ROOTINDEXING
JOURNAL ABSTRACTING AND INDEXING SERVICE
INTERNATIONAL ADVANCE JOURNAL OF ENGINEERING, SCIENCE AND MANAGEMENT (IAJESM)
July-December 2023, Submitted in September 2023, iajesm2014@gmail.com, ISSN -2393-8048
Multidisciplinary Indexed/Peer Reviewed Journal. SJIF Impact Factor 2023 =6.753

Issues of Data Minining: A Review

Raval Chandni Sudhirkumar, Research Scholar, Department of Computer Application, Shri Jagdishprasad Jhabarmal Tibreuniversity, Vidyanagari, Jhunjhunu, Rajasthan
Dr. Ajit Kumar, Assistant Professor, Department of Computer Application, Shri Jagdishprasad Jhabarmal Tibreuniversity, Vidyanagari, Jhunjhunu, Rajasthan

ABSTRACT

In the present time, the World Wide Web has evolved into a distributed data space that contains a few billion pages and 100 million devices. Customers continue to face difficult challenges in their search for the information they want, despite the fact that there is an abundance of data available to them online. As a result of the exponential growth of the Internet, the most important objective for research is to develop a search engine that is both accurate and efficient. Among the many subfields that fall under the umbrella of the more general issue of "data mining," "Web data mining" is among the most well-known of your options. Methods that are both effective and efficient for collecting information from the internet are the major emphasis of this specific topic. After gathering all of this information, it will be simple to divide the data into three distinct categories. A number of factors, including the content, use, and structure of the website, are taken into consideration throughout the process of extracting information from the internet. Within the scope of this essay, we provide one viewpoint on web mining, which is a technique that involves obtaining useful and relevant information from the internet. In addition, we furthermore provide a concise summary of data mining in addition to other resources that cover a wider range of topics.

INTRODUCTION

The practice of obtaining useful information from unstructured server logs is known as "web mining". The website's server is the most important information source for the web logs. The practice of obtaining valuable information from unstructured server web log data is known as web mining. This data is very helpful for maintaining, improving, and protecting websites since it is human-readable and has been classed.

Web mining will use any mathematical techniques for data mining that exist. Web mining, then, is the branch of data mining that makes use of web server logs. The process of finding relevant information from internet servers' databases is called "web mining." Numerous concerns, such as website security access, targeted marketing, personalization system development, server performance, website optimization, and many more, may be clarified by using server log analysis.

REVIEW OF LITERATURE

Anoop Paharia (2017) Because of the internet, we live in a web-based world, and web servers are to thank for the development and ongoing dynamic generation of a vast majority of web pages. Web pages provide a rich environment for data mining; however, there are significant challenges that occur when looking for and analyzing this data because it is more complicated and dynamic than the data that can be found in commercial database management systems. In this article, we are going to explore the various ways in which searching methods can be made more effective on the basis of semantic by employing the principles of data mining, as well as the various ways in which more accurate searching results can be acquired.

In accordance with the definition that was presented by M. Bharati (2010), the term "data mining" refers to a method that is used to recognise significant patterns within large datasets. Data mining is often referred to as "pattern mining." Pattern mining is a name for the process. This page provides an overview of a broad range of data mining strategies, methodologies, and case studies of businesses that have successfully used data mining in order to improve their operations. The usefulness of data mining is shown right here via these case studies. The following list also includes a few companies that have enhanced their operations by using data mining techniques. These companies are featured on this list.

The year 2011's Brijendra Singh Internet mining may be broken down into three distinct subfields: web content mining, web structure mining, and online usages mining. Online data mining is a popular subfield of data mining that focuses on the extraction of useful information

from a range of sources, including the World Wide Web. This information may be used for a variety of purposes. There are three distinct forms of web mining: web usages mining, web structure mining, and online content mining. The purpose of this research is to offer an overview of the field, a criticism of the techniques that are currently being used, and an evaluation of each, respectively. In this research, each and every one of those subjects will be discussed. During this session, participants will also have the opportunity to discuss significant issues about the path that future research will take. This research also compares and analyses a number of different online data mining techniques and the applications of those approaches; more information will be provided in a separate section. A number of important research issues are highlighted, and an overview of the current state of the subject is provided thereafter.

2009 witnessed the A. Fong, SimonThe purpose of this research is to provide an overview of a data mining methodology that we refer to as Business Intelligence-driven Data Mining, or BIdDM for short. Knowledge-driven and method-driven data mining are combined in this approach, which helps to close the knowledge gap that exists between the many data mining approaches that are currently being used in e-Business and business intelligence. Another way to put it is that it combines knowledge-driven data mining with method-driven data mining. The process of constructing a four-layer structure is carried out in the first part of the BIdDM approach, while the process of data mining is carried out in the second portion. In order to begin the installation of the four-layer structure, which is an important component of BIdDM, a procedure must first be devised. For the purpose of illustrating how BIdDM may be used, we provide a case study of a company that participates in direct sales to customers, which is also referred to as a business-to-consumer e-shop during some instances.

RESEARCH METHODOLOGY

During the course of our examination into web usage mining (WUM), we used a research technique that can be broadly classified into four stages:

Information that is being gathered: The first stages towards a successful research venture are the identification of appropriate internet data sources and the collection of the needed data for analysis. The application programming interfaces (APIs) that are provided by the data source and web scraping are two methods that may be used to get the information.

In the first stage of the information processing process: Following the end of the data gathering procedure, the data must be pre-processed in order to eliminate noise, clean and convert the data, and fill in any possible values that may be missing.

Methods and Approaches used in the Process of Data Mining After the data has been pretreated, it is necessary to use data mining techniques such as clustering, classification, association rules, and anomaly detection. The procedures in question ought to be used to the data collection.

Criteria for evaluating the measurement are as follows: In order for researchers to compare and evaluate the effectiveness of the different data mining approaches, they need to come to an agreement on evaluation criteria that are objective and reliable. A number of measures, including the area under the receiver operating curve (AUC-ROC), recall, accuracy, precision, and F1-score, are often utilized in this context.

Python and R are examples of programming languages that are necessary for doing research utilizing data mining approaches.

Investigating new ways of doing things: As part of the research, it will be necessary to conduct tests on the data that was acquired. Additionally, it will be essential to assess the efficacy of the different data mining techniques by making use of the criteria that have been set.

- The techniques utilized to gather the data as well as the features of the web log data.
- The methodological approach.
- The characteristics of the web log data.
- methods to analysis, including assessment and justification of such approaches as well as the strategy or methodology of analysis

INTERNET DATA GATHERING

The extraction of useful patterns from a variety of website components is a significant part of

online content mining. These elements include, but are not limited to, photos, tables, text, audio, video, graphics, and PDFs. The content of websites may be mined using one of two different approaches. The first topic of discussion is the practice of extracting data from web sites. The mining results are sorted in a manner that is determined by the kind of content. Second, you may employ data mining to climb to the top of the pages that display the results of a search engine. Organizing websites into categories according to the content that they contain is the second technique of content mining that may be done online.

METHODS FOR MINING WEB LOGS

Three further steps may be used to further dissect the web usage mining (sometimes called web log mining) process.

- The preliminary processing
- The identification of patterns
- The analysis of patterns

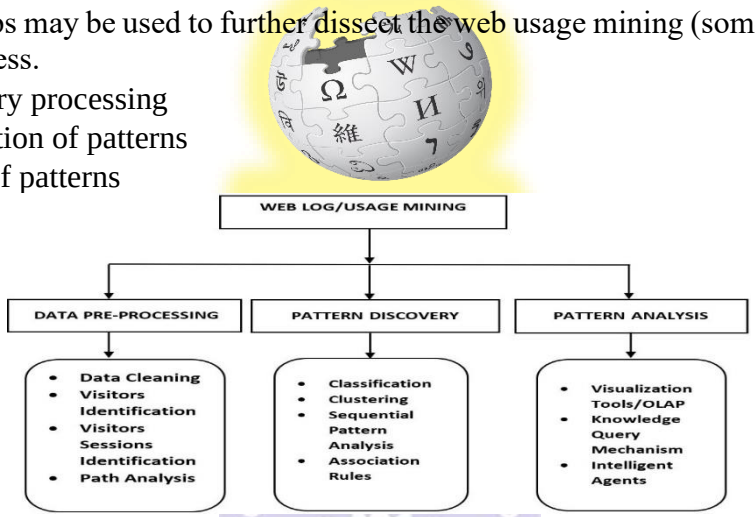

Figure: Types of Web log mining

PRE-PROCESSING

An explanation of the pre-processing stage

The pre-processing stage is the most critical and significant one, and it cannot be skipped when mining for online use. Web logs may provide vast amounts of data, which makes analysis quite challenging. The majority of the data in the server log files is redundant, confused, illogical, and noisy, making it unsuitable for the analysis that was intended to be performed on it. In terms of data mining based on online conduct, this information is not required. This unnecessary data may be eliminated in the internet use mining preparation phase. Superfluous data may be eliminated by software that has the right algorithms or by human interaction. "Data mining" is the process of looking for significant patterns in internet activity by going through site logs.

DISCRETIZATION

In mathematics, the process of transforming continuous data into its discrete equivalents in the form of functions, models, and equations is referred to as discretization. The word "discretization" is used to denote this transformation. After information has been discretized, it is then able to be assessed numerically and employed in applications related to digital computing. The process of mining data is dependent on data discretization, which is the conversion of continuous data into quantized form so that following data mining procedures may be performed with greater ease. Because learning strategies in data mining handle discrete properties of data better than continuous attributes, it is essential to discretize data when it includes both continuous and discrete features.

PROBLEM ON HAND

Before data mining can be utilised to produce online intelligence, there are a number of things that need to be done, some of which are described below as research issues. It might be challenging to create accurate forecasts about the academic accomplishment of youngsters since there are a great number of nuanced elements that need to be taken into consideration. The majority of performance prediction systems are ineffective and include traits or components that are not necessary for the system to function well. The purpose of this survey article is to investigate the most current research that has been conducted in the subject, to

make comparisons and contrasts, and to evaluate which methods are the most successful in forecasting future performance.

WEB LOG MINING AND WEB SEARCH QUERY

As the number of information sources that are available online continues to develop at an exponential pace, the usefulness of automated systems that aid users in identifying relevant information resources and getting a knowledge of their consumption habits is rising. These systems are becoming more important. When this is taken into consideration, it is of the highest significance to develop intelligent systems that are capable of mining data in an efficient way on either the client or the server side.

INTERNET MINING

In the context of the Internet, the term "web mining" refers to the practice of using data mining methods in order to search for patterns. Web mining is the practice of discovering hidden patterns and data in user behaviour or objects on the World Wide Web in order to draw conclusions about those users or artefacts. This is done in order to provide information about the users or artworks.

When it comes to web mining, there are three basic schools of thought: content mining, structure mining, and usage mining.

Content Mining on the Web

The process of translating various types of online material, such as text, photos, and scripts, into formats that are more user-friendly is included in this approach. All of the titles, specific materials, and photos that are now available are all factors that contribute to the grouping and categorizing of the information that can be accessed on various websites. This was quite useful.

Exploiting the Architecture of the Web

This entails doing an investigation of the structure that is present on each individual page that constitutes a website. Due to the fact that different websites do not all have the same organizational structure, the process of mining the online structure may prove to be difficult. As a direct result of this, it is feasible that a unique logic will need to be applied to each new page or site that is created.

COLLABORATIVE FILTERING AND USER PROFILING

As a consequence of the rapid development in popularity and effect of online services such as social bookmarking tools, photo sharing websites, and blog sites, website developers are under pressure to improve the content that they supply to their users (also known as "customers"). Developers of websites have discovered that the availability of metadata has made it simpler for them to construct systems that allow people to contribute their own explanations to content that they have already discovered on the web. When referring to the process of applying descriptive labels on items in an effort to stimulate client demand, the term "tagging" is most often used. Through a system that we refer to as tags, we are able to transmit meta-information about things.

Tags, which are sometimes written as "tagging," are a frequently overused term that are often confused with a number of distinct paraphrases that do the same job. Tags are sometimes written as "tagging."

Customers who shop at the well-known online retailer Amazon have the chance to provide feedback and reviews on a variety of material, including titles of books. Users of Delicious have the opportunity to apply tags to the websites that they have bookmarked and share with other users the tagging style that they like. On the social networking website Facebook, users have the ability to put the faces of their friends in albums that they have produced or that any other user who currently has an account may see. On the well-known website for sharing photographs, Flickr, users have the ability to tag the photos that they upload to the website. Not only are users able to provide labels, but they can also make use of those labels in order to easily discover certain images.

ALGORITHM FOR MODIFIED DATA CLEANING OF VARIABLE WEB LOG DATA

In order to clean and prepare data from internet logs, the researchers working on the project

devised a strategy, and they employed the Java programming language to carry out the data cleaning and preparation process. Additionally, and this was a really exciting discovery, they discovered that the frequency with which each page was viewed varied greatly from one website to the next. In order to accomplish the goal of deleting web log entries that were associated with the movement of a robot throughout the website, Muskan and colleagues created a data cleaning technique with the idea of attaining the goal. During the course of their research endeavours, Shaily Langhnoja and her colleagues developed pre-processing techniques. These methods have shown to be beneficial in the management of online log data. They eliminated the image media by removing the file names from the folders that ended in.gif,.jpeg, and.jpg, respectively. This was done in order to avoid any confusion. The information that was removed from the web log included entries that included the 404 error code as well as log entries that were created by web robots and spiders that visited the website. They even went so far as to erase the information.

SUMMARY

The purpose of this thesis is to collect the online activities of several individuals in order to generate a greater quantity of information than would be possible in situations when only one data source is available. If we do this, we may be able to come closer to rich, individualised online activity that is not just speedy but also relevant and one of a kind. One capability that may be called an intelligent element of a user interface is the ability to ascertain the level of interest that a person has in the page that they are now seeing. One example of an explicit method is to request that readers rate the websites that they have read; however, since this goes against the typical flow of their browsing experience, the majority of consumers choose to disregard it. Users are provided with more free interest signals when implicit techniques are used, despite the fact that they may need more complicated intelligent user interfaces. A large number of implicit interest indicators are identified and presented in this thesis. These indicators may be used to determine the amount of explicit interest in a website or material that is found online. An examination of the accuracy with which these indicators anticipate the level of explicit interest is what determines the usefulness of these various indicators. The primary objective of the thesis is to provide examples of the many uses that may be made of these indicators.

BIBLIOGRAPHY

[1] [Adomavicious and Tuzhilin (1999)] Adomavicious Gediminas and Tuzhilin Alexander, User Profiling in Personalization Applications through Rule Discovery and Validation, in Proceedings of the ACM Fifth International Conference on Data Mining and Knowledge Discovery, KDD-1999, pp. 377-381.

[2] [Grcar (2004-2)] Miha Grcar, User Profiling Collaborative Filtering, in Proceedings of the conference on data mining and warehouses SIKDD-2004, Volume: 10, pp. 803-826.

[3] [Krishnapuram et. al. (2000)] Anupam Joshi, Karuna Joshi and Raghu Krishnapuram, On Mining Web Access Logs, in ACM SIGMOD Work shop on Research Issues in Data Mining and Knowledge Discovery, 2000, pp. 63-69.

[4] [Lin and Ho (2002)] Lin Shin-Hua and Ho Jan-Ming, Discovering Informative content Blocks from Web Documents, in Proceeding of Eighth ACM SI GKKDD conference on Knowledge Discovery and Data Mining, 2002, pp. 588-593.

[5] [Zhao et. al. (2003)] Zhao Kaidi, Liu Bing, a nd Yi Lan, Eliminating Noisy Information in Web pages for Data Mining, in Proceeding of Ninth ACM SIGK KDD conference on Knowledge Discovery and Data Mining, 2003, pp. 296-305.

[6] Paharia, Anoop & Bhawsar, Yachana & Singh, Divakar & Tech, M. (2017). Developing web intelligence using data mining.

[7] Han, Jiawei & Chang, Kevin. (2002). Data Mining for Web Intelligence.. Computer. 35. 64- 70. 10.1109/MC.2002.1046977.

[8] Domingues, Marcos & Jorge, Alípio & Soares, Carlos & Rezende, Solange. (2014). Web Mining for the Integration of Data Mining with Business Intelligence in Web-Based Decision Support Systems. 10.4018/978-1-4666-6477-7.ch007.

[9] Haldorai, Anandakumar & Ramu, Arulmurugan & Suriya, M.. (2019). Web Intelligence and Data Mining in Urban Areas. 10.1007/978-3-030-26013-2_2.

[10] Mishra, Brojo & Hazra, Deepannita & Tarannum, Kahkashan & Kumar, Manas. (2016). Business Intelligence using Data Mining techniques and Business Analytics. 84-89. 10.1109/SYSMART.2016.7894496.

[11] Khder, Moaiad & Abu-AlSondos, Ibrahim & Bahar, Yousif. (2021). THE IMPACT OF IMPLEMENTING DATA MINING IN BUSINESS INTELLIGENCE. International Journal of Entrepreneurship. 25. 1-9.

[12] Azevedo, Ana & Santos, M.F.. (2014). Integration of Data Mining in Business Intelligence Systems. 1-314. 10.4018/978-1-4666-6477-7.

[13] Bharati, M. & Ramageri, Bhavani. (2010). Data mining techniques and applications. Indian Journal of Computer Science and Engineering. 1.

[14] Thuraisingham, Bhavani. (2003). Web Data Mining and Applications in Business Intelligence and Counter-Terrorism. 10.1201/9780203499511.

[15] Singh, Harvinder. (2017). Implementation Benefit to Business Intelligence using Data Mining Techniques.

[16] Prasad, M & Manjula, B & Mohd, Ayesha. (2020). Comparison of Data Mining and Web Mining. 2012.