INVESTIGATION ON SERVICES AND TOOLS FOR WEB SECURITY

Santosh Kumar[1, 2] and Sudhanshu Nigam [3]

[1] Ph.D. Research Scholar, Department of Computer Science and Engineering, School of Engineering and Technology, Shridhar University, Pilani-Chirawa Road, Pilani, Rajasthan- 333031 (India).

[2] Assistant Professor, Department of Computer Science, Lucknow Public College of Professional Studies, Vinamra Khand, Gomti Nagar, Lucknow - 226010 (India).

[3] Associate Professor, Department of Computer Science and Engineering, School of Engineering and Technology, Shridhar University, Pilani-Chirawa Road, Pilani, Rajasthan- 333031 (India).

ABSTRACT: Web usage mining is a type of web data mining that extracts information about users of the web and how they interact with the site. Our goal with web usage mining is to provide web intelligence by automatically and quickly identifying users' web page access patterns, such as frequently visited hyperlinks, frequently accessed web pages, popular web page access sequences, and user groupings, among other things, from massive server access log records. Through web usage mining, we can extract server log data, user registration information, and other related information left by users. This data will be used to start an analysis, which will help the business make decisions and change web pages.

KEYWORDS: DATA MINING, WEB USAGE MINING, WEB PERSONALIZATION, CLUSTERING, CLASSIFICATION. ASSOCIATION

INTRODUCTION: Millions of users engage with websites on a regular basis, visiting a range of sites and leaving a variety of information behind. Web use mining is the process of using this information by website administrators to make their websites better based on what people want to see. Online mining is the process of extracting noticeable and potentially lucrative patterns and implicit information from web site traffic. In the future, this data could be used to make the web more useful. For example, it could be used to predict which website a user will visit next, detect crime in the future, profile users, and understand how people use the web. Web usage mining is necessary to process web data in order to forecast and identify information that has been accessed. Web usage mining is a type of web mining that helps to automatically figure out how people use the web.

BACKGROUND: As we all know, the internet has become the most important way to share, gather, and spread information because of its rapid growth, growing popularity, and easy access to its content. There are a lot of businesses and organisations that use the Internet to give people information and services like computerised customer service, online shopping, and a lot of other resources and apps.Web-based applications and environments for e-commerce, e-banking, distance education, online shopping, online collaboration, news broadcasts, and other purposes are becoming increasingly popular throughout the world. The World Wide Web is becoming a popular tool for ordinary people's daily activities. [1]. To make websites more adaptable and personalized, researchers have used data from web access logs. This means that the websites are more appealing and interesting to people who visit.It is important to keep customers loyal, though. The history of how people have accessed web pages is used to improve web pages in both structure and content. Web usage analysis has been the subject of whole conferences and workshops, mostly for the benefit of e-commerce. The comparison to e-commerce seems simple, but it's not. It's not as easy as it looks. Fact: In e-commerce, the goal is to increase sales and profit by predicting customer access behaviour. In e-learning, the goal might be to improve the web experience of visitors, and this could also be done by predicting how visitors access the web. However, there are a lot of fundamentally different things going on. For example, a session, which is a building block for almost all web usage mining algorithms, is defined by how long it takes for someone to go from visiting the site for the first time to making a purchase or order (i.e., the same access session).

PROPOSED TECHNIQUES: Preprocessing, pattern discovery tools, and pattern analysis tools are three main types of tools which help in finding the patterns in web access log data [2].

PREPROCESSING: Preprocessing of server access log data is a very important task in the mining process. Data is preprocessed in order to improve. Preprocessing the data from the server's access log is a very important part of the mining process. In order to improve the quality of the data and the speed and ease of the mining process, the data is first preprocessed. It's a little difficult and takes a long time. Data cleaning, user identification, session identification, and path completion are some of the things that need to be done.

DATA CLEANING: There are a lot of different ways to clean up data, but the main goal is to get rid of useless data that could hurt the mining process. Helmy and other people came up with a simple algorithm that removes any file extensions from the target URL, like gif, pdf, jpg, and so on. With this algorithm, the data that isn't needed is removed, and the mining process then looks at the results a lot faster.

USER IDENTIFICATION: Clean the HTTP log file. The next step is to figure out who the user is. A user is a person or a client who interacts with the server to get and show resources. Because there are proxy servers, it is hard to figure out who a person is. Web mining is used to solve these problems. However, security and privacy make it hard to do this. Heuristics help you figure out who the user is.

i) Every individual has a distinct IP address.

ii) It's also possible for an IP address to be the same for more than one log entry but for a browser and operating system to change. This means that the combination of IP address, browser, and OS is different for each person.

iii) The URL and referrer fields of the web access log can be used to build a site topology that shows the browsing path for each user. The user's IP address is shown if a page in the browsing sequence isn't directly linked to the page the user was on. This shows that there is another person with that IP address.

SESSION IDENTIFICATION: Once a person has been identified, the click stream is broken up into different groups. Sessionization is the name for this way of splitting things up.

1). Sessions are Web pages that are visited by the same person at the same time. If a person has a lot of different sessions, they can be called "users." This is how we find out what a user is doing at a certain time.

2) If the page the user is looking at in a user session is blank, it means that there will be a new session.

3) If the page the user is looking at in a user session exceeds a certain time limit (around 30 to 25 minutes), it means that the user is starting another session [3].

Because there are proxy servers, local caches, and corporate firewalls, there is a lot of important information that isn't in the web access log data. Path completion is a way to add lost page references to the server access log file.

Users can be identified in the same way that they can be identified for path completion. Web users who act in an unusual way aren't going to need to go through this step. We only want to know if there was any web crime, attack, or suspicious user activity that we could discover using logs from our web servers. On the whole, this step is about how the site looks.

PATTERN DISCOVERY TOOLS: After figuring out what users do, there are a lot of different ways to look for patterns in web log data using mining techniques. There are web logs that show which users are suspected and which web attacks have been carried out. After converting web logs into a relational database, some operations like classification, clustering, and sequence analysis are done on the data to look for patterns in the data [4]. An association rule is used to figure out how web pages that people visit on a website are linked together. An association rule can be used to connect the web pages that are most popular within a single server session to each other. Algorithms like Apriori and Eclat, as well as frequent pattern trees, can be used to find association rules. Each row is for an item, and each column is for a transaction. In this matrix, a bit is set if the item on the row is in the transaction on the column. Otherwise, it's empty, and there are no bits. An algorithm called FP Growth doesn't work well for a large database because Eclat and FP Growth don't work well with that kind of data. Putting things together into groups is called "clustering." Each group is made up of things that are related in some way to each other. Members of one cluster are more like each other than they are like members of another cluster. Clustering can also be used to look for crimes. Afterwards, we might find that some cases don't fit well with any clusters at all. Pattern analysis is used to look for signs that someone who isn't supposed to be there has been a certain link and then to another link in an orderly set of sessions. With this method, we can figure out the suspected user's mental state, which helps us find out about crimes. Apriori All, GSP, SPADE, Prefix Span, and Spam are some of the algorithms that can be used to look for sequential patterns in a group of words. For pattern analysis, the best algorithms are GSP and PrefixSpan, which are both apriori-based and pattern growth-based, which are the two main ways to solve the problem. [5]

THE PATTERN ANALYSIS TOOL: Now, pattern analysis is the last step in the web usage mining method. Its main job is to look for a good model that can be used. According to the document, the most recent visit,

- Who is looking at which document?
- The frequency of use of each hyperlink, and
- The most recent use of each hyperlink

People use visualisation techniques, OLAP (online analytical processing) techniques, data and knowledge querying, and usability analysis to look for patterns in their data and make sure they work well. The visualisation technique is a very good way to help people understand a wide range of things, both real and abstract. It is a way to show data in a picture, so it is a good way to study how people use the web. The OLAP (On-line Analytical Processing) Method For strategic analysis of big data in a big business setting, it is becoming more and more important to use this powerful tool. Some of the main characteristics of strategic analysis are: a very large amount of data; support for different types of information aggregation; and long-term analysis in which overall trends are more important than the specifics of individual data items. This is because the relational database system has a high-level, declarative query language that makes it very good at what it does. This query language lets a person or an application say what must happen for the data they want to get, rather than how to get the data. If there are a lot of patterns in the data, the relational database isn't going to work. It looks like there is a need for a device that can analyse these patterns by mining the big data. For mining big data, first apply constraints to the big data to limit the data and get the data you want. Then, use the mining process to find out what you need to know. As a second step, questions can be asked about the information that has been found during the mining process.

USABILITY ANALYSIS: The first step in this process is to figure out how to get the data you need. Then, with this data, it creates automated models and simulations that show how the data came to be in the first place. In the end, to help an analyst figure out the trend from the data, a lot of different ways of presenting or displaying data are used. Web users' browsing habits can also be shown through this method, but many people don't like them because they're slow, hard to keep up with, and don't have a lot of features. It still needs to be done by both the researcher and the person who makes the tools to make this task more efficient, flexible, and powerful.

FRAMEWORK: It's easy to find simple and simple information from web log data, but it's hard to find complex structural information that can help you learn. It is very important to clean and prepare the data before using any mining method, which is called preprocessing. Many things can go wrong when you're preprocessing your data. For example, you have to get rid of things like image files, crawlers or spiders, and failed requests. After all the preprocessing is done, the relational database is ready to use the data mining techniques. Web usage mining has a general structure that breaks down the whole process into two main parts. Part one is made up of domain-specific processes that convert server access log data into a format that can be used by users. In this process, things like data preprocessing, session identification, and the integration of relevant data are all done. Part two focuses on the use of common data mining and pattern matching techniques that aren't specific to a single field, like the use of association rule techniques and sequential pattern analysis. This is a branch of the system's web mining engine. The whole process of web mining is shown in the figure below.

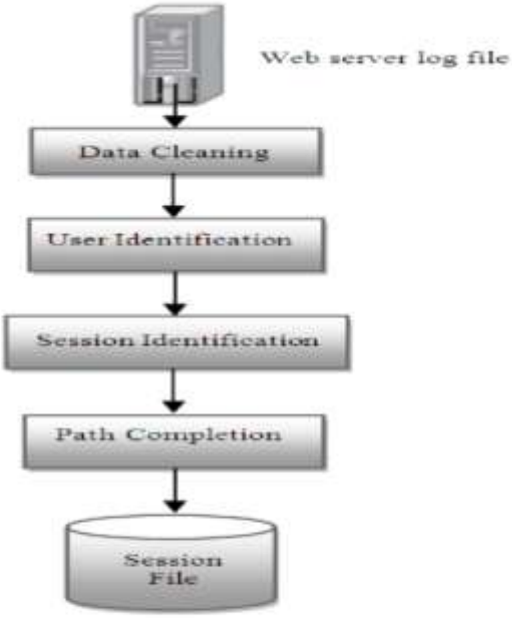

Figure 1: WUM's architecture

In web usage mining, the first thing that is done is to clean up the data. There are a lot of ways to clean up Web log data before you start processing it. Currently, the given system uses a very simple way to check filename suffixes. At this point, some low-level data integration tasks may also be completed. For example, merging several log files into one table, combining referrer logs, and so on [6]. Afterward, web mining techniques need to divide log entries into groups called clusters by one or more user session ID modules. It's possible to clean up a web access log file in two ways: first, in a single session with a lot of web page references, and second, in a set of many sessions, each with a single web page reference. It is the goal of session identification to make it easier for each person to find web pages that make sense to them. So, session identification is a job that either breaks up a big session into smaller ones or makes smaller sessions into a few bigger ones. This process can be broken down into many steps of combining or splitting in order to build sessions that work for a specific web usage mining process. Identifying a user's session can be done by either merging or dividing the session into separate parts. It takes a transaction record, which may have a few other things in it, like i/p.

Then a module does something with the record, and the o/p is a list of transactions from the record that the module made. A number of modules can be put together in any order, as long as the analyst can make sure that the input and output transaction formats for both are the same. The given system has modules for web page reference length and maximum forward reference, time window divide modules, a time window merge module, and a lot of other modules.

The web access log file is not the only thing that can be used to do mining. For example, user registration data plays a very important role in user identification and in safety and privacy alerts. It's also possible for client-side apps to limit server access to a certain range. It must then be added to the web access log file. There are also known or discovered things about reference pages that could be used to make a more detailed database. Such attributes could include the types of pages, their classifications, how often they are used, their meta information, and how they link to other pages. It isn't yet possible to see how a person registered with the system, but in this framework for Web usage mining, many data integration problems are also solved. The domain-dependent data transformation phase calculates transaction data that needs to be changed in a way that fits the data model of the right web-mining task. To give an example, the format of the data to which an association rule is applied may be different from the format required for web mining, which is how the process works.

Finally, a query mechanism will allow the analyst to have more control over the discovery process by setting certain conditions. Because of all of the new web mining tools and systems out there, we need a powerful query language that can be used for data mining. This language can then be used to make many interactive and flexible graphical user interfaces. It can give the user more control over the data mining process and help the user find only useful rules. By adding some basic things to a SQL-like language, WEBMINER makes it easy to search for things. This lets the user give the mining engine a direction by telling it what patterns they want to look for.

BENEFITS: This tool has given e-commerce the chance to make their websites more personalised in marketing, which leads to more sales [7].There are a lot of government agencies that use this tool to find threats and fight terrorism. Mining applications can help society by predicting what criminals will do. A good and healthy relationship can be built by giving the customer the information they want. Companies can better understand the needs of their customers and act more quickly to meet those needs, which mean they can better serve their customers. Companies can keep customers by giving them information that is relevant to them.

LIMITATIONS: When this technology is employed on personal data that could be a source of concern, privacy is considered lost [8]. The information gathered by companies may be used for other purposes, resulting in a breach of users' privacy.

CONCLUSION: Web Usage Mining is a new field of study that is quickly becoming more and more important. This paper gives it a more in-depth look. Useful information about how people use the web needs to be analysed in order to better understand how people use the web and then apply this knowledge to better serve the needs of users. There are so many web-based apps in the world, especially in e-commerce, so this is why. Research paper: In this paper, we tried to make it clear how data is put together and how knowledge is found.

FUTURE SCOPE: Much structured and unstructured data is being shared over the web because more people are using the World Wide Web (www). The reason for this is that people from all walks of life are using the web to share information with each other. Many things, like billboards and patents, can make it hard to find what you want. When you study web mining, you look at how people use the web and what they do with it to learn more about what they like and how web applications can better serve them. Web usage mining (WUM) is one of the main types of web mining research. Finding patterns, analysing patterns, and processing patterns are three steps in web mining. This study looks at web use patterns that can be found in web data. Then, it only shows the information that the user wants to see. This makes for an intelligent, semantics-based web usage mining method [9]. Usually, the Page Rank of a web page is based on how many times the keywords appear in it. This can be used to give websites extra attention by including the keywords again and again, which falsely raises their Page Rank. To make things even better, it's not just important to pay attention to the number of keywords on a web page, it's also important to pay attention to the content of the web pages in terms of how semantics are used [10]. Another way to make web usage mining more intelligent is to add semantic information from domain data to all stages of the process [11]. All of this is done to build an intelligent web usage mining architecture that can find patterns and analyse patterns.

REFERENCES:

[1] Osmar R. Za¨ıane Department of Computing Science,” Web Usage Mining for a Better Web-Based Learning Environment”, University of Alberta Edmonton, Alberta, Canada, 2001

[2] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, “Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data”, 2000

[3] Li Chaofeng School of Management, “Research and Development of Data Preprocessing in Web Usage,” South-Central University for Nationalities, Wuhan 430074, P.R. China, 2005

[4] Amit Pratap Singh, Research Scholar, Dr. R. C. Jain Director Samrat Ashok Technical Institute, Vidisha, Madhya Pradesh, India, “A Survey on Different Phases of Web Usage Mining for Anomaly User Behavior Investigation”, Barkatullah University, Bhopal, M.P, India, 2014

[5] Monika Dhandi, Rajesh Kumar Chakrawarti. "A comprehensive study of web usage mining", 2016 Symposium on Colossal Data Analysis and Networking (CDAN), 2016

[6] Sarita Dalmia, “Wed Mining: survey and Research”, 2016

[7] Anitha Talakokkula Department of Computer Science and Engineering,” A Survey on Web Usage Mining, Applications and Tools”, Stanley College of Engineering and Technology Hyderabad, 2015

[8] Yan Wang, “Web Mining and Knowledge Discovery of Usage Patterns, 2000

[9] Sudheer reddy K, Dep. Of CSE, "Understanding the scope of web usage mining & applications of web data usage patterns" Acharya Nagarjuna Univ., Guntur, India, 2016

[10] Thi Thanh Sang Nguyen, Hai Yan Lu, and Jie Lu," Web-Page Recommendation Based on Web usage and Domain Knowledge", 2014

[11] Raymond kosala, Hendrik blockeel,"Web mining research: A Survey", 2000