effectively[8]. When selling advertisements on the World Wide Web, analysing user access patterns helps in personalising advertisements to certain groups of individuals.

There are three important duties for doing Web Usage Mining or Web Usage Analysis, as shown in Figure 2. This section is about outlines the tasks for each phase and handles any challenges that may arise[6].

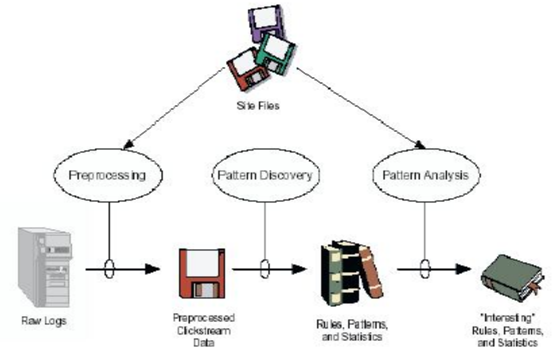
Site Files
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Logs
Preprocessed Clickstream Data
Rules, Patterns, and Statistics
"Interesting" Rules, Patterns, and Statistics

**Fig 2** Process of mining high-level web usages

### 3.1 Step of pre-processing

The inputs to the preparation phase are server logs, site files, and maybe use statistics from a previous study. The user session file, transaction file, site topology, and page classifications are all stored in the user session file are the outputs. Browser and proxy server caching is one of the primary hurdles to establishing a trustworthy user session file. Cookies and cache busting are two current approaches for collecting information about cached references. Cache busting is the process of stopping browsers from using locally saved versions of a website, necessitating a new download from the server each time the page is accessed. None of these approaches are without significant flaws.

The user can erase cookies, Cache busting undermines the speed gain that caching was intended to give, hence it is likely to be removed.

Customer registration is different way to determine who the users are Registration provides the benefit of allowing you to collect extra demographic data beyond what is already available routinely captured in the server log, as well as making user sessions easier to identify. However, many users opt not to visit sites that need registration and logins, or supply false information, due to privacy concerns. The WEBMINER system's preprocessing methods are all designed to work with only the information provided by CERN and NCSA's Common Log Format, which is HTTP protocol is a component of Hypertext transfer protocol.

### 3.2 Cleaning of data

Cleaning a server log to remove extraneous entries is critical not for any kind of Web log analysis only Mining web. Found relationships or provided meaningful accurately represents user accesses to the website. Every file requested from the Web server requires a separate connection, according to the HTTP protocol. Because pictures and scripts are downloaded in addition to the HTML file, a user's request to access a specific page frequently yields multiple data.

In most circumstances, only the hypertext transfer protocal requested file log entry necessary and ought to be saved in and the customer file. This is because, in most cases, a user does not expressly request all of the visuals on a Web page; instead, the HTML elements cause them to be automatically downloaded. It makes no sense to include file requests that the user did not expressly request because the goal The goal of Web Usage Mining is to create a picture of a user's behaviour. Checking the URL name's suffix is a sensible technique to filter out items that are deemed irrelevant. You can delete all log entries with filename suffixes like gif, jpeg, GIF, JPEG, jpg, JPG, and map, for example. You can delete all log entries with filename suffixes like gif, jpeg, GIF, JPEG, jpg, JPG, and map, for example. Moreover, typical scripts such as "count.cgi" can be removed. To remove files, the WEBMINER system uses a standard set of parameters suffixes. However, depending on the situation, sort of site, the list can be altered. being scrutinised For example, if a Web site provides a graphical archive, an analyst is needed. would presumably not want to delete all of the GIF or JPEG files from the server automatically. log. In this situation, graphic file log entries could very well indicate explicit user activities and should be kept for future research. To distinguish between

Publisher: Noida Institute of Engineering & Technology,
19, Knowledge Park-II, Institutional Area, Greater Noida (UP), India.

relevant and irrelevant log entries, a list of actual file names to remove or maintain might be utilised instead of just file suffixes.

### 4. Architecture For Usage of mining & Analysis

We've created a generic architecture for mining web traffic. The WEBMINER is a programme that implements several of the design's features. The technique for mining Web usage is divided into two parts by the design. The domain is introduced in the first part.-specific procedures for converting Web data into transaction-ready formats. This includes preprocessing, transaction identification, and data integration components. The second section of the system's data mining engine consists of the usually domain-independent usage of generic data mining and pattern matching techniques as part of the system's data mining engine (such as the discovery of association rules and sequential patterns). The overall structure of the Internet.

Figure 4 depicts the mining process. The first stage in the Web use mining process is to collect data is data. This is also a good place to do some low-level data integration activities. merging numerous logs, including referrer logs, and so on. Following the data, After cleaning, the log items must be partitioned into logical groups using one or more logical clustering algorithms. Modules for transaction identification The purpose of transaction identification is to construct a database of transactions. for each user, meaningful groupings of references One of the most difficult tasks is detecting transactions. splitting a huge transaction into several smaller transactions or combining tiny transactions into a smaller number of larger ones The input and output transaction formats are identical, allowing any number of transactions to be processed.modules that can be mixed and matched in any order,

The generated transaction data must be prepared to correspond to the data model of the appropriate data mining task when the domain-dependent data transformation phase is done. For example, the data format required for association rule discovery may differ from that required in order to mine sequential patterns Finally, a question method would enable the user to  set various constraints arrange for them additional control over the discovery process.

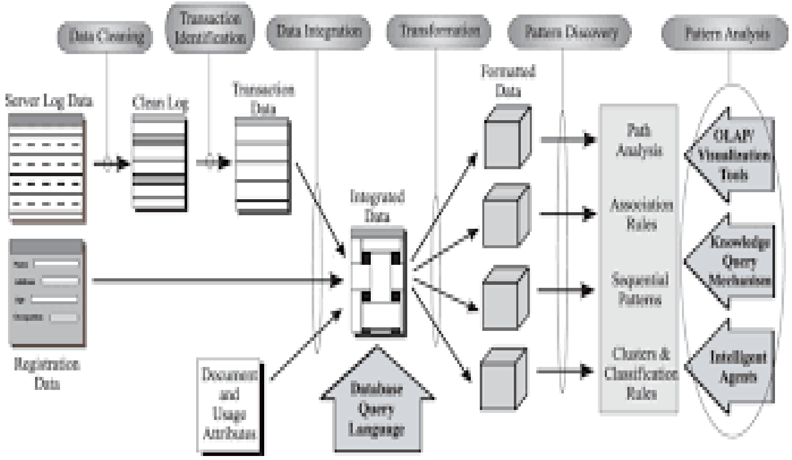

**Fig 4.** Web Usage Mining Architecture in General

## 5. Conclusion

Web mining is a word that has been used to refer to a variety of strategies that deal with a wide range of challenges. However, while this broadness is appealing, it has resulted in the rise of Web mining imply many things to various people individuals, necessitating the development of a standard lexicon. We established a definition of Web mining and developed a taxonomy of the different active projects linked to it in order to achieve this goal. We presented a broad architecture for a Web usage mining system and outlined concerns and problems that require future research and development in this area.

## Reference

[1]. RAYMOND KOSALA, HENDRIK BLOCKEEL, Web Mining Research: A Survey, Sigkdd Expirations, Acm Sigkdd, July 2000.

[2]. M. KOSHER. ALIKE - Archie-Like Indexing In The Web. In Proc. 1st International Conference On The World Wide Web, Pages 91--100, May 1994.

[3]. R. COOLEY, B. MOBASHER, AND J. SRIVASTAVA. Web Mining: Information And Pattern Discovery On The World WideWeb. In Proceedings Of The 9th Ieee International Conference On Tools With Artificial Intelligence (Ictai'97), 1997

[4]. R. KOSALA, H. BLOCKEEL. Web Mining Research: A Survey Data & Knowledge Engineering, Volume 53, Issue 3, June 2005, Pages 225-241

[5]. NASRAOUI, O. ET AL. , A Web Usage Mining Framework For Mining Evolving User Profiles In Dynamic Web Sites, Ieee Transactions On Knowledge And Data Engineering, Volume: 20 Issue:2 On Page(S): 202 – 215, 2008.

[6]. F. MASSEGLIA, ET AL. Web Usage Mining: Extracting Unexpected Periods From Web Logs, Data Mining And Knowledge Discovery Volume 16, Number 1, 39-65, 2007.

[7]. NAVEENA DEVI ET AL. Design And Implementation Of Web Usage Mining Intelligent System In The Field Of ECommerce, Procedia Engineering Volume 30, 2012, Elsevier , Pp 20–27

[8]. MALIK, S.K. ET AL., Information Extraction Using Web Usage Mining, Web Scrapping And Semantic Annotation, In Procd. Of Ieee Cicn, 2011 Pp-465 – 469

[9] Dr. Mohammad Shahid "KNOWLEDGE DISCOVERY ON THE INTERNET (WEB MINING TOOLAND TECHNIQUE" INDIAN JOURNAL OF RESEARCH(2012)6, ANVIKSHIKI ISSN 0973-9777 Advance Access publication 20 July. 2012

[10] Dr. Mohammad Shahid" Taxonomies, Challenge And Approaches To Automotive Web Query Classification "2 ND INTERNATIONAL COFFERENCE ON COMPUTER APPLICATION 2012 ICCA'12 PONDICHERRY, INDIA.