Web Mining

Author: Bhargav Borse, Pradeep Kumar Yadav

Co-Author: Prof. Kaushal Gor,

Department of MCA, Parul University, Vadodara, India.

Abstract: This study has been undertaken to cover some application areas, methodologies, and a few techniques and algorithm. This paper will primarily specialize in the sector of web usage mining, which may be a direct need from the expansion of world Wide Web. Web data processing became a simple and important platform for retrieval of useful information. Users prefer World Wide Web more to upload and download data. As increasing growth of knowledge over the web, it's getting difficult and time consuming for locating informative knowledge and patterns. Digging knowledgeable and user queried information from unstructured and inconsistent data over the online isn't a simple task to perform. Different mining techniques are wont to fetch relevant information from web (hyperlinks, contents, web usage logs). Web data processing may be a sub discipline of knowledge mining which mainly deals with web. Web data processing is split into three different types: web structure, web page and web usage mining. of these types use different techniques, tools, approaches, algorithms for discover information from huge bulks of knowledge over the online.

I. INTRODUCTION

Now the day's data on the internet is huge and increasing more and more day by day. It should manage that great information and display the most relevant information on the user's screen. Analyzing and downloading relevant data from large data bases is not possible manually, since these automated extraction tools are required when the requested user data can download from billions of pages online and download relevant information. Usually, users get data from the WWW of the world through various search engines like Yahoo, Bing, MSN, Google etc. Web Mines is actually a data mining site related to the data available on the Internet. It is the concept of extracting informative data from web pages over the Internet. Users use different search engines to download their required information online, such informative and user-friendly information is available through a mining process called Web Mining. Various tools and algorithms used to extract data from web pages including web documents, images etc. Web mining is fast becoming more important due to the sheer volume of online publications and finding the right patterns, information and informative information is extremely difficult and time-consuming to do by hand. Structure (Hyperlinks), Usage (pages visited, data usage), content (text, pages) is included in the data collected through Web mining. The term World Wide Web is related to the compilation of web texts, videos, audios etc.

II. APPLICATION AREAS

Web mining is an important tool for collecting behavior of web site visitors and thus allows for appropriate adjustments and decisions regarding real Web users and traffic patterns. Along with a description of the processes involved in Web mining claim that Web conversion, System Improvement, Web personalization and Business Intelligence are the four major application areas for Web mining. This is briefly described in the following sections.

III. WEBSITE MODIFICATION

Website content and structure are important to the user’s perception / view of the site and the usefulness of the site. The problem is that different types of users have different options, background, information etc. which makes it difficult (if not impossible) to find a design that is suitable for all users. Web application resources can be used to determine what types of users access the website, and their behavior, information that can be used to design / re-design the websites of the user you have visited.

IV. WEB CONTENT MINING

Content Mining may be a process of Web Mining during which needful informative data is extracted from internet sites (WWW). Content includes audio, video, text documents, hyperlinks and structured record. Web contents are designed to deliver data to users within the sort of text, list, images, videos and tables. Over previous couple of decades the quantity of sites (HTML) increases to billions and still continues to grow. Searching query into billions of web documents is extremely difficult and time-consuming task, content mining extracts queried data by performing different mining techniques and narrow down the search data which become easy to find required user data.

V. BUSINESS INTELLIGENCE

Web companies working with Web mining is a powerful tool for gathering business intelligence to gain competitive advantages. Consumer activity patterns on the Website can be used as 5 important information in the decision-making process, e.g., predicting future customer behavior, hiring new customers and inventing new products are great benefits. There are many companies that provide (among other things) resources in the field of Web Mining and Web traffic analysis to extract business intelligence.

VI. WEB STRUCTURE MINING

Now a day’s massive amount of data is increasing on web. World Wide Web is one of the most loved resources for information retrieval. Web mining techniques are very useful to discover knowledgeable data from web. Structure mining is one of the core techniques of web mining, which deals with hyperlinks structure. Structure mining shows the structured summary of the website. It identifies relationship between linked web pages of websites. Continues growth of data over the internet become a challenging task to find informative and required data. Web mining is just a data mining which digs data from the web. Different algorithmic techniques are used to discover data from web. Structure mining analyses hyperlinks of the website to collect informative data and sort out in categories like similarities and relationship. Intra-page is a type of mining that is performed at document level and at hyperlink level mining is known as inter-page mining. Link analysis is an old but very useful method that is way its value increases in the research area of web mining – Structure analysis is also called as Link-mining.

VII. Web Usage Mining

Web usage mining also called log mining is a process of recording user access data on the web and collect data in form of logs. After visiting any website user leaves some information behind such as visiting time, IP address, visited pages etc. This information is collected, analyzed and store in logs. Which helps to understand user behavior and later can improves website structure. Web usage mining is a technique that automatically archives access patterns of user and this information is mostly provided by web servers which are later collected in access logs. Logs stores much needed information like URL address, visiting time, Internet Protocol addresses etc. which can help an organization to understand their customer's behavior and ensure good service quality. Web usage mining dig and analyze data present in log files which contains user access patterns. Main purpose of web usage mining is to observer user behavior at the time of his interacting with web. There are two types of pattern tracking i.e., general tracking and customized tracking. In general tracking information is collected from web page history. In customized tracking the information is gathered for specific user.

VIII. System Improvement

The functionality and usefulness of Web sites can be improved using Web traffic information to predict current user traffic. This may be used e.g., revenue, measurement or distribution of data to improve performance. Method predictor can be used to detect fraud, burglary, burglary etc.

IX. Web Personalization

Web Personalization is an attractive application site for Web-based companies, allowing recommendations, marketing campaigns etc. That it is created directly by different categories of users, and most importantly to do this in real time, by default, as the user accesses the Website.

X. Tools

1) Screen Scrapper

Screen Scrapper simply replicates what you can do with a web browser, such as clicking links, logging into websites, submitting forms, and downloading files etc. Screen-scraping is a tool for filtering information from web sites which can be used in other situations. Mine data on products and downloading them to a spreadsheet is the one of the most regular usage of this software.

2) Data Miner

Data Miner is a well-known data mining tool. It is hugely effective in extracting data from web pages. It provides the extracted data into CSV file or Excel spreadsheet. You can easily get structured data that you require.

3) Web Content Extractor

Web Content Extractor is a powerful and easy-to-use web scraping software. It allows you to extract specific data, images and files from any website. This tool allows users to extract data from various websites such as online stores, online actions, shopping sites, real estate sites, financial sites, business directories, etc. The extracted data can be exported to a variety of formats, including Microsoft Excel (CSV), Access, TXT, HTML, XML, SQL script, MySQL script and to any ODBC data source.

4) Web Info Extractor

This is a tool for data mining, extracting Web content, and Web content analysis. It can extract structured or unstructured data from Web page, reform into local file or save to database, place into Web server. Difficult templates are not required to be defined.

5) Scrapy

Scrapy is a great web mining tool. It can help you extract data from the websites. It is considered to be a complete solution as a web scraping tool because it can manage requests, preserve user sessions, follow redirects and handle output pipelines.