Overview of Web Mining Techniques and its Application towards Web

***Prof.Pooja Mehta**

Abstract

The World Wide Web (WWW) acts as an interactive and popular way to transfer information. Due to the enormous and diverse information on the web, the users cannot make use of the information very effectively and easily. Data mining concentrates on non trivial extraction of implicit previously unknown and potential useful information from the very large amount of data. Web mining is an application of data mining techniques to extract knowledge from web data, including web documents, hyperlinks between documents, usage logs of web sites, etc. The aim of this paper is to describe the past and current techniques in Web Mining. It also gives the overview of development in research of web mining and few key computer science contributions in the field of web mining, the prominent successful applications and outlines some promising areas of future research. Online resources for retrieval Information on (1) web content mining, extraction and integration of useful data, information and knowledge from Web page content (2) web structure mining, analyze the node and connection structure of a web site and (3) web usage mining, process of extracting useful information from server logs for web. Web mining provides the support for the web site design, providing personalization server and other business making decision etc. It has quickly become one of the most important areas in Computer and Information Sciences because of its direct applications in e-commerce, CRM, Web Analytics, Information Retrieval and Filtering and Web Information System. Web Data mining has immensely penetrated in each and every field of day to day life.

Keywords: World Wide Web, Web Mining, Data mining, Web usage mining, Web structure mining, Web content mining

Introduction

Data Mining

Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. It is the task of discovering interesting patterns from large amounts of data, where the data can be stored in databases, data ware houses, or repositories [1].It involves an integration of techniques from multiple disciplines such as database and data warehouse, statistics, machine learning, high-performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial or temporal data analysis.

Web Mining

Web mining can be broadly defined as the discovery and analysis of useful information from the WWW. Data can be collected at the server side, client side, proxy servers, or obtained from an organizations database. Web mining is the application of data mining. It involves mining logs and the steps that typically have to be gone through to get meaningful data from Web logs - data collection, pre-processing, data enrichment and pattern analysis and discovery. Web-mining is a multi-disciplinary effort that draws techniques from fields like information retrieval, statistics, machine learning, natural language processing. Using Data Mining parameters such as clustering, classification, association, and examination of sequential patterns we evaluate web data. Web mining is most popular among research today. This paper presents web mining concepts, techniques and application

*Assistant Professor, MCA Department,Y.M.T College of Management Email id:poojakm@gmail.com
Contact No:9004380734

associated with different fields like social media, electronic commerce, cloud computing, electronic business.

Comparison of web mining with data mining [3]

	Scale	Structure	Access
Web Mining	Search processing is not a big, 10 million job in web server database	Gets information from structured, unstructured and semi structured from web pages. It fetches information from web.	Access of data publicly.
Data Mining	Search processing is large, a 1 million jobs in data base	Gets information from explicit structure. It does not fetch the information from wide database compares to web mining database.	Access of data privately and authorize user.

2. Web mining task

Web mining can be viewed as the use of data mining techniques to automatically retrieve, extract and evaluate information for knowledge discovery from web documents and services. Web mining is categorized into web content mining, web structure mining and web usage mining. Web content mining studies the search and retrieval of information on the web. Web structure mining focuses on the structure of the hyperlinks within a web. Web usage mining discovers and analyzes user access patterns. According to Etzioni web mining can be divided into four tasks:

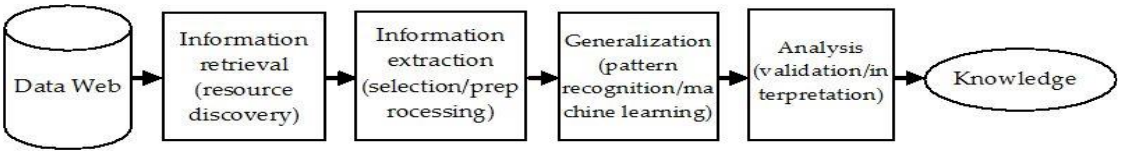

Figure 1. Web Mining Task [4]

1. **Information Retrieval (IR) (Resource Discovery):** It deals with automatic retrieval of all relevant documents. The IR process mainly includes document representation, indexing, and searching for documents.

2. **Information Selection/Extraction (IE) and Preprocessing:** Once the documents have been retrieved, the challenge is to automatically extract knowledge and other required information without human interaction. Information extraction is the task of identifying specific fragments of a single document that constitute its core semantic content.

3. **Generalization:** In this phase, pattern recognition and machine learning techniques are used on the extracted information. A major obstacle when learning about the web is the labeling problem: data is abundant on the web, but it is unlabelled. Many data mining techniques require inputs labeled as positive (yes) or negative (no) examples with respect to some concept.

4. **Analysis:** Analysis is a data-driven problem which presumes that there is sufficient data available so that potentially useful information can be extracted and analyzed. Humans play an important role in the information or knowledge discovery process (KDD) on the web since the web is an interactive medium. Important for validation or interpretation of the patterns. Once the patterns have been discovered, analysts need appropriate tools to understand, visualize, and interpret these patterns.

Category of Web Mining

Two different approaches were taken in defining Web mining. First was a "process-centric view" that defined Web Mining as a sequence of tasks and the second, a "data-centric view" that defined Web Mining in terms of the type of Web data that is used in the mining process [5].Based on the primary kinds of data used in the mining process, web mining tasks can be categorized into three: Web Structure Mining, Web Content Mining and Web Usage Mining.

Web Content Mining

Web Content Mining is the process of extracting useful information from the contents of Web documents. It may consist of text, images, audio, video, or structured records such as lists and tables. Content mining is the scanning and mining of text, pictures and graphs of a Web page to determine the relevance of the content to the search query [5]. This scanning is completed after the clustering of web pages through structure mining and provides the results based upon the level of relevance to the suggested query.

Multimedia data mining is part of the content mining, which is engaged to mine the high-level information and knowledge from large online multimedia sources. Multimedia data mining on the Web has gained many researchers attention recently. Working towards a unifying framework for representation, problem solving, and learning from multimedia is really a challenge, this research area is still in its infancy indeed, many works are waiting to be done.

Web Structure Mining

The goal of Web structure mining is to generate structural summary about the Web site and Web page. Technically, Web content mining mainly focuses on the structure of inner-document, while Web structure mining tries to discover the link structure of the hyperlinks at the inter-document level. Based on the topology of the hyperlinks, Web structure mining categorizes the web pages and generates information, such as similarity and relationship between different Web sites. Web structure mining can also have another direction – discovering the structure of Web document itself. This type of structure mining can be used to reveal the structure of Web pages, which will in turn help in easy navigation through the pages, and also facilitates comparison and integration of web page schemes. This can be further divided into two kinds based on the kind of structure information used.

Hyperlinks: A hyperlink is a structural unit that connects a location in a web page to different location, either within the same or on a different Web page. It connects to a different part of the same page is called an Intra Document Hyperlink, and a hyperlink that connects two different pages is called an Inter-Document Hyperlink.

Document Structure: The contents within a web page can be organized in a tree-structured format, based on the various HTML and XML tags within the page [5].

Web Usage Mining

Web usage mining is the process of extracting useful information from server logs i.e. user's history. Web usage mining process involves the log time of pages. The world's largest portal like google, yahoo, msn etc., needs a lot of insights from the behavior of their users' web visits. Without this usage reports, it will be difficult to structure their monetization efforts.

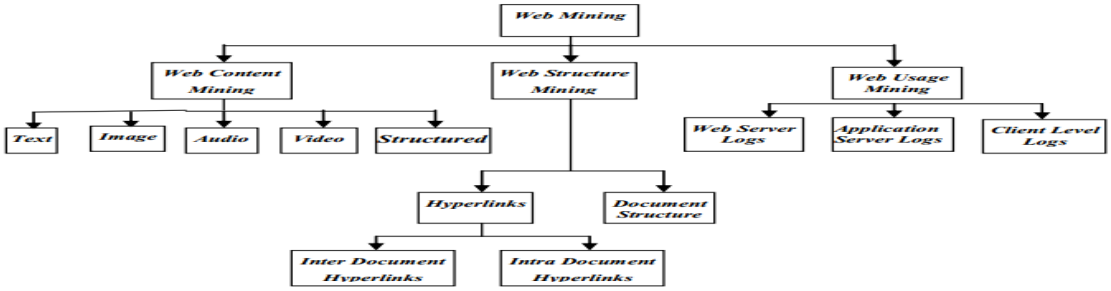

Figure 2 Web Mining classifications

Usage mining has direct impact on businesses [6] .This is the activity that involves the automatic discovery of user access patterns from one or more Web servers.

Web Server Data

User logs are collected by the web server and include IP address, page reference and access time.

Application Server Data

Commercial application servers such as Web logic, Story Server, have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

Application Level Data

New kinds of events can be defined in an application, and logging can be turned on for them generating histories of these events. Many end applications require a combination of one or more of the techniques applied in the above the categories [6].

Web Mining Process

Web mining process divided in four steps:

Source Data Collection

Source data collection includes web log files, records all the behavior of user connected with web

Data Preprocessing

Data collected from web are incomplete, redundant, erroneous and ambiguous. Using Preprocessing we accurate and precise data. Using this technique we clean server log files from irrelevant data. It includes following processes.

- Data Cleaning: removes web log redundant data.
- User Identification: identify user uniquely on web server.
- User Session Identification: basis on user identification, divide user information into separate session.
- Access path Estimation: Path is recorded by access log files on web server.

Pattern Discovery

It is used to find patterns using technique like

- Path Analysis: physical layout of website presented in graphical form.

- Association Rule: focused based on discovery of relation between pages visited by user on website.
- Classification: mapping of a data into one or several predefined data or items using algorithm.
- Clustering: similar attributes or characteristics are grouped together. Using it we develop future marketing strategies.

Pattern Analysis

Main purpose is to find out valuable model using different techniques like

- OLAP (online analytical processing): analysis of databases.
- Visualization: understanding behaviour of web users using graph.
- Data knowledge query: focus on proper analysis of user needs or problems.
- Usability analysis: details of user & software usability on website.

Applications of Web Mining

The following are some of the applications of web mining:

1. **Navigation**: Web mining is used to discover how users navigate a web site and the results can help in improving the site design and making it more visible on the web.
2. **Customer relationship management (CRM)**: Web mining is used to understand customer behavior, evaluate the effectiveness of a particular Web site, and help quantify the success of a marketing campaign.
3. **Digitize images:** web images are usually not annotated using semantic descriptors. To retrieve web images from the internet, web mining is used.
4. **Key phrase extraction**: Key phrases are useful for a variety of purposes, including summarizing, indexing, labeling, categorizing, clustering, highlighting, browsing, and searching.
5. **Social network analysis:** Social network analysis is useful for the Web because the Web is essentially a virtual society, and thus a virtual social network, where each page can be regarded as a social actor and each hyperlink as a relationship.
6. **Security and crime tracing:** web mining used for protection of user system against cyber-crimes as hacking, internet fraud, virus spreading, cyber terrorism.
7. **E-services:** web mining used in e-services like electronic business, E-Politics, E-Democracy, E-Government.

Conclusion

Web data is growing at a significant rate. World Wide Web is very important to carry out our day to day activities like business, education, e-system etc. Web Mining is fertile area of research for developing new technology and algorithm. Web mining enhances user ability to access information so capacity and potential of enterprise information resources can be reflected. It is expected more application of web mining will be developed.

References

- *Web Mining: Day-Today by K. Mohammad Mujahid,Mr. I.S.Raghuram, M. Niranjan Kumar, K.V. Chaitanya Krishna,T. Mohaneshwar.*
- *Web Mining Tasks and Types: A SURVEY by Chintandeep Kaur, Rinkle Rani Aggarwal.*
- *What are the major comparisons or differences between Web mining and data mining? By Michael Jennings, Information Management Online, June 25, 2002.*