SNEH TEACHERS TRAINING COLLEGE, JAIPUR
"Foster Emotional Intelligence in Youth Through Education" (ICFEIYE-2024)
DATE: 15 April 2024
International Advance Journal of Engineering, Science and Management (IAJESM)
Multidisciplinary, Indexed, Double-Blind, Open Access, Peer-Reviewed, Refereed-International Journal, Impact factor (SJIF) = 7.938

Internet Data Gathering Using the Web Mining

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

"Web usage mining" refers to the process of analysing the patterns of user access to web servers. This article is written in response to questions about the limitations of data mining. Our primary focus will be on the many approaches that may be taken to obtain information that is kept on the internet via the use of these technologies. Additionally, we discussed cloud mining, which is a method that involves the utilisation of cloud computing in order to gather information from the internet. The present trend in web mining is a good indicator that web mining will continue to gain momentum in the years to come. The majority of websites on the internet rely on web servers, which are the primary factor in their continuous and dynamic expansion. Web servers are crucial to the majority of websites. As a result of the internet, we now live in a world that is entirely dependent on the online.

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.

Different Categories of Web Mining

It is possible for online logs to include both structured and unstructured data formats. Web mining may be further broken into three distinct sectors, each of which is determined by the kind of data that is being mined.

- Mining Web Content
- Mining Web Structure
- Web Record/Usage Analysis

REVIEW OF LITERATURE

Bhavani Arunachalam during the course of the whole year 2013 Due to the exponential growth of this data, information technology professionals and executives of companies are working hard to develop methods for sifting through the enormous quantity of data that is available on the internet in order to discover insights that may be beneficial to their organisations. This is being done in order to find ways to discover insights that may be beneficial to their organisations. The possibility exists that you will come across these methods in this location. Organisations are finding themselves increasingly entangled in a dizzying maze of techniques and methods for gathering and utilising essential information as a consequence of the expansion of data mining and, more broadly speaking, web mining. This is a result of the fact that web mining has become more widespread. The purpose of this book is to provide a complete definition of web mining, with a specific focus on customer relationship management (CRM). Additionally, for the very first time, it investigates the potential applications of web mining in

International Advance Journal of Engineering, Science and Management (IAJESM)
Multidisciplinary, Indexed, Double-Blind, Open Access, Peer-Reviewed, Refereed-International Journal, Impact factor (SJIF) = 7.938

situations concerning security and counterterrorism. You may be able to find this information in a book that is about to be published with the title Web Data Mining and Applications in Business Intelligence and Counter-Terrorism. It is quite beneficial to have expertise that is not just practical but also hands-on when it comes to the process of developing efficient solutions for internet businesses. Throughout the whole of this essay, the significance of web mining is highlighted, in addition to offering a comprehensive overview of the tools and techniques that are involved. The author, who is the head of a data and applications security project at the National Science Foundation, provides a comprehensive analysis of the many approaches that may be used to uncover and manage the advantages and risks associated with the Internet. This book teaches companies how to gather and evaluate data based on the Internet, which may aid them in creating deeper ties with their clients, growing their sales, and spotting both current and future hazards. The book is a great resource for businesses. The fact that corporations may apply the same strategies for web mining in order to battle the threat of terrorism makes it abundantly evident that web mining is an essential weapon in the intelligence toolbox. This is particularly true when one takes into consideration the fact that web mining is a weapon.

M. In the year 2020, Rajendra Prasad wrote this. Websites that are static are referred to as web1.0, websites that are dynamic are call web2.0, and websites that are semantic are called web3.0. When taken as a whole, these websites make up an ever-increasing amount of online material that is rapidly becoming an invaluable resource for the acquisition of knowledge and the retrieval of information. This accessibility is made possible by each successive version of the web, beginning with web1.0 and continuing through web3.0. Among the many subfields that fall under the umbrella of information technology and computer science, the three that are considered to be the most significant and comprehensive are information and knowledge management, data mining, and web mining. This is a truth that has been studied extensively and is well-documented. Data mining is a rapidly expanding academic area that is gaining popularity as a result of the growing relevance of the internet and the vast amount of data that it stores. It is referred to as "web mining," which is a catch-all name for the activity, and it is the process of applying data mining methodologies to the Internet. Utilisation mining, content mining, and structure mining are the three unique applications of this phrase that may be found in the context of the web.

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.

M. Domingues (2017) provided a definition of the data warehouse for online intelligence in relation to the analysis, design, and maintenance of websites respectively. There are two primary issues that these websites face: the first is the maintenance of the website's content and services, and the second is the dynamic adaptation of the website to the demands made by users. Building a data warehouse and using web mining as a tool for pattern identification and system operations analysis are two of the recommendations made by the authors of this study

International Advance Journal of Engineering, Science and Management (IAJESM)
Multidisciplinary, Indexed, Double-Blind, Open Access, Peer-Reviewed, Refereed-International Journal, Impact factor (SJIF) = 7.938

in order to automate the processes involved in website operations. It is necessary to have data warehouse support for a Web intelligence, recommender, and monitoring framework in order to discover and record authoritative connections, traversal patterns, and semantic structures that will intelligently guide and enhance our interactions with the internet. The enigma may be solved by using this key, which holds the key. Due to the fact that the internet, commonly referred to as the World Wide Web, is home to a massive and ever-changing collection of websites, online mining is an essential process. Additionally, these websites include a substantial quantity of backlinks, in addition to a wealth of information on access and use. When it comes to data mining, this information can prove to be highly useful. The most significant challenge that web mining faces in the modern day is design. A multi-layered and multi-dimensional web may be developed with the assistance of web intelligence, which is responsible for performing a variety of web mining activities. One example of this is the dynamic mining of data from online search engines.

RESEARCH METHODOLOGY

METHODOLOGICAL APPROACH

1. Type of research

For the sake of clarity and order, our research has been separated into two distinct sections. Improvements to methods for pre-processing data obtained from site logs were the primary topic of discussion during the first half of our meeting. We will be focused on the creation of preprocessing algorithms in order to aid with the identification of visitors and sessions, as well as the purification of data. A number of our earlier algorithms are going to be revised in order to take into account the new circumstances that have arisen as a result of the proliferation of websites that cover topics that are pertinent to the topic at hand and the enhancement of data services. The proliferation of websites that investigate topics that are connected has led to the emergence of these new situations.

Using a variety of online log mining techniques, the next section will investigate the information that was extracted from the logs of the internet services. Through the use of reports and data, we will make an effort to recognise a variety of existing patterns. These reports illustrate the manner in which people interacted with the website that they were visiting throughout their sessions. In order to establish a website, it is necessary to carry out this approach in the same order that it is shown here. There is also the possibility of using these reports to enhance the security standards of the website, which is both advantageous and practicable. Through the use of observable patterns that may be identified in analysis reports, it is possible to improve the appearance of the website's pages and structure.

2. Methodology

Before any information extraction or pattern identification may take place during web log mining (WLM), the following tasks need to be completed without fail:

The data are selected and cleaned in the following manner:

Based on our findings, we have established that using the information acquired from site logs would make the analytical process easier. We have been successful in obtaining the essential information that we want from the hosting server ofgeeksforgeeks.org.

Transforming the data or doing preliminary processing on the data

The web server keeps a record of each and every visit to the website, including and not limited to information on bot visits, script and HTML use, and page loads that were failed. At any moment, you are free to access this information from our website. It is necessary for us to filter the data in order to reduce the amount of time that is required to finish the remaining phases of the operation. Through meticulous planning and preparation that takes a significant amount of time, it is possible to reduce the amount of data by removing data that is not essential. To do this, careful preparation is required. It is possible to increase the preparedness of the data for

International Advance Journal of Engineering, Science and Management (IAJESM)
Multidisciplinary, Indexed, Double-Blind, Open Access, Peer-Reviewed, Refereed-International Journal, Impact factor (SJIF) = 7.938

pattern analysis and identification by pre-processing, which also results in more accurate findings during those phases.

- **Mining of websites using various methods and tools for data mining**
  There is a vast array of approaches available for pattern analysis and pattern discovery. There are many approaches to pattern analysis, the most popular ones being sequential pattern analysis, association rule mining, clustering, and classification.
- **Evaluation of patterns (including Analyses and Representations of Knowledge)**
  The process of developing new patterns involves the use of a variety of various approaches to pattern assessment. The ones that are most often used are the ones that include online analytical processing, usability analysis, data and information searching, and visualization.

3. Properties of data

In point of fact, "web log data files" are nothing more than a special sort of text file that is intended to record information on website visitors. Data files that are used for web logs are one example of this kind of file. Spend some time exploring the website; you should be able to get all the information that you are looking for. In the part that comes after the introduction to this essay, we are going to take a more in-depth look at some of the defining traits that it has.

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.

Web Structure Mining

"Web structure mining" is a field of research that looks at the network of linkages that are present on different websites. Understanding a website's structure and linkage may be achieved using a technique called web structure mining, which can provide insightful information on how to improve the website. Web structure mining is essentially an analysis of the networked web of websites. Interstice connections may be analyzed thanks to link structures that are designed based on topology. Rearranging website content and raising page ranks are only two of the numerous useful applications of web structure mining.

Extracting Information from Web Logs and User Conduct Finding patterns in user activity on a website and using those similarities to develop globally applicable rules is the aim of web log mining. Analyzing this data for patterns in user behavior on the website is the first step towards utilizing them to create rules. The technique of identifying valuable use patterns in data recorded in web server logs is called "web usage mining". Web log mining and web usage mining are interchangeable terms. The technique of examining web server logs to look for trends and patterns in online behavior is known as "web use mining." This technique is known as "web usage mining," or just "web mining."

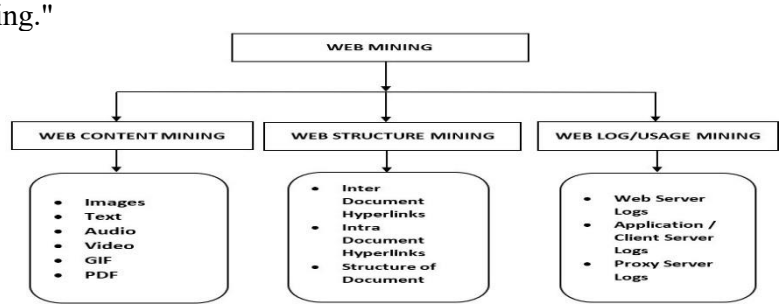

Figure: Various forms of online mining

International Advance Journal of Engineering, Science and Management (IAJESM)
Multidisciplinary, Indexed, Double-Blind, Open Access, Peer-Reviewed, Refereed-International Journal, Impact factor (SJIF) = 7.938

The whole clickstream of a person is captured in web logs. This implies that a plethora of information about website visitors may be obtained from web logs. The visitor's IP address, login credentials, browser type, operating system, data transmitted, HTTP status code, referring URL, requested resource, and data transfer volume are all included in this data. The Cookie that was used to start the request is also supplied. Webmasters should investigate this data if they have any issues about site visits, maintenance, personalisation, or security.

SCOPE OF THE STUDY

The World Wide Web is becoming into a repository for information that is always moving forward. Even though there are a great number of people who use the internet, there are surprisingly few websites that really deliver information that is significant or beneficial. Within the scope of the work that has been emphasized, the question "How can a search identify that portion of the web that is truly relevant to one user's interests?" is addressed. We also provide an answer to a question that is connected to this one: "How can a search find high-quality web pages on a particular subject?" There are three primary methods that have developed throughout time that users may use to access content that is housed in digital repositories. To get started, you should investigate the materials that are available on the internet. You may also do searches using keywords and random terms. It is clear from their accomplishment that the internet has the potential to someday supplant all other routes for the delivery of knowledge. This is the exhibit.

OBJECTIVE OF THE STUDY

1. To study on Analyzing the benefits of web intelligence.
2. To study on Comparing different web intelligence applications.
3. To study on Assessing the challenges of developing web intelligence.
4. To study on Examining the impact of web intelligence on business and society.

DATA COLLECTION METHODS AND WEB LOG DATA PROPERTIES

Data type: web log data

For the purpose of our investigation, we made use of data files, which included online logs, that were accessible from a number of different web servers. Web logs are text files that are used to record and store information about users' visits to websites. This information is kept in the text files. The things in question are those that are produced by the server in an automated fashion and are already present on it. Not only may this data be useful for determining the efficiency of websites, but it can also be put to use for research and development in the future.

- **Which type of data, main or secondary, do you have?**

The cornerstone of our inquiry was comprised of this information, which we collected by establishing a direct connection to the server that is responsible for hosting the websitegeeksforgeeks.org. The government makes use of this website as a resource in order to establish the most suitable places for classroom teachers and other members of the educational staff.

- A site is an assortment of many pages, and pages are computerized records that are composed utilizing HTML (HyperText Markup Language)

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

International Advance Journal of Engineering, Science and Management (IAJESM)
Multidisciplinary, Indexed, Double-Blind, Open Access, Peer-Reviewed, Refereed-International Journal, Impact factor (SJIF) = 7.938

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.

Analyzing Web Use

Web Usage Mining is a method that analyses the many ways in which people use the internet in order to make predictions about what they will do in the future. In this way, the meaning of the term is completely understood. There are occasions when Web Usage Mining is just intended to cover a single website, or at most a few websites. Pattern recognition is one of the outcomes that might be achieved by using this strategy. Because we have a limited amount of data available to us, including IP addresses, essential user information, and site clicks, this stage provides a set of obstacles that are not found anywhere else. When just this little quantity of information about a person is made public, it becomes more difficult to follow that person throughout a website. This is because the information is more difficult to ascertain.

COLLABORATIVE FILTERING AND USER PROFILING

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.

EVALUATION OF THE LEVEL OF SECURITY OF WEBSITE AND MODIFIED DATA CLEANING ALGORITHM

The phase in the method for preparing the online log data that is referred to as "data cleaning" is thus the one that receives the most significance and attention. In order to do this, it removes unnecessary information from massive online log files, which in turn improves the results of operations for pattern building and discovery. Through the usage of web log files, the web server is able to preserve a record of all the activities that users do while they are on its website. Each and every activity that is carried out is recorded in these directories. It is necessary for this information to be very specific and compacted in order for it to be relevant for analysing the behaviour of visitors. By deleting any information from the database that is either

superfluous or does not provide anything of value to the analysis of user activity, it is feasible to declutter the database.

CONCLUSION

The approaches that are presented in this article provide a mechanism to gather implicit interest indicators on the client side by making use of a web browser that has been configured appropriately. In the course of this inquiry, this approach was used. On the other hand, it is also feasible to gather implicit interest indicators on the server side of a web server, mostly via the logs. This is a possibility. This may be accomplished in a number of different ways. The results that are provided by the client-side implicit interest indicators are more accurate than those provided by the server-side indicators, despite the fact that they do not need specialised client software. When it comes to identifying explicit interest, the server-side signals deliver better results by providing superior outcomes. An additional advantage of the arrangement is that the firm often ends up with a higher profit than the person does in the end. Furthermore, they make a significant effort to guarantee that the privacy of each and every user is never compromised in any way.

BIBLIOGRAPHY

[1] [Abrams et. al. (1998)] David Abrams, Ron Baecker and Mark Chignell, Information Archiving with Bookmarks - Personal Web Space Construction and Organization, ACM SIG- CHI-1998, pp. 41-48.

[2] [Ahn et. al. (2007)] Jae-wook Ahn, Peter Brus ilovsky, Jonathan Grady, Daqing He and Sue Yeon Syn, Open user profiles for adaptive news systems: help or harm?, in Proceedings of the 16th international conference on World Wide Web, WWW-2007, pp.11-20.

[3] [Ahn et. al. (2008)] Jae-wook Ahn, Peter Brusilovsky, Daqing He, Jonathan Grady and Qi Li, Personalized Web Exploration with Task Models, in Proceedings of WWW-2008, Refereed Track: Browsers and User Interfaces, Beijing, April 21-25, 2008, pp. 1-10.

[4] [Anderson et. al. (2001)] Corin Anderson, Pedro Domingos and Daniel S. Weld, Proteus - Adaptive Web Navigation for Wireless Devices, In Proceedings of the 17th international joint conference on Artificial intelligence - (IJCAI-01) Volume 2, pp. 879-884.

[5] [Baeza-Yates (2005)] Ricardo Baeza-Yates, Applic ations of Web Query Mining, in LNCS- 3408, Springer Berlin / Heidelberg, 2005, pp. 7-22.

[6] [Baeza-Yates and Poblete (2006)] Ricardo B aeza-Yates and Barbara Poblet, A Website Mining Model Centered on User Queries, in Semantics, Web and Mining, EWMF/KDO 2005, Springer LNAI-4289, 2006, pp. 1-17.

[7] [Balabanovic and Shoham (1997)] Marko Balabanovic and Yoav Shoham, Fab: content-based, collaborative recommendation, in Communications of the ACM, Volume 40, Number 3, March 1997, pp. 66-72.

[8] [Baxendale (1958)] Baxendale Phyllis, Man-made index for technical literature - An Experiment, IBM Journal of Research and Development, 1958.

[9] [Begelman et. al. (2006)] Grigory Begelman, Philipp Keller and Frank Smadja, Automated Tag Clustering: Improving search and exploration in the tag space, WWW-2006, May 2006, pp. 22-26.

[10] [Berners-Lee (1998)] Tim Berners-Lee, Cool URIs don't change, 1998.

[11] Available at - http://www.w3.org/Provider/Style/URI Last Cited on - 27-April-2011.

[12] [Berners-Lee (2001)] Tim Berners-Lee, James Hendler and Ora Lassila, The Semantic Web, in The Scientific American Magazine, May 17, 2001. Available at - http://www.scientificamerican.com/article.cfm?id=the-semantic-web Last Cited on - 27-April-2011.

International Advance Journal of Engineering, Science and Management (IAJESM)
Multidisciplinary, Indexed, Double-Blind, Open Access, Peer-Reviewed, Refereed-International Journal, Impact factor (SJIF) = 7.938

[13] [Billsus and Pazzani (1998)] Daniel Billsus and Michael J. Pazzani, Learning Collaborative Information Filters, in Proceedings of the Fifteenth International Conference on Machine Learning, ICML-1998, pp. 46-54.
[14] [Bodenreider (2001)] Olivier B odenreider and Anita Burgun, Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System, NAACL'2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations, 2001, pp. 77-82.
[15] [Boer and Bosselaers (1993)] Bert den Boer and Antoon Bosselaers, Collisions for the Compression Function of MD5, Springer, pp. 293-304.