iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** I **Month of publication:** January 2024

DOI: https://doi.org/10.22214/ijraset.2024.57962

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue I Jan 2024- Available at www.ijraset.com

A Survey on Machine Learning Based Classification Algorithms for Web Usage Mining

G. Punithavathi[1], Dr. R. Sankarasubramanian[2]
[1]*Ph.D Research Scholar, Erode Arts and Science College, Erode-9, Tamilnadu*
[2]*Principal, Erode Arts and Science College, Erode-9, Tamilnadu*

Abstract: ***Machine learning as well as data mining is research areas of computer technology whose quick expansion is due to the advances in data analysis research, growth in the database industry as well as the resulting market requirements for methods that are accomplished of extracting valuable knowledge from large data stores. It is a powerful platform that stores as well as retrieves mass information besides it becomes a time consuming uncomfortable task to search the information due to unstructured as well as heterogeneous nature of data on the web development. Web usage mining is one of the popular techniques of data mining that is utilized to discover besides quotation useful information from web document besides its service. This survey paper, conduct a comprehensive overview of the state of the art methods, clustering based machine learning based algorithms and its advantages and disadvantages of web usage mining.***
Keywords: Machine Learning, Web Crawling, Web Usage Mining, Hierarchical, DBSCAN, KNN, BIRCH.

I. INTRODUCTION

The rapid growth of data in social media and databases has led to increasing demands for more sophisticated information. Simple queries and structured query languages are no longer to support the requirements. Data Mining is mainly utilized for analysing large volumes of data in a semi-automated manner to find hidden functional patterns like unusual records called as anomaly detection and cluster analysis like collection of records and mining the rule association. Subsequently these patterns can work as the input data. Web usage mining is utilized to derive most preferable data, information and knowledge from the weblog data as well as helps to identifying the user access designs for web pages. In Data mining the management of web resources and the individual is monitor about the data which are requested from users of a website that are composed as web server logs. The content and mechanism of the set of web pages follow the intentions of the authors of the pages as well as single request displays the users view point of the web pages. Web usage mining can disclose relationships that are not suggested by the designer of the pages. The web server generally enters the weblog for each access by the designer of the pages. A web server generally login the application or web log entry for each access of a web page. This contains the URL requested the IP address from the request of the timestamp. A large number of web access log data are being collected with the famous websites can register weblog records in the order of thousands of megabytes of every hour. Web log records in the order of websites can register weblog records in the order of thousands of megabytes of every second. A Weblog database supports rich databases of rich set of web dynamics. This is essential to create sophisticated weblog mining approaches.

A. Association Rule

Association rules in web usage mining are utilized to find the relationships between the pages that are frequently visible to the next in the user sessions. Association Rule is one the basic Data Mining method utilized frequently by developers for Web usage Mining. This method helps to enable the website to track utilizers information and provide recommendations based on their search history and behaviours. The basic rule and principle of the Association Rule contains the two parts an antecedent and a consequent means condition. An antecedent is an item found within data and the consequent is the item found in combination with the Antecedent and is consider the consumer search behaviour on the e-commerce website for protein powder(Antecedent).

B. Sequential Patterns

The sequential patterns are utilized in a large volume of sequential data. The user navigational patterns are discovered utilizing sequential patterns in Web usage mining. Basically the sequential patterns are built over time that the sequence of events is defined in two basic types of algorithms utilized to generate the sequential patterns. Association Rule Mining GSP and Apriori All for instance are two established Apriori algorithms for extracting association rules. These are helps to identify the sequential pattern utilizes tree structure as well as Markov chains to represent the sequential patterns.

C. Clustering

Clustering is a technique helps to create a cluster or arrange similar items in a group among the high amounts of data. These types of clustering can be done by utilizing the distance function that helps to calculate the degree of similarities between different items. Clustering is a method of grouping comparable encounters in web usage mining.

Two different types of clustering methods available:

1) User Clustering

2) Page Clustering

D. Classification Mining

Classification Mining is depends on developing in a profile of items that belong to a particular group classified according to their common attributes. The profile can classify any new items are added as well as update the database accordingly [13]. Classified mining allows the developer to build a profile for clients who can able to access particulate web usage mining link page based on the demographic information available.

II. RELATED WORKS

Web usage mining is one of the research techniques utilizing traditional educational research data collection methods like as survey, observation, interview, assessment are regularly enhanced with computer technology. The computer aided learning environments such as online courses, intelligent tutors, course management systems have traditional assessments as well as participation measuring are built in. Computer enhancement cause these traditional learning and research mediums and to have a wider reach and simplify the collection as well as archiving of data have made available to fine grained, non-traditional data not initially collected for educational purposes.

The web usage mining is one of the data mining techniques to usage logs of large web data repositories in order to produce results and this can be utilized in the design tasks. The server logs cannot be utilized directly for pattern discovery and for analysis purposes. There are various preprocessing prior methods helps to applying data mining algorithms to the data collected from the server logs.

The clustering technology has important applications in data mining, pattern recognition, machine learning and other fields[4]. However, with the explosive growth of data, traditional clustering algorithm is more and more difficult to meet the needs of big data analysis. How to improve the traditional clustering algorithm and ensure the quality and efficiency of clustering under the background of big data has become an important research topic of artificial intelligence and big data processing. The density-based clustering algorithm can cluster arbitrarily shaped data sets in the case of unknown data distribution. DBSCAN is a classical density-based clustering algorithm, which is widely used for data clustering analysis due to its simple and efficient characteristics.

In [5] the author described a detailed report on the technique, architecture and its types of Web crawler that is utilized for structured data mining and also author explained that web crawler plays an important role in search engines so that is must be adequate, scalable, robust, extensible and quality fetcher of content. The general architecture besides the types of web crawlers is also discussed in depth in the article. This study also affords an overview of eight types of web crawlers, named as Web Crawler, Deep Web Crawlers, namely customary web crawler, Rich Internet Application, focused web crawler, Parallel Crawler and Distributed Web Crawler. The dissimilar Wrapper approaches utilized to excerpt information from unstructured data remained to perform[9]. An Automated Wrapper Generation, Semi-Automatic Generation, Wrapper Induction, Wrapper Maintenance and information extraction approaches are explored in this report. These comparative studies were performed on the algorithms utilized in wrapper approach for data extraction.

Clustering is an analytical technique that involves dividing data into groups of similarity of the objects. For every group is called as a cluster and it is formed from objects that have affinities within the cluster but are significantly different from other objects. The main aim of this paper is to look the main concepts of hierarchical algorithms and compare two various types of algorithms related to hierarchical. Partition and hierarchical clustering are the two main types of clustering techniques.

Data clustering is one of the main influential branches of machine learning and data analysis as well as Gaussian Mixture Models (GMMs) are regularly adopted in data clustering due to their easy of implementation. There are various limitations to this approach that necessary to be acknowledged [19]/ GMMs necessary to determine the cluster numbers manually as well as they may fail to extract the information that may fail to extract the information and in the dataset during initialization as well as address issues a new clustering algorithm called as PFA-GMM has been proposed. PFA-GMM is mainly based on GMMs and the pathfinder algorithm (PFA) and its aims to overcome the shortcomings of GMMs.

The algorithm automatically determines the optimal number of cluster based on the dataset. The clustering problem as a global optimization issues for getting trapped in local convergence during the initialization.

Web usage mining focuses on a website's whole network as well as it might be the links of the same webpage or between the web pages of various websites. There are two possible categories for web structure mining is document structure besides the hyperlink structure. The ranking systems heavily rely on web mining techniques due to the incorrect data and lack of mining tools and other difficulties in classification as well as clustering approaches that has some problems that need to be resolved [6].

The importance of location prediction utilizing machine learning methods as well as its potential applications in various fields. The main accuracy of these models is very high but they face challenges such as data quality, model complexity, privacy concerns and limited data availability. Despite these challenges and the future scope of location prediction is vast and Machine learning techniques plays a vital role improving these models. The shades of the modern techniques for better performance as well as highlights are difficulty of predicting a utilizer's position in real time that limits the utility of location based on services [20]. This study proposes a novel approach for predicting complete utilizer trajectories utilizing a Balanced Iterative Reducing as well as clustering using Hierarchies depends on a scalable architecture that utilizes clustering to reduce the search space. Bidirectional Long Short Term Memory (BiLSTM) with random forest classifier models are utilized for analysing temporal data as well as predicting trajectories.

III. WEB USAGE MINING

Web Mining can be widely be viewed as the application of the adapted data mining methods to the web mining and data mining is represented as the application of the algorithm helps to discover patterns. Mostly structured data has fixed into the knowledge discovery process and it has a distinctive property to support a collection of multiple data types. The web mining has several aspects that yield multiple approaches for the mining process including text manages in web pages with hyperlinks connected and helps to identify the utilizers current activity can be more mentioned and monitored via web server logs. The important aim of web usage mining is to discovering interesting patterns by exploiting usage data stored during the interactions of utilizers with the website that generally characterize the navigational behaviour of the users. Web usage mining consists of large collections of data deriving from several sources like as web server access logs, proxy server logs, registration form data as well as mouse movements or clicks.

Web usage mining is a process of extracting useful information readily available on the Internet of the World Wide Web helps to analyse user activities on different web pages as well as track them over a period of time to understand customer's behaviour and also surfing patterns. Web usage mining is broadly categorized into three main sub categories.

Web Content Data

- Web Structure Data
- Web Usage Data

A. Web Content Data

The data from web usage mining in HTML, web pages, images etc. The main layout for the Internet/ Web content is HTML with a slight difference depending upon the utilization of the web browser, but the basic layout structure is the same everywhere.

Mining Techniques Using Agents and Databases

1) Agent-Based Approaches

a) Intelligent Search: The type of search basically refers to a particular goal of the user and will return the results bases on the conclusion of that goal.

b) Information Filtering/ Categorization: This type of search basically deals with the filtering of data that is the removal of unwanted information or redundant information utilizing certain Artificial Intelligence based methods like Pursuit, BO Bookmark Organizer. In fig 1.1 describes the growth of sophisticated AI systems replacing utilizers in an automated or automated manner. One of these are referred to as Deep learning and the system is trained by feeding it with certain kinds of data.

2) Database Approaches

The unstructured data is utilized for transforming data into a more structured and high level collection of resources like as in relational databases as well as utilizing standard deviation database querying mechanisms and datamining techniques to access and clearly analyse the information.

a) Multilevel Databases

- Lowest level- semi structured information is kept.
- High Level- generalization from lower levels organized into the relations and objects.

b) Web Query Systems:

- Web-query systems are developed such as SQL and Natural Language Processing for extracting data.

B. Web Structure Data

On a typical web page and the contents are arranged within the HTML tags. These pages are hyperlinked allowing the users to navigate back and forth to find related information and web structure data is simply relationships links that describe the connection between web pages.

C. Web Usage Data

The web server and the data are generated by the web server and application server on a typical web page. Web/Application server collects the log data including information about the users like geographical location time and the content they interacted with the log files are categorized into three types based on the source it appear from server side, client side, proxy side.

The user visits a web page and they leave a lot of information that the web servers can collect in logs. There is geographical information the path through the pages they have accessed on the web page.

IV. CLUSTERING ALGORITHMS

In machine learning, classification is a supervised learning approach utilized to analyse a given data set besides to build a model that separates data into a desired as well as distinct number of classes.

A. Hierarchical Clustering

Hierarchical algorithm is one of the unsupervised machine learning algorithm, that is utilized to group the unlabelled data sets into a cluster besides is known as hierarchical cluster analysis or HCA.

The hierarchy of clusters in the form of a tree as well as this tree shaped structure is known as the dendrogram. The results of K-means clustering as well as hierarchical clustering may be look similar but they are different based on the work. There is no need to predetermine the number of clusters as in the K-Means algorithm. The hierarchical clustering technique has two approaches:

- *Agglomerative:* Agglomerative is a bottom up approach and this algorithm starts with taking all data points as single clusters and merging them until one cluster is left.
- *Divisive:* Divisive algorithm is the reverse of the agglomerative algorithm as it is a top down approaches.

Hierarchical based clustering is typically utilized on hierarchical data that would to get from a company database or taxonomies. This builds a tree of clusters so that everything is organized from the top –down.This method is more restrictive than the other clustering types it is perfect for specific kinds of datasets.

1) Advantages of Hierarchical Clustering Algorithm are

- Robustness
- Easy to interpret
- Flexible
- Scalable
- Visualization

2) Disadvantages of Hierarchical Clustering Algorithm

- Hierarchical clustering is computationally expensive. The time required to run the algorithm increases exponentially as the number of data points increases, making it difficult to use for large datasets.
- Hierarchical clustering methods require a predetermined number of clusters before they can begin clustering, which may not be known beforehand. This makes it difficult to use in certain applications where this information is not available.
- Agglomerative or divisive clustering is prone to producing overlapping clusters, where different groups of data points may share common characteristics and thus be grouped together even though they should not belong to the same cluster.

B. DBSCAN Clustering Algorithm

Clustering analysis is an unsupervised learning method that separates the data points into several specific bunches or groups that the data points in the same groups have different properties in some sense. It comprises of many different methods based on different distance measures. All clustering methods utilize the same approach first we calculate similarities and then we use it to cluster the data points into groups or batches. The Density based spatial clustering of applications with noise (DBSCAN) clustering method. Density based clustering method refers to unsupervised learning methods that identify distinctive groups/clusters in the data based on the idea that a cluster in data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density[11]. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a base algorithm for density-based clustering. It can discover clusters of different shapes and sizes from a large amount of data, which is containing noise and outliers.

1) Advantages of DBSCAN

- Handles irregularly shaped and sized clusters. One of the main advantages of DBSCAN is its ability to detect clusters that are irregularly shaped.
- Robust to outliers.
- Does not require the number of clusters to be specified.
- Less sensitive to initialization conditions.
- Relatively fast.

2) Disadvantages of DBSCAN

- While DBSCAN is great at separating high-density clusters from low-density clusters, DBSCAN struggles with clusters of similar density.
- Struggles with high dimensionality data. If given data with too many dimensions, DBSCAN suffers.

C. Gaussian Mixture Model Algorithm

Gaussian Mixture Model is a Mixture of Gaussian as it is sometimes called, is not so much a model as it is a probability distribution. It is a universally used model for generative unsupervised learning or clustering. It is also called Expectation-Maximization Clustering or EM Clustering and is based on the optimization strategy. It allows the model to learn the subpopulations automatically. This constitutes a form of unsupervised learning. A Gaussian is a type of distribution, and it is a popular and mathematically convenient type of distribution. A distribution is a listing of outcomes of an experiment and the probability associated with each outcome.

1) The main Advantages of Gaussian Mixture Model can be summarized as

a) *Flexibility:* Gaussian Mixture Models have the ability to model a wide range of probability distributions, as they can approximate any distribution that can be represented as a weighted sum of multiple normal distributions.
b) *Robustness:* Relatively robust to the outliers which are present in the data, as they can accommodate the presence of multiple modes called “peaks” in the distribution.
c) *Speed:* Fast to fit a dataset, especially when using an efficient optimization algorithm such as the expectation-maximization (EM) algorithm.
d) *To Handle Missing Data:* Ability to handle missing data by marginalizing the missing variables, which can be useful in situations where some observations are incomplete.

2) The limitations of Gaussian Mixture Mdoel are

a) *Sensitivity To Initialization:* Sensitive to the initial values of the model parameters, especially when there are too many components in the mixture.
b) *Assumption of Normality*: Gaussian Mixture Models assume that the data are generated from a mixture of normal distributions, which may not always be the case in practice
c) *Number of Components*: Components may overfit the data, while using too few components may underfit the data. The extremes of both points result in a challenging task, which becomes tough to be handled.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue I Jan 2024- Available at www.ijraset.com

D. K-NN

K Nearest Neighbor (KNN) is a good classifier that can predict the class of an instance based on its nearest neighbors. The value of k determines the number of nearest neighbors to consider. The value of k is chosen to be an odd number to avoid race conditions. kNN can be a good model to remove all of the extreme values. K-Nearest Neighbours is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining, and intrusion detection. It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data). We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.

1) Advantages of the KNN Algorithm

- Easy to implement as the complexity of the algorithm is not that high.
- Adapts Easily as per the working of the KNN algorithm it stores all the data in memory storage and hence whenever a new example or data point is added then the algorithm adjusts itself as per that new example and has its contribution to the future predictions as well.
- *Few Hyperparameters:* The only parameters which are required in the training of a KNN algorithm are the value of k and the choice of the distance metric which we would like to choose from our evaluation metric.

2) Disadvantages of the KNN Algorithm

- *Does Not Scale:* The KNN algorithm is also considered a Lazy Algorithm. The main significance of this term is that this takes lots of computing power as well as data storage. This makes this algorithm both time-consuming and resource exhausting.
- *Curse of Dimensionality:* There is a term as the peaking phenomenon according to this the KNN algorithm is affected by the curse of dimensionality that implies the algorithm faces a hard time classifying the data points properly when the dimensionality is too high.
- *Prone to Overfitting:* The curse of dimensionality is a prone to the problem of overfitting as well as the problem of feature selection and dimensionality reduction techniques are applied to deal with this problem.

E. BIRCH Algorithm

Balanced Iterative Reducing and Clustering using Hierarchies**,** or BIRCH for short, deals with large datasets by first generating a more compact summary that retains as much distribution information as possible, and then clustering the data summary instead of the original dataset. BIRCH actually complements other clustering algorithms by virtue if the fact that different clustering algorithms can be applied to the summary produced by BIRCH. BIRCH can only deal with metric attributes.

BIRCH incrementally and dynamically clusters incoming multi-dimensional metric data points to try to produce the best quality clustering with the available resources (i.e., available memory and time constraints). BIRCH can typically find a good clustering with a single scan of the data, and improve the quality further with a few additional scans.

BIRCH is also the first clustering algorithm proposed in the database area to handle "noise" effectively. A metric attribute is one whose values can be represented by explicit coordinates in a Euclidean space. A metric attribute is one whose values can be represented by explicit coordinates in a Euclidean space and BIRCH is most used integrated hierarchical clustering algorithm.

BIRCH is often used to complement other clustering algorithms by creating a summary of the dataset that the other clustering algorithm can now use. A metric attribute is an attribute whose values can be represented in Euclidean space, i.e., no categorical attributes should be present.

The BIRCH clustering algorithm consists of two stages:

- *Building the CF Tree:* BIRCH summarizes large datasets into smaller, dense regions called Clustering Feature (CF) entries.
- *Global Clustering:* Applies an existing clustering algorithm on the leaves of the CF tree. A CF tree is a tree where each leaf node contains a sub-cluster. Every entry in a CF tree contains a pointer to a child node, and a CF entry made up of the sum of CF entries in the child nodes.

1) Advantages of the BIRCH Algorithm

- An advantage of BIRCH is its ability to incrementally and dynamically cluster incoming, multi-dimensional metric data points in an attempt to produce the best quality clustering for a given set of resources.
- It is local in that each clustering decision is made without scanning all data points and existing clusters. It exploits the observation that the data space is not usually uniformly occupied, and not every data point is equally important.
- It uses available memory to derive the finest possible sub-clusters while minimizing I/O costs. It is also an incremental method that does not require the whole data set in advance.

2) Disadvantages of the BIRCH Algorithm

- BIRCH has one major drawback it can only process metric attributes.
- Sensitive to input order of objects.
- Number of node entries nodes does not represent natural clusters.

Table 1.1 Pros and Cons of Clustering Algorithms

ALGORITHMS	AUTHORS	ADVANTAGES	DISADVANTAGES
Hierarchical Clustering	Pranav Shetty et.al	Easy to interpret	Predetermined number of Cluster
DBSCAN	Deng et.al	Robust to outliers	Struggles with high dimensionality data
Gaussian Mixture Model Algorithm	Huang et.al	Flexibility Speed	Sensitivity To Initialization
K-NN	Shukla et.al	All the data in memory storage	Lots of computing power
BIRCH	Madhur et.al	Summarize large Datasets	Process only Metric Objects

The above table 1.1 gives the various machine learning algorithms based on clustering and with their advantages and disadvantages.

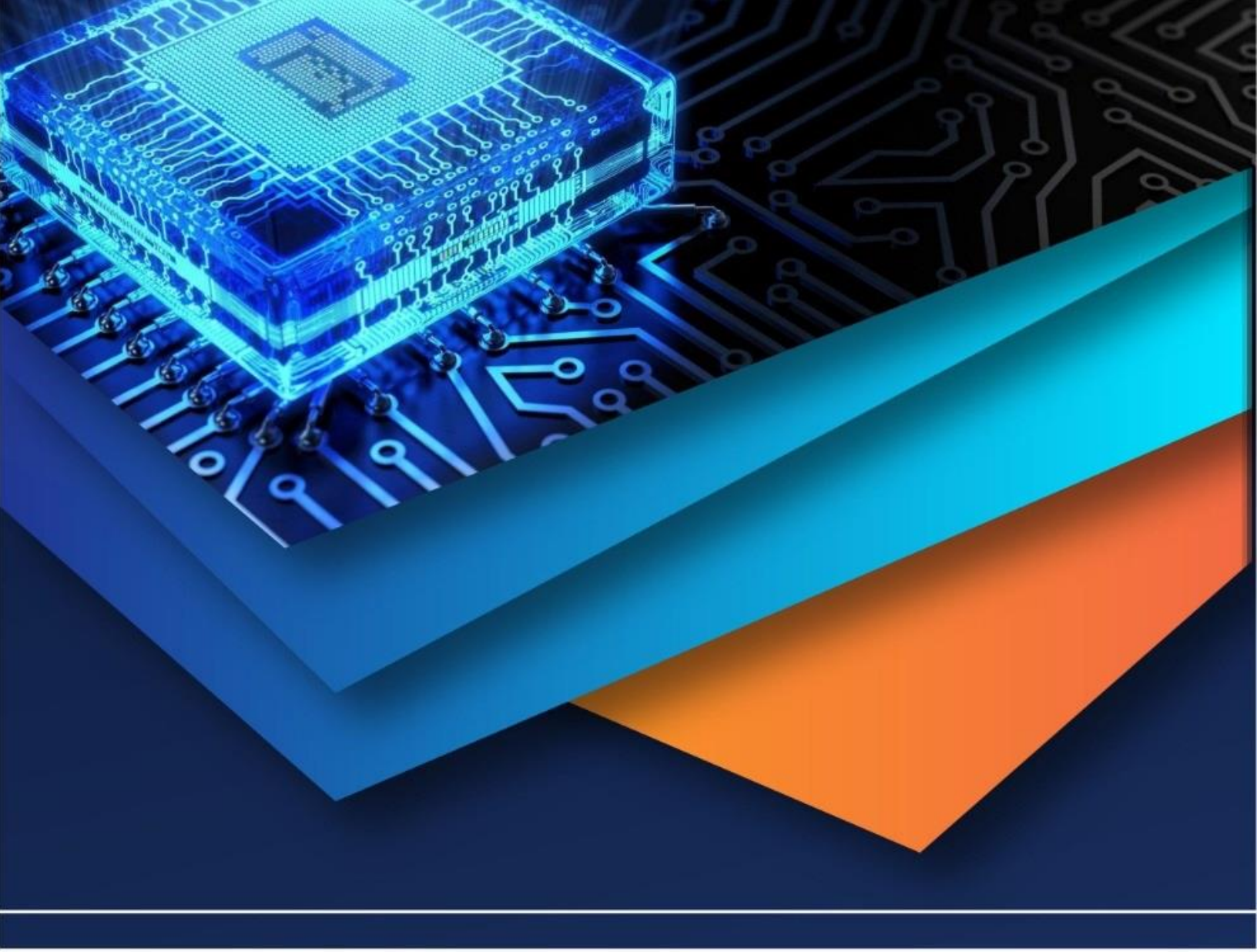
VII. CONCLUSION

This paper provides an survey of rapidly growing area of Web Usage mining. Machine-learning-based text classification is one of the leading research areas and has a wide range of applications, which include spam detection, hate speech identification, reviews, rating summarization, sentiment analysis, and topic modelling. Web mining is faced with challenges such as incorrect and inaccurate data, lack of tools, custom tools, lack of the required resources, management and so on. This paper explains the important clustering research algorithms using machine learning concepts and its issues of web usage mining applications.

REFERENCES

[1] Aggarwal, H. (2017). Review of web usage of datamining in web mining. International Journal of Advanced Research in Computer Science, 8, 2742-2746.
[2] A. J., et al, (2019a) Portable image based moon date detection and declaration: system and algorithm code sign. In 2019 IEEE international conference on computational intelligence and virtual environments for measurement systems and applications (CIVEMSA) (pp. 1–6).
[3] Dr.S.Brindha, Dr.S.Sukumarn, Relevance Pattern Discovery for Text Classification Using Taxonomy Methods" in International Journal for Science and Advance Research in Technology (IJSART)" Volume 4 Issue 11 –November 2018 ISSN [online]:2395-1052.
[4] Dingsheng Deng, DBSCAN Clustering Algorithm Based on Density, 2020 7th International Forum on Electrical Engineering and Automation (IFEEA), IEEE.
[5] Jalili, A. (2019). A new SDN-based framework for wireless local area networks. International Journal of Nonlinear Analysis and Applications, 10, 177-183.
[6] Jalili, A., Keshtgari, M., &Akbari, R. (2018). A new set covering controller placement problem model for large scale SDNs. Information Systems & Telecommunication, 25.
[7] Jalili, A., Keshtgari, M., Akbari, R., &Javidan, R. (2019). Multi criteria analysis of controller placement problem in software defined networks. Computer Communications, 133, 115-128.
[8] Moghadasi, M. N., Safari, Z., &Zhuang, Y. (2020). A sentimental and semantical analysis on Facebook comments to detect latent patterns. In 2020 IEEE international conference on big data (Big Data).IEEE.Moshayedi,

[9] Moshayedi, A. J., et al. (2022). Sunfa Ata Zuyan machine learning models for moon phase detection: algorithm, prototype and performance comparison. TELKOMNIKA (Telecommunication, Computing, Electronics and Control), 20, 129-140.
[10] Nural, M. V., Cotterell, M. E., & Miller, J. A. (2015). Using semantics in predictive big data analytics.In 2015 IEEE international congress on big data.IEEE.
[11] Olson, D. L., & Shi, Y. (2007). Introduction to business data mining, Boston: McGraw- Hill/Irwin. Patel, B. R., &Rana, K. K. (2014). A survey on decision tree algorithm for classification.IJEDR, 2. https://www.ijedr.org/ papers/IJEDR1401001.pdf
[12] PranavShetty and Suraj Singh, Hierarchical Clustering: A Survey, Internaational Journal of Applied Research 2021, Volume 7, Issue 4, Pp: 178-181.
[13] Phyu, A. P., & Thu, E. E. (2021). Short survey of data mining and web mining using cloud computing. International Journal of Advanced Networking and Applications, 12, 4725–4731.
[14] Sasi Kumar, A. and Sasi Kumar, A. and Aithal, P. S., DeepQ Based Heterogeneous Clustering Hybrid Cloud Prediction Using K-Means Algorithm (June 30, 2023). International Journal of Management, Technology, and Social Sciences (IJMTS), 8(2), 273-283. ISSN: 2581-6012, 2023,
[15] Sapountzi, A., &Psannis, K. E. (2018). Social networking data analysis tools & challenges. Future Generation Computer Systems, 86, 893-913. https://doi.org/10.1016/j.future.2016.10.019
[16] Safari, Z., Mursi, K. T., &Zhuang, Y. (2020). Fast automatic determination of cluster numbers for high dimensional big data. In Proceedings of the 2020 the 4th international conference on computer and data analysis.
[17] Xia Xie WC, Fu Y, Jin H, Zhao Y (2020) A novel text mining approach for scholar information extraction from web content in Chinese, future generation computer systems. In: Future generation computer systems, vol 111, pp 859-872.
[18] NS, Shukla MKRK, Sharma P (2020) Web usage mining-a study of Web data pattern detecting methodologies and its applications in data mining. In: 2nd international conference data, engineering application,pp
[19] Huang, H.; Liao, Z.; Wei, X.; Zhou, Y. Combined Gaussian Mixture Model and Pathfinder Algorithm for Data Clustering. Entropy 2023, 25, 946. https://doi.org/10.3390/e25060946
[20] MadhurArora, Sanjay Agrawal, RavindraPatell, User Location prediction using Hybrid BIRCH clustering and Machine Learning approach, Journal of Integrated Science & Technology, Volume 12, Issue2, 2021.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY