J. Integr. Sci. Technol. 2023, 11(2), 479 Article

Journal of Integrated
SCIENCE & TECHNOLOGY

Web server access log pattern analysis using GWO-based clustering

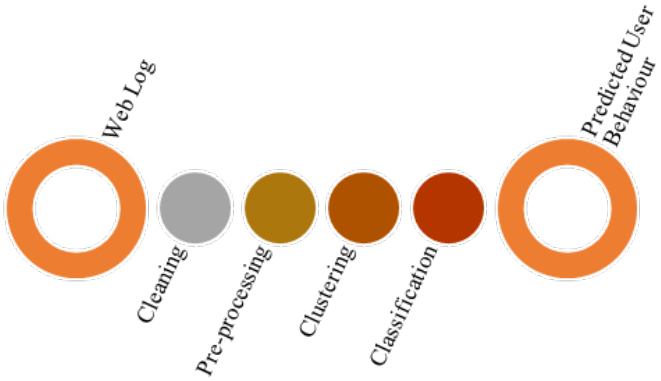
Anshu Dixit,* Shailja Sharma

Department of CSE, Rabindranath Tagore University, Bhopal, Madhya Pardesh, India

Received on: 24-Oct-2022, Accepted and Published on: 14-Dec-2022

ABSTRACT

There is a great increase in the amount of information that is available at any given moment, and the internet is a significant source of information. Every day, more people are using the internet than there was the day before. Numerous research is being conducted to reduce the amount of time individuals spend browsing the internet. In the process known as Web usage mining, mining strategies are carried out on a dataset provided by a proxy server to identify the actions of web users. Clustering is an important technique that has many different uses, including the analysis of data from online logs, customer relationship management (CRM), advertising and marketing, scientific diagnostics, and computational biology, amongst many others. Clustering is the collection of data objects that are connected. The problem with clustering is the fact that there are multiple types of measurements that might determine whether or not two things are similar or unique. The paper presents the clustering and classification of web server log access. The clustering is performed using the grey wolf optimization algorithm and classification is performed using five classifiers i.e., support vector machine, random forest, k-NN, decision tree, and gradient boosting. The performance was also compared with existing state of art models and the proposed model has achieved better results.



Keywords: Web Usage, Data Mining, Pattern Recognition, Clustering, Classification.

INTRODUCTION

The amount of material available presently through the internet is tremendous and constantly growing. It is essential to organize this enormous amount of data and present the most pertinent results on the user's screen. It is impossible to analyze and retrieve significant data from big data sets directly; therefore, automatic extracting techniques are needed to get user-requested information from the World Wide Web's billions of web pages and find pertinent data. Typically, people utilize search engines including Yahoo, Bing, MSN, Google, etc. to obtain information on the WWW (World Wide Web).[1] The practice of extracting information from database warehousing is called data mining. This information can be divided into several regulations and trends that can assist a consumer or corporation in the analysis of gathered information and anticipated decision-making procedures.[2] Data warehouses are concentrated databases where all of an organization's information systems are kept in one big dataset. Organizations employ data extraction as a technique to extract informational value from unstructured data. Software is used to search through huge amounts of information (a "data warehouse") for the necessary trends that might help businesses better understand their clients, forecast behaviors, and develop more effective advertising plans. In reality, web mining is a branch of data mining that deals with data that is accessible online. The idea behind it is to take advantage of the useful information that is present on online websites.[3] To find the information they seek on the world wide web, consumers utilize a variety of search engines. Web mining is a strategy for finding this information.

Various techniques and methods are employed to retrieve information from internet sites, including web documents, photos, and so forth. The number of textual information on the internet is growing, and personal research to identify significant trends, information, and useful data is quite time-consuming and difficult. This is why web mining is quickly getting progressively significant.

Anshu Dixit, Department of CSE, RNTU Bhopal (M.P.), India
Email: dixitanshu88@gmail.com

Cite as: *J. Integr. Sci. Technol., 2023, 11(2), 479.*
URN:NBN:sciencein.jist.2023.v11.479

 ISSN: 2321-4635
http://pubs.thesciencein.org/jist

Information acquired using Web mining includes structure (hyperlinks), utilization (accessed webpages, information utilization), and contents (text documents, pages).[4,5] Term The mix of web papers, movies, audio files, and other content is known as the "World Wide Web."

Web mining involves a variety of procedures, including:

- Information Retrieval: Finding pertinent and helpful data on the internet is a practice known as data extraction. The selection of pertinent facts from a sizable database gathering and the discovery of novel information from a sizable amount of information is now the main goals of data extraction. Browsing, sorting, and comparing are IR stages.[5,6] An automated method of obtaining data from the analysis is called information extraction (structured). While IE functions similarly to information extraction, it places a greater emphasis on obtaining pertinent facts.[5]
- Machine Learning: Machine learning is a supporting mechanism that aids in internet data mining. By understanding customer behaviors, machine learning can enhance internet search (interest). Search engines employ a variety of machine-learning techniques to deliver sophisticated web services. Compared to the conventional method of data extraction, it is significantly more effective. It is a method that can improve efficiency on a given activity and understand consumer conduct.

WEB MINING CATEGORIES

Web mining generates enormous amounts of data that are largely unorganized, dynamic, and diversified. The web's rapid expansion causes various issues, including the difficulty of accessing pertinent information online and tracking consumer behaviour. An attempt was performed to present pertinent information in a structure (table) that is simple to grasp and helpful for enterprises to anticipate client requirements to address this type of issue.[7] Web mining is divided into three categories as illustrated in Figure 1.

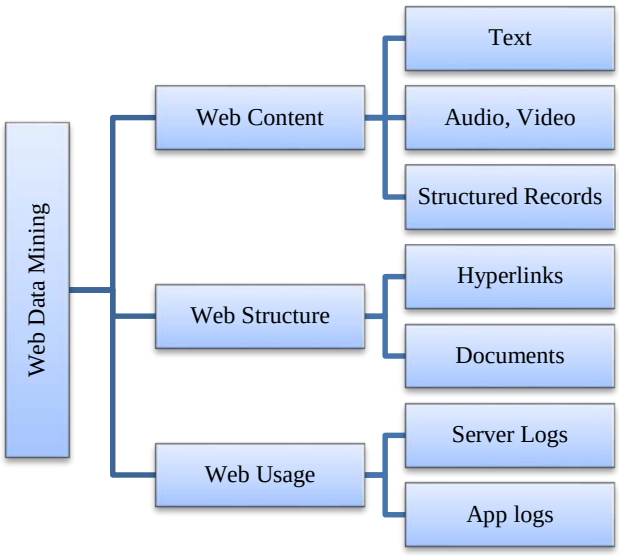


Figure 1. Web mining taxonomy.[8]

Web Usage Mining: web usage mining, also known as log mining, is the technique of capturing consumer usage information from the internet and collecting it in the form of logs. After accessing any website, the consumer gives certain data including the time of access, IP addresses, pages accessed, and so forth. This data is gathered, processed, and saved in logs. This aids in understanding consumer conduct and, as a result, may optimize the architecture of a website.[9]

The goal of web usage mining is to find secondary information that results from consumer engagement while browsing the internet. It collects pertinent utilization information in-depth, eliminates extraneous usage information, establishes the actual utilization information, identifies intriguing navigational trends, and displays those trends straightforwardly and understandably.

The information from the proxy server logs, browser logs, web server logs, consumer accounts, sessions, inquiries, registration information, bookmark information, mouse movements and scrolling, cookies, and any additional information that is the outcome of these interactions are all considered secondary information. Web usage mining is a method that automatically tracks consumer accessing trends, and this data is typically supplied by web browsers and subsequently gathered in client data. Data that is crucial for a business to comprehend its users' conduct and guarantee high-quality services are stored in logs, such as the URL address, accessing duration, Internet Protocol addresses, etc. The logging data can assist us in determining answers to queries like "from what search engines are people arriving? Which pages are the most and least visited pages? Which operating systems and browsers are the most popular among viewers? Mining web usage delves into and examines log file information that includes consumer accessing trends. Connection regulations, path analysis, sequence analysis, grouping, and categorization are a few methods for data mining that are used to analyze and understand user behaviors.[10] The primary utilization of web usage information is to record consumer surfing habits from a certain webpage. Depending on the types of user information analyzed, web usage mining could be categorized. In this situation, the user data consists of web log information, which preserves details of the consumer movement. As the focus of the research is web usage mining, this is the process of using data mining tools to identify utilization trends in web data. Internet server logs, for example, are commonly used to capture information gathered from consumer interactions with the internet.[10] To assist the web developer, enhance the webpage, draw visitors, or provide devoted consumers with tailored and adaptable services, utilization mining techniques identify and forecast customer behavior. The primary goal of web usage mining is to track consumer conduct as they communicate with the website. Generalized and tailored trend monitoring are the two categories of trend monitoring. The historical background of web pages is often where monitoring data is gathered. Data is collected specifically for each consumer during personalized monitoring.[11] Three stages make up Web Utilization Mining: pre-processing, trend finding, and pattern recognition (Figure 2).

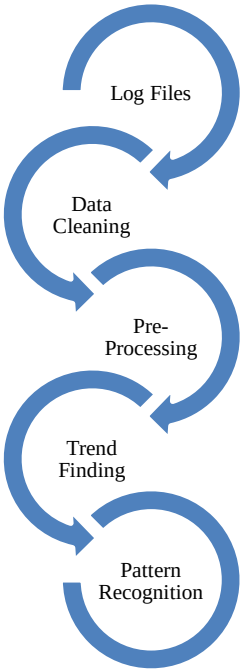


Figure 2. Web Usage Mining Phases

Data Collection: The main source of information for Web Utilization Mining is the weblog records on the web application. These three places are where information can be gathered.

- Web Servers
- Web proxy servers
- Client browsers

Various fields could be written into the log file in various forms by configuring web servers. The fields IP Address, Login Names, Customer Name, Query Types, Status, Bytes transmitted, Referrer, Visiting Pathway, Path travelled, Timestamp, page last viewed, rate of achievement, User agent, URL, etc. are the ones that web servers utilize the most frequently.

Data pre-processing: Real-world data as well as some datasets are incomprehensible, imprecise, and lacking in detail. Information pretreatment is a mining approach that combines datasets and transforms unprocessed information into uniform, intelligible information.[9] Web records are treated during dataset preprocessing due to their sparse and chaotic character. In the initial stage, superfluous, pointless, mistaken, inadequate, and inconsistent information is removed from the original information.[11] Cleaning, correcting, and preparing the input information for mining are all preprocessing tasks. Web usage mining uses a variety of e-sources from which information can be gathered and analyzed, including data logs, websites, consumer login details, web access logs, cache, cookies, etc. Web access logs are regarded as a trustworthy source for utilization mining since they employ standardized log formats (Common LF and Extended CLF) for recording.[12] Data preparation techniques include those for cleansing information and identifying users and sessions.

Data cleaning is crucial for various analysis approaches in addition to utilization mining. The goal is to purge logs of unnecessary and unnecessary details. Videos and images should be taken out of site logs because use mining doesn't require them.[13] Numerous entries are recorded in the log file when a consumer asks the webserver to deliver a specific website. Information that is useless for usage mining need to be deleted.

Finding consumer activities from an admission log file requires the usage of the Consumer and Session Identifying approach. Identifying consumers comes after cleansing the information. For consumer recognition, many methods are employed, including cookies and consumer login data to identify users with IDs for particular web pages. Session identification is the process of counting the number of pages one person views consecutively during a single website visitation. Each IP means a new consumer, while a session is a collection of web pages that users browse. When a proxy server is utilized and many consumers' IP addresses are recorded in the log file, this is a challenging step. The referrer approach is proposed as a remedy for this issue. If IPs are identical, various browsers or operating systems can detect new customers just as various IP signals new customers. The referrer technique considers URL account data if OS, IP, and browsers are identical. If the URL account has not been visited previously, it will be treated as a new client.[9]

Pattern Discovery: It is challenging to find required trends and to derive information that is comprehensible from pre-processed data. Here, we'll give a brief overview of a few methods for recognizing trends in data that have been analyzed.

Association Rules: It is possible to link pages that are frequently cited collectively within a single session using connection rules generation. In a single client session or consumer interaction, it can identify the relationships between sites that are frequently referred to in combination.

Statistical analysis: It is a potent tool used to gather information about website users. While evaluating utilizing various factors, analysts explain statistical study on session logs. The information gained through statistical analysis of the results can help to boost safety systems, increase functionality, and optimize commercial approaches.[14] Three statistical assessments, frequency, median, and mode, are frequently employed on visits to display the duration of the site, the more current pages visited, and the amount of time spent viewing it.[11]

Sequential Patterns: Web advertisers can forecast upcoming user behavior and target specific user categories with adverts by employing this strategy. The drawback is that if there is a large quantity of information, it can be challenging to identify intriguing trends.

Clustering: A process called clustering brings together a collection of things with related traits. Two different types of intriguing clusters can be found in the Web Use realm: utilization groups and page groups. Consumers are grouped when they demonstrate identical surfing behaviors, a process known as clustering. The clustering of sites, on the contrary hand, identifies groupings of pages with relevant information. Web service companies and Internet search engines can both benefit from this knowledge.

Classification: To determine the traits that distinguish the category to which every case pertains, categorization is used. This

trend can be utilized in both scenarios, i.e., to comprehend the current information and forecast how subsequent cases would act. Three methods—statistical, machine learning, and neural networks—are used in categorization techniques.

Pattern Analysis: The procedure of Web Usage Mining is completed with a phase called Trend Evaluation.

The last and most important phase in mining techniques is trend evaluation. All uninteresting or unnecessary regulations or trends found in the previous stages are segregated in this stage, and then the useful or pertinent regulations or trends are retrieved. This may contribute to better system functionality.[14] These are the methods used for pattern recognition:

SQL is the language that is more frequently utilized for querying mechanisms. SQL, which stands for Structured Query Language, is used to glean usable data from trends identified.[9]

After trend detection, the information enters the OLAP phase. OLAP operations (roll up, etc.) are carried out during this stage, during which information is stored in data cube format.

An agent is an individual who can aid a consumer with certain activities by acting as their helper. An intelligent agent can sense incoming elements, identify them, and decide which activity needs to be carried out. The patterns found in earlier stages are analyzed by utilization mining agents [9].

CURRENT TRENDS

Web usage mining's primary objective is to understand how users navigate the web and how they utilize its assets. Eventually, with the help of this data, the webpage can be made better from the consumers' perspective. Recent trends for online usage mining may be seen in a variety of fields. Some of these patterns are:

- **Personalization of web content** A variety of web-use mining methods can be employed to customize websites based on individual characteristics and conduct. Customization has various benefits, some of which include automating things for the clients, developing a deep connection, and being utilized to develop new industry tactics. To help them create better websites, it is also employed to supply additional user-related data.[15]
- **E-learning environment** Web usage mining is particularly significant in the e-learning atmosphere since it keeps track of the actions of the webpages and attempts to collect behavior and trends that could then be utilized to create changes to the webpage. The mining software assists us in comprehending the most often accessed sites, links, and so on.[16]
- **Security** Web usage mining can be employed to detect safety problems and offer trends that are relevant and helpful in detecting fraud, invasion, and other threats.[17]
- **Site design support** web developers face a lot of challenges when creating it. Web usage mining helps web developers by giving them more assistance and making their tasks simpler by monitoring consumer conduct and patterns.[17] It gives them more crucial data about how visitors behave, allowing them to revamp the website or adjust the contents.
- **Business Intelligence** Web usage mining is effective in enhancing marketing tactics, assisting businesses in maintaining their position in the market and helping them make choices that will enhance their performances. It takes details from the databases that have been recorded on client action and use that data to generate advertisement plans for the business.

RELATED WORKS

The process of finding patterns is essential to Web usage mining. It incorporates methods and formulas from machine learning, trend identification, and data mining. To find patterns in information, a variety of techniques including connection rule mining, categorization, grouping, etc. are utilized.

Clustering is the process of extracting and grouping altogether comparable web pages. This makes it easier to identify persons or groups of individuals who identically utilize the web and allows us to derive information trends. It also aids in the understanding of consumer demographics so that one can offer each consumer individualized web content. The majority of web service operators and search engines utilize grouping. Web clustering is one of those applications kinds, where diverse item kinds could be grouped into distinct groupings for varied reasons. Clustering can take many different forms, including progressive grouping, hierarchical clustering, and partitional grouping.

Gangadwala et al.[18] reported numerous validity metrics, such as the partitioning coefficient and partition entropy. Several fuzzy clustering methods, including Modified Fuzzy Possibilistic C-Means Clustering, Fuzzy Possibilistic C-Means, and Fuzzy C-Means Clustering, are tested in experimentation. These techniques are created and tested on the proxy log database. Analysis of algorithmic experiment results follows last. Empirical results clearly show that the clusters created utilizing the suggested MFPCM approach are significantly superior to those created utilizing the k-Means, Fuzzy C-Means, Fuzzy Possibilistic C-Means, and Reconfigured Fuzzy Possibilistic C-Means Clustering algorithms in terms of different validity parameters.

Abbasimehr et al.[19] proposed a methodology that visualizes every conduct of consumers as a time-series sequence of the Recency, Frequency, and Financial variables, and then uses time-series grouping methods to carry out customer fragmentation. Modern grouping methods, such as hierarchical, spectral, and k-shape algorithms, are used in the architecture to classify consumers into uniform groupings and to apply point-of-sale gadget transactional information from grocery and appliance merchants.

Dafir et al.[20] summarized the most recent parallel clustering methods and grouped in terms of the horizontal and vertical scaling systems, which are the computational platforms utilized to manage big information. Peer-to-peer networks, MapReduce, and Spark platforms fall under the first group, whereas multicore processors, graphics processing units, and field programmable gate array platforms fall under the second. It also compares how well the assessed methods performed depending on certain established benchmarks for validating clusters in the context of big information. The viewer is thus given a comprehensive picture of the methods for parallel clustering currently in use.

Inuwa-Dutse et al.[21] presented a unique identification technique to find connected groups in a network which is centered on a scalable architecture. They provided a multilevel clustering method (MCT) that makes use of textual and structural data to pinpoint

small groups of people they call microcosms. The effectiveness of the method is demonstrated through an empirical assessment of reference models and sets of data. This research provides a fresh perspective on the identification of strong groups on social networking sites.

Jung et al.[22] suggested a clustering method for big information merging that is based on social mining. The suggested approach creates accurate consumer modeling by combining data from traditional static models and data mined from social networks, and it imposes a varied level of weight dependent on the relationships between consumers. It is feasible to forecast the health risk for illness survivors by looking at the clustering processes for their medical issues. Their health circumstances can be enhanced depending on the likelihood and expectations of a medical incident.

Djenouri et al.[23] addressed the issue of retrieving the most pertinent thing from a grouped collection of objects to answer user inquiries. It aims for quick, high-quality extraction of data and tackles frequent problems with cluster-based methods. A unique cluster-based data retrieving method called Cluster-based Extraction utilizing Pattern Mining is suggested for this aim (CRPM). Several grouping and trend-mining methods are integrated into this method.

Križanić et al.[24] explained how methods of data mining were used to analyze the academic records of a Croatian institution of higher education. Activity logs retrieved from an actual e-learning course's environment are the information utilized for the study. Cluster assessment and decision trees are two methods of data mining used in the study. By grouping sets of trends depending on how similarly students used course materials, the cluster assessment was carried out. The approach of choice for creating a representation of decision-making that permitted identifying groups of entities for the aim of conducting a more thorough investigation of how children learn was to use decision trees.

El Aissaoui et al.[25] suggested a general method for automatically identifying learning forms by a provided framework of learning types. This method is independent of any particular LSM. There are two main stages to this study. Utilizing web usage data mining methods, they initially recovered learning patterns from learners' log data. The retrieved learners' patterns are then categorized utilizing clustering algorithms by a particular training type model. This method is carried out by the LSM Felder-Silverman Models and the clustering algorithm Fuzzy C-Means. They used a real-world database in an empirical investigation. The findings gathered demonstrate that this method surpasses the conventional method and yields encouraging outcomes.

The quantitative study of human identification is called pattern identification, and it makes use of information technology to let computer systems simulate human identification conduct. In other terms, pattern identification is the investigation of how to teach a machine to observe its surroundings, understand to distinguish important characteristics from the backgrounds, and accurately determine the type of pattern it is. Recognizing objects and recognizing abstract objects are the two basic categories that best describe identification behaviours. The recognition of spatiotemporal data is a requirement for the recognition of objects.

Laskari et al.[26] characterized heating behavior in housing developments by offering a data mining tool. Principal Component Analysis (PCA) and cluster analysis are the two multimodal statistical assessment techniques used in the approach. The techniques were used on information from five homes in Italy where gas expenditure was being tracked.

With the help of pattern recognition techniques applied to web log information, Singh et al.[27] evaluated web usage mining in this work. The input of unprocessed data and intervention depending on the 'class' of the patterns is what is meant by the term "pattern recognition." Web usage mining is broken down into three stages: preprocessing, pattern finding, and pattern evaluation. Additionally, the investigation in this research uses web log information and was supposed to be empirical. Using "Web Log Explorer," researchers examined the web log information that they had obtained from the "NASA" web server. One tool that is essential to this task is Web Log Explorer, which is used to mine web usage data.

Abd El-Aziz et al.[28] presented a framework for customer and session preprocessing and grouping with the Hidden Damage Data algorithm (HDD) to identify visit layouts in web log information. In addition, it analyzes customer navigation systems' behavior using the Improved Conviction Frequent Pattern Mining Algorithm (CFPMA). The experiment's outcome demonstrates that, in comparison to alternative strategies, the presented method achieves a shorter processing time and more accuracy.

PROBLEMS IDENTIFICATION

The algorithms used for mining web usage are more precise and useful. But there is a problem that needs to be considered. According to studies, information pre-processing takes up 70% of the time, making web cleansing the most crucial operation. However, varied information makes data cleansing challenging. It is important to focus on retaining precision when categorizing the data. Even though there are numerous categorization approaches, the effectiveness of clustering remains unclear. The dataset is massive and comprises a lot of information; therefore, mining useful regulations contribute to the enormous number of useless regulations. These are brought on by the enormous item sets, which inevitably reduce the effectiveness of mining approaches. It also gets difficult to extract regulations from semi-structured and unstructured data, such as that found in the semantic web.

METHODOLOGY

Mining information from the use of the web can make use of one of three distinct types of log files. Log files are saved in three different locations: on the client, on the proxy servers, and the server. The mining procedure is made more complicated when there is more than one site for keeping the knowledge of navigation patterns of the users. It is only possible to generate results that are truly credible if one obtains data from all three different types of log files. This is because records of Web page accesses that are cached on the client side or the proxy servers are not stored on the server side. The reason for this is that the server side does not contain these records. Additional information can be found in the log file that is kept on the proxy server in addition to the log file

that is kept on the main server. On the other hand, the page requests that were saved on the client side are not there. On the other hand, collecting all the information from the client side presents some challenges. As a result, the algorithm relied almost exclusively on the data stored on the server. Association rule mining, sequence mining, and clustering are only a few of the data mining methods that are frequently used for mining Web usage data. In this paper, we have adopted optimization-based clustering for web usage pattern mining.

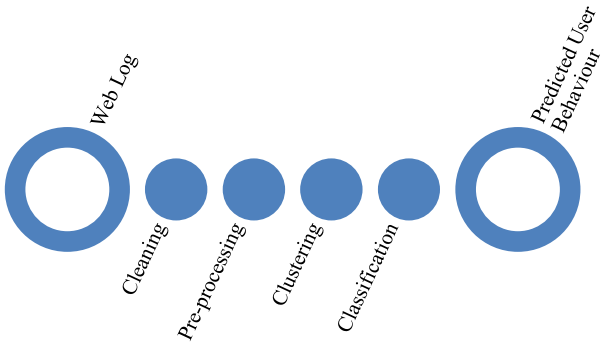


Figure 3. Proposed Framework

Figure 3 describes the designed framework to identify user pattern mining from web server logs.

The mining of web users can be broken down into three distinct categories of activity:

- Activities known as preprocessing, involve doing a review of the web log data before it is processed.
- Discovery Activity Pattern, also known as Pattern Mining, accounts for the majority of time spent on all mining activities. This is because these activities search to uncover concealed patterns in the data log.
- Pattern Analysis, also known as "Analyzing Pattern," is a method that involves studying and doing an analysis of the findings acquired from observing recurring patterns of search behavior.

Data cleaning

The process of turning raw web log data into processed Log data based on successful responses and user interest is referred to as "Data Cleaning." This procedure also determines the successful response that was obtained from the raw web log data. If a website's status code is between 200 and 400, it is considered to have a successful response; if the status code is lower than 200, it is considered to have an unsuccessful response. The identification of a successful response is determined by the successful response that can be identified by the status code on websites.

Pre-processing

In the Preprocessing stage, the data log or record of the use of web pages is typically not in a format that the mining apps might exploit. If data is going to be utilized by mining applications, the format of the data will need to be converted into a different form that may be utilized by the mining program.

Clustering

In this stage of the procedure, the one-hot encoding method is utilized to convert the pre-processed data, which have previously undergone statistical analysis, for them to produce more accurate prediction results. This technique is used because it yields more accurate results from machine learning predictions when applied to data sets that do not appear to have any connections with one another. The numerical order of integers is taken into account as a major characteristic of the algorithms that make up machine learning. To put it another way, it might be said that a larger number will be perceived as being superior to or more significant than a smaller number. After encrypting the data, a modified version of the grey wolf optimization technique is used to cluster the data.

With just one hot encoding, our training data becomes more valuable and detailed, and it can be easily scaled up or down as necessary. Utilizing numerical values allows us to build a probability for our values in a more straightforward manner. Because it provides more comprehensive predictions than single labels, we have decided to employ one hot encoding for the values that are generated by our system.

To formulate a data clustering problem numerically a set of M objects is denoted by $O = \{O_1, O_2, \ldots . O_M\}$. Then, $O_p = O_p^1, O_p^2, O_p^3 \ldots \ldots O_p^M$ is a vector representing p[th] data object and o_p^j denotes the j[th] attribute of o_p. The idea of clustering is to a lot each object O to one of K clusters $D = D_1, D_2, D_3, D_4, \ldots \ldots \ldots D_K,$ such that each cluster contains at least one object.

Investigating the search space is the method of exploration. An algorithm's initial iterations examine the search area in pursuit of better answers. Throughout this method, search agents can skip local optima while thoroughly traversing the search space. For the algorithm to converge to an optimal solution, referred to as the exploiting phase of an algorithm, investigation decreases gradually, and exploitation increases. For the algorithm to work properly, these two steps must be balanced properly. Therefore, suggesting a novel strategy is preferred. These two steps in GWO are managed by the variable $\vec{b}$. This value is linearly reduced, as indicated in the preceding section.

Early iterations of the algorithm place a greater emphasis on investigation, whereas succeeding iterations focus more on exploiting. We can modify the power of exploration as well as exploitation as well as create a balancing strategy among these 2 phases by altering the linear behavior of $\vec{b}$. The following is our novel control parameter:

$$\vec{\boldsymbol{b}}(\boldsymbol{i}) = \mathbf{2} - \mathbf{2}(\frac{\boldsymbol{i}}{\boldsymbol{i_{max}}})^{\boldsymbol{K}} \quad (1)$$

Where i is the present iteration, i_{max} is the overall number of iterations and K is the constant. Here the parameter $\vec{b}$ is still decreased from 2 to 0 but in a non-linear method. The emphasis will be placed on exploitation for k values ranging from 0 and 1, but the efficiency of exploring capacity may degrade. The search space will be fully investigated for values greater than 1 before the algorithm proceeds to exploitation. The trial, as well as error, should be employed to determine an acceptable value for k.

The non-linear lowering of the control factors is expected to enhance GWO's performance, although there is still an opportunity for growth. We employ a mapping method to conduct a local search

all around the best answer to compensate for the reduced number of iterative steps that explore the search space. The gray wolf will be relocated to the new destination if it produces a higher level of fitness after using this strategy to the best gray wolf position. The revised position is determined as:

$$\overrightarrow{\boldsymbol{O_n}} = \overrightarrow{\boldsymbol{O_\alpha}} + \boldsymbol{r}(\boldsymbol{u_b} - \boldsymbol{l_b})(\boldsymbol{z} - \boldsymbol{0.5}) \quad (2)$$

Where the upper boundary and lower boundary are represented as u_b, l_b, r is the center and z is the mapping parameter which gets updated at each iteration:

$$\boldsymbol{z_{i+1}} = \boldsymbol{4} \times \overrightarrow{\boldsymbol{z_i}} \times (\boldsymbol{1} - \overrightarrow{\boldsymbol{z_i}}) \quad (3)$$

The modified GWO is expressed by the following pseudo-code.
Randomly generated grey wolves:$O_n(n = 1,2,3 \ldots . . N)$
Establishing initial values of $\vec{b}$, $\vec{A}$ and $\vec{D}$
i_{max} *is the total number of iterations*
Fitness calculation of grey wolves
$\overrightarrow{O_\alpha}$, $\overrightarrow{O_\beta}$ and $\overrightarrow{O_\delta}$ are the first, second, and third-best solutions
While $i < i_{max}$ do
For each $\overrightarrow{O_n}$ in $\vec{O}$ do
Update the position for $\overrightarrow{O_n}$
End for
Update $\vec{b}$, $\vec{A}$ and $\vec{D}$
Calculate the fitness of all grey wolves
If there is any improvement, make a new position of $\overrightarrow{O_\alpha}$
Update $\overrightarrow{O_\alpha}$, $\overrightarrow{O_\beta}$ and $\overrightarrow{O_\delta}$
i=i+1
end while
return $\overrightarrow{O_\alpha}$ as the solution

A crucial stage in every meta-heuristic approach is solution encoding. All the cluster centers are represented by each solution (gray wolf). These answers are first produced at random. The best options (alpha, beta, and delta) at every iteration of the grey wolf Optimization, however, serve as a guide for the remaining gray wolves. Every solution is an array with the dimensions d * K, wherein d is the overall number of characteristics of every data object as well as Ki is the overall number of clusters in the given dataset.

The objective function is the total intra-cluster distance. The objective function must be reduced to discover the best cluster centers using MGWO. It is preferred to reduce the sum of intra-cluster distances. Cluster defines the center, and the ranges among cluster members as stated below:

$$\boldsymbol{p_j} = \frac{\boldsymbol{1}}{\boldsymbol{b_j}} \sum_{\forall \boldsymbol{o_p} \in \boldsymbol{D_j}} \boldsymbol{o_p} \quad (4)$$

$$distance\ (o_p - p_j) = \sqrt{\sum_{n=1}^{\alpha} (o_{pn} - p_{jn})^2} \quad (5)$$

Here, p_j is the center of cluster j, o_p is the p[th] member of the cluster. The total number of attributes is denoted by a, $b_j$is the quantity of members of clusters in j and D_j are the members of the clusters j.

Classification

In this step, the optimized data is fed into the classifier for learning the pattern. For classification, the paper analyzed the performance of five classifiers. These are discussed below:

Linear support vector machine (LSVM): A support vector machine is a type of machine learning model that can simplify the distinction between two distinct classes if a certain amount of data that has been categorized is provided to the algorithm as part of its training set. The fundamental objective of the support vector machine (SVM) is to pattern for that hyperplane that can differentiate between the two classes. A two-dimensional database is said to be linearly separable if it can distinguish between positive and negative objects using a line. Such a database is known as a linearly separable 2D database. It makes no difference if there is more than one line in this category. If data cannot be categorized, linearization will not be able to fully distinguish between the two groups. The line separator will still be "good enough" for the majority of non-linear databases, allowing it to accurately segment several cases.

Decision tree: A predictive method that is used in machine learning is called a decision tree. It is laid out like a flowchart, and every other block has the property "test" in it. The progression of results from one block to the next constitutes a representation of the classification rules. There are three different ways in which nodes can be categorized: The decision nodes are represented by squares in this diagram. Triangles are used to represent ending nodes, whereas circles are used to represent chance nodes in a network. The operation of a decision tree can be summarized as follows: Start at the root node, which is responsible for holding the entire dataset. Find the attribute in the dataset that best fits your needs by making use of the Attribute Selection Measure. Create subgroups within the root node that include the potential values of the best property, and then divide the node. Build the branch of the decision tree that contains the most favorable characteristic in its node. Recursively create brand new decision trees by using subsets of the dataset. Continue this process until the nodes in the tree can no longer be classified, at which point the final node will be labeled as a leaf node. Research about data mining makes use of it. This is the most accurate method that may be used for making forecasts. The decision tree offers several advantages, such that it is simple to use and calls for very little in the way of data preparation. The following is a list of some of the drawbacks: This may result in overfit trees, which are overly intricate.

Random forest: It is a popular machine-learning method that falls under the category of supervised learning. It is possible to apply it in ML as a classifier, as well as make use of it for regression data. It is predicated on the idea of supervised techniques, which refers to the practice of integrating several different classifications to solve a challenging problem and raise the individual's level of effectiveness. This strategy is made up of several decision trees, each of which can be constructed off of datasets acquired from a training dataset and are together referred to as the bootstrap sample. The RF considers the prediction made by each tree and then provides a result based on the combination of all of the different forecasts. Because there are so many trees in the forest, we can achieve a higher level of precision, which helps us avoid the problem of overfitting.

Gradient boosting: Boosting methods for machine learning include the gradient boosting technique. It assumes that reducing the overall estimation error will occur when the best subsequent modeling is combined with the previous model. To cut down on errors, the most important thing to do is to identify target outcomes for each of the following concepts. The GB Algorithm is applied quite frequently these days to cut down on bias mistakes. The gradient boosting strategy could be useful for improving prediction and classification methods simultaneously. In a problem involving regression, the MSE is used as the cost function, while the function Loss is used in a problem involving classification. To generate a conclusive prediction, GB Machine combines the findings from several different decision trees. Take into consideration the fact that each learning rate in a gradient-boosting machine is a decision tree.

k-nearest neighbor: It is an ML approach that references for k-nearest neighbor method. It is referred to as a non-parameterized technique because, rather than making assumptions about the data, it implies that the cases are similar to one another. It monitors all of the other major datasets and classifies new data sets according to how similar they are to the ones it already knows about. The following describes how the system functions: Pick the number belonging to the Kth neighbor. Analyze the distance between K and its nearest neighbors using the Euclidean method. Taking into consideration the K neighbors who are the closest to us according to the Euclidean distance. Determine the total number of points that can be earned inside each category by competing against these k neighbors. Once the data points have been assigned to the categories that have the greatest number of neighbors, the k procedure will be finished. According to the K Nearest Neighbors idea, the relevant data that are considered to be the closest neighbors are those that have the shortest distance in the feature set from the new data value. And K is the total number of such data values that are utilized during the operation of the system. As a consequence of this, the distance measure and the corresponding values are two aspects that play an essential role in the KNN classifier. Euclidean distance is by far the most common and widely used metric for distance. Think about the 'K' data points that are closest to each other in a new dataset's feature space, together with the labels or value systems that correspond to them.

RESULT ANALYSIS

This paper has implemented and trained the models in the Keras framework with TensorFlow. The proposed model was trained using GPU on google colab. The following performance parameters are used to evaluate the model's efficiency in terms of Accuracy, precision, and recall.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

Where, TP= True Positive, FP= False Positive, FN= False Negative, and TN = True Negative

Figure 3 presents the most popular Web URLs accessed by users. Figure 4 presents the optimized clustering of server logs for user access.

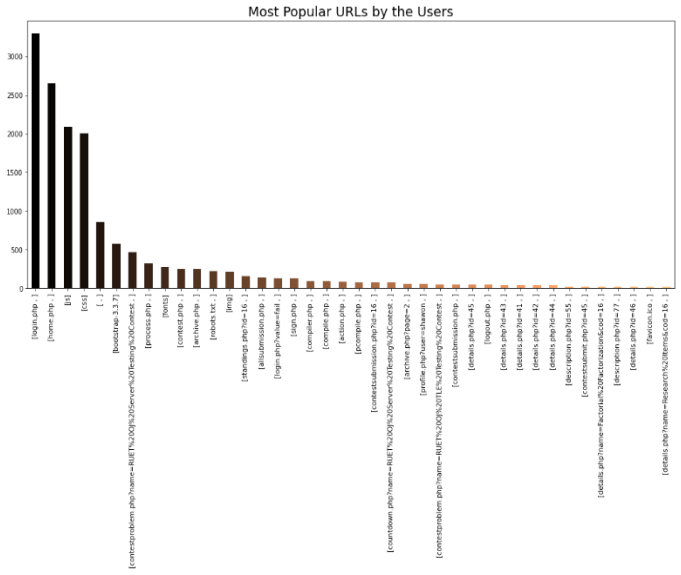


Figure 3. Server Log Data Visualization

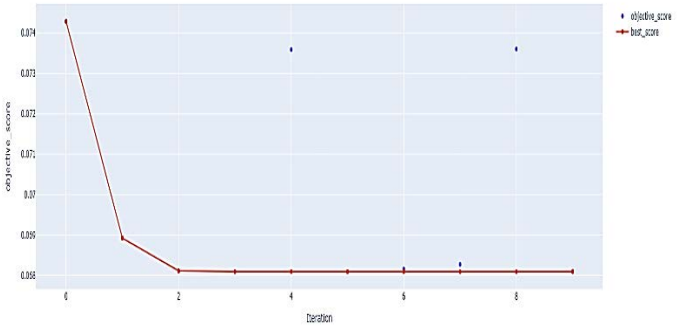


Figure 4. Cluster Optimization Performance

Figure 5 shows the ROC curve of SVM the graph plotted between the true positive rate and false positive rate where the true positive rate increases at a rate of 0.9 after that it remains constant with an increase in false positive rate and the AUC of the ROC curve remains at 0.75.

Figure 6 shows the ROC curve of the Random Forest the graph plotted between the true positive rate and false positive rate where the true positive rate increases at a rate of 0.9 after that it remains constant with an increase in the false positive rate and the AUC of Random forest remains at 0.78.

Figure 7 shows the ROC curve of the Decision Tree the graph plotted between the true positive rate and false positive rate where the true positive rate increases at a rate of 0.9 after that it remains constant with an increase in false positive rate and the AUC of Decision Tree remains at 0.78.

Figure 8 shows the ROC curve of K-NN the graph plotted between the true positive rate and false positive rate where the true positive rate increases at a rate of 0.9 after that it remains constant with an increase in false positive rate and the AUC of k-NN remains at 0.75.

Figure 9 shows the ROC curve of Gradient Boosting the graph plotted between the true positive rate and false positive rate where the true positive rate increases at a rate of 0.9 after that it remains

constant with an increase in false positive rate and AUC of Gradient Boosting remains at 0.78.

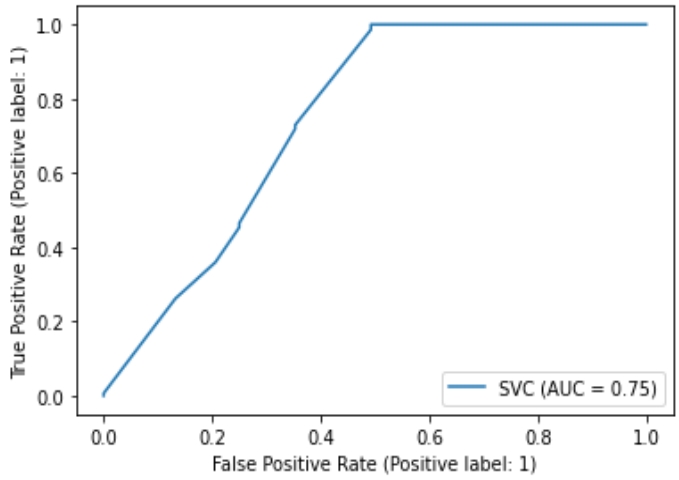


Figure 5. ROC Curve of SVM

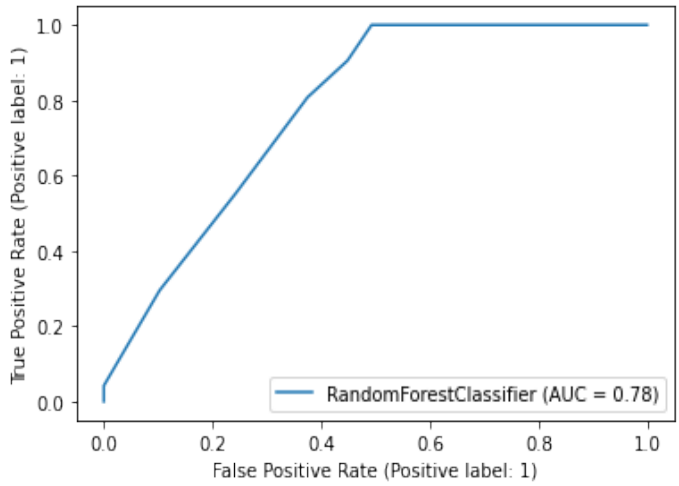


Figure 6. ROC Curve of Random Forest

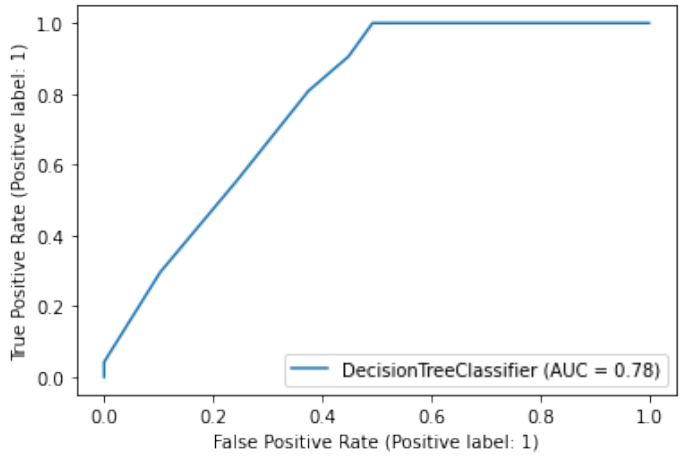


Figure 7. ROC Curve of Decision Tree

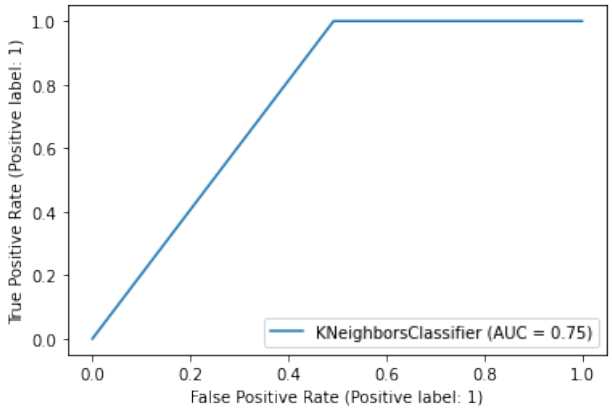


Figure 8. ROC Curve of k-NN

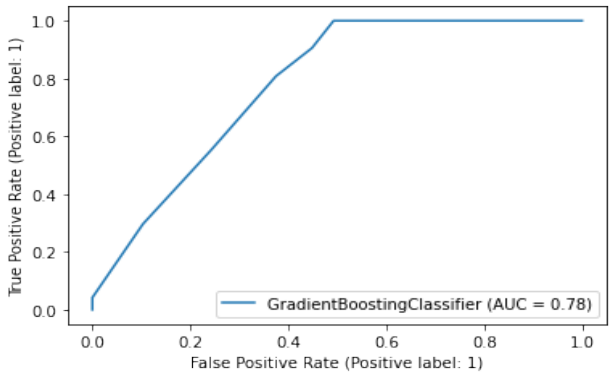


Figure 9. ROC Curve of Gradient Boosting

Below table 1 presents the result of different classifiers implemented for web server access log analysis. In this table, the paper presented the implementation result of different classifiers used. The highest accuracy was achieved by decision tree, random forest and SVM classifiers. Along with these comparision, the paper presented the performance comparision of proposed model with existing state-of-art models. Figure 10 presents the comparative state-of-art of presision rate. In ref [28], the author presented ACO based clustering approach for web server logs and achieved approx. 97% of the precision rate. Similarly, in ref [29], the author presented a fuzzy c-mean algorithm for clustering server log analysis and achieved a precision rate of 88%. Whereas in our approach we have achieved a precision rate of 98%. This shows better precision concerning state of art models. The proposed model have achieved best result because model have clustered the best features that would result in better classification.

Table 1. Performance of Classifiers for Web Log Analysis

Classifiers	Accuracy	Precision
LinearSVM	0.98605	0.985847
RandomForest	0.98605	0.985847
DecisionTree	0.98605	0.985847
KNeighbors	0.984801	0.985639
Gradient Boosting	0.985842	0.984599

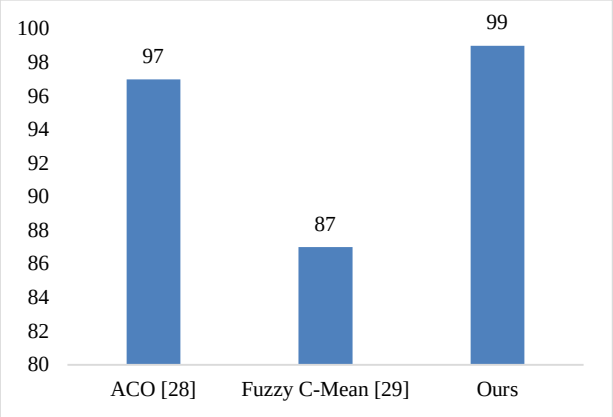


Figure 10. Comparative State-of-Art

CONCLUSION

In various fields, including e-business, e-CRM, e-services, e-education, advertisement, marketing, and bioinformatics, additional research is needed, and web usage mining is utilized. Sequential patterns, association rules, classification, clustering, and path analysis are the basic methods for finding patterns. To aid web usage mining, a variety of solutions are already on the marketplace. The problems with current methods of research and their shortcomings have been explored in this paper. Since web mining will likely take over the web in the coming future, additional research must be conducted in this area. The goal of this study was to present a clustering and classification approach based on GWO for Web usage mining. The model has achieved an accuracy of 98% and precision rate of 99% which shows enhancement over existing state-of-art mdoels. In future this work will be extended to predict the web usage pattern according statistical approaches.

REFERENCES

1. P. Bachhal, S. Ahuja, S. Gargrish. Educational data mining: A review. In *Journal of Physics: Conference Series*, **2021**, 1950(1), 012022.
2. M.J. Hamid Mughal. Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview. *Int. J. Adv. Comput. Sci. Appl.* **2018**, 9 (6), 208–215.
3. R. Mittal, V. Malik, J. Singh, V. Singh, A. Mittal. Web Usage Mining - Process, Tools and Practices. *Lecture Notes in Electrical Engineering*, 2022, 832, 449-457.
4. Z.A. Adhoni, D. Lal N. Framework, semantic and standard approaches in multi-clouds to achieve interoperability: A survey. *J. Integr. Sci. Technol.* **2022**, 10 (2), 67–72.
5. B. Kotiyal, A. Kumar, B. Pant, et al. User behavior analysis in web log through comparative study of Eclat and Apriori. *7th Int. Conf. Intell. Syst. Control. ISCO 2013* **2013**, 421–426.
6. K. Pol, S. Patankar, N. Patil, C. Das. A survey on web content mining and extraction of structured and semi-structured data. *Proc. - 1st Int. Conf. Emerg. Trends Eng. Technol. ICETET 2008* **2008**, 543–546.
7. M.K. Khribi, M. Jemni, O. Nasraoui. Automatic recommendations for e-learning personalization based on Web usage mining techniques and information retrieval. *Proc. - 8th IEEE Int. Conf. Adv. Learn. Technol. ICALT 2008* **2008**, 241–245.
8. C. Romero, P.G. Espejo, A. Zafra, J.R. Romero, S. Ventura. Web usage mining for predicting final marks of students that use Moodle courses. *Comput. Appl. Eng. Educ.* **2013**, 21 (1), 135–146.
9. M.M. Sharma, A. Bala. An approach for frequent access pattern identification in web usage mining. *Proc. 2014 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2014* **2014**, 730–735.
10. M. Kumari, S. Soni. A Review of classification in Web Usage Mining using K- Nearest Neighbour. **2017**.
11. M. Koutri, N. Avouris, S. Daskalaki. A Survey of Web-Usage Mining. *Adapt. Adapt. Hypermedia Syst.* **2011**, 125–149.
12. G. Neelima, S. Rodda. An overview on web usage mining. *Adv. Intell. Syst. Comput.* **2015**, 338, 647–655.
13. S. Asadianfam, H. Kolivand, S. Asadianfam. A new approach for web usage mining using case based reasoning. *SN Appl. Sci.* **2020**, 2 (7).
14. G. Castellano, A.M. Fanelli, M.A. Torsello. Web usage mining: Discovering usage patterns for web applications. *Stud. Comput. Intell.* **2013**, 452, 75–104.
15. S. Chaudhuri, U. Dayal. An Overview of Data Warehousing and OLAP Technology. *SIGMOD Rec. (ACM Spec. Interes. Gr. Manag. Data)* **1997**, 26 (1), 65–74.
16. M.G. Da Costa, Z. Gong. Web structure mining: An introduction. *ICIA 2005 - Proc. 2005 Int. Conf. Inf. Acquis.* **2005**, 2005, 590–595.
17. A. Chemchem, F. Alin, M. Krajecki. Deep learning and data mining classification through the intelligent agent reasoning. *Proc. - 2018 IEEE 6th Int. Conf. Futur. Internet Things Cloud Work. W-FiCloud 2018* **2018**, 13–20.
18. H.A. Gangadwala, R.M. Gulati. Analysis of Web Usage Mining Using Various Fuzzy Techniques and Cluster Validity Index. *2022 1st Int. Conf. Electr. Electron. Inf. Commun. Technol. ICEEICT 2022* **2022**.
19. H. Abbasimehr, A. Bahrini. An analytical framework based on the recency, frequency, and monetary model and time series clustering techniques for dynamic segmentation. *Expert Syst. Appl.* **2022**, 192, 116373.
20. Z. Dafir, Y. Lamari, S.C. Slaoui. A survey on parallel clustering algorithms for Big Data. *Artif. Intell. Rev. 2020 544* **2020**, 54 (4), 2411–2443.
21. I. Inuwa-Dutse, M. Liptrott, I. Korkontzelos. A multilevel clustering technique for community detection. *Neurocomputing* **2021**, 441, 64–78.
22. H. Jung, K. Chung. Social mining-based clustering process for big-data integration. *J. Ambient Intell. Humaniz. Comput. 2020 121* **2020**, 12 (1), 589–600.
23. Y. Djenouri, A. Belhadi, D. Djenouri, J.C.W. Lin. Cluster-based information retrieval using pattern mining. *Appl. Intell.* **2021**, 51 (4), 1888–1903.
24. S. Križanić. Educational data mining using cluster analysis and decision tree technique: A case study. *Int. J. Eng. Bus. Manag.* **2020**, 12.
25. O. El Aissaoui, Y. El Alami El Madani, L. Oughdir, Y. El Allioui. A fuzzy classification approach for learning style prediction based on web mining technique in e-learning environments. *Educ. Inf. Technol.* **2018**, 24 (3), 1943–1959.
26. M. Laskari, S. Karatasou, M. Santamouris, M.N. Assimakopoulos. Using pattern recognition to characterise heating behaviour in residential buildings. *Advances in Building Energy Research*, **2022**, 16 (3), 322–346.
27. N. Singh, Achin Jain, and Ram Shringar Raw. Comparison analysis of web usage mining using pattern recognition techniques. *Int. J. Data Mining & Knowledge Management Process* **2013**, 3(4), 137.
28. A.A. Abd El-Aziz, Abd El-Aziz, P.S. Pandian, S.N. Almuayqil, A.S. Alruwaili. A Framework for Clustering & Enhanced Approach for Frequent Patterns in Web Usage Mining. *Int. J. Advanced Sci. Technol.*, **2020**, 29(8), 1484-1495.
29. V. Malik, R. Mittal, J. Singh, V. Rattan, A. Mittal. Feature Selection Optimization using ACO to Improve the Classification Performance of Web Log Data. *Proc. 8th Int. Conf. Signal Process. Integr. Networks, SPIN 2021* **2021**, 671–675.
30. P.P.G. Om, S. Ananthakumaran, M. Sathishkumar, R. Ganeshan. Analyzing the user navigation pattern from web logs using maximum frequent pattern approach. *Proc. 6th Int. Conf. Inven. Comput. Technol. ICICT 2021* **2021**, 877–883.