Int. j. inf. tecnol. (May 2022) 14(3):1515–1523
https://doi.org/10.1007/s41870-022-00867-3

ORIGINAL RESEARCH

An efficient and scalable dynamic session identification framework for web usage mining

H. K. Sowmya[1] · R. J. Anandhi[1]

Received: 30 August 2021 / Accepted: 10 January 2022 / Published online: 9 February 2022

Abstract Session identification is an essential component for determining usage patterns from web trace log files. Several papers have reported dynamic log session detection incorporating various optimization techniques, however the significant challenges exists considering the volume of Data, usage of different types of logging mechanisms and masking of Internet Protocol (IP) addresses during proxy access. This paper presents a novel online dynamic session identification framework based on the weblogs trace. The novel approach uses a user-defined schema with unique IPs, a unique Uniform Resource Locator (URL), the number of sessions, and the average session length to detect sessions. The objective is to develop a scalable user and session identification paradigm for weblog data. The outcome is dependent on a scalable association rule mining technique. Some rules are applied, such as frequent page visits based on the amount of time spent on the page. To reduce the number of times the entire log is scanned, a better association is defined. The identification efficiency of the proposed framework is measured. The obtained minimum Mean Absolute Percentage Error (MAPE) in identification is 7%. For a weblog with 1 million log entries, MAPE is 11% which is significant. The framework could achieve an hourly Root Mean Square Error (RMSE) of 0.0500 and is better than published paradigms using Artificial Neural Network (ANN) 0.0639.

H. K. Sowmya
hk.sowmyakiran@gmail.com

R. J. Anandhi
rjanandhi@hotmail.com

[1] Department of Information Science and Engineering, New Horizon College of Engineering, Affiliated To VTU, Bangalore, India

Keywords Mean absolute percentage error (MAPE) · Artificial neural network (ANN) · URL · IP · Python · Web mining · Root mean square error (RMSE)

1 Introduction

Since the inception of web technology, there is a need to understand usage patterns and this had been an area of research and addressed by several studies. Over a while due to changes in technology, we found that the methods published earlier are no more relevant. The essence of new research is to make the developed tool meaningful. Make the mining rely on data instead of user-submitted credentials. When a user connects to the server, it supplies the IP, user ID, timestamp, mode of request, status and several other information. An access protocol such as Hypertext Transfer Protocol (http) is responsible for content access. When a user wants to visit the site using http, they can use several devices. The visitor provides a Uniform Resource Locator (URL) or uses a page link. There are several off-the-shelf software's available that could provide web browsing statistics. The information is retrieved using such software for getting an idea of the server traffic. As the volume of traffic increases, the complexity of data mining also increases. Increased traffic provides an opportunity for malicious attacks and requires a more advanced technique for web log mining [1].

During data mining, weblog data are collected from the server, client, or other sources. The collected data usually consists of different content. The data about the history of visits received from several sources give information about the browsing patterns of user during the overall usage. This data may include several types of users and corresponding visit patterns [2]. Weblog of a server does not portray sufficient insight about the visit behavior at the client side as they reflect the pages accessed. There is a need for preprocessing and cleaning web log data for further analysis. The essential component of pattern discovery and pattern analysis is to identify association rules. Many research works are available on this topic with several improvised techniques for data cleaning.

The web-based applications are growing at an alarming speed as a data exchange hub and a source for meeting day to day operations in all spheres of life. Web mining is the extraction of meaningful information from the logs generated during web surfing [3]. In general, weblog mining is divided mainly into content-based, usage based and structure based. The literature survey concerning the current work is limited to web mining based on the weblog's page visit, dwelling time and unique IP's. When a user accesses a web server, a log of visited sites gets generated. Analysis of the web logs can provide the user access history and about the content. The information gathered from the visit history could be used for planning e-commerce, m-commerce and other e-services such as banking, property renting and bill payment [4]. Accurate usage data could invite new users, also helps to maintain current customers, suggest new services, assess the impact of promotion alerts. Most of the time, marketing teams use this weblog mining information to create potential user profiles using navigation history, viewing time and page content [5].

The objective of the current work is to present a novel online dynamic session identification framework depending on the weblogs trace. The approach selected includes session detection with a user-defined schema using unique IPs, unique URLs, several sessions and average session length. The overall intent is to develop a scalable user and session identification paradigm for weblog data. The expected outcome is dependent on a scalable association rule mining technique. The mining objectives of the current work are as follows:

- Distinguish one pattern from another.
- Facilitate proper choice of the target segment of web content.
- Facilitate effective tapping of the segmentation.
- Crystallize the visit pattern of the target visitors.
- Make the mining effort more efficient than published work.
- Spot less used segments and succeed in reducing the impact on overall mining by such segments.
- Brings benefits not only to the mining usage pattern but also to the infrastructure scaling decision.

Further, we will discuss the literature survey, the proposed framework, implementation details, results obtained and analysis of results. Comparison based on the results of current work with the published result from other researchers also included.

In the next section, a detailed literature survey for web usage mining is discussed. The objective is to identify the gaps in implementation that can be improvised thorough investigation.

2 Literature survey

The use of data exploration methods over the weblog is one of the focus areas of research for information aggregation [6]. This paper provided a detailed approach for discovering patterns in web log data. The authors used a transformation and interpretation technique to extract information from user activities. The published work on i-Miner has an optimized solution using a fuzzy clustering system to capture the web access details [7]. For fuzzy rules, it used chromosomal structure modelling and representation. For the best results with the lowest RMSE, it used an efficient hierarchical distribution structure.

Next study organises web pages into a two-dimensional map using Kohonen's self-organizing map [8]. Rather than the content of the web pages, the organisation of the web pages is dependent only on the navigation behaviour of the users. The generated map serves as a visual analysis tool for webmasters to better understand the characteristics and navigation habits of web users visiting their pages. Another study [9] looks at various data processing methods. The limits of many web usage mining techniques applicable at various phases have been explored by the authors. Apriori, FP-growth, and Single-scan algorithms were also studied and compared.

Another approach is on analysis of real use patterns of the web site visitors. The objective here is to extract data from the weblogs and then use sophisticated algorithms for predictive modelling [10]. To extract relevant information from massive amounts of web traffic, the author adopted the soft computing paradigm. Further, to improve the trend analysis, the Fuzzy Clustering Method and Self-Organizing Map clustering approaches are used. It produced the best results and a high correlation coefficient.

To assess the performance of the clustering algorithm, the authors conducted a rigorous investigation on partitioning based and hierarchical based clustering algorithm [11]. Extensive experiments were carried out to investigate these clustering strategies. In anatomized trials, internal and external validity indices are utilised to evaluate the effectiveness of these two algorithms. Based on their research, they found that the K-means algorithm produces more promising results than the hierarchical approach. To detect navigation-related usability difficulties and improve usability, the authors developed a cost-effective ideal user interactive path (IUIP) model [12]. They also gave a comparison of real usage patterns and the IUIP models that match to them. The method's applicability and effectiveness were also assessed in this study.

The next study [13] included a time and referrer component to a session reconstruction algorithm in order to generate actual sessions and avoid excessively long sessions. The authors provided a framework for creating a semantic based session creation using heuristic approach to achieve better results. Recent research study presented a new graph based approach to cluster a data by constructing a graph from data with Markov Stability [14]. This approach is evaluated and tested for its robustness and performance with other graph based approaches. It is also compared with other clustering algorithms to measure the quality of the partitions based on Normalized Mutual Information (NMI) and Adjusted Rand Score (ARI).

Another method uses weblog analysis to generate trends and track user behavior [15]. Based on the web structure, each URL in the web log data is parsed into tokens. To investigate the navigation patterns, sessions from various users are aggregated using the hierarchical agglomerative clustering technique. To extract and evaluate user behavior patterns, the authors proposed a weblog analysis using the Pyspark (WAP) algorithm [16]. For efficient cluster processing, large weblog data has been separated and dispersed over numerous parallel nodes. Another study involves exploration and development of a utility for log files using Big Data platforms. The data gathered were utilized to classify the websites visited [17]. Using the Hadoop approach, this study constructed a web log analysis tool. In addition, this research developed an improved method for analyzing terms and phrases found in Google searches.

The scalable weblog mining method uses a tree-based clustering algorithm. This work intended to explore the suitable elements from the knowledgebase and predict the users browsing patterns [18]. An improvised approach implemented for creating web browsing logs. This approach could permit to validate and enhance the applications with a low cost [19]. The authors in an article segregated the webserver log file, selected new data patterns, and analyzed the content over tree-based classification algorithms [20]. Web use exploration can be segregated into following phases as: (i) data discovery, (ii) usage pattern analysis, (iii) pattern presentation and visualization [21].

Going over the publications, we could understand that a high performing framework for web mining is required. Different authors have taken different approaches. But this study provided the necessary direction for our research. Based on this study, we could understand the importance of the user sessions and capturing every detail. One gap that came up is the dependency of size in defining the framework. The current work is to bridge that gap. Hence one of the focus areas is to find the mining outcome depending on the size of the weblog. Another area added to the investigation is the optimization of the weblogs to get rid of unwanted data. Based on the literature survey, we could find that different researchers have used different development environments. The implementation of current work is completed with a standard state of the art development platform.

3 Proposed framework

Figure 1 shows the dynamic session identification framework proposed in the present work. The overall process of the framework implementation involves five stages.

- Identify Data: Knowing and accessing the data is the critical step for high-quality web usage mining. Knowledge about the business goal provides insight into finding the data source. Data is generally a weblog with millions of logs. It is essential that the data must be aligned to the goal of the investigation.
- Prepare Data: This step involves removing the unwanted data, make formatting consistent, remove redundant data. This step is very important as the final data analysis is done based on the cleaned data. Any unwanted data or wrong formatting of the data type may lead to wrong analysis.
- Prepare Model: A model is evolved by analyzing the trend and patterns of data. The essential component of the model is to fit new data into the model to provide a prediction of a future trend. Businesses can derive insight into the related activities and impact of changes

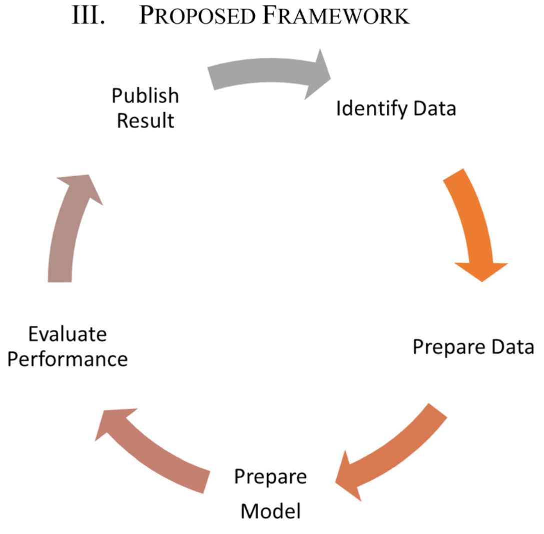


Fig. 1 Dynamic session identification framework

by using the developed model. The models can be descriptive or predictive depending on the use. The developed models can be further refined with new data trends.

- Evaluate Performance: Performance evaluation can be done using several data analysis algorithms, such as Decision Tree, Random Forest, Naive Bayes, AdaBoost, multilayer perception neural network. However, we used a simple statistical Sum of Squared Error (SSE) Clustering method.
- Publish Result: Once the performance is evaluated the result is communicated. A comparison is done with the previously published result and compared. The parameters considered for measurement are Mean Absolute Percentage Error (MAPE) and hourly Root Mean Square Error (RMSE).

4 Implementation

Table 1 shows the details of data used from available weblogs. The files Wblog1, Wblog3, Wblog4, and Wblog5 are the weblog files used for training and Wblog2 is used for performance measurement of the algorithm. The proposed model is tested on cleaned weblog data. Cleaning of data is done based on conventional data cleaning processes.

The source of data is from University of California, Irvine (UCI), USA [22] and Kaggle.com [23]. Kaggle provides data to do data science work.

Figure 2 shows the steps to run the developed algorithm. Once the data has been cleaned, the algorithm will run on the predefined parameters for analysis and segmentation reporting.

In the present work, an implementation is proposed to build a knowledge base using a Python-based application. The algorithm then uses the improved knowledge base to perform analysis based on Algorithm 1.

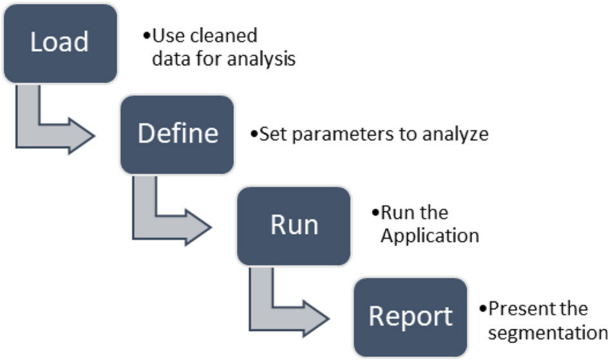

Load
•Use cleaned data for analysis
Define
•Set parameters to analyze
Run
•Run the Application
Report
•Present the segmentation

Fig. 2 Algorithm running steps

Algorithm 1 Segmentation and Clustering Analysis

```
Initialize: The path and environment directory

For Each Weblog
        If first time is 1 Then
                Run the cleaning routine
                Create the cleaned file.
        Else
                Do segmentation with cleaned file
        End
End

For     Each cleaned file
        Set range
        Check segmentation
        If segmentation is Yes Then
        Plot Clustering Analysis & Centroid Analysis
        Else
        Alert: No segment found
        End
End
```

Figure 3 shows the progress bar of the developed application using Python. In current work, SSE is used. SSE is a metric that helps to choose the appropriate number of divisions for segmentation. Clustering is a mathematical approach evolved to figure out the “best fit” for the identified cluster (segment). This involves multiple iterations to

Table 1 Data source and pre cleaning and post cleaning parameters

Aliases	Entries		Average	Number of unique			Size before cleaning (KB)
	Before cleaning	After cleaning	Session (s)	IPs	URLs	Sessions	
Wblog1	9874	9183	37.83	1344	195	4902	2343
Wblog2	2653	1940	43.01	568	456	1293	610
Wblog3	30,969	27,972	3.39	2310	1759	17,190	5681
Wblog4	15,789	11,382	88.36	5	263	6367	2764
Wblog5	1,031,567	922,450	72.84	51,364	6110	543,350	255,672

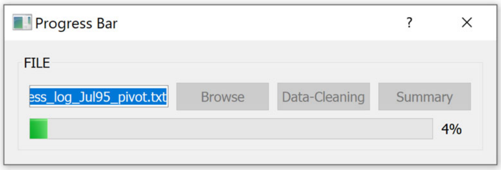


Fig. 3 Application interface for showing data claning status

bring the segments in proximity. If the respondents matched the segment scores exactly, then SSE would be zero = no error = a perfect match.

With real-world content, however, this is a difficult prospect to achieve. Further, the investigation continued with segmentation using a lower SSE. This is due to the fact that a low SSE represents large number of similar users in the identified segment. A higher SSE indicates that the users within the segment have sizable differences in browsing pattern.

5 Results

The subsequent subsection detailed the result obtained for data cleaning and segmentation performance of the developed framework based on the predefined matrices.

5.1 Application cleaning performance

Table 2 shows the performance of cleaning of raw data from weblogs. There is a minimum of 4% optimization and a maximum of 27% optimization, according to the data. Figure 4 shows the performance for optimization during cleaning. The optimization is an essential part as this reduces unwanted data volume hence reduces processing time during segmentation.

5.2 Selection of data for segmentation

Figure 5 shows the regression analysis of average session. It shows that Wblog2's average session is the best fit, hence Wblog2 data was chosen for clustering. The analysis resulted in average session computation predictions as given by Eq. 1.

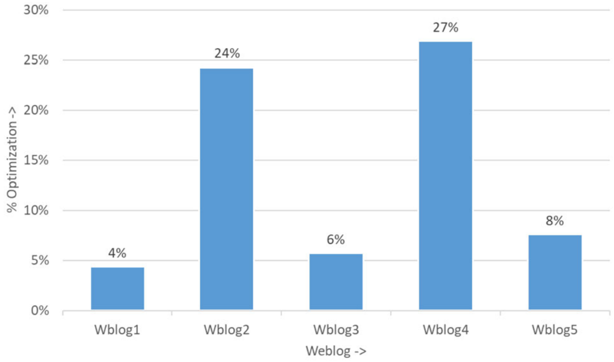


Fig. 4 Data cleaning performance

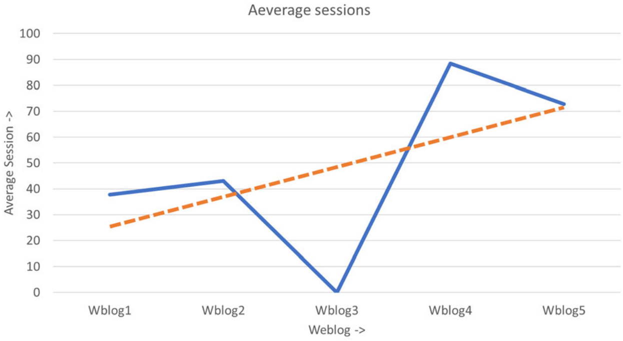


Fig. 5 Selection of testing data from average session

$$y \ = 14.472 + 11.538x \tag{1}$$

5.3 Developed model performance

There is a clear possibility for the developed framework to adopt a web mining technology that will help to solve critical data issues. A modelling framework using web mining is essential to escape from manual processes and get more accurate insight. There are many different file types on Kaggle Datasets, including CSV files. To select the list of datasets that are available as CSV files, select from the drop-down menu towards the top of the screen: [File types] > [CSV]. Similarly, UC Irvine Machine Learning Repository provides a dataset. Table 3 presents the derived performance indicators using data from UCI [22] and Kaggle[23].

Table 2 Data source and pre cleaning and post cleaning parameters

Aliases	Before cleaning (KB)	After cleaning (KB)	% Optimization (%)
Wblog1	2343	2241	4
Wblog2	610	462	24
Wblog3	5681	5356	6
Wblog4	2764	2021	27
Wblog5	255,672	236,264	8

Table 3 Key performance indicators

Aliases	Actual	Optimized	Absolute optimization	Deviation (from the mean)	Mean absolute percentage error (MAPE) (%)	Absolute deviation	Cumulative Abs optimization	Cumulative Abs Dev
Wblog1	9874	9183	691	208,296	7	208,296	691	208,296
Wblog2	2653	1940	713	215,517	27	215,517	1404	423,814
Wblog3	30,969	27,972	2997	187,201	10	187,201	4401	611,015
Wblog4	15,789	11,382	4407	202,381	28	202,381	8808	813,397
Wblog5	1,031,567	922,450	109,117	− 813,397	11	813,397	117,925	1,626,793
Total	1,090,852	972,927						

The different size of weblog chosen to analyze the variation of data exploration with volume of the weblog.

Set of dashboards are available in number of existing data analysis tools, and these can be used to identify solutions to a variety of problems. These platforms provide a big picture about the usage patterns and many more other relevant information. However, during the current research work, we developed a Python tool for statistical analysis. Table 4 shows the key measures obtained using the developed tool. The focus is to measure only the relevant information.

Results obtained using above parameters are plotted.

6 Analysis of result

The intercept is 14.472 and the coefficient is 11.538 based on available data using Eq. 1. This analysis provides a prediction of average session time for a weblog from the same web server.

This analysis is for planning a server's scaling capacity. As the number of users increases, the requirement for computer resources also increases. Scaling planning can be done using any commercial software based on the findings of Eq. 1.

From Fig. 6. MAPE is initially inconsistent with the log's session patterns, with high fluctuations (28%) for low volume. Therefore, we conclude that more data points stabilize the framework. Further, since the MAPE is 11% for 1 million records and RMSE 0.05 (approx.) is low, we have a stable paradigm for accurate rule-based mining.

Table 4 Key measure

Arithmetic mean of actual	218,170
Mean absolute deviation	23,585
Standard deviation	454,821.87
MAPE	16.41%
Mean square error	2,983,977,299
Average RMSE	0.050076265

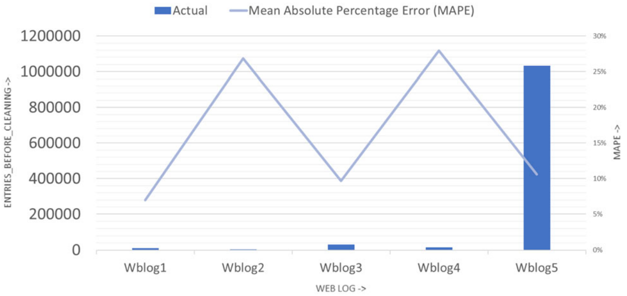


Fig. 6 MAPE for file Weblogs

From Fig. 7, it is evident that the errors computed from the model and deviation are initially not aligned, but by the end of the time, they are nearly aligned. Therefore, we can follow the Mean ± 1 standard deviation formula after carefully factoring in all the segments, IP and URL trends.

7 Comparison analysis

The performance of the identical data set from the 10th and 17th of June has now been compared. Table 5 provides the details about the segments. Each segment has 62 entries.

SgCL: Allocated Segment for Cluster, #: Number of Entries, Clustering is done as per Table 5 and plotted as in Fig. 8.

Figure 9 shows centroid of the clustering, Fig. 10 shows the SSE by number and segments. The obtained minimum MAPE is 7%. For a weblog with 1 million log entries, MAPE is 11% which is significant. The framework could achieve an hourly RMSE of 0.0500 and is better than the published 0.0639 [10].

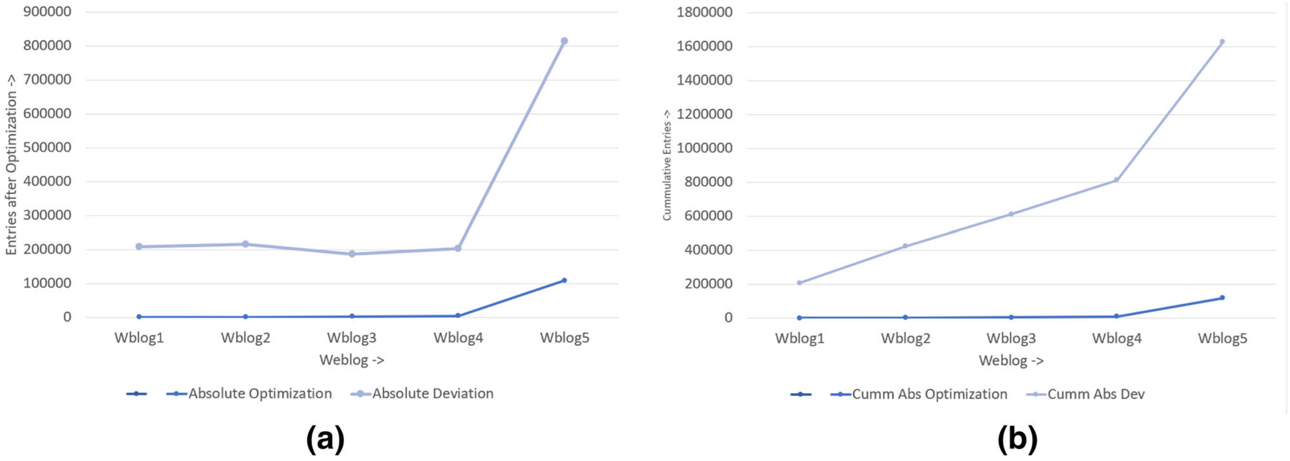


Fig. 7 Optimization vs. deviation **a** absolute, **b** cumm absolute

Table 5 Data set segmentation for 10th and 17th June

Respondent (Case)	SgCL1	SgCL2	SgCL3	SgCL4	SgCL5
# Segment 1	62	54	23	19	19
# Segment 2		8	12	12	10
# Segment 3			27	17	13
# Segment 4				14	14
# Segment 5					6
TOTAL	62	62	62	62	62

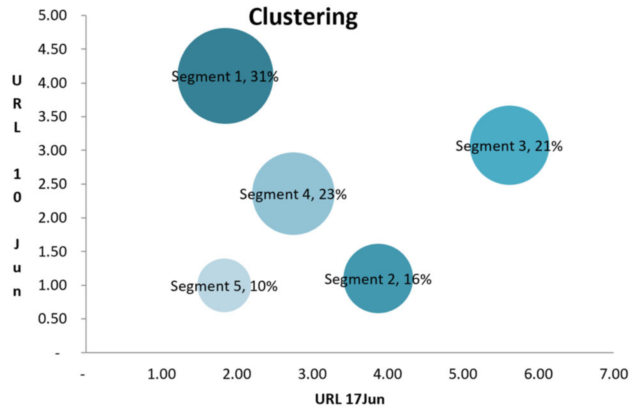


Fig. 8 Segmentation for 10th and 17th June

8 Conclusion

The authors presented a method for obtaining high-quality data all through preprocessing phase. A cleaning procedure was carried out by the algorithm. The algorithms employ this data to uniquely identify users, which aids in the discovery of user sessions. In terms of identification, the obtained minimum MAPE is 7%. MAPE is 11% for a weblog with 1 million log entries, which is significant. The framework achieved an hourly RMSE of 0.0500, which is better than published ANN (0.0639) paradigms [10].

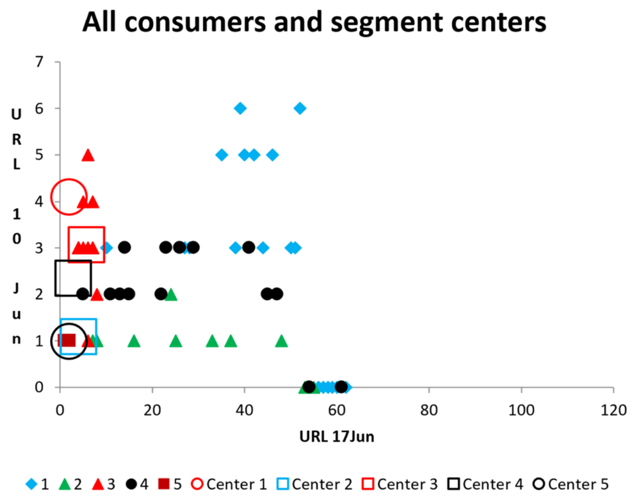
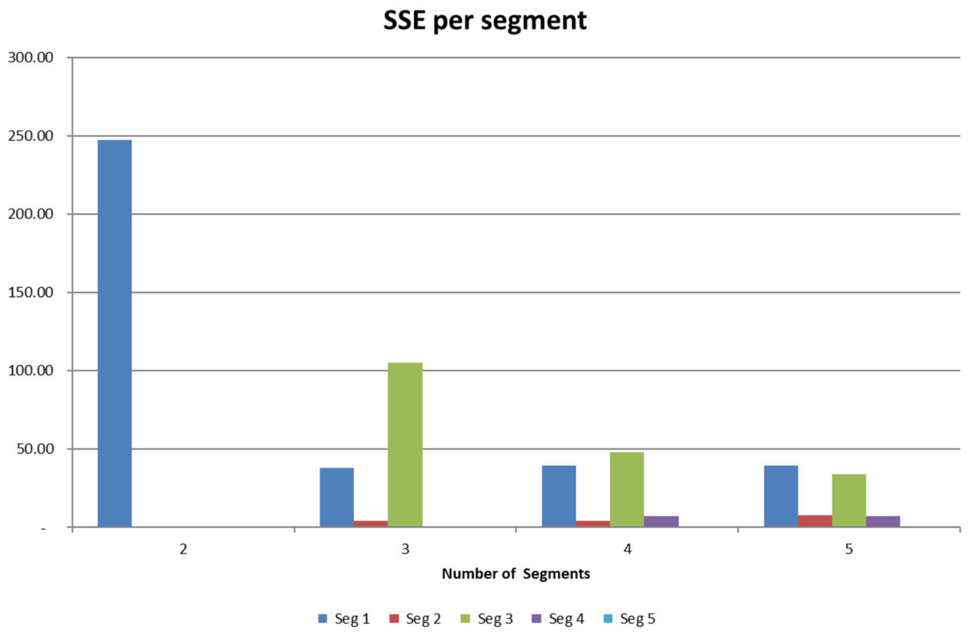


Fig. 9 Centroid analysis for 10th and 17th June segments

9 Future scope

The proposed revolutionary algorithm is more efficient and can be developed into a tool in future for administering exhaustive formulas during the preprocessing step. This would aid in the automation of visitor behaviour analysis based on the number of visitors, bandwidth consumed, and their interests in order to forecast future e-service patterns [24].

Fig. 10 SSE per segment (Seg)



References

1. Jespersen SE, Thorhauge J, Bach T (2002) A hybrid approach to web usage mining, data warehousing and knowledge discovery. LNCS 2454:73–82
2. Svec P, Benko L, Kadlecik M (2020) Web usage mining: data pre-processing impact on found knowledge in predictive modelling. Proc Comput Sci. https://doi.org/10.1016/j.procs.2020.04.018
3. Kumar H, Anuradha (2020) Progressive machine learning approach with WebAstro for Web usage mining. Proc Comput Sci 167:1400–1410
4. Orit R, Anat G, Lior F (2017) Analyzing online consumer behavior in mobile and PC devices: a novel web usage mining approach. Electron Commer Res Appl 26:1–12. https://doi.org/10.1016/j.elerap.2017.09.003
5. Nigam C, Sharma AK (2020) Experimental performance analysis of web recommendation model in web usage mining using KNN. Mater Today Proc. https://doi.org/10.1016/j.matpr.2020.09.364
6. Nina SP (2009) Pattern discovery of web usage mining, computer technology and development, 2009. ICCTD'09. In: International conference on. Vol. 1. IEEE, 2009
7. Abraham A (2003) i-Miner: a web usage mining framework using hierarchical intelligent systems. In: The 12th IEEE international conference on fuzzy systems, FUZZ'03, vol 2, pp 1129–1134. https://doi.org/10.1109/FUZZ.2003.1206590
8. Smith KA, Ng A (2003) Web page clustering using a self-organizing map of user navigation patterns. Decis Support Syst 35(2):245–256
9. Sowmya HK, Anandhi RJ (2019) Web usage mining algorithms: a survey, alliance international conference on artificial intelligence and machine learning (AICAAM), April 2019
10. Ajith A (2003) Business intelligence from web usage mining. J Inf Knowl Manag 02(04):375–390
11. Hassan SI, Samad A, Ahmad O, Alam A (2019) Partitioning and hierarchical based clustering: a comparative empirical assessment on internal and external indices, accuracy, and time. Int J Inf Technol 12(4):1377–1384. https://doi.org/10.1007/s41870-019-00406-7
12. Ruili G, Jeff T (2015) Improving web navigation usability by comparing actual and anticipated usage. IEEE Trans Hum Mach Syst 45(1):84–94
13. Navjot K, Himanshu A (2017) A novel semantically-time-referrer based approach of web usage mining for improved sessionization in pre-processing of web log. Int J Adv Comput Sci Appl 8(1):2017. https://doi.org/10.14569/IJACSA.2017.080122
14. Liu Z, Barahona M (2020) Graph-based data clustering via multiscale community detection. Appl Netw Sci. https://doi.org/10.1007/s41109-019-0248-7
15. Anupama DS, Gowda S (2015) Clustering of web user sessions to maintain occurrence of sequence in navigation pattern. Proc Comput Sci 58:558–564. https://doi.org/10.1016/j.procs.2015.08.073
16. Bakariya B (2021) Efficient approach of analyzing and generating intrinsic information from weblog. Natl Acad Sci Lett. https://doi.org/10.1007/s40009-020-01042-7
17. Namahoot CS, Brückner M, Lekkam W (2020) System for analysing big weblog data. In: Kim K, Kim HY (eds) Information science and applications. Lecture notes in electrical engineering, vol 621. Springer, Singapore. https://doi.org/10.1007/978-981-15-1465-4_53
18. Kousik NV, Sivaram M, Yuvaraj N, Mahaveerakannan R (2021) Improved density-based learning to cluster for user web log in data mining. Lecture notes in networks and systems, vol 173. Springer, Singapore. https://doi.org/10.1007/978-981-33-4305-4_59
19. Pavanetto S, Brambilla M (2020) Generation of realistic navigation paths for web site testing using recurrent neural networks and generative adversarial neural networks. Lecture notes in computer science, vol 12128. Springer, Cham. https://doi.org/10.1007/978-3-030-50578-3_17
20. Mittal R, Malik V, Rattan V, Jhamb D (2021) Performance comparison of tree-based machine learning classifiers for web usage mining. Lecture notes in electrical engineering, vol 728. Springer, Singapore. https://doi.org/10.1007/978-981-33-4866-0_47
21. Baeza-Yates R (2005) Applications of web query mining. In: ECIR 2005. Volume 3408. Lecture Notes in Computer Science. The Intention Behind Web Queries 109

22. Dua D, Graff C (2019) UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science
23. https://www.kaggle.com/datasets
24. Hassan Umair M, Dar S, Kamran Niu D, Sundas M, Ma Y, Zhao X, Muhammad Shabir A (2018) Web-logs prediction with web mining. https://doi.org/10.1109/IMCEC.2018.8469256