WEB MINING

Web is a collection of inter-related files on one or more Web servers. Web mining is the application of data mining techniques to extract knowledge from Web data. Web data can be considered as

Web content - text, image, records, etc.

Web structure – hyperlinks, tags, etc.

Web usage – http logs, app server logs, etc.

Based on this web mining can involve mining data from contents, structure as well as web usage patterns.

Web Content Mining - Extract "snippets" from a Web document

Web Structure Mining - Identifying interesting graph patterns or preprocess the entire web graph to come up with metrics such as PageRank.

Web Usage - User identification, session creation, robot detection and filtering, and extracting usage path patterns.

WEB CONTENT MININIG

Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. Research activities in this field also involve using techniques such as Information Retrieval (IR) and natural language processing (NLP).

The more basic and popular data mining techniques include:

Classification

Clustering

Associations

Significant usage if Web Content Mining are : Document Classification and Clustering, Topic Identification, Tracking and drift analysis Concept hierarchy creation Relevance of content.

Applications of Web Content Mining

Document Classification :

It is a "Supervised" technique . In this technique Categories of the documents are identified and documents are assigned to one or more existing categories. The "definition" of a category is usually in the form of a term vector that is produced during a "training" phase. Training is performed through the use of documents that have already been classified (often by hand) as belonging to a category.

Document Clustering :

It is a "Unsupervised" technique. In this Documents are divided into groups based on a similarity metric. No pre-defined notion of what the groups should be is known. Most common similarity metric is the dot product between two document vectors.

Topic Identification and Tracking:

It is a Combination of Clustering and Classification. As new documents are added to a collection an attempt is made to assign each document to an existing topic (category). The collection is also checked for the emergence of new topics.

Concept Hierarchy Creation:

Creation of concept hierarchies is important to understand the category and sub categories a document belongs to.

Key Factors to be considered while generating concept hierarchy are –

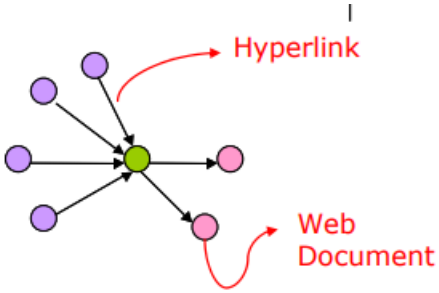
Organization of categories; e.g. Flat, Tree, or Network

Maximum number of categories per document.

Category Dimensions; e.g. Subject, Location, Time, Alphabetical, Numerical

WEB STRUCTURE MINING

The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages



Web Graph Structure

Web Structure Mining is a process of discovering structure information from the Web This type of mining can be performed either at the (intra-page) document level or at the (inter-page) hyperlink level. The research at the hyperlink level is also called Hyperlink Analysis.

Hyperlinks serve two main purposes - Pure Navigation and Points to the pages with authority* on the same topic of the page containing the link. This can be used to retrieve useful information from the web.

Web Structure Terminology –

Web-graph: A directed graph that represents the Web.

Node: Each Web page is a node of the Web-graph.

Link: Each hyperlink on the Web is a directed edge of the Web-graph.

In-degree: The in-degree of a node, p, is the number of distinct links that point to p.

Out-degree: The out-degree of a node, p, is the number of distinct links originating at p that point to other nodes.

Directed Path: A sequence of links, starting from p that can be followed to reach q.

Shortest Path: Of all the paths between nodes p and q, which has the shortest length, i.e. number of links on it.

Diameter: The maximum of all the shortest paths between a pair of nodes p and q, for all pairs of nodes p and q in the Web-graph.

Web Structure Mining is helpful in improving the performance of search engines.

Eg: Page Rank and HITS techniques.

Applications of Web Structure Mining

Web Structure is a useful source for extracting information such as

Quality of Web Page - The authority of a page on a topic - Ranking of web pages

Interesting Web Structures - Graph patterns like Co-citation, Social choice, Complete bipartite graphs, etc.

Web Page Classification - Classifying web pages according to various topics

Which pages to crawl - Deciding which web pages to add to the collection of web pages

Finding Related Pages - Given one relevant page, find all related pages

Detection of duplicated pages - Detection of neared-mirror sites to eliminate duplication

WEB USAGE MINING

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

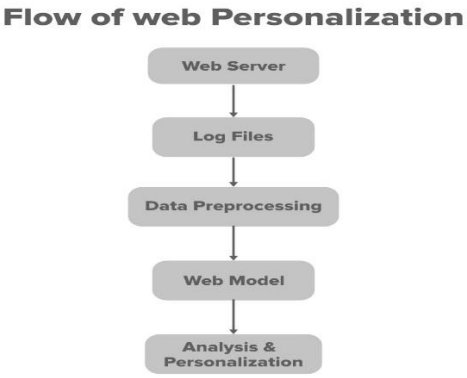
Privacy stands out as a major issue. Analyzing data for the benefit of customers is good. But using the same data for something else can be dangerous. Using it within the individual's knowledge can pose a big threat to the company.

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.

Applications of Web Usage Mining

1. **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.

Flow-of-Web-Personalization



2. **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.

3. **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.