WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Study on Web Data Mining – Tools and Techniques

Valliappan B [a], Xavier [b]

[a] *Graduateship / Associate Membership, Bachelor of Computer Science, Indian Institute of Industry Interaction Education and Research*
[b] *Professor / Project Coordinator / Indian Institute of Industry Interaction Education and Research*
DOI: https://doi.org/10.55248/gengpi.5.0124.0316

A B S T R A C T

Web data mining became an easy and important platform for retrieval of useful information. Users prefer World Wide Web more to upload and download data. As increasing growth of data over the internet, it is getting difficult and time consuming for discovering informative knowledge and patterns. Digging knowledgeable and user queried information from unstructured and inconsistent data over the web is not an easy task to perform. Different mining techniques are used to fetch relevant information from web (hyperlinks, contents, web usage logs). Web data mining is a sub discipline of data mining which mainly deals with web. Web data mining is divided into three different types: web structure, web content and web usage mining. All these types use different techniques, tools, approaches, algorithms for discover information from huge bulks of data over the web.

Keywords: Web content mining, web structure mining, and web usage mining.

1. Introduction of various grinding process

Now a day's data over the internet is enormous and increasing frequently day by day. It is must to manage that massive information and display most related queried information on User's screen. Analyzing and fetching relevant data from large data bases is not possible manually, for this automated extraction tools are required through which user queried data can be fetch from billions of pages over the internet and discovers relevant information. Usually users find data from world wild web WWW by using different search engines like Yahoo, Bing, MSN, Google etc. Data mining is a process of analyzing usable information and extract data from large data warehouses, involving different patterns, intelligent methods, algorithms and tools. This process can help business to analyze data, behavior and predict future trends. Data mining includes four strategies steps for relevant data extraction. Data source is a set on data in large data base which can have problem definition in it. Data exploration is a step of investigation true information from bulks of unfamiliar data. Third step is modeling, in these different models are designed and then evaluate. At the end tested models are deployed, that occurs in final step of data mining strategies. Organizations can use data mining techniques to change raw data into convenient information. It can also help business to improve their marketing strategies and increase the profit by learning more about customer's behavior.

Web mining is one of the types of techniques use in data mining. The main purpose of web mining is to automatically extract information from the web. For discovering useful data (videos, tables, audio, images etc.) From the web different techniques and tools are used. Information over the internet is huge and increasing with passage to time due to which size of data bases are also growing. Digging knowledgeable information and analyzing the data sets for relevant data is much difficult because data over the internet in not in plain text. It could be unstructured data, multimedia, table, tag.

2. Web Mining

Web Mining is the process of Data Mining techniques to automatically discover and extract information from Web documents and services. The main purpose of web mining is discovering useful information from the World-Wide Web and its usage patterns.

2.1 Applications of Web Mining

Web mining is the process of discovering patterns, structures, and relationships in web data. It involves using data mining techniques to analyze web data and extract valuable insights. The applications of web mining are wide-ranging and include:

Personalized marketing:

Web mining can be used to analyze customer behavior on websites and social media platforms. This information can be used to create personalized marketing campaigns that target customers based on their interests and preferences.

E-commerce:

Web mining can be used to analyze customer behavior on e-commerce websites. This information can be used to improve the user experience and increase sales by recommending products based on customer preferences.

Search engine optimization:

Web mining can be used to analyze search engine queries and search engine results pages (SERPs). This information can be used to improve the visibility of websites in search engine results and increase traffic to the website.

Fraud detection:

Web mining can be used to detect fraudulent activity on websites. This information can be used to prevent financial fraud, identity theft, and other types of online fraud.

Sentiment analysis:

Web mining can be used to analyze social media data and extract sentiment from posts, comments, and reviews. This information can be used to understand customer sentiment towards products and services and make informed business decisions.

Web content analysis:

Web mining can be used to analyze web content and extract valuable information such as keywords, topics, and themes. This information can be used to improve the relevance of web content and optimize search engine rankings.

Customer service:

Web mining can be used to analyze customer service interactions on websites and social media platforms. This information can be used to improve the quality of customer service and identify areas for improvement.

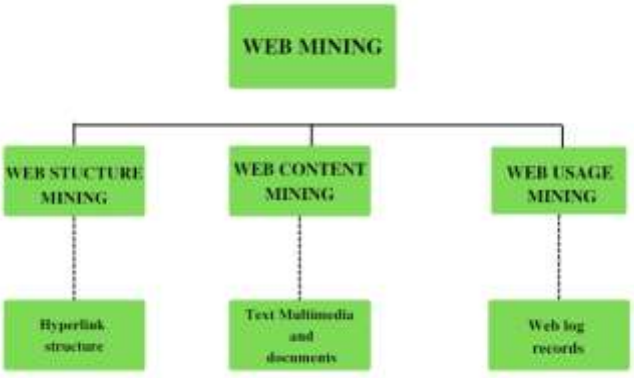
Healthcare:

Web mining can be used to analyze health-related websites and extract valuable information about diseases, treatments, and medications. This information can be used to improve the quality of healthcare and inform medical research.

2.2 Process of Web Mining

Web Mining Process

Web mining can be broadly divided into three different types of techniques of mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. These are explained as following below.



Categories of Web Mining

1. **Web Content Mining:** Web content mining is the application of extracting useful information from the content of the web documents. Web content consist of several types of data – text, image, audio, video etc. Content data is the group of facts that a web page is designed. It can

provide effective and interesting patterns about user needs. Text documents are related to text mining, machine learning and natural language processing. This mining is also known as text mining. This type of mining performs scanning and mining of the text, images and groups of web pages according to the content of the input.

2. **Web Structure Mining:** Web structure mining is the application of discovering structure information from the web. The structure of the web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Structure mining basically shows the structured summary of a particular website. It identifies relationship between web pages linked by information or direct link connection. To determine the connection between two commercial websites, Web structure mining can be very useful.

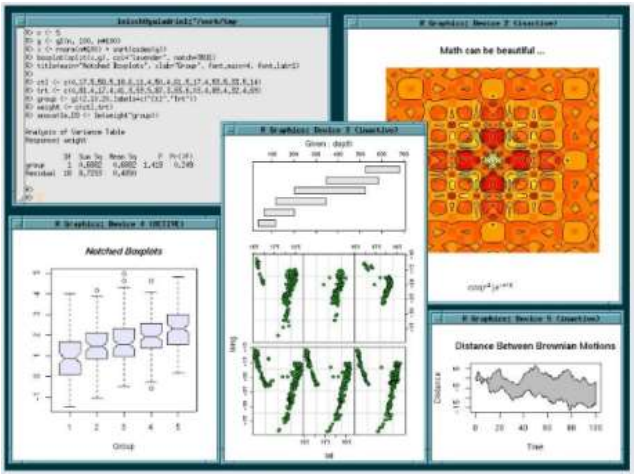
3. **Web Usage Mining:** Web usage mining is the application of identifying or discovering interesting usage patterns from large data sets. And these patterns enable you to understand the user behaviors or something like that. In web usage mining, user access data on the web and collect data in form of logs. So, Web usage mining is also called log mining.

3. Web Mining Tools

A web data miner is computer software that uses data mining techniques to identify or discover patterns from large data sets. Data is money in today's world, but the information is huge, diverse, and redundant. Having the tools for mining is going to be a gateway to help you get the right information. In this post, you can learn the list of the 7 most popular web mining tools around the web.

1. Octoparse is a simple but powerful web data miner that automates web data extraction. It allows you to scrape data from any website with its easy auto-detecting function and preset templates. With Octoparse, you can finish the data mining process within a few clicks. However, it also provides advanced functions like AJAX, pagination, loop, IP proxies, cloud service, etc., to get more and accurate data.

2. R is a language or a free environment for statistical computing and graphics. It has been made accessible from scripting languages like Python, Ruby, Perl, etc.



3. Oracle Data Mining is a data mining software by Oracle. Oracle Data Mining is implemented in the Oracle Database kernel, and mining models are first-class database objects. Oracle Data Mining processes use built-in features of Oracle Database to maximize scalability and make efficient use of system resources.

4. Tableau offers a family of interactive data visualization products focused on business intelligence. Tableau allows instantaneous insight by transforming data into visually appealing, interactive visualizations called dashboards. This process takes only seconds or minutes rather than months or years and is achieved through the use of an easy-to-use, drag-and-drop interface.

5. Scrapy is an open-source framework for collecting data from websites. It is written in Python and you can write the rules to extract web data.
6. HITS, short for Hyperlink-Induced Topic Search, also known as hubs and authorities, is a link analysis algorithm that rates Web pages. In the HITS algorithm, the first step is to retrieve the most relevant pages to the search query. This set is called the root set and can be obtained by taking the top pages returned by a text-based search algorithm. A base set is generated by augmenting the root set with all the web pages that are linked from it and some of the pages that link to it. The web pages in the base set and all hyperlinks among those pages form a focused sub graph.
7. PageRank Algorithm is a popular Web structure Mining Algorithm. PageRank is a link analysis algorithm and it assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set. The algorithm may be applied to any collection of entities with reciprocal quotations and references.

4. Techniques in Web Data 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- Support vector machines, K-Nearest Neighbors, 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 also in other cases 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.

5. Conclusion

Data mining is a concept that helps to find information which is needed from large data warehouses by using different techniques. It is also used to analyze past data and improve future strategies. Web data mining is considered as sub approach of data mining that focuses on gathering information from web. Web is a large domain that contains data in various forms i.e.: images, tables, text, videos, etc. As size of web is continuously increasing; it is becoming very challenging task to extract information. In this paper we described three important types of web data mining that can help in finding informative data. Each type has different algorithms, tools and techniques that are used for data retrieval. Various algorithms, tools and techniques for each type are described. Table I summarizes all types and Table II shows comparison for web usage mining techniques. Web content mining is useful in terms of exploring data from text, table, images etc. Web structure mining classifies relationships between linked web pages. Web usage mining is also an important type that stores user access data and get information about specific user from logs.

References

1. Anurag Kumar and Ravi Kumar Singh, "Web Mining Overview, Techniques, Tools and Applications: A Survey," International Research Journal of Engineering and Technology (IRJET), vol. 03, no. 12, pp. 1543-1547, December 2016.
2. Simranjeet Kaur and Kiranbir Kaur, "Web Mining and Data Mining: A Comparative Approach," International Journal of Novel Research in Computer Science and Software Engineering, vol. 2, no. 1, pp. 36-42, January - April 2015.
3. Ahmad Tasnim Siddiqui and Sultan Aljahdali, "Web Mining Techniques in E-Commerce Applications," International Journal of Computer Applications, vol. 69– No.8, pp. 39-43, May 2013.
4. Kshitija Pol, Nita Patil, Shreya Patankar, and Chhaya Das, "A Survey on Web Content Mining and extraction of Structured and Semistructured data," Emerging Trends in Engineering and Technology, pp. 543-546, July 2008.
5. R. Malarvizhi and K Saraswathi, "Web Content Mining Techniques Tools & Algorithms – A Comprehensive Study," International Journal of Computer Trends and Technology (IJCTT), vol. 4, no. 8, pp. 2940-2945, Augest 2013.
6. Raymond Kosala and Hendrik Blockeel, "Web Mining Research: A Survey," SIGKDD Explorations, vol. 2, no. 1, pp. 1-15, July 2000.
7. Faustina Johnson and Kumar Santosh Gupta, "Web Content Mining Techniques: A Survey," International Journal of Computer Applications (0975 – 888), vol. Volume 47– No.11, pp. 44-50, June 2012.