# A Review Paper on Web Usage Mining and Pattern Discovery

[1]RACHIT ADHVARYU

[1]*Student M.E CSE, B. H. Gardi Vidyapith, Rajkot, Gujarat, India.*

**ABSTRACT: -** *Web Technology is evolving very fast and Internet Users are growing much faster than estimated. The website users are using a wide range of websites leaving back a variety of information. This information must be used by the websites administrator to manipulate their websites according to the users of the websites. This actually is Data Mining. Thus the following paper focuses on the concept of 1) Web Usage Mining which is a mining of usage of websites and the information used and delivered on the websites. This also specifies the importance of Web Usage Mining. 2) Brief Details on the Pattern Matching and other Different Functionalities of Data Mining used in Web Usage Mining***.**

**Keywords : Web, Mining, Pattern**

## INTRODUCTION

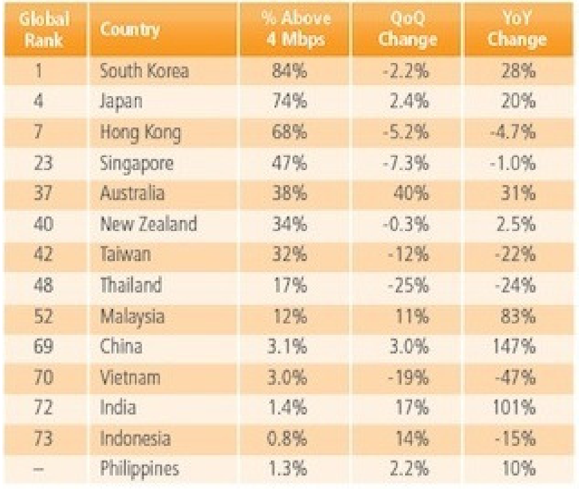
Web technology is not evolving in comfortable and incremental steps, but it is turbulent, erratic, and often rather uncomfortable. It is estimated that the Internet, arguably the most important part of the new technological environment, has expanded by about 2000 % and that is doubling in size every six to ten months. In recent years, the advance in computer and web technologies and the decrease in their cost have expanded the means available to collect and store data. As an intermediate consequence, the amount of information (Meaningful data) stored has been increasing at a very fast pace. Traditional information analysis techniques are useful to create informative reports from data and to confirm predefined hypothesis about the data. It is reasonable to believe that data collected over an extended period contains hidden knowledge about the business or patterns characterizing customer profile and behavior. With the rapid growth of the World Wide Web, the study of knowledge discovery in web, modeling and predicting the user's access on a web site has become very important.

From the administration, business and application point of view, knowledge obtained from the Web usage patterns could be directly applied to efficiently manage activities related to e-Business, e-CRM, e-Services, e-Education, e-Newspapers, e-Government, Digital Libraries etc. Web is becoming the necessity of the businesses and organizations because of its demand from the clients. With the large number of companies using the Internet to distribute and collect information, knowledge discovery on the web has become an important research area. With the explosive growth of information sources available on the World Wide Web, it has become necessary for organizations to discover the usage patterns and analyze the discovered patterns to gain an edge over competitors.

Many approaches have been proposed for analyzing the visitor click stream sequences. Some researchers proposed the web personalization system, which consists of offline tasks related to the mining if usage data and online process of automatic Web page customization based on the knowledge discovered. Some used relational online analytical processing approach for creating a Web log warehouse using access logs and mined logs. Web mining can be divided into three areas, namely web content mining, web structure mining and web usage mining.

**Global Internet Usage Average Usage shows the current usage around the globe and in INDIA.**

Number of internet users in India is growing by 150K every month or 1.8 Million new users every year. India is the fastest growing online market in the world with 75% of the users being below the age of 35. Unfortunately the sex ratio (like the population) is very skewed with only 39% of the users being women.



| Global Rank | Country | % Above 4 Mbps | QoQ Change | YoY Change |
|---|---|---|---|---|
| 1 | South Korea | 84% | -2.2% | 28% |
| 4 | Japan | 74% | 2.4% | 20% |
| 7 | Hong Kong | 68% | -5.2% | -4.7% |
| 23 | Singapore | 47% | -7.3% | -1.0% |
| 37 | Australia | 38% | 40% | 31% |
| 40 | New Zealand | 34% | -0.3% | 2.5% |
| 42 | Taiwan | 32% | -12% | -22% |
| 48 | Thailand | 17% | -25% | -24% |
| 52 | Malaysia | 12% | 11% | 83% |
| 69 | China | 3.1% | 3.0% | 147% |
| 70 | Vietnam | 3.0% | -19% | -47% |
| 72 | India | 1.4% | 17% | 101% |
| 73 | Indonesia | 0.8% | 14% | -15% |
| – | Philippines | 1.3% | 2.2% | 10% |

Broadband (>4 Mbps) Connectivity, Asia Pacific Countries

## APPLICATION OF WEB USAGE MINING

Each of the applications can benefit from patterns that are ranked by subjective interesting.

Web usage mining is used in the following areas:

- Web usage mining offers users the ability to analyze massive volumes of click stream or click flow data, integrate the data seamlessly with transaction and demographic data from offline sources and apply sophisticated analytics for web personalization, e-CRM and other interactive marketing programs.
- Personalization for a user can be achieved by keeping track of previously accessed pages. These pages can be used to identify the typical browsing behavior of a user and subsequently to predict desired pages.
- By determining frequent access behavior for users, needed links can be identified to improve the overall performance of future accesses.
- In addition to modifications to the linkage structure, identifying common access behaviors can be used to improve the actual design of Web pages and to make other modifications to the site.
- Web usage patterns can be used to gather business intelligence to improve Customer attraction, Customer retention, sales, marketing and advertisement, cross sales.
- Mining of web usage patterns can help in the study of how browsers are used and the user’s interaction with a browser interface.

## WEB USAGE AND PATTERN DISCOVERY

Web usage mining is the application of data mining techniques to discover usage pattern from Web data, in order to understand and better serve the needs of Web-based applications. Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis. A high level Web usage mining Process is presented in Figure 1. It proposes that the web mining process can be divided into two main parts. The first part includes the domain dependent processes of transforming the Web data into suitable transaction form. This includes preprocessing, transaction identification, and data integration components. The second part includes some data mining and pattern matching techniques such as association rule and sequential patterns.

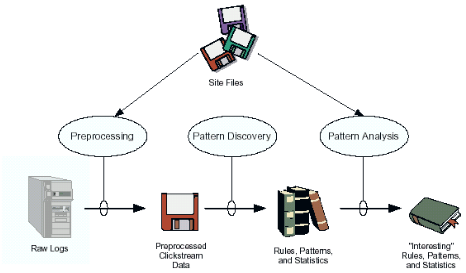

Site Files
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Logs
Preprocessed Clickstream Data
Rules, Patterns, and Statistics
"Interesting" Rules, Patterns, and Statistics

Figure 1: High Level *Web Usage Mining* Process

The first is preprocessing state in which user sessions are inferred from log data. The second searches for patterns in the data by making use of standard data mining techniques, such as association rules or mining for sequential patterns. In the third stage an information filter bases on domain knowledge and the web site structures is applied to the mining patterns in search for the interesting patterns.

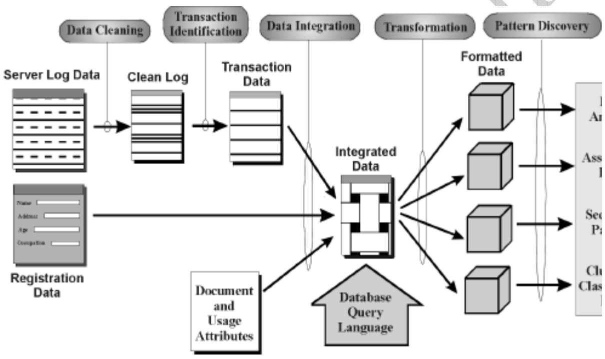


**Figure2: Architecture for Web Usage Mining**

In this case, episodes are either all of the page views in a server sessions that the user spent a significant amount of time viewing, or all of the navigation page views leading up to each content page view. The click-stream or click-flow for each user is divided into sessions based on a simple thirty-minute timeout. Four dimensions used to classify interestingness measures are pattern-form, representation, scope, and class. Pattern-form defines what type of patterns a measure is applicable to, such as association rules or classification rules.

Data preprocessing consists of data filtering, user identification, session/transaction identification, and topology extraction. Data filtering filters out some noise, i.e., unsuccessful requests, automatically downloaded graphics, or requests from robots, to get more compact training data. Now people use some heuristic rules to identify user, such as IP address, cookies, etc. Preprocessing consists of converting the usage, content, and structure information contained in the various available data sources into the data abstractions necessary for pattern discovery.

**Usage preprocessing**: Usage preprocessing consists of Web pages, such as IP addresses, page references, and the date and time of accesses. Typically, the usage data comes from an Extended Common Log Format (ECLF) Server log.

**Content Preprocessing:** Content preprocessing consists of converting the text, images, scripts, and multimedia data into forms that are useful for the web usage mining process. Often this consists of performing content mining such as classification or clustering. In the context of web usage mining, the content of Web sites can be used to filter the input to the pattern discovery algorithms.

**Structure Preprocessing:** Web structure mining analyses the link structure of the web in order to identify relevant documents. The structure of a site is created by the hypertext links between page views. Intra-page structure information includes the arrangement of various HTML or XML tags within a given page. The principal kind of inter-page structure information is hyper-links connecting one page to another. The Google Search engine makes use of the web link structure in the process of determining the relevance of a page. The Google search engine achieves good results because while the keyword similarity analysis ensures high precision the use of a probability measure ensures high quality of the pages returned.

The usage data collected at the different sources such as Server level, Client Level and Proxy Level represent the navigation patterns of different segments of the overall Web traffic.

**Server-level Collection:** A Web server log records the browsing behavior of site visitors. The data recorded in server logs reflect the concurrent and interleaved access of a Web site by multiple users. These log files can be stored in various formats such as Common Log Format (CLF) or Extended Common Log Format (ECLF). ECLF contains client IP address, User ID, time/date, request, status, bytes, referrer, and agent. Tracking of individual users is not an easy task due to the stateless connection model of the HTTP protocol. In order to handle this problem, Web servers can also store other kind of usage information such as cookies in separate logs, or appended to the CLF or ECLF logs. Cookies are tokens generated by the Web server for individual client browsers in order to automatically track the site visitors. Packet sniffing technology (also referred to as "network monitors") is an alternative method for collecting usage data through server logs.

**Client level collection:** Client-side collection can be implemented by using a remote agent (such as Java scripts or Java applets) or by modifying the source code of an existing browser (such as Mosaic or Mozilla) to enhance its data collection capabilities.

**Proxy Level Collection:** The Internet Service Provider (ISP) machine that users connect to through a model is a common form of proxy server. A web proxy acts as an intermediary between client browsers and Web servers. Proxy-level caching can be used to reduce the loading of time of a Web page experienced by users as well as the network traffic load at the server and client sides.

**Pattern Discovery:** Pattern discovery uses methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition. The knowledge that can be discovered is represented in the form of rules, tables, charts, graphs, and other visual presentation forms for characterizing, comparing, predicting, or classifying data from the web access log. Visualization can also be used in web usage mining, and it presents the data in the way that can be understood by users more easily.

**Statistical Analysis:** Statistical techniques are the most common method to extract knowledge about visitors to a web site. By analyzing the session file, one can perform different kinds of descriptive statistical analyses (frequency, mean, median, etc) on variables such as page views, viewing time and length of a navigational path. For example e-Trade developed a website in German language for Germany and scrapped it because German people were visiting the English site rather than the German site. Many web traffic analysis tools produce a periodic report containing statistical information such as the most frequently accessed pages, average view time of a page or average length of a path through a site. This type of knowledge can be potentially useful for improving the system performance, enhancing the security of the system, facilitating the site modification task, and providing support for marketing decisions.

**Association Rules:** Association rule generation can be used to relate pages that are most often referenced together in a single server sessions. In the context of web usage mining, association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. Association rule mining has been well studied in Data Mining, especially for basket transaction data analysis. Aside from being applicable for e-Commerce, business intelligence and marketing applications, it can help web designers to restructure their web site. The association rules may also serve as heuristic for pre fetching documents in order to reduce user-perceived latency when loading a page from a remote site.

**Clustering:** Clustering is a technique to group together a set of items having similar characteristics. Clustering can be performed on either the users or the page views. Clustering analysis in web usage mining intends to find the cluster of user, page, or sessions from web log file, where each cluster represents a group of objects with common interesting or characteristic. User clustering is designed to find user groups that have common interests based on their behaviors, and it is critical for user community construction. Page clustering is the process of clustering pages according to the users' access over them. clustering of pages will discover groups of pages having related content. This information is useful for the Internet search engines and Web assistance providers.

**Sequential Patterns:** The technique of sequential pattern discovery attempts to find inter-session patterns such that the presence of a set of items is followed by another item in a time-ordered set of sessions or episodes. A new algorithm MiDAS (Mining Internet data for Associative Sequences) for discovering sequential patterns from web log files has been proposed that provides behavioral marketing intelligence for e-commerce scenarios. MiDAS contains three phases: 1. A priori phase is the input data preparation, which consists of data reduction and data type substitution. 2. Discovery Phase discovers the sequences of hits and generates the pattern tree. 3. A posteriori Phase filters out all sequences that do not fulfill the criteria laid in the specified navigation templates and topology network and also pruning is done in this phase. By using this approach, Web marketers can predict future visit patterns, which will be helpful in placing advertisements aimed at certain user groups.

**Dependency modeling:** Dependency modeling is another useful pattern discovery task in web mining. The goal here is to develop a model capable of representing significant dependencies among the various variables in the web domain. As an example, one may be interested to build a model representing the different stages a visitor undergoes while shopping in an online store based on the actions chosen (ie, from a casual visitor to a serious potential buyer. There are several probabilistic learning techniques that can be employed to model the browsing behavior of users. Modeling of Web usage patterns will not only provide a theoretical framework for analyzing the behavior of users but is potentially useful for predicting future Web resource consumption.

**Deviation/Outlier Detection:** It contains techniques aimed at detecting unusual changes in the data relatively to the expected values. Such techniques are useful, for example, in fraud detection, where the inconsistent use of credit cards can identify situations where a card is stolen. The inconsistent use of credit card could be noted if there were transactions performed in different geographic locations within a given time window.

**Pattern analysis:** Pattern analysis is the last step in the overall Web Usage mining process as described in Figure 2. The motivation behind pattern analysis is to filter out uninteresting rules or patterns from the set found in the pattern discovery phase. The exact analysis methodology is usually governed by the application for which Web mining is done. The most common form of pattern analysis consists of a knowledge query mechanism such as SQL. Another method is to load usage data into a data cube in order to perform OLAP operations. Visualization techniques, such as graphing patterns or assigning colors to different values, can often highlight overall patterns or trends in the data. Content and structure information can be used to filter out patterns containing pages of a certain usage type, content type, or pages that match a certain hyperlink structure.

## CONCLUSION

This paper has attempted to provide an up-to-date survey of the rapidly growing area of Web usage mining, which is the demand of current technology. In this paper a general overview of Web usage mining is presented in introduction section. Web usage mining is used & more research must be made in many areas such as e-Business, e-CRM, e-Services, e-Education, e-Newspapers, e-Government, Digital Libraries, advertising, marketing, bioinformatics and so on. The main techniques for pattern discovery are sequential patterns, association rules, Classification, Clustering, and path analysis. We need a systematic web-site design methodology to create new web pages, or modify existing web pages, such that different user's navigation patterns could be better mapped to answers to a set of specific questions. There is a need to develop tools, which incorporate statistical methods, visualization, and human factors to help better understand the mined knowledge. One of the open issues in data mining is the creation of intelligent tools that can assist in the interpretation of mined knowledge. These tools need to have specific knowledge about the particular problem domain to do any more than filtering based on statistical attributes of the discovered rules or patterns.

## BIBLIOGRAPHY

[1] J.W. Han, M .Kamber, Data Mining--Concepts and Techniques, Elsevier Science & Technology Books, 2006.

[2] J. Srivastava, R. Cooley, M. Deshpande, P. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations, 1(2), 20001, pp.2-23.

[3] Jaideep Srivastava and Robert Cooley ˈUsage Mining: Discovery and Applications of Usage Patterns from Web Data. ACM SIGKDD Explorations Newsletter .2000,1(2):12-23.

[4] Bamshad Mobasher .Web Usage Mining .Springer Berlin Heidelberg, 2007, 449-483.

[5] Navin Kumar Tyagi, A.K. Solanki and Sanjay Tyagi: "An Algorithmic Approach to Data Preprocessing in Web Usage Mining". International Journal of Information Technology and Knowledge Management, Volume 2, No. 2, July-December 2010, pp. 279-283.

[6] José Roberto de Freitas Boullosa. "An Architecture for Web Usage Mining".

[7] Yan Wang." Web Mining and Knowledge Discovery of Usage Patterns". CS 748T Project. February, 2000.

[8] Cyrus Shahabi, Amir M. Zarkesh, Jafar Adibi, and Vishal Shah, Knowledge Discovery from Users Web-page Navigation, IEEE RIDE 1997.