Web Usage Mining

Web usage mining, a subset of Data Mining, is basically the extraction of various types of interesting data that is readily available and accessible in the ocean of huge web pages, Internet- or formally known as World Wide Web (WWW). Being one of the applications of data mining technique, it has helped to analyze user activities on different web pages and track them over a period of time. Basically, Web Usage Mining can be divided into 2 major subcategories based on web usage data.

Web Usage Data: The main source of data here is-Web Server and Application Server. It involves log data which is collected by the main above two mentioned sources. Log files are created when a user/customer interacts with a web page. The data in this type can be mainly categorized into three types based on the source it comes from:
- Server-side
- Client-side
- Proxy side.

There are other additional data sources also which include cookies, demographics, etc.

Types of Web Usage Mining based upon the Usage Data:

1. Web Server Data: The web server data generally includes the IP address, browser logs, proxy server logs, user profiles, etc. The user logs are being collected by the web server data.

2. Application Server Data: An added feature on the commercial application servers is to build applications on it. Tracking various business events and logging them into application server logs is mainly what application server data consists of.

3. Application-level data: There are various new kinds of events that can be there in an application. The logging feature enabled in them helps us get the past record of the events.

Advantages of Web Usage Mining

- Government agencies are benefited from this technology to overcome terrorism.
- Predictive capabilities of mining tools have helped identify various criminal activities.
- Customer Relationship is being better understood by the company with the aid of these mining tools. It helps them to satisfy the needs of the customer faster and efficiently.

Disadvantages of Web Usage Mining

- Privacy stands out as a major issue. Analyzing data for the benefit of customers is good. But using the same data for something else can be dangerous. Using it within the individual's knowledge can pose a big threat to the company.
- Having no high ethical standards in a data mining company, two or more attributes can be combined to get some personal information of the user which again is not respectable.

Techniques for Web Usage Mining

1. Association Rules: The most used technique in Web usage mining is Association Rules. Basically, this technique focuses on relations among the web pages that frequently appear together in users' sessions. The pages accessed together are always put together into a single server session. Association Rules help in the reconstruction of websites using the access logs. Access logs generally contain information about requests which are approaching the webserver. The major drawback of this technique is that having so many sets of rules produced together may result in some of the rules being completely inconsequential. They may not be used for future use too.

2. Classification: Classification is mainly to map a particular record to multiple predefined classes. The main target here in web usage mining is to develop that kind of profile of users/customers that are associated with a particular class/category. For this exact thing, one requires to extract the best features that will be best suitable for the associated class. Classification can be implemented by various algorithms – some of them include- Logistic Regression, Decision Trees, etc. For example, having a track record of data of customers regarding their purchase history in the last 6 months the customer can be classified into frequent and non-frequent classes/categories. There can be multiclass classification possible too.

3. Clustering: Clustering is a technique to group together a set of things having similar features/traits. There are mainly 2 types of clusters- the first one is the usage cluster and the second one is the page cluster. The clustering of pages can be readily performed based on the usage data. In usage-based clustering, items that are commonly accessed /purchased together can be automatically organized into groups. The clustering of users tends to establish groups of users exhibiting similar browsing patterns. In page clustering, the basic concept is to get information quickly over the web pages.

Applications of Web Usage Mining

Personalization of Web Content: The World Wide Web has a lot of information and is expanding very rapidly day by day. The big problem is that on an everyday basis the specific needs of people are increasing and they quite often don't get that query result. So, a solution to this is web personalization. Web personalization may be defined as catering to the user's need-based upon its navigational behavior tracking and their interests. Web Personalization includes recommender systems, check-box customization, etc. Recommender systems are popular and are used by many companies.

E-commerce: Web-usage Mining plays a very vital role in web-based companies. Since their ultimate focus is on Customer attraction, customer retention, cross-sales, etc. To build a strong relationship with the customer it is very necessary for the web-based company to rely on web usage mining where they can get a lot of insights about customer's interests. Also, it tells the company about improving its web-design in some aspects.

Prefetching and Catching: Prefetching basically means loading of data before it is required to decrease the time waiting for that data hence the term 'prefetch'. All the results which we get from web usage mining can be used to produce prefetching and caching strategies which in turn can highly reduce the server response time.