Web Mining Techniques and Technologies: A Landscape View

1 Kareem K. Ibrahim - 2 Ahmed J. Obaid
1,2 University of Kufa College of Computer Science and Mathematics Department of Computer Science

karimk.aljabri@student.uokufa.edu.iq, ahmedj.aljanaby@uokufa.edu.iq

Abstract: Web mining is the process of discovering important information and patterns in a weblog file. And contains loads of information. It consists of pages of various formats (HTML documents, images, etc.). And the data for this record is still increasing at a very high rate. So the process of retrieving information from it is a very difficult task. In order to obtain the required information from the network. This requires mining the weblog file. In this paper, the researcher conducted a review of previous studies and what researchers have dealt with in web prospecting. Explain web mining concepts. And how to obtain information and the visitors' patterns through the techniques and algorithms used in Web usage mining.

Keywords: Web Usage Mining, website, Web mining.

1. Introduction

The data On the Internet is enormous and is growing every day. The website is a cluster of web pages. Text, images, and videos may be included on web pages. And its they are connected through hyperlinks through which navigation takes place. Log files are created whenever users visit any website. The log file documents the full information about each user's access to the website. The size of log files is growing day by day because of the increased use of web pages[1]. The method for finding confidential information from a weblog file is named Web Mining. It is a branch of data mining. The purpose of it is to obtain information about the navigational activity and recall useful information from very enormous raw data, which can be expressed by several millions of event records in a log file. Weblog data includes various kinds of information so including weblog, web layout, and user profiles.[2] Web mining has been classified into three large topics of focus depend on which part with Web to be exploited which include: Web content mining, Web structure mining, and Web usage mining. The Previous three mining jobs can be used separately or can be combined with other jobs as they may contain web document links. Web mining is classified under data mining technology and its main purpose is to retrieve and extract knowledge from many documents and web services mechanically. As well to extract the required and exciting patterns from a series of huge data sets in the latest trend as well as used for traditional data mining. [3]

Web content mining its mining of this type is a very important area in analyzing users and the behavior of their web file content. Web content may contain very large amounts of multimedia data (images, video, animation, text, and document data in different styles such as pdf, etc., hypertext).
WCM is also used with a focus on visitor opinion extraction and sentiment analysis. Through which it is possible to locate comments in social media, blogs, and web forums. This type of mining is also used in education by extracting valuable information. Such as identifying consumer needs and etc. WCM is text mining and user opinion. It is one of the techniques that can be used to find important information from textual data. [4] .

Web structure mining this type of web server mining uses a web file connection structure diagram and analysis of nodes in that file. The primary goal of mining the web structure is to extract the previously lost relationship between the web file pages. It is also used to generate a structural summary of the website and any page of the internet. It is also a mining method that will categorize web pages and show relationships between other websites. This technique is implemented either on a single page or on multiple pages (hyperlink). This technique produces structural information from a web server file consisting of a web graph with pages as nodes and hyperlink as edges [5].

Web Usage Mining this type of mining focuses on customer knowledge and their collaboration with websites. This type is very important in extracting important information that explains the visitor's use of web pages [6]. That is, it is the process of discovering very important patterns of web usage data. Such as visitor access patterns and various other requests made by one or more users. The main objective of WUM is to understand navigation and browse through web pages. This is very important in enhancing the quality of services provided by websites. Business services, in the field of information and network security, fraud detection. And improve the structure and performance of web pages and servers [2].
2. Related Work

In this section, we uncover relevant work related to daily web use for big data results from web usage exploration. Which is a popular research area in prospecting for web usage, data analysis is essential to track user behavior in order to effectively serve users. Either in terms of security or website development and others.

According to [1] The authors described mining on the web as the application of several techniques to extract data from web records. This log contains large amounts of data, so reprocessing is required to remove useless data. Two stages of pre-processing were proposed to clean up the data and identify the user using the proposed algorithms. A test was performed on a log file consisting of 500 records, after data cleaning it found 441 record. Note the difference is not a little because it URLs used are few compared to the size of big the weblog file. And this is the record used doesn't contain much needless content like jpeg, video, etc. After the user identification stage, it found 52 unique visitors out of 441 URL user records.
G. Shiva Prasad et al. [7] suggested that the neuro fuzzy typical collecting bins the neural network and the customary (fuzzy notion). Where they used neuro based hybrid typical is effected to determine concealed patterns in the Web Log server special of (polytechnic web site). Web Log Pre-processing methods based on dimensionality reduction methods and collective methodologies were employed. The preprocessing phase eliminates all irrelevant, and blaring data, with a resultant Web Log size of 20% of the original log size. The (neuro fuzzy) grouped the users taking the same browsing patterns into set clusters. And the knowledge that has been collected after the analysis will can after that used from before the website for effective management and Personalize their website. This model was used to discover deviations in user behavior in many applications that require high security and high data privacy.

According to [8] the authors described the web mining process in three imperative types (web content mining, website architecture mining, and web usage mining) that assist users to find valuable data. Each category has different algorithms and techniques needed to retrieve the information used in some applications such as fraud detection etc. Web content mining is helpful in terms of discovering the data from the (images, table, text, etc.). Web structure mining categorizes dealings between related Internet pages. Web usage mining is a too imperative kind which stores the user entree data and acquires info about a candid user of records. everybody methods may have some benefits and drawbacks. But they can be drawbacks better by additional studies.

R. Roy et al. [9] they conducted a comprehensive survey of mining methods on the web log document initial with weblog data sources which contain of only a text file. The information stored in the web log consists of a huge amount of useless data and noise, so this huge amount of data requires a pre-processing process for this record in a distinctive way to identify the problems in the web log document and also clarified the process of identifying the customer through (Client Identification IP address, by cookies, user data, site topology, and by authentication information).

According to [10] the authors presented the technique of Web Usage Mining containing phases: Data Collection, Pre-processing, Pattern Discovery, and Pattern Analysis. They suggested a web history pre-processing algorithm where each page was assigned a specific token. According to this symbol and frequency, a technique can be applied to extract data (Classification, Association Rules, and Clustering) and reach that the highest and lowest value can be detected according to the frequency of accessing the page. Can simply learn the highest and lowest value.

Several researchers are shown in [11]. They used a method to improve the prediction of the next webpage Based on visited web pages by assigning top web browsing profiles for interested visitors. And it is recommended for current users interested in similar sites visited. They used a window sliding method with size N over the navigation session. Using the CTI dataset, The experimental results show a higher prediction accuracy for the pages of the next visit.

C. E. Dinuca et al. [12] Had explained that the session has a very important role in the reprocessing process. Also, most session delimitation algorithms use constant values to define the end of the session. He mentioned why using constant values causes errors in defining sessions. They propose a new method for identifying sessions based on average visit time for web pages that they implemented in Java programming language using NetBeans, IDE, which are two algorithms for identifying sessions. The former uses a constant value of 30 minutes to indicate the end of the session and the second using the
average time spent by users on their website pages. Was chosen the NASA log file available online to apply the suggestion.

Smriti Pandya et al. [13] They proposed a system interested in investigating serial mining techniques for access patterns with high efficiency and effectivity for web usage data and then using the pattern mined to match and create web links across the Internet. This usage data affords the trails leading to edited Web Pages Where this system has been taken advantage of in predicting the needs of the user and thus the use has been improved through the appropriate guidance of the web visitor to the important pages that are similar to the sites he is trying to reach.

Mr. Jitendra and et al. [14] presented a study showing reviews of different data to several researcher's processing methods. Such as the compilation and cleaning of data, Identification of users., route completion, and session determination. In addition to the advantages and disadvantages of these methods, which will be beneficial to society. To choose one or more of the available techniques for an effective pre-treatment. For more specific and consistent outcomes. As well proposed a complete preprocessing method that will enable analysts to convert any web server record set into an organization file (table, text) in the database. After comparing them with the preprocessing techniques used by other researchers, it showed more accurate results and Minimum preprocessing log file size.

Sucheta V. Kolekar et al. [15] they Submitted a proposal for the user's learning styles, by describing the learning behavior in the e-learning portal via Web Log Mining. The learning styles are then assigned to FSLSM classes. Provided that Each class is provided a learner with the appropriate contents and interface for the group He used the algorithm Fuzzy C Means (FCM) is used for collect educational, behavioral data captured from FSLSM classes. This method processes previous web data and converts it into (FCM) format, then defines unique sequences for each learner according to their sessions. Where it divides the sequences into eight categories based on the learning objects of the class, the work describes a methodology to automatically detect and identify the learning styles of learners using the Web Log Analysis approach where the validity of the algorithm was verified and compared with other, and the result was shown that it performs better in providing the contents and the adaptive interface for a new learner.

K.SELLAMY et al. [16] they provided a preliminary study of web mining, from several aspects (tools, techniques, applications). They proposed an approach to analyze and compare the skills taught in university or other studies, and the skills required at work. The aim of the study is to apply the latest methods of web data mining in the field of education and work.

The authors in the paper. [17] presented they a two algorithm that analyzes visitor behavior, And preprocessing the website visitors, Where Used the (HDD and CFPMA) algorithms, the first is used to filter and organize the appropriate information. And the second to mining the repeating pattern(CFPMA) to obtain recurring patterns from weblog data. In the present algorithms, repeated patterns mining is performed depending on the support threshold limit so that the accuracy is low, and the implementation period is high. But with this method. Based on the conviction threshold value, the session, and clustering in the weblog file. Focused the job on the weblog file format, pre-processing, and assembly, and pattern creation techniques.
The authors in the paper [18] presented a new approach to web usage mining using (Case Based Reasoning), by using clustering algorithms for titular web data. And it is founded on the reprocessing of previous labor experiments. Thus, the previous experience could be an effective guide for solving new problems. Web customization which has the ability to adapt the next set of pages visited with individual users according to their interests and navigational behaviors has been proposed. A number of components of the proposed architecture are, Basic log preprocessing, methods of pattern discovery (By Reasoning Based on Case and similarity from peer to peer, Clustering, Association Rules Methods for Mining). In this study, a new predictive approach was introduced using CBR Method and mining of web use, which gives the user a choice to navigate to the next page. This technique gave high results compared with previous methods.

The authors in the paper [19]. Very big information the internet and tremendous resources. It has become imperative to determine the behaviors of website visitors through web history mining. To improve website performance and predict which page the user is likely to visit. And by searching, first in the similar cluster, then in similar sessions in the same cluster. This, in turn, causes an increase in the time of comparison with previous sessions when the data block is very large. In the study, they proposed three strategies (Least Frequent Ones Leave, First come First Leave, Timeframe Leave) for disposing of old sessions in order to replace them with new sessions (candidate sessions for prediction). When implementing the proposed methodology, the result was lower at the time of comparison and higher prediction accuracy.

| No. | Year | Authors | Technologies used | Focused on | Remarks |
|-----|------|---------|-------------------|------------|---------|
| 1   | 2014 | P. Verma and et al. [1] | Two stages of preprocessing were proposed to clean up the data and identify the user using the proposed algorithms. | Data cleaning User identification | After completing the preprocessing stages, patterns will be detected that help in inferring user access patterns. |
| 2   | 2015 | G. Shiva Prasad et al. [5] | A hybrid model by on neural noise is used to discover knowledge in weblogs. | Neuro-fuzzy model to discover hidden patterns in the weblog. | This model was used to discover deviations in user behavior in many applications that require high security and high data privacy. |
| 3   | 2018 | Hamid et al. [6] | The three types of web mining (Structure, Content, Usage). | Web usage mining (tools, techniques). | Web exploration is very useful (explore data, find out about visitor patterns ...). |
| 4   | 2019 | Roy et al. [7] | Use detailed preprocessing to improve the quality and effectiveness of weblog data. | Transaction Identification. | Pre-processing of the log document is important to improve the productivity and ability to retrieve the log document information. |
|   | Year | Author(s)                          | Methodology and Technology | Description                                                                 |
|---|------|-----------------------------------|-----------------------------|-----------------------------------------------------------------------------|
|5  | 2018 | Mehra and et al. [8]              | Use web usage mining technology with web history pre-processing algorithm.| Implemented the algorithm in java programming language. Be funded that most popular page and least popular page. |
|6  | 2011 | Y. Almurtadha et al. [9]          | They used the N-size window sliding method on the mobility session using a CTI dataset. | Improved prediction for the next visited web Pages is recommended for the current anonymous user. A system with higher predictive accuracy was found on the following pages compared to previous systems. |
|7  | 2012 | Dinuca et al. [10]                | Using (NetBeans, IDE) which are two algorithms for identifying sessions. | It proposed a new method for identifying sessions based on average visit time for web pages that they implemented in the Java programming language. In the first algorithm, they used 30 minutes for the session. In the second algorithm, the time depends on the page because it used the average visit. |
|8  | 2015 | Pandya et al. [11]                | They proposed a system interested in investigating serial mining techniques for efficient and effective access patterns. | Mainly focused on investigating the accuracy and effectiveness of mining technology for web usage data. The resulting patterns are used to match web links to create links to recommendations. |
|9  | 2015 | Jitendra et al. [12]              | Use preprocessing technology that enables users to convert a web server into a database in the form of an organized table or text. | Pre-processing methods consisting of data collection, cleaning, visitor identification, session, path, etc. The web server contains irrelevant data, which is important to preprocess. |
|10 | 2017 | Kolekar et al. [13]               | They used the (FCM) algorithm to collect the educational behavior data and the (GSBPNN) algorithm to predict the learning patterns of the visitor. | This method focused on processing previous web data and converting it into a format (FCM), then defining the sessions for each learner based on their sessions. Description of a methodology for automatic detection and identification of learning patterns through web history analysis. |
|11 | 2018 | Sellamy et al. [14]               | Use web server data mining algorithms and techniques. | The study focused on extracting web data from several aspects (techniques, tools, applications). The main goal of the study is to understand and apply web prospecting and prospecting for education and employment information. |
|12 | 2018 | El-Aziz et al. [15]               | Presented they a two algorithm that analyzes visitor behavior, And focused the job on the weblog file format, pre-processing, and assembly, and pattern. | The algorithm (HDD) is shown to filter and organize the appropriate information, either |
12. 2020. Preprocessing the visitors.

13. 2020. Asadianfam et al. [16]. Introduce a new predictive method, using (CBR) to explore web usage mining. (CFPMA) is used to obtain repeated data patterns.

14. 2019. P. M. Bharti et al. [19]. -Least Frequent Ones out
    -First Come First out,
    -Timeframe Leave. The study focused on getting rid of previous sessions.

3- WEB USAGE MINING

As the number of internet users increased, the use of websites increased to get the required information. This, in turn, led to more data usage of these sites. This data is stored in various formats in the weblog file in an unordered format. And to understand the contents and user patterns of these records. The web history is mined. One of the necessary mining methods is web usage mining. This method consists of serial stages [20].

**Fig (2) Stages of WUM:[6]**

1) Data Collection:
The web history data is in the form of a document file. The information is stored in various types of documents. It is available in the following:[6]
A- Web server log file: These logs contain server-side data. This data consists of the IP address, URL, byte count, etc. In most cases, this data is displayed in a standard configuration. The most prominent of them is the Common Log Format CLF [6].

B- Web Proxy Server Log file:

It is an average web server that is between the client and web server. When the webserver receives the visitor request through the proxy server. The log file entries will be the proxy server information and not the original client information. Proxy servers keep a separate record for all information user [20].

C- User Browser Log file:

Browser logs are collected from the devices of visitors who access the websites. This user data is found by proxies that are imposed by some applications such as (JavaScript programs). Attached to a web page is used to take information from the visitor as a record of their navigation. This data requires the cooperation of the customer, who most often restricts the operation of (JavaScript programs) for security reasons of the user [6].

2) Data Pre-Processing:

It is a very important step in web usage mining. After gathering huge amounts of weblog data from various sources it appears to contain a lot of outliers that must be removed. To obtain consistent and integrated data for use in later stages (pattern discovering, pattern analysis), stages on preprocessing the as follows: [6]

- Data Cleaning

The log file contains a lot of non-essential records that are not related to our work. Like server error messages. Specified by the status code that the server sends when the visitor requests certain content, a field is recorded for each state of the record on the web. Also, the entries that contain graphics files, images, and others of the extension (jpeg, jpg, gif) are deleted. All codes greater than 299 and less than 200 are removed. Because it is considered invalid and is removed from the log file as in fig (3). Also, deleted records that were previously browsed are removed. Browsing records for the main pages are also deleted. This is because their links are found in most web server logs [2].

| Status Code | Code Range |
|-------------|------------|
| SUCCESS     | 200 series |
| REDIRECT    | 300 series |
| FAILURE     | 400 series |
| SERVER      | 500 series |

Fig (3) Classes of status code.

- User and Session Identification:

It is not important to know the identity of the visitor. But there is a need to uncover and characterize visitor behavior. The server records multiple client sessions. And as a visitor can visit certain sites frequently. This is done without any authentication mechanisms. In many web servers. Because some
users disable the cookies feature. View its privacy content. This results in the IP address alone is insufficient to identify the unique visitor. Therefore, other criteria are used with the IP address such as user agent and referrer [2]. As for defining the user's session, define it. A group of pages visited by the same visitor in a certain period of time. On some web servers the session time is limited to 30 minutes. After this period, the second session begins. It is possible for a single visitor to have one or more individual sessions on the same page [21]. Visitor activities are numerous and controlling them is important in facing many issues, such as detecting visitor behavior and others. Among the most important of these activities are the following [22].

**Table (2) Visitor activities**

| NO. | Activity                          | Clarify the activity                                                                 | Example                      |
|-----|---------------------------------|-------------------------------------------------------------------------------------|------------------------------|
| 1   | The most active visitors identified | The visitors with the highest activity are limited by IP address, number of visits, and country. | 192.168.0.110 IRAQ NO. V 13 |
| 2   | Time to spend a visitor          | It represents the time that the visitor spent at a particular site and can be represented by a diagram. |                              |
| 3   | Daily visitor activity           | It shows the differences in visitor traffic on the site during a specified period. It is measured in bytes transferred to each user | Sunday 15 Hits 200 KB        |
| 4   | Hourly transfer rate             | It shows the amount of data transferred in units of time                              | Hours of day Hit D. T        |
|     |                                 |                                                                                      | 15 35 244 KB                 |
| 5   | Number of visits per visitor     | It shows the number of visits of each visitor and deduces from it what the visitor’s interests are in certain sites without others. | NO. of visitors’ NO. visitors Visits |
|     |                                 |                                                                                      | 3 45                         |
| 6   | Countries with the most visits   | Most countries ranked according to the number of unique visitors.                      | Country Unique V             |
|     |                                 |                                                                                      | United stated 23             |
| 7   | Browsers                        | It ranks your visitors according to the number of hits per browser. It also shows the byte size transferred from each browser. | Firefox No. Hits D. T        |
|     |                                 |                                                                                      | 78 200KB                     |
| 8   | Operating system                 | A visitor’s entries is determined by the number of arrival the O.S is installed on the visitor's | Windows 10, No. Hits D. T    |
|     |                                 |                                                                                      | 66 205KB                     |
computer. And the byte moved in the session.

### Table (3) User identification methodologies:

| methodologies    | Description of the methodology                                                                 | Privacy risks | Advantages                       | Disadvantages                                                                 |
|------------------|------------------------------------------------------------------------------------------------|---------------|----------------------------------|-------------------------------------------------------------------------------|
| IP address & Agent | Assume when IP add & agent is unique this signifies that the user is unique                  | LOW           | Available without technology     | There is no guarantee that this user is unique                                |
| Embedded session ID | The dynamically created pages are used to link identifiers in the hyperlink                  | Low to medium | Since it is independent of IP address this makes it always available | Difficulty detecting duplicate users                                          |
| Registration      | The user enters the site directly                                                                | Medium        | It can track the visitor and the browser | Visitors cannot register if it is not available                               |
| Cookies           | By saving the identifier on the visitor's machine                                               | Medium to high| Ease of tracking serial visits if it is for the same browser | It may be turned off by the user                                               |
| Software agent    | By downloading the program to the browser then it sends the used data                           | High          | The data is accurate as it is for a specific location | It may not be accepted by the visitor                                          |

The process of controlling access resources and website security. Figure (4) shows the highly accessed or downloaded files arranged from highest to lowest. This is done by specifying the following [22].

- Define files with higher access
- The facts that have been reached
- Search engines
- Referring websites
- Security alert
Fig (4) The most important directories reached.

- Path Completion
  It is an important step in pre-treatment. Mostly it takes place after completing the session. The agent or client is mostly due to temporary caching due to the loss of access references for some pages [23]. Also when the real URLs are more than recorded in the server log. This indicates a loss of access to references to these pages. It is possible to reveal this when a visitor requests a specific page that is not related to the previous page (the previous request) for the same visitor. The referrer can refer to the register to find out which page the above application contains. If the missing page is in the visitor's last click log. It is a page not registered in the registry. This state indicates that the visitor browsed again using the (Back) button [20].

3) Pattern Discovery
   It is one of the most important parts of Web Usage Mining extracting data from the web history. After performing the data cleaning process, selecting the user and the session. The main goal of this part is to discover interesting patterns [8].
   
   A – Statistics:
   It is an important technique for finding useful information for any web history. And to know the content of the web history and the number of visits to clients in that record. The number of visits is calculated on the basis of each valid entry in the weblog. These restrictions may be posting, browsing, downloading. Because it helps improve system performance. Such as monitoring visitor activities, monitoring and checking pages and sites, and aggregating visitors based on their behavior [24].

   B - Association rule:
   This technique is used to find recurring rules and patterns in the data generated from the preprocessing stage of the weblog data. Such as the number of users' frequent visits to certain pages. The task of this technology is to understand the visitor’s requirements. This is done by discovering the relationships between the pages visited by a particular visitor to a specific website. Several algorithms are used as Apriori algorithm, To find recurring association rules [6].

   C -Clustering:
   Clustering is a method used to group certain elements (pages, users, etc.). Based on similar characteristics. Such as a grouping of webpages with similar content. Or Clustering a group of visitors with similar browsing behavior. Or collecting many users who visit similar sites and others. It is also possible to use the normalization process with Clustering. This will give better combination results. Because there are different ranges in the data point for each area. Clustering technology helps in deducing customer stats in e-commerce market operations. And provide customized web content based on individual visitors. Also, Clustering is useful in making indexes of websites on the Internet [6].
D - Classification:

This technique categorizes data elements into distinct, predefined categories. Related to a specific category. This technology requires extracting and selecting the distinct classes on which the classification is made first. Then, the classification process is performed.[6] The main goal of categorizing weblog data is to develop the log for visitors belonging to a specific category as opposed to aggregation. Because classification is a directed learning method (supervised learning). Of the algorithms used by this technique naïve Bayesian classifiers, decision tree algorithm [20].

E - Sequential pattern:

It is the conduct of analysis to find patterns in sequence through serial sessions. And by applying several algorithms such as SPADE, Apriori, etc. For example, a specific visitor visited link A and then link B one by one at the same time. Using analysis for such a pattern, we can predict the suspected visitor. When visited in a pattern similar to the previous one. Through psychology used to uncover crime, predict shopping, advertisements, etc. [6].

4) Pattern Analysis:

It is the last stage of web usage mining. In it, knowledge is found in the detected patterns. Interesting patterns. And that by getting rid of inappropriate patterns. This can be done by applying a validation rule. To get rid of inappropriate patterns and uncover appropriate patterns [20].

A commonly used technique in pattern analysis is the OLAP (Online Analytical Processing Technique). Visualization methods, administrative advertising deals, It uses graphic patterns to interpret the file results in an easier way. As well as (the mechanism of using the knowledge query SQL). Which is used to analyze several reasons for the abnormal patterns of your visitors[6].

4- Web Usage Mining Advantages and Disadvantages:

A) Advantages:

WUM It has many important advantages that make it more attractive to many, including governments, organizations, and companies. This technology appeared in electronic commerce to develop private shopping, the results of which showed large volumes in trade and profits. In the field of security, government agencies and others use this technology to classify and combat terrorist threats. Through the ability to predict Internet fraud detection and sort and identify people and pages with a high-security risk. Helps companies find the best relationship with the customer by identifying exactly what they need. Through WUM companies can understand well the requirements of their visitors. And provide a faster response to the requests of the visitors. It is also very beneficial for companies in attracting and retaining useful customers for the company who can save the best production costs [20].

B) Disadvantages:

When using WUM on personal information shows some concerns and negative consequences including. Violating the privacy of the user's information if it is obtained or published, especially when the user is not aware of it. It is the main concern for some users when some companies collect data from people for a specific purpose, such as a job or work, and others. These entities may use the data for other purposes such as fraud through e-mail or other personal information. Some techniques may be used to classify individuals based on controversial characteristics such as race, sexual orientation, religion, or gender. These practices can be in opposition to the Anti-Discrimination Act. Some companies also sell personal data obtained from people's websites in various ways. [20].

5- CONCLUSION
Data mining is the process of discovering needed information. Web mining is a branch of data mining which is the discovery of the knowledge and patterns required of the sheer volume of web data. This enormous amount of data is reflected negatively in the web mining process. Because it contains a Data outlier. Therefore, a pre-processing and cleaning stage must be carried out before mining. The web mining process Classify into three main types (Web Structure Mining, Web Content Mining, Web Usage mining). Each type is used according to the user's desire and what is part to be mined. This paper focused on conducting a survey of previous studies in web mining and types. And explained web usage mining and algorithms and techniques used in this genre. Starting with data collection, preprocessing, Discovery, and Analysis of Patterns. And what are the Advantages and Disadvantages of Web Usage mining.

References

[1] P. Verma and N. Kesswani, “Web Usage mining framework for Data Cleaning and IP address Identification,” 2014.

[2] T. A. Al-Asadi and A. J. Obaid, “Discovering similar user navigation behavior in web log data,” International Journal of Applied Engineering Research, vol. 11, no. 16. pp. 8797–8805, 2016.

[3] M. Gayatri, “Review of Current Trends in Web Usage Mining,” no. October 2018, 2020, doi: 10.14419/ijet.v7i3.202297.

[4] Yehia Helmy, “Empirical-Study-of-Data-and-Web-Mining-in-Education.pdf.” International Journal of Scientific & Engineering Research, p. 8, 2020.

[5] K. Velkumar and P. Thendral, “a Survey on Web Mining Techniques,” pp. 167–174, 2020.

[6] B. Bhavani, V. Sucharita, and K. V. V. Satyanararana, “Review on techniques and applications involved in web usage mining,” International Journal of Applied Engineering Research, vol. 12, no. 24. pp. 15994–15998, 2017.

[7] G. Shivaprasad, N. V. S. Reddy, U. D. Acharya, and P. K. Aithal, “Neuro-Fuzzy Based Hybrid Model for Web Usage Mining,” Procedia Comput. Sci., vol. 54, pp. 327–334, 2015, doi: 10.1016/j.procs.2015.06.038.

[8] M. J. Hamid Mughal, “Data mining: Web data mining techniques, tools and algorithms: An overview,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 6. pp. 208–215, 2018, doi: 10.14569/IJACSA.2018.090630.

[9] R. Roy and G. Appa Rao, “Survey on pre-processing web log files in web usage mining,” Int. J. Adv. Sci. Technol., vol. 29, no. 3 Special Issue, pp. 682–691, 2020.

[10] J. Mehra and R. S. Thakur, “An Effective method for Web Log Preprocessing and Page Access Frequency using Web Usage Mining,” Int. J. Appl. Eng. Res., vol. 13, no. 2, pp. 1227–1232, 2018, [Online]. Available: http://www.ripublication.com.

[11] Y. Almurtadha, N. Mustapha, and N. I. Udzir, “Improved Web Page Recommender System Based on Web Usage Mining,” no. 079, pp. 8–9, 2011.

[12] C. E. Dinuca and D. Ciobanu, “Improving the session identification using the mean time,” Int. J. Math. Model. Methods Appl. Sci., vol. 6, no. 2, pp. 265–272, 2012.

[13] S. Pandya, I. Mtech Scholar, Department Of CSE, RGVP, Bhopal, and M. R. Nigam, “REVIEW PAPER ON WEB PAGE PREDICTION USING DATA MINING,” Comput. Eng. Intell. Syst., vol. 7, no. 2222–1719, pp. 3369–3401, 2015.

[14] Mr. Jitendra B. Upadhyay and Dr. S. V. Patel, “A Review Analysis of Preprocessing Techniques in Web usage Mining,” Int. J. Eng. Res., vol. V4, no. 04, pp. 1160–1166, 2015, doi:
10.17577/ijertv4is041348.

[15] S. V. Kolekar, R. M. Pai, and M. M. Manohara Pai, “Prediction of Learner’s Profile Based on Learning Styles in Adaptive E-learning System,” Int. J. Emerg. Technol. Learn., vol. 12, no. 6, pp. 31–51, 2017, doi: 10.3991/ijet.v12i06.6579.

[16] K. Sellamy et al., “Web mining techniques and applications: Literature review and a proposal approach to improve performance of employment for young graduate in Morocco,” 2018 Int. Conf. Intell. Syst. Comput. Vision, ISCV 2018, vol. 2018-May, pp. 1–5, 2018, doi: 10.1109/ISACV.2018.8354043.

[17] A. A. A. El-aziz, P. S. Pandian, S. N. Almuayqil, and A. S. Alruwaili, “A Framework for Clustering & Enhanced Approach for Frequent Patterns in Web Usage Mining A Framework for Clustering & Enhanced Approach for Frequent Patterns in Web Usage Mining Assistant Professor , Department of Information Systems , College of Computer,” no. June, 2020.

[18] S. Asadianfam, H. Kolivand, and S. Asadianfam, “A new approach for web usage mining using case based reasoning,” SN Appl. Sci., vol. 2, no. 7, pp. 1–11, 2020, doi: 10.1007/s42452-020-3046-z.

[19] P. M. Bharti and T. J. Raval, “Improving Web Page Access Prediction using Web Usage Mining and Web Content Mining,” Proceedings of the 3rd International Conference on Electronics and Communication and Aerospace Technology, ICECA 2019. pp. 1268–1273, 2019, doi: 10.1109/ICECA.2019.8821950.

[20] R. S. Rao and J. Arora, “A Survey on Methods used in Web Usage Mining,” Int. Res. J. Eng. Technol., vol. 4, no. 5, pp. 2627–2631, 2017.

[21] P. V Patil and M. E. Scholar, “Preprocessing Web Logs for Web Intrusion Detection,” pp. 11–15, 2012.

[22] A. K. Kassem, B. Daya, and P. Chauvet, “A proposed methodology on predicting visitor’s behavior based on web mining technique,” International Journal of Advanced Computer Science and Applications, vol. 9, no. 12, pp. 245–255, 2018, doi: 10.14569/IJACSA.2018.091236.

[23] R. Gopikaramanan, T. Rameshkumar, B. Senthil Kumaran, and G. Ilangovan, Novel control methodology for H-bridge cascaded multi level converter using predictive control methodology, vol. 11, no. 5. 2015.

[24] T. A. Al-Asdi and A. J. Obaid, “An efficient web usage mining algorithm based on log file data,” J. Theor. Appl. Inf. Technol., vol. 92, no. 2, pp. 215–224, 2016.