WEBSTUR: A TOOL FOR ANALYZING WEB USER BEHAVIOR ON PROXY SERVER USING WEB MINING

DR. JAI PRAKASH [1]

[1] ASSOCIATE PROFESSOR, CHATTERED INSTITUTE OF TECHNOLOGY ABU ROAD.

ABSTRACT:

The growing reliance on data mining techniques across industries necessitates a comprehensive evaluation of their performance, particularly when applied to sparse datasets, a common challenge in real-world applications. This study focuses on a comparative analysis of four widely used data mining methods Apriori, FP Growth, Web Log, and Modified Web Log-evaluated across varying sparsity levels (20%, 40%, 60%, 80%, and 100%) to assess their accuracy and robustness. Employing a methodical approach, the research analyzes how these techniques adapt to the complexities of sparse environments, identifying significant variations in their performance. The Modified Web Log algorithm consistently demonstrated superior performance, achieving accuracy levels exceeding 90% across all sparsity levels, thereby proving its effectiveness in managing sparse datasets with minumal loss of information. FP Growth and Web Log showed moderate performance, maintaining accuracy between 80% and 90%, while Apriori consistently underperformed with accuracy levels below 80%, revealing its limitations in sparse conditions. These findings underscore the critical importance of algorithmic optimization for improving data processing and predictive modeling in industries such as e-commerce, healthcare, and social media, where sparsity is a prevalent issue. By highlighting the effectiveness of the Modified Web Log approach, this study contributes to the ongoing development of efficient data mining methodologies, offering valuable insights for organizations seeking to enhance decision-making and operational efficiency through improved data-driven strategies. The research also provides a foundation for further exploration into advanced algorithmic designs tailored to tackle sparsity and other challenges inherent in large-scale data environments.

KEYWORDS:

WEB MINING, WEB USAGE, USER BEHAVIOR.

PAPER ACCEPTED DATE:	PAPER PUBLISHED DATE:
3rd January 2026	5th January 2026

1. INTRODUCTION

The rapid and exponential expansion of information on the Internet has transformed the World Wide Web into a formidable platform for storing, disseminating, retrieving, and extracting valuable knowledge. Web data research has several obstacles due to the vast, diversified, dynamic, and unstructured characteristics of Web data, including scalability, multimedia, and temporal concerns [1]. Consequently, web users are always inundated with an "ocean" of information, encountering the challenge of information overload throughout their online interactions. The web mining system [2] highlights the magnitude at which data mining methods are used to extract vast amounts of data from the online. Web mining focuses on techniques for collecting important insights from online data. Data mining involves using technical skills to analyze records from unconventional viewpoints, therefore generating novel and intriguing patterns to uncover meaningful connections and trends within massive data, as practiced in network statistics to identify patterns of specialized tools. Web use mining is an emerging subject of research that is increasingly attracting attention in contemporary society. Study web mining methods to understand Internet user behavior [3]. It pertains to predicting user behavior on a website, and anticipating the subsequent page might serve as a significant distraction. For the online consumer, conduct searches for web clients and provide a more engaging and customized site throughout their online experience, among other aspects. Consequently, online use prospecting reveals its relevance in web cache design, network allocation, website restructure and development, website traffic management, and business intelligence. The significant increase in Internet users is drawing the attention of several scholars to investigate the topic of networking via the exploration of intriguing and complex methodologies. Identifying appealing and effective patterns may serve as a crucial asset in enhancing the network environment for users, so drawing new clientele and retaining existing consumers without diminishing their excitement. The process of extracting intriguing data about diverse elements from the internet is referred to as net mining. Web mining first applies to electronic operator-based websites. Examples of e-services include e-commerce, finance, advertising, hanking, education, and government [4]. The research examines the efficacy of several data mining algorithms-Apriori, FP Growth, Web Log, and Modified Web Log-across different degrees of sparsity. Sparsity, a prevalent trait of real-world datasets, profoundly influences algorithmic performance and precision. This study elucidates the comparative efficacy of different algorithms, notably emphasizing the resilience of the Modified Web Log method. This research is noteworthy for its contributions to data mining and algorithm optimization. Analyzing algorithm performance across varying sparsity levels yields significant insights for sectors dependent on sparse datasets, including e-commerce, healthcare, and social media analysis. The results underscore the significance of Modified Web Log algorithms in managing sparse data with precision, providing a means for enhanced decision-making and predictive modeling. The paper solves the essential gap in assessing algorithmic resilience under sparse situations, facilitating future developments in data mining technology.

1.1 WEB MINING

The recent increase in online users has prompted several academics to investigate the web domain in order to uncover intriguing and non-trivial patterns [5]. Identifying intriguing and beneficial patterns may enhance the online experience for users, therefore attracting additional new users and maintaining current users without diminishing their enthusiasm. Web mining is the process of extracting valuable information related to different facets of the web [6]. Web mining is predominantly used in e-service oriented websites [7]. Some of the e-services include ecommerce, e-learning, e-marketing, e-finance, e-banking, and e-governance [8]. Web mining is primarily classified into three categories: (i) web structure mining, (ii) web content mining, and (iii) web usage mining.

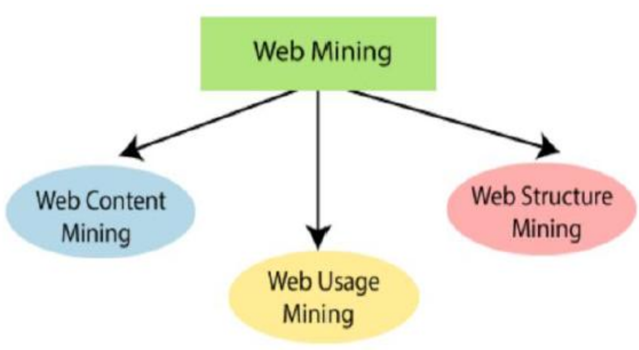

FIGURE 1: TYPES OF WEB MINING [9]

Web structure mining [10][11] seeks to uncover significant patterns from the hyperlink structure at the inter-document level. Web pages are categorized based on the topology of the link structure, allowing for the inference of intriguing information on the similarities and relationships across websites. Relationships that may be established include synonymous, analogous material, co-location on the same server, and others. Web structure mining is used to reveal the hierarchical nature of linkages inside the websites of a certain domain. This facilitates the modeling of effective database systems related to these websites.

A further aspect of web mining is web content mining [12][13]. It analyzes web page text to extract intriguing and non-trivial patterns. A web page may include several types of data, including text, photos, audio, video, metadata, and hyperlinks. Certain data are semi-structured as HyperText Markup Language (HTML), while others are more structured, presented as tables and database-generated HTML pages.

Another avenue in web mining is web use mining [14][15]. It emphasizes the use of data mining methods to identify valuable patterns from secondary data generated by user interactions with the web [16]. A record is kept of the web pages accessed by a person when browsing the internet. Additionally, several additional system-related information is also kept. This record is maintained on the web servers that provide different web pages upon request by online users.

1.2 WEB USAGE MINING TECHNIQUES

In Web usage mining, pattern discovery plays a crucial role. It combines the techniques and algorithms taken from data mining, pattern recognition and machine learning. In data mining. various methods like association rule mining, classification, clustering etc. are used for pattern discovery.

- **ASSOCIATION RULE MINING**

Association rules may be established to identify which web sites were browsed concurrently on the internet. Pages visited concurrently within a single server session may also be recovered. In association rule mining, minimum support is defined as the frequency with which web pages are viewed concurrently. If the count exceeds the minimal support, we assert that the web sites are often browsed in conjunction. This facilitates the prefetching of web pages, hence decreasing time delay and latency. The webpages may also be restructured based on the access logs. The disadvantage of using association rules is that a lower minimum support may result in the generation of several irrelevant rules.

Numerous algorithms have used association rule mining in online usage mining, including the following: Association rules were derived using the scaled Apriori method in [17]. This approach establishes a novel protocol suite that incorporates confidence as a metric for constructing association rules. In [18], the common use patterns were derived from the log data gathered in log files. Apriori and FP-Growth are integrated to identify consumption trends for a certain website. In [19], the Apriori method was augmented to include set size and frequency for the computation of meaningful web sites. The suggested approach enhances memory efficiency.

CLUSTERING ALGORITHM

In clustering, similar web pages are identified and aggregated. This facilitates the identification of persons or groups of users exhibiting analogous navigational behaviors on the web, while also enabling the extraction of data patterns. It also aids in understanding user demographics, enabling the provision of tailored online content to specific users. Clustering is mostly used by scarch engines and online service providers. Clustering encompasses many types, including incremental clustering, hierarchical clustering, and partitional clustering. Several methods that have used clustering in online usage mining include: In [2], the global Fuzzy-C means technique was used for clustering and optimizing the objective function. This technique employs an incremental approach and is independent of beginning circumstances. This enhances clustering and is achieved using a deterministic global search method. The fuzzy-C means approach was used for clustering in [21]. The primary emphasis was placed on enhancing the algorithm's performance. Selecting an appropriate starting centroid cluster is crucial since it significantly impacts the algorithm's performance. This research examines a method for picking an optimal starting centroid to enhance clustering quality. In [22], the author compares clustering techniques, namely the K-Means method with representative object-based algorithms. They are evaluated according to the quantity of dataset points and the number of clusters.

1.3 APPLICATIONS OF WEB USAGE MINING

The primary objective of Web Usage Mining is to gather pertinent and intriguing information on users' navigation habits, primarily to characterize web users. This collected data serves as a foundation for enhancing the website from the user's perspective. Results acquired via the web may be used in several ways:

- User navigation can be improved by pre-fetching and change
- Improving the websites
- Improving the user needs
- Useful in e-commerce
- Personalization of web content

1.3.1 PERSONALIZATION OF WEB CONTENT

Various types of web use mining may be employed to tailor webpages based on user profiles and behavior. Personalization offers several benefits, including the automation of goods for clients, the cultivation of deep relationships, and the formulation of new market strategies. It is used to provide more information on users, aiding them in enhancing their website designs [23].

1.3.2 E-LEARNING ENVIRONMENT

Web Usage Mining is crucial in the e-learning environment, mostly for monitoring website activity and collecting behavioral patterns, which may then inform website improvements. The mining tool enables us to analyze commonly accessed sites, links, and other related data [24].

1.3.3 SITE DESIGN SUPPORT

When it comes to developing the website, the designers face a substantial number of challenges. Through the process of web usage mining, consumers are provided with more crucial information about their behavior, which enables them to modify the content of the website or redesign it. In addition to providing further assistance to the designers, it also makes their task much simpler after observing the behavior and pattern of the users [25].

1.3.4 BUSINESS INTELLIGENCE

Web usage mining enhances market strategies, assists organizations in sustaining their market position, and facilitates decision-making to increase performance. It pulls information on client behavior, which may be used from the stored database to formulate the company's marketing strategy.

2. REVIEW OF LITERATURE

Benova et al., (2024)[26] introduced an innovative machine learning framework for web server anomaly detection that integrates the Isolation Forest method with expert assessment, concentrating on individual user actions in NGINX server logs. In this study, the method overcomes the constraints of conventional techniques by proficiently isolating and examining small abnormalities within large datasets. The Isolation Forest approach was first used on massive NGINX server logs, effectively detecting anomalous user actions that traditional methods often miss. Subsequently, they used DBSCAN for comprehensive clustering of these anomalies, classifying them according to user request timings and categories. An essential advancement of our process is the integration of post-clustering expert evaluation. Cybersecurity experts assessed the discovered clusters, including an essential layer of qualitative evaluation. This enabled the precise differentiation between benign and possibly detrimental behaviors, resulting in targeted interventions such as access limitations or web server configuration modifications. The method signifies a substantial improvement in network security, providing a more nuanced comprehension of user behavior. They provided a detailed and sophisticated method for improving cybersecurity by combining algorithmic accuracy with expert insights. This study enhanced anomaly detection approaches and underscores the essential need for a comprehensive strategy to safeguarding web server infrastructures.

Win et al., (2024) [27] examined that the extraction of data from the Internet involves using data mining approaches and techniques to acquire information via hyperlinks, server logs, online publications, web-based services, and more sources. via the web mining is the act of collecting and analyzing data to identify patterns in web pages that may provide insights into customer behavior, market trends, and general human behavior. It is the process of retrieving information from the Internet that individuals need. The information provided by the Web not only aligns with the user's exact requirements but also proposes similar stuff. The three primary responsibilities included under the domain of web mining are web content mining, web structure mining, and web use mining. Web mining include a component known as "web usage mining," which collects data from web log files on the online behaviors of users and customers. To identify individuals exhibiting similar behavior on a particular website, an analysis of web log data is required. The strategies of web use are examined, including the preprocessing stages fundamental to web mining. This study utilized click-stream data derived from a web use mining technique to ascertain client interest in goods, specifically for items offered on e-commerce platforms. This study further analyzes K-Means clustering. DBSCAN, and Hybrid Density-Based K-Means Clustering, after an introduction to the ideas and methods of data mining, along with the classification and techniques of Web data mining for web log files. The system displayed a comparison of successful clustering outcomes for mean squared errors in the K-Means and Hybrid Density-Based K-Means clustering algorithms.

Leenas et al., (2023)[28] stated that contemporary businesses was shifting from manual processes to web-based solutions across all sectors. Consequently, the number of Internet users is expanding dramatically. Consequently, there is significant traffic, which elevates the strain on the server and the delay of server responses for web objects. The proxy server caching technique is a method to improve the speed of accessing web assets via the Internet. Due to the finite capacity of the cache, a replacement approach is necessary to determine which cached web object should be removed to accommodate new web objects. Proxy servers use many cache replacement techniques, including Least Recently Used (LRU), Least Frequently Used (LFU), and SIZE. The web items in the proxy cache are affected by factors such as recency, frequency, retrieval time, and size. Conventional caching strategies evaluate just a single element at a time, resulting in the inefficient use of cache memory for unpopular web items (cache pollution), which diminishes the performance of the proxy cache. To enhance speed, they recommended use the proxy server log file to identify specific users and their sessions, while categorizing web items into three classifications: high priority, moderate priority, and low priority. The generated log file is used for training the classifiers. Future requests are categorized as high, moderate, or low priority using classifiers, after which a decision is made about their storage in the proxy cache. The aim of this study is to improve the proxy caching mechanism via the use of the aforementioned strategies. They evaluated the efficacy of the proposed strategy against conventional caching strategies via a trace-driven simulation technique. The analysis used two performance metrics: Hit Ratio (HR) and Byte Bit Ratio (BHR). The testing results indicated that the suggested method surpasses conventional caching strategies.

Canay et al., (2023) [29] studied that the primary data source for web use mining (WUM) is often considered to be server logs. Nevertheless, access log files provide only restricted information about the customers. Extracting sessions from this disorganized data requires significant work, and the procedures conducted for this aim do not consistently provide optimal outcomes. This data is not suitable for effective web analytics. This study introduced a method for user monitoring. session management, and the collection of web usage statistics. The technique primarily relies on a novel strategy for using gathered data as the data source in online use mining for web analytics extraction. An application-based API has been created using an alternative approach to traditional client-side techniques for acquiring and processing log data. The log data has been effectively collected by incorporating the approach into a corporate web application. The findings indicate that the uniformly acquired and stored data using this technique is more efficient for browsing. filtering, and processing compared to web server logs. This structured data serves as a dependable source for high-performance internet usage mining, real-time web analytics, machine learning algorithms, or an effective recommendation system.

Benova et al., (2023) [30] aimed to user abnormalities within large datasets of web server queries. They offered a method for monitoring user behavior in NGINX logs, using a cybersecurity company's network of web servers as a case study. The suggested technique does not need a labeled dataset and can effectively detect various user abnormalities in big datasets including millions of daily requests. The analysis results enhanced comprehension of user behavior in requesting updates via internet searches and facilitated the interpretation of the findings. The clustering of anomalies facilitated the formation of typical clusters and enhanced the understanding of the findings. This study offered significant insights into user behavior inside web server networks and underscores the need of effective anomaly detection methods in extensive datasets. The results had significant practical implications in cybersecurity, especially in offering network security experts an automated and more impartial method for threat analysis. This study highlighted the significance of automated techniques for assessing user behavior in web server networks, offering a more objective and effective means of identifying user abnormalities within extensive datasets. This approach enhanced the efficacy and accuracy of cybersecurity systems, hence augmenting the safeguarding of network infrastructures against harmful incursions.

Bommi et al., (2022)[31] analyzed that the web servers experienced information overload, making it difficult for users to grasp the correct concept. The study in web mining is focused on enhancing the information retrieval process. Web mining has a significant position as it assists firms and website analysts in making informed selections about their clients. Web use mining gathers the surfing information of website visitors, which is recorded as plain text files on web servers. This information aids in identifying the navigation activities of website visitors. Web suggestion, a type of web customization, promotes web sites to users based on their previous browsing experience. Web recommendation systems assist website organizations in enhancing page navigation, swiftly reaching their objectives, and acquiring relevant insights. This study developed an effective via the web page recommendation strategy using web use mining and pattern mining algorithms to enhance the accuracy of current web page suggestion methods. The objective is to classify user behavior in identifying browser designs and navigation data of online users, as well as to assess the efficacy of the Frequent Pattern (FP) Growth algorithm, the Apriori algorithm, and a modified web log search algorithm. The association rule mining technique was used to the clustered web logs to determine the frequently visited web sites by users. Real-time datasets were assessed. The suggested recommendation approach is evaluated against current strategies to demonstrate its enhancement in web page suggestion accuracy.

Gangadwala et al., (2022) [32] stated that the internet serves as a substantial repository of information, with an exponential increase of data occurring continuously. The substantial population of internet users is growing daily. Numerous research is being conducted to reduce the surfing time of web users. In web use mining, techniques are executed on a proxy server dataset to analyze the behavior of web users. Clustering is essential in several applications, including online log dataset analysis, customer relationship management, advertising and marketing. scientific diagnostics, and computational biology, among others. Clustering refers to a collection of associated data pieces. The primary concern in clustering is the many metrics that may determine if two things are similar or dissimilar. This study delineated many validity metrics, including the partition coefficient and partition entropy. Experiments conducted on several fuzzy clustering algorithms, including Fuzzy C-Means, Fuzzy Possibilistic C-Means, and Modified Fuzzy Possibilistic C-Means Clustering. These approaches are formulated and evaluated for the proxy log dataset. The experimental outputs of the algorithms are ultimately assessed. The experiential results clearly indicated that the clusters generated by the suggested MFPCM technique significantly outperform those produced by k-Means, Fuzzy C-Means, Fuzzy Possibilistic C-Means, and Modified Fuzzy Possibilistic C-Means clustering algorithms in terms of several validity parameters.

Sowmya et al., (2022) [33] examined that session identification is a crucial element for analyzing use patterns from online trace log files. Numerous studies have shown dynamic log session identification using diverse optimization strategies; nonetheless, substantial obstacles persist due to the data volume, the use of multiple logging technologies, and the obfuscation of Internet Protocol (IP) addresses during proxy access. This study introduced an innovative online dynamic session identification system using weblog trace data. The innovative method employs a user-specified schema including distinct IP addresses, a unique Uniform Resource Locator (URL), session count, and average session duration to identify sessions. The aim is to provide a scalable framework for user and session identification in weblog data. The result is contingent upon a scalable association rule mining method. Certain regulations are implemented, including frequent page views contingent upon the duration of time spent on the page. To minimize the frequency of comprehensive log scans, an improved relationship is established. The identification efficacy of the suggested framework is assessed. The achieved minimal Mean Absolute Percentage Error (MAPE) in identification is 7%. The weblog has 1 million log entries, and the MAPE is 11%, indicating a considerable value. The framework attained an hourly Root Mean Square Error (RMSE) of 0.0500, surpassing the performance of existing paradigms using Artificial Neural Networks (ANN) at 0.0639.

Roy et al., (2021) [34] presented several methods in online use mining to minimize user latency and enhance web performance. The server-side web log data assists in determining the most suitable pages according to the user's request. Analyzing web log data presents challenges due to its comprehensive nature. This study introduced a novel approach for preprocessing web log data to uncover sequence and navigation patterns that are beneficial for prediction. This study clusters people into groups identified by their most frequently visited websites to uncover preferences and inform website reorganization. The sequence of URLs accessed by a user during a session. A session indicates a user's surfing behavior. Sessions of several users are aggregated using the hierarchical clustering method to examine the sequence of their navigation habits. The two predictive models are Adaptive Mahalanobis Distance (AMD) and Hidden Markov Model (HMM), which provide a list of relevant web sites for the user. They assessed the propos architecture using web log files from the NASA, Clarknet, and SERC datasets.

Bhuvaneswari et al., (2021)[35] recognizing people with similar interests is essential for developing the recommendation model. The web server log serves as a repository from which information necessary for identifying users and sessions (pagesets) is retrieved. The Sparse ID list and Vertical ID list are used to identify closed frequent page sets, which is advantageous for memory and processing efficiency. Thes user's surfing behavior is determined by calculating the similarity among the pages inside the user's set. In this study, a novel measure for assessing intra-user similarity is suggested. This technique is innovative in that it considers only users with consistent behavior over time for clustering. Consistent users are then grouped using several clustering methodologies, including Agglomerative clustering, Clustering Large Applications Using RANdomized Search (CLARANS), and the suggested Density-Based Community Detection (DBCD). The clusters generated by DBCD demonstrate superior quality for user clustering. The results indicated substantial improvements in both the quality and speed of the clustering.

Ali et al., (2020)[36] studied an extracting relevant and valuable information from the vast quantities of data generated by online users is a formidable challenge. This study pertained to applications using Web Usage Mining (WUM). Clickstream, transaction data, and user profile data provide distinct sources of online use information. Preprocessing is a crucial procedure in WUM for comprehending the user's overall experience during a session on a website. It denoted the accumulation of successive page views performed by an individual person navigating a website. Consequently, it is a vital duty that entailed transforming raw web log data to derive an appropriate pattern for effective analysis. The significance of these operations is attributed to the intricate structure of web architecture, constituting over 80% of the mining process. This study presented a thorough framework for data preparation, accompanied by an extensive demonstration using a real dataset.

Choi et al., (2020) [37] studied that proxy servers function as intermediaries and gateways between users and other Internet servers, offering several advantageous applications that enhance user privacy, such as circumventing server-side restrictions and geographical limitations. Although proxies have advantageous uses, they are also used by adversaries to conceal their identities and perpetrate various assaults. Consequently, several websites impose restrictions on proxy access, leading to the creation of blacklists to identify and facilitate the banning of such proxies. This study examines the ecology of proxies by analyzing their affinities and distributions in a comparable manner. They analyzed residential and open proxies through country-level and city-level assessments to elucidate their geospatial distributions, similarities, and differences, utilizing an extensive array of blacklists and associated categories, such as spam and maliciousness analysis, to comprehend their characteristics and attributes. They concluded that, although pursuing the same objective, residential and open proxies possess unique properties that need their independent consideration within the broader Internet ecosystem. Furthermore, we emphasize the relationship between proxy location distribution and five national characteristics, including Internet censorship, political stability, and Gross Domestic Product (GDP).

3. PROBLEM STATEMENT

The exponential increase in web usage has resulted in a vast quantity of data produced by users engaging with websites and web apps, presenting difficulties for enterprises to comprehend and improve user behavior efficiently. Proxy servers, acting as mediators between users and the internet, generate extensive log data that is mostly underused for meaningful information. Although web mining methods has the capacity to analyze data and reveal trends in user activity. current solutions sometimes lack the ability to interact smoothly with proxy servers and deliver thorough, real-time analysis. This disparity leads to lost possibilities for augmenting user experience, optimizing website performance, and identifying abnormalities, including harmful actions. Consequently, there is an urgent want for a comprehensive, intuitive solution that utilizes sophisticated web mining methodologies to examine user activity on proxy servers. This solution should provide an in-depth analysis of user navigation patterns, content preferences, and use trends, while maintaining data privacy and scalability, therefore enabling enterprises to make informed choices and enhance their digital strategy.

4. RESEARCH METHODOLOGY

The section employed a structured methodology to analyze the performance of four data mining algorithms Apriori, FP Growth, Web Log, and Modified Web Log-under varying sparsity levels of 20%, 40%, 60%, 80%, and 100%. These logs served as the primary input for analysis. In online use mining, connection criteria are employed to identify the relationships among user sessions that often occur simultaneously. Each association rule presented above is predicated on the correlation between the Support and Confidence metrics. The quantity of transactions containing the specified ordered pair corresponds to the Support of that item set. Web use mining commences with the configuration of factors such as support and confidence. The quantity of database scans required for item set construction, the data preprocessing approach used, and the method's sensitivity to variations in user settings, such as assistance, have all been analyzed. The Apriori method employs an efficient procedure for reducing candidate item sets; yet, it is time-consuming and requires many dataset scans to generate contender item sets. The FP-growth approach needs just two scans of the database for the development of common patterns. This method employs a compact tree structure to represent the whole database. It eliminates the need for candidate generation, hence reducing computation time. The web log search technique analyzes log data derived from user involvement over a specified duration. Figure 2 illustrates the proposed methodology. Real-time datasets from Sci-Hub, PUBG, Amazon, and Bitcoin are evaluated to ascertain user behavior. Nevertheless, the proposed web log algorithms exhibit superior overall speed and memory utilization across diverse data structures. The suggested Modified Web log search strategy requires less effort compared to standard methods.

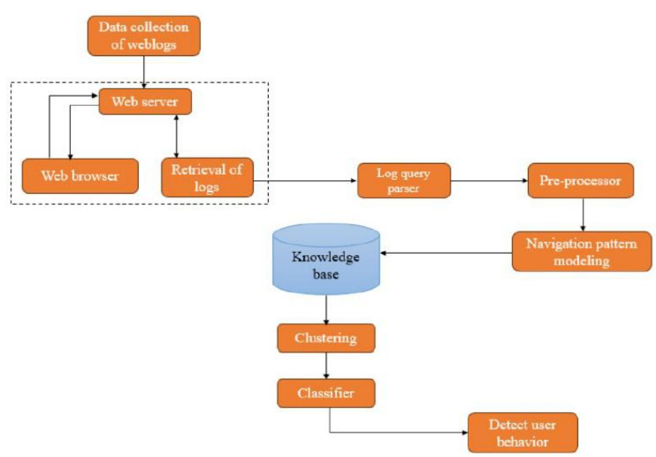

FIGURE 2: FLOWCHART OF METHODOLOGY

The steps of the methodology are described as follow:

STEP 1: DATA COLLECTION OF WEBLOGS

The process begins with gathering data from weblogs. These logs are generated from user activities such as browsing, searching, or interacting with web applications. The data is collected through web servers and web browsers, which act as intermediaries for recording user interactions. This raw log data serves as the primary input for further processing and analysis.

STEP 2: WEB SERVER AND RETRIEVAL OF LOGS

The web server captures all user requests and responses, storing them in log files. These files contain valuable information such as timestamps, requested URLs, IP addresses, and user agents. The retrieval of logs involves extracting this information for subsequent processing. This step ensures that all relevant data points from user interactions are systematically recorded.

STEP 3: LOG QUERY PARSER

After retrieving the logs, the log query parser breaks down the raw data into structured components. It filters out irrelevant information and organizes the logs into meaningful categories such as user sessions, query patterns, and navigation paths. This step ensures that the data is in a format suitable for analysis.

STEP 4: PRE-PROCESSOR

The pre-processor cleanses and transforms the parsed data. Tasks include removing duplicates, handling missing values, normalizing data formats, and converting raw logs into a standardized structure. This step also segments user sessions based on activity timelines and eliminates noise to improve the accuracy of the analysis.

STEP 5: NAVIGATION PATTERN MODELING

Using the pre-processed data, navigation patterns are modeled to understand user behavior. This step identifies trends in user navigation, such as frequently visited pages, paths taken through the website, and time spent on specific sections. The insights gained here are crucial for identifying user preferences and behavior trends.

STEP 6: KNOWLEDGE BASE

The processed data and derived navigation patterns are stored in a knowledge base, a centralized repository for organizing and managing data. This knowledge base acts as a reference for clustering and classification algorithms, enabling them to efficiently access relevant patterns and metrics.

STEP 7: CLUSTERING

In the clustering step, the data is grouped into clusters based on similarities in user behavior. For instance, users with similar browsing habits or preferences may be grouped together. Clustering algorithms such as K-means are applied to discover these groups. This step is essential for segmenting users and tailoring services to specific clusters.

STEP 8: CLASSIFIER

After clustering, a classifier is applied to assign labels or categories to the data. This step uses supervised learning techniques to identify specific user behaviors, such as distinguishing between casual browsers and frequent buyers. The classifier refines the understanding of user behavior by predicting categories based on past patterns.

STEP 9: DETECT USER BEHAVIOR:

The final step is to detect and interpret user behavior. By analyzing the classified data, insights are drawn about how users interact with the system. For example, patterns such as frequent product searches or prolonged page visits may indicate purchase intent. These insights help in personalizing user experiences, improving website design, and optimizing content delivery.

4.1 CLUSTERING TECHNIQUE

- **K-MEANS**

K-Means Clustering is an unsupervised learning process that categorizes an unlabeled dataset into distinct clusters. K denotes the quantity of predetermined clusters to be established in the procedure; for instance, K-2 results in two clusters, K-3 yields three clusters, and so on. It is an iterative approach that partitions the unlabeled dataset into k distinct clusters, ensuring that each dataset is assigned to a single group with analogous characteristics. It enables the clustering of data into distinct groups, facilitating the autonomous identification of categories within an unlabeled dataset without requiring any prior training. It is a centroid-centric method, whereby each cluster is linked to a centroid. The primary objective of this technique is to reduce the aggregate distances between data points and their respective clusters. Figure 3 depict the working process of k-means algorithms as shown below.

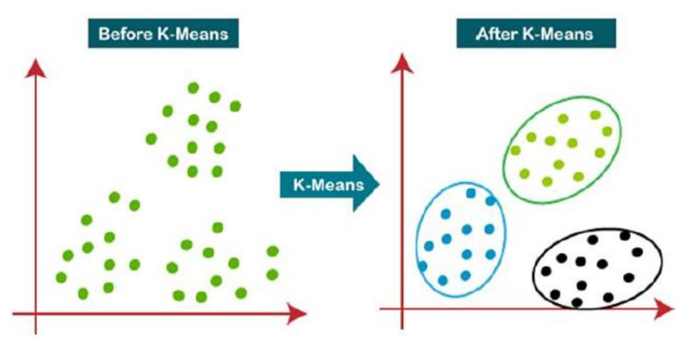

FIGURE 3: WORKING PROCESS OF K-MEANS [38]

4.2 DATASET

The real-life datasets chosen are shown in Table 1. The real-time datasets of Sci-Hub, PUBG, Amazon, and Bitcoin are analyzed to ascertain user behavior.

TABLE 1: DATASET

Dataset	Website	Number of logins	Number of items	Average number of transactions per items
SCI-HUB	Search.dataone.org	49301	98	18
PUBG Games	Figshare.com	128342	106	14

Amazon e-commerce	data.world	1134202	108	10
Bitcoin Block chain historical data	www.kaggle.com	261268	112	15

5. RESULT AND DISCUSSION

The findings reveal distinct performance patterns for the four algorithms across the tested sparsity levels. An analysis of precision and recall metrics further emphasized the advantages of the Modified Web Log algorithm, which maintained a strong balance between retrieving relevant data and minimizing false positives. This resulted in the highest Fl-scores among all algorithms.

Figure 4 depict the comparison of the precision of four algorithms Apriori, FP Growth, Web Log, and Modified Web Log-across different sparsity levels (20%, 40%, 60%, 80%, and 100%). The precision of all methods increases with higher sparsity levels. The Modified Web Log algorithm consistently shows the highest precision across all sparsity levels, while Apriori and FP Growth generally exhibit lower performance. This indicates the Modified Web Log's superior accuracy in sparse data environments.

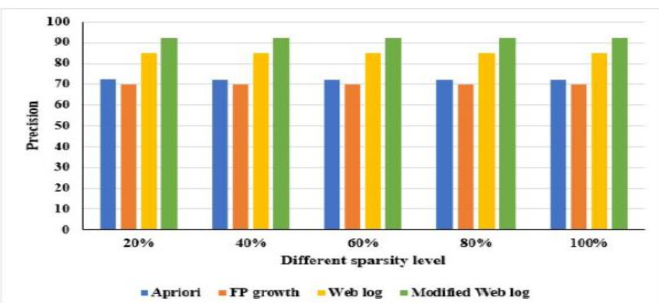

FIGURE 4: COMPARISON GRAPH OF PRECISION

Figure 5 depicts the recall performance of four algorithms-Apriori, FP Growth, Web Log, and Modified Web Log-across various sparsity levels (20%, 40%, 60%, 80%, and 100%). The Modified Web Log algorithm consistently achieves the highest recall at all sparsity levels, indicating its superior ability to retrieve relevant results. In contrast, Apriori and FP Growth exhibit relatively lower recall values, with Web Log performing moderately. The trend shows stability in recall across sparsity levels, highlighting the robustness of the algorithms in sparse data, particularly the Modified Web Log.

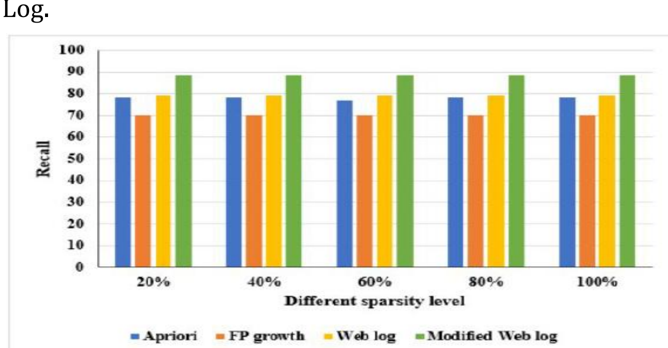

FIGURE 5: COMPARISON GRAPH OF RECALL

Figure 6 illustrates the Fl-score performance of four algorithms-Apriori, FP Growth, Web Log, and Modified Web Log-across different sparsity levels (20%, 40%, 60%, 80%, and 100%). The Modified Web Log consistently achieves the highest FI-scores across all sparsity levels, indicating its superior balance between precision and recall. The Web Log algorithm follows with moderate performance, while Apriori and FP Growth show relatively lower Fl-scores. The results demonstrate that the Modified Web Log is the most effective in handling varying sparsity levels, maintaining high accuracy and reliability in sparse data contexts.

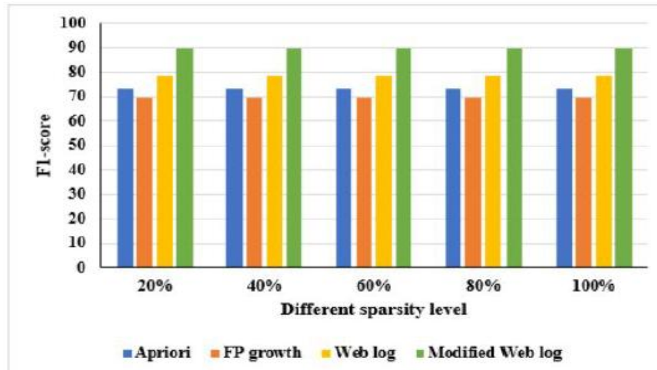

FIGURE 6: COMPARISON GRAPH OF FL-SCORE

Figure 7 depict the comparison of the accuracy of four methods Apriori, FP Growth, Web Log, and Modified Web Log at different sparsity levels (20%, 40%, 60%, 80%, and 100%). The accuracy of the methods generally increases as the sparsity level rises. Modified Web Log consistently achieves the highest accuracy across all sparsity levels, reaching near 90% or higher. FP Growth and Web Log perform better than Apriori, with FP Growth showing slightly higher accuracy than Web Log. Apriori exhibits the lowest accuracy among the four methods, remaining consistent across sparsity levels. The figure highlights the superior performance of Modified Web Log for sparse datasets.

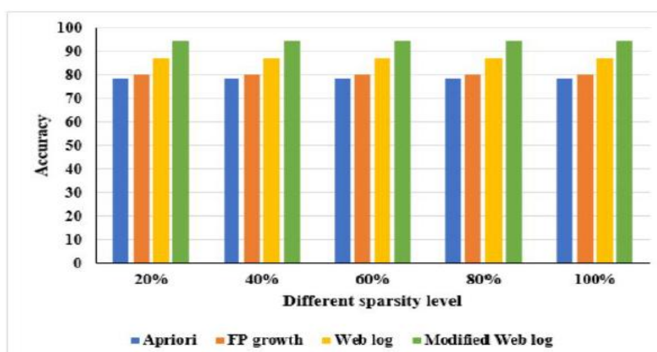

FIGURE 7: COMPARISON GRAPH OF ACCURACY

6. CONCLUSION

The findings of this study underscore the critical importance of selecting appropriate algorithms for data mining tasks, particularly when dealing with sparse datasets. The Modified Web Log method emerged as the most effective, maintaining accuracy levels above 90%

across all sparsity scenarios. FP Growth and Web Log methods demonstrated competitive performance, achieving accuracies between 80% and 90%, while the Apriori method lagged behind, with accuracy remaining below 80%. This performance disparity illustrates the superiority of optimized web log algorithms in addressing sparsity challenges. The research concludes that the Modified Web Log method offers the most reliable and scalable solution for applications requiring high accuracy in sparse data environments. These insights provide a foundation for improving decision-making processes in industries reliant on large, sparse datasets and underscore the need for continued development of advanced data mining algorithms to meet the demands of complex, data-driven tasks.

REFERENCES

1. Varnagar, Chintan R., Nirali N. Madhak, Trupti M. Kodinariya, and Jayesh N. Rathod. "Web usage mining: a review on process, methods and techniques." *In 2013 International Conference on Information Communication and Embedded Systems (ICICES)*, pp. 40-46. IEEE, 2013.

2. Cooley, Robert, Bamshad Mobasher, and Jaideep Srivastava. "Web mining: Information and pattern discovery on the world wide web." In Proceedings ninth IEEE international conference on tools with artificial intelligence, pp. 558-567. IEEE, 1997.

3. Mughal, Muhammd Jawad Hamid. "Data mining: Web data mining techniques, tools and algorithms: An overview." *International Journal of Advanced Computer Science and Applications* 9, no. 6 (2018).

4. Suguna, R., and D. Sharmila. "An overview of web usage mining" International Journal of Computer Applications 39, no. 13 (2012): 11-13.

5. Sharma, Kavita, Gulshan Shrivastava, and Vikas Kumar. "Web mining: Today and tomorrow." *In 2011 3rd International Conference on Electronics Computer Technology*, vol. 1, pp. 399-403. IEEE, 2011.

6. Kosala, Raymond, and Hendrik Blockeel. "Web mining research: A survey." *ACM Sigkdd Explorations Newsletter* 2, no. 1 (2000): 1-15.

7. Vidya, S., and K. Banumathy. "Web mining-concepts and application." *International Journal of Computer Science and Information Technologies* 6, no. 4 (2015): 3266-3268.

8. Gautam, Dr SS, and Manish Kumar Tiwari. "Web Mining Concepts and its Applications." *International Research Journal of Computer Science (IRJCS)* 01 (2014): 8-13.

9. https://www.javatpoint.com/data-mining-world-wide-web

10. da Costa, Miguel Gomes, and Zhiguo Gong. "Web structure mining: an introduction." *In 2005 IEEE International Conference on Information Acquisition*, pp. 6-pp. IEEE, 2005.

11. Baeza-Yates, Ricardo, and Paolo Boldi. "Web structure mining." *In Advanced Techniques in Web Intelligence-I*, pp. 113-142. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.

12. Xu, Guandong, Yanchun Zhang, Lin Li, Guandong Xu, Yanchun Zhang, and Lin Li. "*Web content mining." Web Mining and Social Networking: Techniques and Applications* (2011): 71–87.

13. Johnson, Faustina, and Santosh Kumar Gupta. "Web content mining techniques: a survey." *International journal of computer applications* 47, no. 11 (2012).

14. Liu, Bing. "Web usage mining." *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data* (2007): 449-483.

15. Román, Pablo E., Gastón L'Huillier, and Juan D. Velásquez. "*Web usage mining." Advanced techniques in web intelligence*-1 (2010): 143-165.

16. Srivastava, Jaideep, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. "Web usage mining: Discovery and applications of usage patterns from web data*." Acm Sigkild Explorations Newsletter* 1, no. 2 (2000): 12-23.

17. Wang, Chuan, Yancheng Liu, Qinjin Zhang, Haohao Guo, Xiaoling Liang, Yang Chen, Minyi Xu, and Yi Wei. "Association rule mining based parameter adaptive strategy for differential evolution algorithms." *Expert Systems with Applications* 123 (2019): 54-69.

18. Srivastava, Jaideep, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. "Web usage mining: Discovery and applications of usage patterns from web data." *Acm Sigkdd Explorations Newsletter* 1, no. 2 (2000): 12-23.

19. Rathee, Sanjay, Manohar Kaul, and Arti Kashyap. "R-Apriori: an efficient apriori based algorithm on spark." *In Proceedings of the 8th workshop on Ph. D. Workshop in information and knowledge management*, pp. 27-34.2015.

20. Wang, Weina, Yunjie Zhang, Yi Li, and Xiaona Zhang. "The global fuzzy c-means clustering algorithm." *In 2006 6th World Congress on Intelligent Control and Automation*, vol. 1, pp. 3604-3607. IEEE, 2006.

21. Cai, Weiling, Songcan Chen, and Daoqiang Zhang. "Fast and robust fuzzy e-means clustering algorithms incorporating local information for image segmentation." *Pattern recognition* 40, no. 3 (2007): 825-838.

22. Yuan, Chunhui, and Haitao Yang. "Research on K-value selection method of K-means clustering algorithm." *J*2, no. 2 (2019): 226-235.

23. Md Amin, Mohd Afandi. "A user acceptance model of web personalization systems." PhD diss., Queensland University of Technology, 2012.

24. Mobasher, Bamshad, Robert Cooley, and Jaideep Srivastava. "Automatic personalization based on web usage mining." *Communications of the ACM* 43, no. 8 (2000): 142-151.

25. Abalı, Secil. "An implementation of web usage mining for an e-commerce web site." Master's thesis, Marmara Universitesi (Turkey), 2006.

26. Benova, Lenka, and Ladislav Hudec. "Comprehensive Analysis and Evaluation of Anomalous User Activity in Web Server Logs." *Sensors* 24, no. 3 (2024): 746.

27. Win. Tin Nilar, and Nang Khine Zar Lwin. "Analysis of Customers Interest for Web Log Clustering. *In 2024 IEEE Conference on Computer Applications (ICCA)*. pp. 1-6. IEEE, 2024.

28. Leenas, Thasan, and H. A. Caldera. "Performance Improvement of Proxy Server Cache Management Using Web Usage Mining." *In 2023 8th International Conference on Information Technology Research (ICITR)*, pp. 1-6. IEEE, 2023.

29. Canay, Özkan, and Ümit Kocabıçak. "An innovative data collection method to eliminate the pre-processing phase in web usage mining." *Engineering Science and Technology, an International Journal* 40 (2023): 101360.

30. Benova, Lenka, and Ladislav Hudec. "Using Web Server Logs to Identify and Comprehend Anomalous User Activity." *In 2023 17th International Conference on Telecommunications (ConTEL)*, pp. 1-8 IEEE, 2023.

31. Bommi Harika, Dr T. Sudha. "Identification of User Behaviour by Web Usage Mining." *Mathematical Statistician and Engineering Applications* 71, no. 4 (2022): 678-692.

32. Gangadwala, Hardik A., and Ravi M. Gulati. "Analysis of Web Usage Mining Using Various Fuzzy Techniques and Cluster Validity Index. *In 2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, pp. 1-7. IEEE, 2022.

33. Sowmya, H. K., and R. J. Anandhi. "An efficient and scalable dynamic session identification framework for web usage mining" *International Journal of Information Technology* 14, no. 3 (2022): 1515-1523.

34. Roy, Rita. "Predicting User's web navigation behaviour using AMD and HMM approaches." *In IOP Conference Series: Materials Science and Engineering*, vol. 1074, no. 1, p. 012031. IOP Publishing, 2021.

35. Bhuvaneswari, M. S., and K. Muneeswaran. "User community detection from web server log using between user similarity metric." *International Journal of Computational Intelligence Systems* 14, no. 1 (2021): 266-281.

36. Ali, Noaman M., Ahmed M. Gadallah, Hesham A. Hefny, and Boris Novikov. "An integrated framework for web data preprocessing towards modeling user behavior." *In 2020 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon)*, pp. 1-8. IEEE, 2020.

37. Choi, Jinchun, Mohammed Abuhamad, Ahmed Abusnaina, Afsah Anwar, Sultan Alshamrani, Jeman Park, Dachun Nyang, and David Mohaisen. "Understanding the proxy ecosystem: A comparative analysis of residential and open proxies on the internet." *IEEE Access* 8 (2020): 111368-111380.

38. https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning