International Journal of Engineering & Technology, 7 (3.24) (2018) 438-444

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Behavioral Pattern Mining for User Identity and Access Control A Cluster based Ensemble Model

Gokulapriya R[1]*, Ganesh Kumar R

[1]Research Scholar, Computer Science and Engineering, CHRIST(Deemed to be University).
[2]AssociateProfessor, Computer Science and Engineering,CHRIST(Deemed to be University).
**Corresponding author E-mail:r.gokulapriya@res.christuniversity.in*

Abstract

Web usage mining extracts user's behavior patterns from the internet. Behavior of web users services are monitored and controlled to authenticate the user identity and access. Several data mining techniques are presented to analyze the web user behavioral patterns. But complex activity pattern discovery is not performed to maintain the decision making. To improve the complex user behavior pattern mining, Ensemble of Fuzzy K-Means with Logit Boost Clustering (EFK-LBC) technique is developed. EFK-LBC technique extracts the web user behavioral patterns from web logs. First, preprocessing is exploited to clean the unwanted data and select consistent web patterns from the original web log files. Next, fuzzy k means clustering technique is employed as a base learner to group the frequent web user behavioral patterns based on the objective function. To improve clustering performance, Logit Boost clustering technique is designed to make strong cluster by combining the several base learners. Experimental evaluation of proposed EFK-LBC technique and existing methods are carried out with the web server log files. The results reported that the proposed EFK-LBC technique obtains high clustering accuracy of user identity with minimum time and space complexity. Based on the observations, EFK-LBC technique is more efficient than the existing methods.

***Keywords**: Web usage mining; complex user behavior pattern mining; preprocessing; web log files; fuzzy k means clustering; objective function; Logit Boost Clustering; frequent behavioral patterns of web users.*

1. Introduction

Web usage mining focuses on the determining of potential knowledge from the browsing behavioral patterns of the users. Web usage mining is the task by using various data mining techniques to find out usage patterns (i.e. behavioral patterns) from web data. For different session, the group of similar activities performed by a user is grouped to find the user identity. The different data mining technique are developed for frequent user behavioral patterns analysis. A Linear-Temporal Logic (LTL) Model Based Checking technique was introduced in [1] for discovering the more complicated behavioral patterns from Web logs. The web server logs include the information about users' behaviour. The analysis of such information has concentrated on applying web usage mining techniques where a rather static classification is used to model users' behaviour and the flow of the actions performed by them is not frequently considered. Also, this technique failed to analyze more behavioral patterns and to facilitate their automatic discovery. A cluster-based method, named MiND (Mining Neubot Data) was developed in [2] to discover groups of users with related Internet activities. The large volume of captured data measurements over time is used to verify whether the service received by the users is rational with the one of other users with the same subscription or if there are anomalies. It is unable to discover some anomalous patterns in the web access performance that directs to reduce the user behaviour pattern detection efficiency. The MiND framework failed to effectively finds the frequent user behaviors with less complexity. A learning action pattern (LA-Pattern) was developed in [3] to determine the each user's activities patterns taken from sequences of actions performed among a group of users. The learning actions were not predefined, which also limit the scale of the discovery of learning patterns. In [4], a statistical technique like classification, association rule mining discovery and statistical correlation analysis was carried out for determining the groups of web pages that are generally accessed by the web users. These web pages collection was not identify the user for web access control. An investigation of semantic information on the patterns generated in [5] for web usage mining in the form of repeated sequences. The investigation failed to analysis the more user behavior patterns. In [6], a hybrid sequence alignment measure (HSAM) was handled to discover the access patterns through distance estimation for clustering the user sessions. The HSAM method failed to collect the frequent web pages accessed by the user for reducing the latency.

Apriori algorithm and frequent-pattern tree was introduced in [7] for extracting frequent usage patterns of users from large databases. But, the algorithm was not grouping the web frequent patterns for user identification. A Hybrid Markov model and Hidden Markov Model was developed in [8] to predict the list of web pages of user interest. The hybrid model has high complexity.

An efficient approach was developed in [9] to create the user profiles for user identity. The approach was not reduced the complexity while identifying the user. An online navigational behavior prediction was presented in [10] with mining process namely session identification, navigational pattern discovery and online prediction. The performance of clustering time remained unsolved.

The above said literature provides the certain issues such as failed to analyze more behavioral patterns, high time and space complexity, failed to group the frequent patterns for reducing the space complexity and so on. In order to overcome above said issues, an efficient technique called EFK-LBC is introduced. The contribution of the paper EFK-LBC technique is described as follows,

An efficient EFK-LBC technique is introduced for web user identification and access control with high clustering accuracy with minimum time. The number of user behavioral patterns is collected from the server log files. Then preprocessing is done for removing the redundant data from the extracted user behavioral information's to reduce space complexity.

The fuzzy k means clustering is applied to group the frequent web patterns accessed by the same user using distance measure. The minimum distance between web patterns and centroid are used to group the patterns. The logit boost technique is applied to increase the fuzzy k means clustering performance with minimum error. The base cluster makes a strong cluster by summing all base clusters according to their weight value. As a result, the corresponding user ID is correctly identified with the help of IP address. This process increases the clustering accuracy with less false positive rate.

The rest of this paper is ordered as follows. Section 2 briefly discusses the EFK-LBC technique with neat architecture and flow process diagram. Experimental evaluation of proposed and existing state-of-art methods are described with web server log files in section 3. Section 4 provides the results and discussion of certain parameters with table and graphical representation. The works related to the research is described in section 5. Finally, Section 6 provides the conclusion of the research work

2. Ensemble Fuzzy K means with Logit Boost clustering for web user behavioral pattern mining

Web usage mining is the discovery of user access pattern from web servers to verify the user identity. With the rapid growth of web user, the identity verification plays a vital role to access attempt by the authorized user from internet. Web user behaviour pattern mining is a considerable way to detect the access behaviour of web users. During the user behavioural analysis, large number of user access patterns is available in web logs. A web log record contains user access patterns to reduce the time on handling the complex patterns. In order to handle relatively less number of patterns instead of dealing with more patterns in web log, clustering technique is employed for partitioning the patterns for user identification. In order to achieve such motivation, Ensemble of Fuzzy K-Means with Logit Boost Clustering (EFK-LBC) technique is introduced. The EFK-LBC technique is designed for user identification with the help of user behavioural pattern mining based on dynamic characteristics of web users. The architecture diagram of EFK-LBC technique for user behavioural pattern mining is shown in Figure 1.The number of users is denoted as $User_1, User_2, \ldots, User_n$ to access the web services (i.e. internet). Based on the web access, the behavioural information about the web users are collected from server log files. The EFK-LBC technique includes two processes for user identification. Due to large volume of data on the web in the form of text, image, video, audio and so on. It is very hard to discover relevant patterns for a web user. Therefore, pre-processing is carried out in first step to clean the log files by removing the unwanted data to select standardized user patterns from the original log files for user identification. Followed by, clustering is applied for grouping the frequent user behavioural patterns. EFK-LBC technique uses ensemble of fuzzy k means with Logit boost clustering to group the similar user behavioural patterns for identifying the user ID. Brief description about the EFK-LBC technique is presented in following sections.

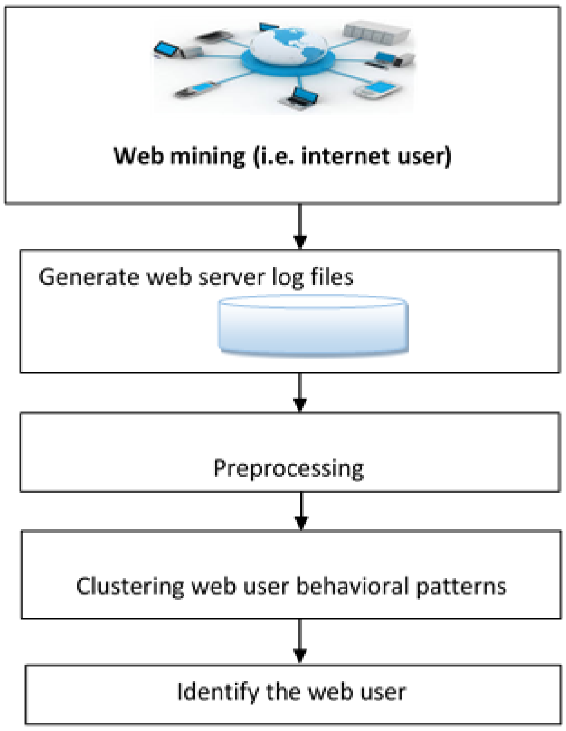
Web mining (i.e. internet user)
Generate web server log files
Preprocessing
Clustering web user behavioral patterns
Identify the web user

Fig 1: Architecture diagram of EFK-LBC technique

2.1. Clustering of frequent web user behavioral patterns

After pre-processing, an ensemble of fuzzy k means with logit boost clustering is applied to group the frequent web user behavioural patterns (i.e. web patterns) for user identification. Web user behaviour is defined as the reliable observations of a sequence of actions performed by the same user under certain specified time interval (i.e. session). The boosting is the ensemble of weak prediction into strong for improving the clustering performance. Therefore, an ensemble of weak clusters are combined them into a final strong cluster to provide the final results. In EFK-LBC technique, the weak learner is considered as fuzzy k means clustering. The performance of fuzzy k means clustering is improved by applying logit boost ensemble technique. Generally, clustering is the procedure for grouping the patterns of web user within a group is similar to one another and dissimilar from the other groups.
In figure 2, ensemble of fuzzy k means with Logit boost clustering is described. Let us consider the training samples as web user behavioural patterns P_i, i=1,2,3..n. The set of weak learner output is $\{h_1, h_2, h_3, \ldots h_t\}$.

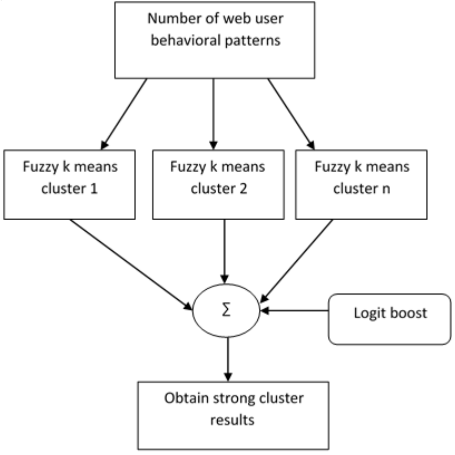

Fig 2: Ensemble of fuzzy k means with Logit boost clustering

All the clusters are summed to obtain final strong cluster results for identifying user ID. Initially, base fuzzy k means clustering is described to obtain strong cluster results.

Figure 3 demonstrates the flow process of fuzzy k means clustering to group the frequent web patterns accessed by the web user. The reason for choosing fuzzy concept is that it takes the imprecision encountered when analysing real-life data. Thus, the user identity is provided with more information about the structure in the data. Based on the clustering results, the corresponding user is identified through IP address for access control. Let us consider 'n' number of web user behavioural patterns P_i, i=1,2,3..n and it is partitioned into k number of clusters c_1,c_2, c_3,….c_k. The objective is to design a cluster for each web user behavioural patterns. Fuzzy k-means is a statistical method and determines soft clusters where a particular point fit into number of cluster with exact probability. Fuzzy k-means works on web user behavioural patterns which are represented in two-dimensional vector space and a distance measure is measured.

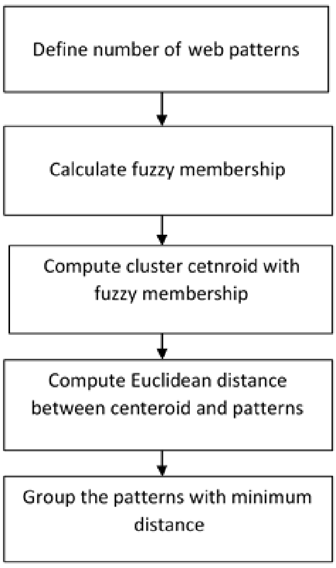
Define number of web patterns
Calculate fuzzy membership
Compute cluster cetnroid with fuzzy membership
Compute Euclidean distance between centeroid and patterns
Group the patterns with minimum distance

Fig3 Flow process of fuzzy k means clustering

Any data point (i.e. web user behavioral patterns) comprises coefficients providing the degree in the k^{th} cluster. By applying fuzzy concept, membership function is allocated to every data point related to each cluster centeroid on the basis of distance measure. The cluster centroid is the mean distance of every data points weighted by their degree of membership function. The fuzzy membership of each data point is measured as follows,

$$w_{ij} = \frac{1}{\left(D_{ij}\right)^{\left(\frac{1}{l-1}\right)} * \sum_{k=1}^{c}\left(\frac{1}{D_{ik}}\right)^{\frac{1}{l-1}}} \quad (1)$$

From (1), w_{ij} denotes a fuzzy membership of each data points, D_{ij} represents the Euclidean distance between the 'i^{th}' data points and to their j^{th} cluster centre. 'l' represents a fuzzifier measures level of cluster fuzziness. In general, membership functions that defines the value between 0 and 1. Depending on the fuzzy membership function, centroid of the cluster is defined for grouping the frequent web patterns. Centroid is a mean distance of entire data points in a cluster. Therefore, the fuzzy cluster centroid is formulated as follows,

$$c_{ij} = \frac{\sum_{i=1}^{n}\left(w_{ij}\right)^{l} p_i}{\sum_{i=1}^{n}\left(w_{ij}\right)^{l}} \quad (2)$$

From (2),C_i denotes a fuzzy centroid of cluster and w_ij represents a membership function and p_i denotes web user behavioral patterns. The distance among the data points and cluster centroid is computed by using Euclidean distance measure.

$$D_{ij} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{k}\left(c_i - p_i\right)^2} \quad (3)$$

From (3), D^{ij} denotes distance between cluster centroid and web patterns. Fuzzy k-means clustering is employed to cluster the frequent web user behavioural patterns. Therefore, the fuzzy k-means clustering minimize the objective function (i.e. distance between cluster centroid and web user behavioural patterns). The argument of minimum function is used for reducing the objective function.

$$h_t = arg\,min \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}{}^{l} D_{ij} \quad (4)$$

From (4) the minimum distance between the centroid and web patterns is grouped into that particular cluster. The pattern which is close to the centroid is called as frequent patterns accessed by same web users. The web user is correctly identified by frequent pattern clustering analysis. Each web patterns is assigned to some cluster depending on distance measure. After that, cluster centroid is updated by taking the weighted mean of the entire web patterns in that cluster. This recalculation of cluster centres results in better clustering results. This process is continuous until there is no alteration in cluster centroid. After training the patterns with k-means clustering, the EFK-LBC technique uses boosting technique to improve the performance of clustering. The output of strong cluster is a summation of weak learner which is mathematically denoted as,

$$f = \sum_{t=1}^{n} h_t \quad (5)$$

From (5),f denotes a strong cluster output and h_t denotes a base learner output. Initialize the weight value to each base cluster. For each iteration, the weight of the cluster is normalized. During the cluster centroid updating, the overlapping the data points between clustered are occurred and it causes the error. Therefore, EFK-LBC technique utilizes Logit boosting technique for selecting the base classifier with minimum error. Logit boosting technique minimizes the loss (i.e. error) which is expressed as,

$$h_t = \arg\min \alpha_E \quad (6)$$

By applying logit boosting algorithm, logit loss (i.e. error) is measured as follows,

$$\alpha_E = \sum_{i=1}^{n} \log(1 + \exp\,(-y_i h_t)) \quad (7)$$

From (7), α_E denotes an error of the base learner. After calculating the error, the cluster with the minimum error is selected. Then the cluster weight is updated. The weight of the cluster is increased if the web patterns are incorrectly clustered using base learner. The weight is decreases if the base leaner correctly clustered the web patterns. Therefore the weight is assigned according to the error value. Depending on the weight value, the strong cluster is constructed by combining the entire base cluster which is expressed as follows,

$$f = \sum_{t=1}^{n} \delta_t h_t \quad (8)$$

From (8), δ_t denotes a updated weight of the cluster h_t. The output of final strong cluster improves the clustering accuracy of frequent web user behavioural patterns with minimum time. From the clustering results, the user is correctly identified through their IP address.

Algorithm 1 describes the algorithmic description of ensemble of Fuzzy k means and Logit boost clustering. For each web patterns, the number of cluster is defined and randomly selects cluster centroid. Then the fuzzy membership and a centroid are measured. The Fuzzy k means clustering is used to group web user frequent patterns and updates the centroid. Then the algorithm verifies if all the patterns are moved to group then the process is completed otherwise it is repeated. Moreover, the Logit boost clustering is applied to boost the base cluster into a strong for improving clustering performance. By applying Logit boosting, the weight of base cluster is initialized. Train the patterns with base learner and error of base learner is computed. The base cluster is selected with minimum error and updates the weight value. Finally, all the base learners are summed with their weight value. Based on the clustering results, the user is identified effectively for access control in web server.

```
Input: web server log, number of patterns
$P = \{p_1, p_2, p_3 \ldots\ldots, p_n\}$ and number of cluster
$C_k = \{c_1, c_2, c_3, \ldots . c_k\}$
Output: Identify web user
Begin
1:  For each pattern $P_i$
2:  Define  number of cluster $C_k$ and randomly select cluster centroid
3: Calculate the fuzzy membership $w_{ij}$ based on distance using (1)
4:  Calculate fuzzy centroid $c_{ij}$ using (2)
5:  Compute Euclidean distance $D_{ij}$ using (3)
6:  Group the web patterns based on minimum distance using (4)
7:  Update cluster centeroid and group the web patterns
8:   Repeat the process step 3 to step 8
\\ Apply logit boosting technique
9: Initialize weight to all base clusters $h_t$
10: Calculate error $\alpha_E$ using (7)
11: Select base learner with minimum error using (6)
12: Update the weight of cluster
13: Combine all cluster with weight value using (8)
14: Obtain strong clustering results $(f)$
15:  End for
End
```

Algorithm1: Ensemble of Fuzzy k means and logit boost clustering

3. Experimental Settings

An experimental evaluation of EFK-LBC technique is implemented using java language. For the experimental consideration, (http://www.monitorware.com/en/logsamples/apache.php) apache web log dataset is used for user identity and access control by clustering the web user behavioural patterns. The dataset contains the column fields are user IP address, date, time, method (i.e. HTTP, GET), URL, response code, bytes. Based on these fields, the web user id is identified through the IP address by clustering the frequent web user behavioural patterns. Number of users is frequently access the web. The information about users' behaviour is stored in the apache web server logs. Based on stored information in log files about the user, different activities are predicted that consider the various operations done by a web user in a session. The frequent user behavioural patterns are grouped to identify the particular user through user IP address.

The performance of EFK-LBC technique is compared with existing linear-temporal logic (LTL) model based checking approach [1], MiND framework [2]. The efficiency of EFK-LBC technique is measured in terms of clustering accuracy, clustering time, false positive rate, space complexity.

4. Results And Discussion

Results and discussion of EFK-LBC technique is described in this section with various performance metrics such as clustering accuracy, clustering time, false positive rate and space complexity. With the help of these parameters, comparison between three methods namely EFK-LBC technique and existing LTL-based model checking technique [1], MiND framework [2].The comparison results of three methods and their experimental results are explained.

4.1 Impact of Clustering Accuracy

Clustering accuracy is defined as the ratio of number of frequent web patterns is correctly grouped to the total number of behavioral patterns. The formula for clustering accuracy is formulated as follows,

$$CA = \frac{Number\ of\ web\ patterns\ are\ grouped}{n} * 100 \quad (9)$$

From (9), where ‘CA’ denotes a clustering accuracy and n’ signifies a number of web patterns. It is measured in terms of percentage (%).

Table 1: Tabulation for clustering accuracy

Number of web patterns	Clustering accuracy (%)		
	EFK-LBC	LTL-based model checking technique	MiND framework
20	80	61	72
40	81	63	72
60	82	65	73
80	83	68	76
100	86	70	78
120	88	71	79
140	89	73	80
160	90	78	84
180	94	80	89
200	96	85	91

Table 1 describes a clustering accuracy of three methods namely, EFK-LBC technique and existing LTL-based model checking technique [1] MiND framework [2]. The number of web patterns is taken as input for user identification. The number of web patterns is varied from 20 to 200. From the table value, the accuracy for clustering the web patterns is considerably improved using proposed EFK-LBC technique than the existing methods. Let us consider the web server log files for identifying the user logged activities belonging to the same user. Based on logged activities, the user behavioral patterns are grouped for access control in Web based services. The user behavioral characteristics are grouped by applying ensemble clustering technique. An ensemble of fuzzy k-means with logit boost clustering is applied. Here the base cluster is considered as a fuzzy k-means. This clustering is employed to group the web patterns which are frequently accessed by the same web user at a particular session. Initially, the number of cluster and cluster centroid are defined. Then the fuzzy membership function is assigned to web patterns and cluster center based on distance measure. Based on the distance calculation, the data point (i.e web patterns) which is close to centroid is grouped with high clustering accuracy. Furthermore, the performances of clustering results are increased by applying boosting technique. The weights of base clusters are initialized. Then the error of base learner is computed for each base learner. Followed by, the cluster with minimum loss function is selected. Finally, all the base clusters are combined along with their weight value. As a result, user behavioral web patterns are correctly clustered. With the help of these clustered results, the corresponding user is identified. Based on the above process, initially 20 web patterns are considered. Each user’s behavioral patterns are grouped through ensemble clustering technique and identify the user through their IP address. Therefore the clustering accuracy of proposed EFK-LBC technique is 80% whereas the clustering accuracy of LTL-based model checking technique [1] MiND framework [2] are 61% and 72%. Similarly, all the nine runs are carried out to obtain the compared performance results. As a result, clustering accuracy of proposed EFK-LBC technique is considerably improved by 22% and 10% when compared to existing LTL-based model checking technique [1] MiND framework [2] respectively.

4.2 Impact of Clustering Time

Clustering time is defined as the amount of time required for clustering user behavioral web patterns. Clustering time is measured as follows,

$$CT = n * time\ (clustering\ the\ web\ patterns) \quad (10)$$

From (10), where CT denotes a Clustering time and ‘n’ indicates a number of web patterns. Clustering time is measured in terms of milliseconds (ms). Experimental results of Clustering time with three different methods are illustrated in figure 4.

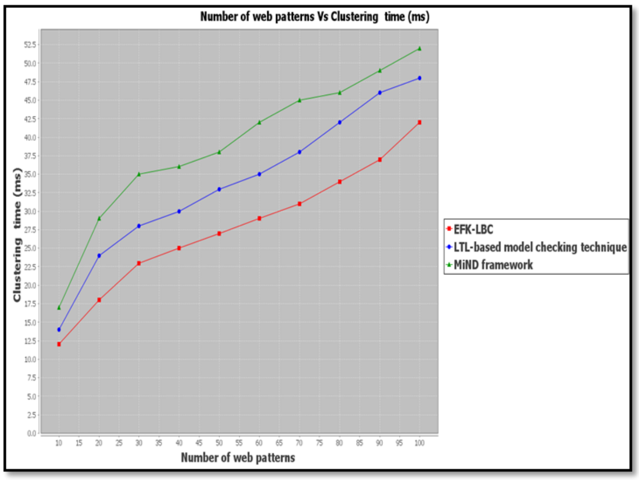

Fig 4: Performance results of Clustering time

Figure 4 depicts the performance results of clustering time with respect to number of web patterns. The number of web patterns is taken as input in ‘X’ direction and the clustering time is taken in ‘Y’ direction. As shown in figure, the red color curve indicates the clustering time of EFK-LBC technique where blue color and green color indicates the clustering time results of LTL-based model checking technique [1], MiND framework [2] . From the above graphical representation, it is clearly evident that the clustering time using EFK-LBC technique is considerably reduced than the existing LTL-based model checking technique [1] MiND framework [2].

The identification of user plays a significant part in behavioral biometrics to exhibit the uniqueness of the user. With the growth of World Wide Web, large number of access events performed by the users was recorded in server log files. In web, many users distribute, send, post and download lot of things from Web Sites, this manner very difficult to much organization for monitoring and controlling their access. Therefore, log files provided many events information according to user activities. These log analysis used for identifying the user IP address by monitoring user’s behavior in web. The user is correctly identified with minimum time. This is achieved by grouping the user frequent access behavioral patterns using ensemble of fuzzy k means and Logit boost clustering algorithm. The logit boost clustering algorithm effectively groups the web patterns by the means of combing all base clusters which is used to obtain from base fuzzy k means clustering. This process helps for EFK-LBC technique to reduce the clustering time in an effective manner.

Let us consider the number of web patterns is 20, the clustering time of proposed EFK-LBC technique is 12ms and 14ms, 17ms of clustering time is obtained by using LTL-based model checking technique [1] MiND framework [2]. This shows the significant improvement of the proposed EFK-LBC technique. Similarly, nine various runs are performed for all three methods. Finally, comparison results of proposed and existing techniques are calculated. From the comparison results, proposed EFK-LBC technique improves the clustering accuracy with minimum time by 18% and 29% than the existing methods.

4.3 False Positive Rate

False positive rate is defined as the ratio of number of web patterns are incorrectly grouped to the total number of web patterns. The false positive rate is measured as follows,

FPR= (Number of web patterns incorrectly grouped)/n*100 (11)

From (11), FPR denotes a false positive rate and 'n' denotes a number of web patterns. FPR is measured in terms of percentage (%). Experimental results of false positive rate with number of web patterns are illustrated in table 2.

Table 2: Tabulation for False positive rate

Number of web patterns	False positive rate (%)		
	EFK-LBC	**LTL-based model checking technique**	**MiND framework**
20	20	32	41
40	23	35	44
60	25	36	48
80	29	38	50
100	30	40	52
120	32	44	53
140	33	48	56
160	34	51	58
180	38	52	64
200	42	57	69

As shown in table 2, experimental results of false positive rate versus number of web patterns are described. There are 10 different runs are carried out to show the performance of proposed EFK-LBC technique and existing LTL-based model checking technique [1] , MiND framework [2]. From the experimental results it is clearly reported that the false positive rate of proposed EFK-LBC technique is considerably reduced than the existing methods. The LTL-based model checking technique [1] discovers different behavioral patterns that include activities presented by a user during a session. Depending on behavioral patterns analysis, it does not identify the corresponding web user for biometric identification. In addition, MiND framework [2] applied to find similar internet access performance of the user through cluster analysis. MiND framework was not correctly discovering the user ID who performs frequent access in web at a particular session. In order to overcome such kind of problems, EFK-LBC technique utilizes efficient clustering technique to group the similar patterns and identifying the corresponding user identity. The different user behavioral information’s are collected from apache server log files. The dataset contains the different user information such as user IP address, data, time, request field, URL, response code, bytes. Based on this information, the user frequent similar accessed web patterns are grouped with the help of fuzzy k means clustering. The distance among the web patterns and cluster centroid are employed to cluster the web patterns into those particular clusters. This clustering performance enhanced by applying boosting techniques with the help of combining all base clusters to make a strong cluster. The boosting technique calculates logit loss (i.e. error) for all base clusters. Based on error value, the weight for each base cluster is assigned. Finally, the weak cluster combined with their weight value to provide strong cluster results. This helps to minimize the false positive rate. As a result, the false positive rate of EFK-LBC technique is considerably reduced by 30% and 43% when compared to existing LTL-based model checking technique [1], MiND framework [2] respectively.

4.4 Impact of Space Complexity

Space Complexity is defined as the amount of memory space taken for storing the clustered web user behavioral patterns for user identification. The formula for measuring the space complexity is defined as follows,

$$\mathit{Space\ complexity} = \mathit{Number\ of\ web\ patterns} * \mathit{space\ (storing\ the\ web\ patterns)} \tag{12}$$

Space complexity is measured in terms of Kilobytes (KB). The graphical representations of the space complexity with number of web user behavioral patterns are illustrated in figure 5.

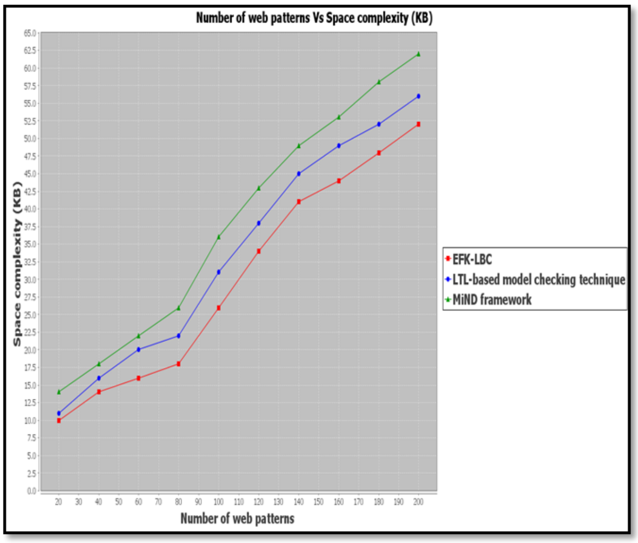

Fig 5: Performance results of space complexity

Figure 5 shows the performance results of space complexity versus number of web patterns. The space complexity after clustering the user behavioral web patters are calculated for obtaining the efficient results. Figure 5 illustrates the space complexity of EFK-LBC technique and existing LTL-based model checking technique [1], MiND framework [2]. Initially, the web user activities are collected from server log files. After that, preprocessing is carried out for removing the unwanted web log data and considered the relevant user patterns from the original log files for detecting the user ID. This process reduces the amount of space complexity. Moreover, by applying the ensemble clustering technique, the user frequent behavioral patterns are clustered from the large volume of data collected from server log files. This process takes minimum storage space and reduces the complexity while maintaining the user behavior profile in web. Let us consider the 20 web patterns for conducting experimental process, the proposed EFK-LBC technique obtains 10KB of space complexity whereas LTL-based model checking technique [1] and MiND framework [2] attains 11KB and 14KB respectively. Similarly, all the ten runs are performed and calculated the average comparison results. After performing the ten runs, the EFK-LBC technique significantly reduced the space complexity by 12% and 22% when compared to existing LTL-based model checking technique [1] MiND framework [2] respectively.

From the above said discussion, the web user is correctly identified through clustering the user behavioral patterns with less complexity

5. Related Works

A Markov model and Kth Markov model was developed in [11] for predicting the user's web-browser behaviors. The markov model was not identifying the number of web patterns. A neuro-fuzzy based hybrid approach was introduced in [12] for clustering the users having related browsing patterns into clusters. This approach failed to identify the user based on their clustering of similar user behavior patterns.

A fuzzy approach was designed in [13] for clustering the human behavior activities using Levenshtein distance-based fuzzy C-medoids (L-FCMd) algorithm. The algorithm was not increase the clustering performances. In [14], an algorithm was designed to analysis unobserved information contents in Log files and determining patterns in web server. The algorithm was not analyzed more patterns visited by user and many others activities.

A web usage mining technique was developed in [15] to find association rules which is used for pattern analysis with the help of RapidMiner tool. The technique does not find the exact user identity through clustering the frequent web patterns. An effective Web service ranking approach based on collaborative filtering (CF) approach was developed in [16] by estimating the user behavior with their history to collect the potential user behavior. The clustering of frequent similar user interested data was not carried out for user identification.

The k means clustering algorithm was introduced in [17] for grouping the web users with the pattern of user navigation. This clustering algorithm failed to consider the effect of users' navigation behavior during collaborative and sharing activities. K-means clustering algorithm was presented in [18] to make clusters through web pages accessed by targeted user and other users. The clustering algorithm was not considered memory employed to store the data.

An enhanced approach of Gap-BIDE algorithm was developed in [19] for determining sequential user patterns in web log data. The Gap-BIDE algorithm failed for mining the closed gap constraint sequential patterns. A map reduced model was introduced in [20] for predicting the navigation patterns of web users to reduce the size of the input database. The model was not minimizes the space complexity.

6. Conclusion

An efficient technique called Ensemble of Fuzzy K-Means with Logit Boost Clustering (EFK-LBC) is introduced for grouping the similar user web patterns extracted from server log files. At first, the unwanted data are eradicated from the extracted patterns using preprocessing. Then the base clustering technique namely fuzzy k means is applied to group the frequent user behavioral web patterns based on distance measure. The web pattern which is close to centroid is grouped correctly with minimum time. Then the fuzzy k means clustering is enhanced by applying Logit boosting technique. This boosting technique makes a strong cluster by summing the entire base cluster with their weight value. The weigh value of cluster is assigned based on the error rate. As a result of similar web patterns clustering, the web user ID is correctly determined with minimum complexity. Experimental evaluations of proposed EFK-LBC technique and existing methods are conducted with sever log files. The performance results show that the proposed EFK-LBC technique significantly improves clustering accuracy and reduces the clustering time, false positive rate as well as space complexity than the state -of –the- art methods.

References

[1] Sergio Hernández , Pedro Álvarez , Javier Fabra ,Joaquín Ezpeleta, "Analysis of Users' Behavior in Structured e-Commerce Websites", IEEE Access, Volume 5, 2017, Pages 11941 - 11958

[2] Tania Cerquitelli, Antonio Servetti, Enrico Masala, “Discovering users with similar internet access performance through cluster analysis”, Expert Systems With Applications, Elsevier, Volume 64, 2016, Pages 536–548

[3] Xiaokang Zhou, Jian Chen, Bo Wu, and Qun Jin, “Discovery of Action Patterns and User Correlations in Task-Oriented Processes for Goal-Driven Learning Recommendation”, IEEE Transactions on Learning Technologies, Volume 7, Issue 3, 2014, Pages 231-245

[4] Zakaria Suliman Zubi , Mussab Saleh El Raiani, “Using Web Logs Dataset via Web Mining for User Behavior Understanding”, International Journal Of Computers and Communications, Volume 8, 2014, Pages 103- 111

[5] Pinar Senkul and Suleyman Salin, “Improving pattern quality in web usage mining by using semantic information”, Knowledge and Information Systems, Springer, Volume 30, Issue 3, 2012, Pages 527–541

[6] G. Poornalatha, S. Raghavendra Prakash,” Web sessions clustering using hybrid sequence alignment measure (HSAM)”, Social Network Analysis and Mining, Springer, Volume 3, 2013, Pages 257–268

[7] Rahul Mishra, Abha choubey, “Comparative Analysis of Apriori Algorithm and Frequent Pattern Algorithm for Frequent Pattern Mining in Web Log Data”, International Journal of Computer Science and Information Technologies, Volume 3, Issue 4, 2012, Pages 4662 – 4665

[8] Meera Narvekar and , Shaikh Sakina Banu, “Predicting User’s Web Navigation Behavior Using Hybrid Approach”, Procedia Computer Science, Elsevier, Volume 45 , 2015 , Pages 3 – 12

[9] Yinghui (Catherine)Yang, “Web user behavioral profiling for user identification”, Decision Support Systems, Volume 49, Issue 3, June 2010, Pages 261-271

[10] Abdelghani Guerbas, Omar Addam, Omar Zaarour, Mohamad Nagi, Ahmad Elhajj, Mick Ridley, Reda Alhajj, “Effective web log mining and online navigational pattern prediction”, Knowledge-Based Systems, Elsevier, Volume 49, 2013, Pages 50–62

[11] Mamoun A. Awad and Issa Khalil, “Prediction of User's Web-Browsing Behavior: Application of Markov Model”, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), Volume 42, Issue 4, 2012, Pages 1131 – 1142

[12] G. Shivaprasad, N. V. Subba Reddy, U. Dinesh Acharya and Prakash K. Aithal, “Neuro-Fuzzy Based Hybrid Model for Web Usage Mining”, Procedia Computer Science, Volume 54, 2015, Pages s 327 – 334

[13] Pierpaolo D’Urso and, RiccardoMassari, “Fuzzy clustering of human activity patterns”, Fuzzy Sets and Systems, Elsevier, Volume 215, 2013, Pages 29–54

[14] Tawfiq A. Al-asadi and Ahmed J. Obaid, “Discovering similar user navigation behavior in Web log data”, International Journal of Applied Engineering Research, Volume 11, Issue 16, 2016, Pages 8797-8805

[15] Amit Dipchandji Kasliwal and Girish S. Katkar, “Web Usage mining for Predicting User Access Behaviour”, International Journal of Computer Science and Information Technologies, Volume 6 , Issue 1, 2015, Pages 201-204

[16] Guosheng Kang , Jianxun Liu, Mingdong Tang , Buqing Cao , Yu Xu, “An Effective Web Service Ranking Method via Exploring User Behavior”, IEEE Transactions on Network and Service Management , Volume 12, Issue 4, , 2015 , Pages 554 - 564

[17] Marios Belk, Efi Papatheocharous, Panagiotis Germanakos, George Samaras,” Modeling users on the World Wide Web based on cognitive factors, navigation behavior and clustering techniques”, The Journal of Systems and Software, Elsevier, Apr 2013

[18] Vedpriya Dongre, and Jagdish Raikwal, “An Improved User Browsing Behavior Prediction Using Web Log Analysis”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 4 Issue 5, May 2015, Pages 1838- 1842

[19] Xiuming Yu,Meijing Li, Taewook Kim, Seon-phil Jeong and Keun Ho Ryu, “An Application of Improved Gap-BIDE Algorithm for Discovering Access Patterns”, Hindawi Publishing Corporation, Applied Computational Intelligence and Soft Computing, Volume 2012, May 2012, Pages 1-7

[20] Meijing Li, Xiuming Yu, Keun Ho Ryu, “MapReduce-based web mining for prediction of web-user navigation”, Journal of Information Science, Volume 40, Issue 5, Pages 557-567, 2014