Received: September 18, 2020. Revised: October 31, 2020. 244

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Mutual Information Pre-processing Based Broken-stick Linear Regression Technique for Web User Behaviour Pattern Mining

Gokulapriya Raman[1]* Ganesh Kumar Raj[1]

[1]*Department of Computer Science and Engineering, School of Engineering and Technology, CHRIST (Deemed to be University), Bangalore,560074 India.*
* Corresponding author's Email: r.gokulapriya@res.christuniversity.in

Abstract: Web usage behaviour mining is a substantial research problem to be resolved as it identifies different user's behaviour pattern by analysing web log files. But, accuracy of finding the usage behaviour of users frequently accessed web patterns was limited and also it requires more time. Mutual Information Pre-processing based Broken-Stick Linear Regression (MIP-BSLR) technique is proposed for refining the performance of web user behaviour pattern mining with higher accuracy. Initially, web log files from Apache web log dataset and NASA dataset are considered as input. Then, Mutual Information based Pre-processing (MI-P) method is applied to compute mutual dependence between the two web patterns. Based on the computed value, web access patterns which relevant are taken for further processing and irrelevant patterns are removed. After that, Broken-Stick Linear Regression analysis (BLRA) is performed in MIP-BSLR for Web User Behaviour analysis. By applying the BLRA, the frequently visited web patterns are identified. With the identification of frequently visited web patterns, MIP-BSLR technique exactly predicts the usage behaviour of web users, and also increases the performance of web usage behaviour mining. Experimental evaluation of MIP-BSLR method is conducted on factors such as pattern mining accuracy, false positives, time requirements and space requirements with respect to number of web patterns. Outcomes show that the proposed technique improves the pattern mining accuracy by 14%, and reduces the false positive rate by 52%, time requirement by 19% and space complexity by 21% using Apache web log dataset as compared to conventional methods. Similarly, the pattern mining accuracy of NASA dataset is increased by 16% with the reduction of false positive rate by 47%, time requirement by 20% and space complexity by 22% as compared to conventional methods.

Keywords: Broken-stick linear regression analysis, Frequent access, Mutual information based pre-processing, Usage behaviour, Web patterns, Web user.

1. Introduction

WWW is growing fast and massive amount of information is generated owing to user's communications with web sites. Rapid growth of the digital technology, increased the usage of WWW in different fields like e-commerce, digital valet, e-health monitoring services, online education, food ordering, transport service rendering, tourism and hotel management, cloud based storage support, etc., for anything and everything people rendering the world wide web support. Discovering usage pattern of web users is very significant. Analysing the web access data interpreting useful information is identified as web usage mining. Web usage mining is the branch of data mining techniques. It includes three stages namely pre-processing of data, discovery of patterns and analysis of patterns.

Data pre-processing is a method used to produce accurate functional data from the raw data. Data cleaning, data transformation and data reduction are three distinct features of data pre-processing. The missing information can be filled and remove the noisy information by applying various techniques namely regression, clustering, etc., under data cleaning. Data transformation is helps in attribute selection, normalization. Data reduction is works for refining the data after cleaning and transformation process. By which the required data with intensive

attributes were acquired. Analysis of such data will help in retrieving use full information as patterns.

User access patterns can be derived by mining the web access logs. Server log, Proxy server log, Client/Browser log are various forms of access data logs available for interpreting user web access behaviour patterns. In the web server log, web usage mining techniques are used to extract a user activity, i.e. regular users, synthetic users and potential users. Pre-processing is the first step we need to implement to analyse and predicting the usage behaviour of the web user. Pre-processing of data serving as a major feature of web mining, it is used to shift and systematize only suitable information. The web access server logs are cleaned, sorted and grouped into key sessions before they are used for analysing to interpret useful information under web usage mining.

Extracting user behaviour from the web access is a continuous evolving research domain. So many researchers devise various techniques to extract the behaviour of users from the web access logs. Sequence based analysis and clustering was designed in [1] to analysis the behaviours of online users. To find out the recurrent search patterns, Maximal Repeat Patterns (MRPs) and Lag Sequential Analysis (LSA) were used, however the accuracy of behaviour analysis was not focused. Frequent pattern mining based cross-social network user identification algorithm (FPM-CSNUIA) was introduced in [2] to study the user behaviour in social networks. To analyse the relationship between the user data, SimFunc () function was utilized where it considers the vector.

The model checking approach based on Linear-Temporal Logic (LTL) was designed in [3] to obtain user activity patterns from structured e-commerce weblogs. However, pattern mining accuracy using LTL technique was not sufficient which increases the false positive rate. Fuzzy clustering algorithm was applied for predicting user behaviour in web [4]. But, the amount of time taken for web usage mining was very higher. A general framework was introduced in [5] for targeting audience behaviours using change points to fetch the user behaviour. However, mining accuracy using this framework was very lower.

The relationship between the object is measured in [6] using their attributes from different entity set is used to fetch behavioural patterns. The space density was not reduced using similarity score. In [7] author proposed a model to collect the behaviour patterns of a host network named Log Mining for Behaviour Pattern (LogM4BP) using nonnegative matrix factorization algorithm. But, LogM4BP model does not reduces the false positive rate.

Assisted pattern mining was performed in [8] for determining interactive behaviours from the web with a lower time complexity. However, mining patterns requires more storage space. Sequential Association rule based web usage mining approach was introduced in [9] to find out e-commerce behaviour of user in hand held device and computers with minimal false positive rate. But, this approach requires more time for mining the user behaviour. Hybpmine was presented in [10] to increases the performance of mining the behavioural patterns over a data stream. However, accuracy of behavioural pattern mining was lower.

More number of research was performed earlier in the field of pattern mining using web logs, however, pattern mining accuracy and time requirements using conventional works show that there is a scope of improvement in performance. To overcome the conventional concerns identified from the study a new approach is designed namely MIP-BSLR using Mutual Information based pre-processing (MI-P) method and Broken-Stick Linear Regression analysis (BLRA).

The objectives of MIP-BSLR technique is described below,

1. To improve the pre-processing performance of mining web user behaviour patterns when compared to conventional works, Mutual Information based Pre-processing (MI-P) method is proposed in MIP-BSLR technique. MIP-BSLR technique considers web patterns from Apache web log dataset and NASA dataset. MI-P method compute mutual dependence between the two web patterns to efficiently perform pre-processing with lower amount of time for both Apache web log dataset and NASA dataset. Based on the computed mutual dependence value, relevant web patterns are selected and irrelevant web patterns are removed in MIP-BSLR technique. This aids to decreases the space complexity as well as time complexity in MIP-BSLR technique compared to existing works.
2. To get better web user behaviour patterns mining performance while related to contemporary works, Broken-Stick Linear Regression analysis (BLRA) is applied in MIP-BSLR technique. As, BLRA can model any complex, non-linear web patterns. Furthermore, BLRA can seamlessly handle any irregularly sampled web log files which supports for MIP-BSLR technique for accurately extracting the web user behaviour patterns. During BLRA process, the relationship between two web patterns is analysed. From that, the web pattern with maximum relationship is selected to determine the web user behaviour with higher

accuracy. Thus, the most frequently visited web patterns are identified. Besides, hit ratio is also measured to reduce the error rate.

The article is structured as follows: Similar works are reviewed in section 2. The proposed MIP-BSLR technique with a neat diagram was discussed in Section 3. In section 4 and section 5, simulation scenarios and output results of the proposed MIP-BSLR technique are discussed. Conclusion is given in section 6.

2. Related work

Sequential search pattern analysis and clustering was designed in [1] to investigate the online user behaviour. Results confirmed that the sequential pattern analysis and webpage clustering boost the method to examine the customers shopping behaviours. However, the customer behaviour analysis accuracy was poor. User identification algorithm was designed in [2] according to the behaviour. The weight of user data was not only dependent on the information entropy, but was also based on the previous likelihood. The weight of each user's data is, therefore, more accurate. The implementation of the algorithm maximizes the precision rate, recall rate, as well as the F-Measure. However, the time requirement was not decreased.

An analysis of structured e-commerce web logs was carried out using LTL based checking model [3]. Among the wide set of possible behaviours, the designed model finds different website sections by visiting and buying actions of navigational patterns. User category and navigation pattern detection was presented through clustering and classification methods [4]. The designed method has the advantages of causal relations amount events of a user trace in contrast to provide global view of the whole session. Besides, is also avoiding the need of tagging the web pages. However, time taken to mining the similar behaviour was increased.

A detection of reliable interaction depends on the unique mouse behaviour patterns of web users were studied in [11]. The method takes physiological and psychological features of user for discovering the mouse behaviour characteristics. This helped to recognize the reliable access behaviour of web users. Besides, the abnormal behaviour in the web users also detected. But, the accuracy was not improved while identifying the user emotions. To increase the accuracy when identifying the user behaviour, the Broken-Stick Linear Regression Analysis is employed in proposed method. Role based approach was applied in [12] for extracting user roles by mining the usage patterns of web application. However, space and time requirements were not minimized.

A pattern-growth-based mining method was designed in [13] to detect the repeated behaviour of the user. The designed method greatly lessens the behavioural variability and lead to enhance the performance of user identification. Besides, mouse-behaviour features were analysed to obtain diverse kind of user activity. On the other hand, the time requirements were not minimized. To handle this issue, mutual information based pre-processing is employed in proposed MIP-BSLR technique.

A user focused review selection mechanism was developed in [14] by means of find outing significant profiling user features. Through the significant features, the online reviews are effectively identified. User-based review selection filter was employed to gather and pre-process the data. Then the entire history of users and businesses reviewed was processed to provide the new features for review selection. However, the time requirements was not minimized. Therefore, MIP-BSLR technique uses the novel pre-processing and regression analysis technique to lessen the space and time requirements problem. Interesting pattern-based parallel FP-growth (MIP-PFP) algorithm was implemented in [15] to get better mining accuracy for finding frequent patterns from big weblogs. But, ratio of overall count of web patterns that are mistakenly mined as frequent was higher using the MIP-PFP algorithm.

A novel method was introduced in [16] for extracting the social collective behaviour of Twitter users concerning a group of brands based on the users' temporal activity. These data are pre-processed and apply temporal clustering where it groups the data depends on the interest or activity. Then the hidden Markov model was utilized for obtaining temporal behaviour patterns at a certain time period. However, the error rate was not decreased. In order to manage this problem, mutual information based pre-processing and broken-stick linear regression analysis is designed in proposed technique. A remote assessment method was utilized in [17] to increases the accuracy of mining the web usage patterns. But, the computational complexity was higher.

A novel Web Usage Mining method was presented in [18] to observe the online customer behaviour in mobiles and computers. The designed method includes data pre-processing, pattern identification and analysis process. During the pattern identification, the sequence of online user activity is segmented into diverse group. But, the accuracy was not increased. Thus, MIP-BSLR technique is proposed to handle this limitation.

Posting behaviour of online users was identified in [19] depends on the user modelling and profiling. Hidden Markov Model was used to provide the individual level of User posting behaviour. But, false positive rate was remained unaddressed.

From the above described literature, some of the issues such as minimal pattern identification accuracy, higher time complexity, space complexity, error rate, false positive rate are occurred during the web user behaviour pattern mining. In order to handle such issues, in this work Mutual Information Pre-processing based Broken-Stick Linear Regression (MIP-BSLR) technique is designed with the implementation of Mutual Information Based pre-processing and Broken-Stick Linear Regression analysis where it effectively increases the performance of the web user behaviour pattern mining in terms of accuracy, space and time requirements. In this work, proposed MIP-BSLR technique is discussed in section 3 Simulation sceneries are showcased in section 4. MIP-BSLR technique performance result is analysed in Section 5. Conclusion of this work is discussed in section 6.

3. Mutual information pre- processing based broken-stick linear regression technique

The activity of web users is stored as a web pattern in the weblog archives. The gathered data from web log file is incomplete, noisy and not suitable for mining initially. Pre-processing is required to convert the weblog files into relevant form for pattern finding. Besides to that, extracting the web user behaviour is a key problem in web mining using different data mining techniques. To enhance the web usage behaviour patterns mining performance with minimal time requirements and false positive rate, MIP-BSLR technique is designed. The proposed MIP-BSLR technique designed by combining Mutual Information based Pre-processing (MI-P) method and Broken-Stick Linear Regression analysis (BLRA). This approach find outs the usage behaviour of web users via extracting frequent web patterns from input web log files.

The structural design of the proposed MIP-BSLR approach is depicted in Figure 1 shows the overall processes of MIP-BSLR technique for effective web usage behaviour patterns mining. MIP-BSLR technique at first gets web log files i.e. Apache web log dataset as input. After taking input, MIP-BSLR technique applies Mutual Information based Pre-processing (MI-P) method with objective of removing unwanted, redundant and irrelevant web

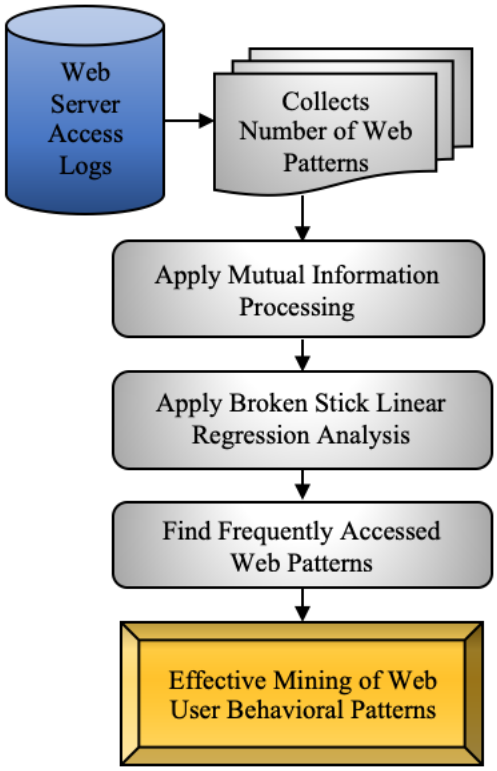


Figure. 1 Proposed MIP-BSLR technique architecture diagram

patterns with lower amount of time utilization. Followed by, MIP-BSLR technique applies Broken-Stick Linear Regression analysis (BLRA) with aiming at identifying frequent web patterns with improved accuracy. With help of mined frequent web patterns, finally MIP-BSLR technique accurately determines the usage behaviour of web users with minimal time requirements when compared to conventional works. Detailed processes of MIP-BSLR method is described in subsections.

3.1 Mutual information based pre-processing

Pre-processing is an essential activity while mining the web usage data. Information collected from WWW activity is incomplete, noisy, and inconsistent. Duplicate or missing data may cause incorrect identification of web user behaviour patterns. Therefore, a Mutual Information Based pre-processing (MI-P) method is introduced in MIP-BSLR technique. The MI-P method eliminates the unwanted, redundant information from the web logs and thereby performing an effective pattern mining. The accuracy and quality of web user behaviour pattern mining algorithms are improved with the help of MI-P method. MI-P method determines mutual dependence between web patterns that are sampled simultaneously. Based on the evaluated mutual information value, MI-P method removes unwanted,

redundant and irrelevant web patterns with minimal amount of time. Figure 2 shows the process involved in Mutual Information Based pre-processing method. As presented in figure 2, web log files are considered as input from Apache web log dataset and NASA dataset. These dataset contains number of web patterns represented as '$\beta_1, \beta_2, \beta_3, \ldots . \beta_n$'. Here, '$n$' denotes total number of web patterns in given log files. Mutual dependence determines the relation between the two web patterns. From that, Mutual dependence between two web patterns *'β_i'* and *'β_j'* is mathematically determined using Eq. (1),

$$MI_{ij} = \int_{\beta_i} \int_{\beta_j} p(\beta_i, \beta_j) \log \frac{p(\beta_i, \beta_j)}{p(\beta_i) p(\beta_j)} d\beta_i d\beta_j \quad (1)$$

From the mathematical expression Eq. (1), '$p(\beta_i, \beta_j)$' indicates the joint probability density function between the two web patterns. Here, *'$p(\beta_i)$'* and *'$p(\beta_j)$'* represents the marginal density functions. The mutual dependence defines how similar the joint distribution *'$p(\beta_i, \beta_j)$'* is to the products of the factored marginal distribution. In MI-P method, mutual dependence 'MI_{ij}' in yields values in the range of '0' (no mutual information – two web patterns β_1 and β_2 are independent) to '$+\infty$'. According to the measured mutual dependence value, MI-P method selects web patterns that are more related and removes irrelevant, redundant web patterns with minimal time as compared to conventional works. The algorithmic processes of Mutual information based pre-processing is shown in Algorithm 1.

By using the algorithm1 steps, MI-P method initially calculates mutual dependence for each input web patterns. If the estimated mutual dependence value is equal to zero, then MI-P method considered that the two web patterns are independent and therefore selects both web patterns for behaviour mining process. If the determined mutual dependence value is not equal to zero, then MI-P method considered that the two web patterns are dependent and therefore choose one web patterns for predicts usage behaviour and consequently eliminates another one as irredundant web patterns. From that, MI-P method enhances the pre-processing performance for efficient mining of web usage behaviour.

3.2 Broken-stick linear regression analysis

The Broken-Stick Linear Regression Analysis (BLRA) is intended to increases the accuracy of web usage behaviour mining in which the input web patterns are partitioned into intervals and a separate line segment is fit to each interval. The BLRA is performed on multivariate data through partitioning

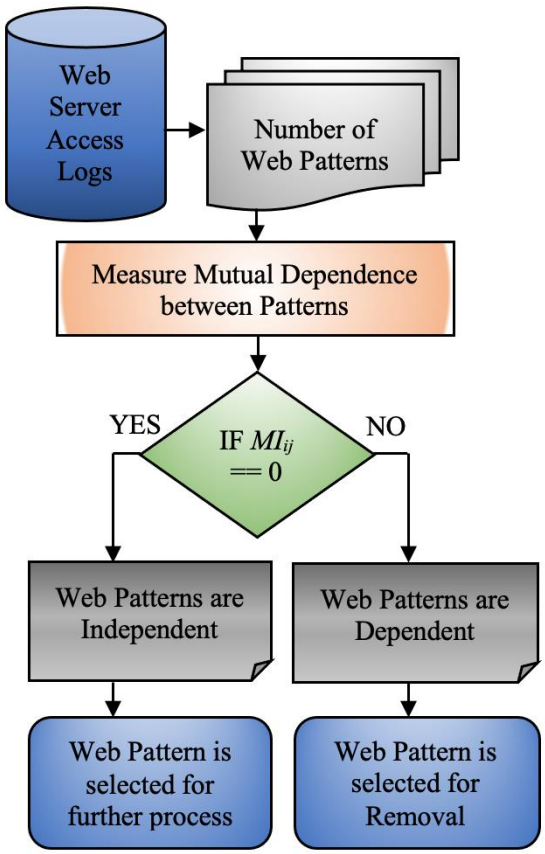


Figure. 2 Block diagram of mutual information based pre-processing

// Mutual Information Based Pre-processing Algorithm
Input: web server log files, Number Of Web Patterns '$\beta_1, \beta_2, \beta_3, \ldots . \beta_n$'
Output: Remove irrelevant, redundant and unwanted web patterns
Step 1: Begin
Step 2: **For** each input web pattern 'β'
Step 3: Determine mutual dependence between web patterns 'MI_{ij}' using (1)
Step 4: **If** ($MI_{ij} == 0$), **then**
Step 5: Two web patterns are independent
Step 6: Choose web patterns for further processing
Step 7: **Else**
Step 8: Two web patterns are dependent
Step 9: Select one web pattern and remove another one as irrelevant
Step 10: **End**
Step 11: **End For**
Step 12:End

Algorithm 1. Mutual Information Based Pre-processing

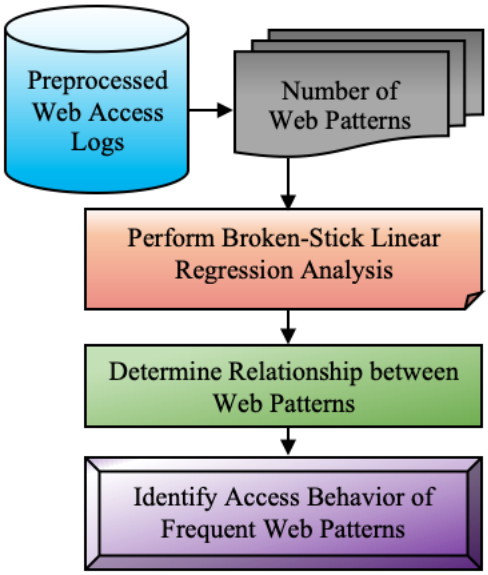


Figure 3. Flow process of BLRA for mining web user behaviour atterns

the various independent variables i.e. web patterns. When the autonomous variables clustered into dissimilar sets, the BLRA is suitable to exhibit different relationships between the variables in these regions. The boundaries between segments called as breakpoints. The process involved in BLRA is demonstrated in Fig. 3.

BLRA initially acquires pre-processed web patterns as input and then performs regression analysis. BLRA is a form of linear regression which arises when a single line isn't sufficient to model the data set. Piecewise regression breaks the domain into numerous "segments" and fits a separate line through each one. From that, Broken-Stick Linear Regression Analysis is mathematically defined in Eq. (2) as,

$$y = \alpha_0 + \alpha_1 \beta + \alpha_2 (\beta - c)^+ + \varepsilon \qquad (2)$$

From the equation (2), 'c' denotes the value of breakpoint and 'β' represents input web pattern whereas '$\alpha_0, \alpha_1, \alpha_2$' indicates regression coefficients (represents the slope of the line segments). Here, 'ε' refers error vector and 'y' represents the predicted output. Through carried outing a regression analysis process, BLRA determines relationship between the input web patterns with help of hit ratio using Eq. (3),

$$y \leftarrow H_R = \frac{\mathrm{c}\,(\beta)}{\mathrm{n}} \qquad (3)$$

From the mathematical formula (3), n indicates the total number of web patterns in weblog database whereas $\mathrm{c}\,(\beta)$ represents the count of total occurrences of the particular web patterns found in weblog file. In BLRA, hit ratio measures the number of times that the particular web patterns found in weblog files to the total number of web patterns. Based on estimated relationship (i.e. hit ratio), BLRA breaks input web log files into individual segments (i.e. frequent web patterns or non-frequent web patterns) and fits a linear regression within each segment. Break point is the beginning and end location of a segment. Here, break point represents the threshold value for hit ratio. BLRA reduces sum of square error by finding optimal breakpoint. From that, BLRA significantly find outs the frequent web patterns with less time requirements and higher accuracy.

// Broken-Stick Linear Regression Analysis Algorithm
Input: Pre-processed Web Patterns '$\beta_1, \beta_2, \beta_3, \ldots . \beta_n$'
Output: Mine web user behaviour with higher accuracy
Step 1: Begin
Step 2: **For** each pre-processed web pattern 'β_i'
Step 3: Apply Broken-Stick Linear Regression Analysis using (2)
Step 4: Determine relationship between web patterns using (3)
Step 5: Find frequent web patterns
Step 6: Identify web usage behaviors
Step 7: **End For**
Step 8:End

Algorithm 2. Broken-Stick Linear Regression analysis

4. Investigational settings

MIP-BSLR technique is implemented in Java language using Apache web log dataset [20] and NASA dataset [21] to analyse the proposed technique performance. Apache web log dataset contains numerous numbers of web patterns with user IP address, date, time, method (i.e. HTTP, GET), URL, response code and bytes. NASA dataset includes web patterns with the attributes such as host, log name, time, method, url, response and bytes. By taking a log file as input, MIP-BSLR technique initially performs pre-processing and regression analysis in order to identify web user behaviour through mining a frequent web patterns. The efficiency of MIP-BSLR technique is measured in terms of pattern mining accuracy, false positives, space and time requirements with respect to various number of web patterns as input. For determining the proposed performance, four existing methods namely Sequential search pattern analysis and clustering [1], FPM-CSNUIA [2], LTL-based model checking

technique [3] and fuzzy clustering algorithm [4] are chosen from literature survey according to research objective during the experimental process for simplicity. We also take additional number of existing methods for performance analysis during the validation process to analyse the proposed performance. The conventional Sequential search pattern analysis and clustering [1] was designed depends on the need states to find out online user behaviour. MRPs and LSA were developed to discover the sequence of search paths and repeated search patterns. Besides, the web pages related to the recommendation functions and non-recommendation functions were investigated with the help of clustering process. This aids to connect the evaluation results of search patterns with page traversal behaviours. With this, four types of users who browse for information, adopt recommendations, consult reviews, and conduct searches were obtained. But, the accuracy of behaviour analysis was not enhanced. To increase the accuracy while identifying the user behaviour, FPM-CSNUIA was developed in [2]. Information entropy weight allocation depends on the posterior probability was designed to discover the user by means of weight allocation. Though the designed method increased the accuracy and precision for user identification, time requirements was not reduced effectively. However, time requirement was remained high. LTL-based model checking technique was presented in [3] to provide better interpretation of users' behaviour. But, the rate of false positives was remained high. Also, fuzzy clustering algorithm was introduced in [4] to predict the diverse types of users. But, the time consumption was remained high. These existing methods are selected for result analysis because the proposed MIP-BSLR technique resolves the issues of conventional works.

5. Results

Experimental outcome of proposed MIP-BSLR performance is deliberated. The performance outcome of MIP-BSLR technique is matched with Sequential search pattern analysis and clustering [1], FPM-CSNUIA [2], LTL-based model checking technique [3] and fuzzy clustering algorithm [4] respectively using different parameters. Outcomes were visualized using tables and graphical representation.

5.1 Case 1: Performance measure of pattern mining accuracy

In MIP-BSLR technique, Pattern Mining Accuracy 'PMA' is determined as the proportion of the frequently accessed web patterns by the user is correctly mined (FWPCM) to the overall web patterns accessed (NWP). Mathematical model of pattern mining precision is represented as,

$$PMA = \frac{FWPCM}{NWP} \times 100 \quad (4)$$

From the mathematical Eq. (4), pattern mining accuracy is evaluated in percentage (%).

Table 1. (a) Pattern mining accuracy result using apache web log dataset

Number of web patterns	Pattern Mining Accuracy (%)				
	MIP-BSLR technique	Sequential search pattern analysis and clustering	FPM-CSNUIA	LTL-based model checking technique	fuzzy clustering algorithm
25	92	85	80	76	72
50	94	87	82	78	74
75	93	86	81	76	73
100	89	84	80	75	73
125	92	86	82	78	77
150	91	85	81	80	79
175	91	86	82	78	77
200	94	89	84	82	81
225	95	91	87	84	83
250	94	90	86	77	76

Table 1. (b). Pattern mining accuracy result using NASA dataset

Number of web patterns	Pattern Mining Accuracy (%)				
	MIP-BSLR technique	Sequential search pattern analysis and clustering	FPM-CSNUIA	LTL-based model checking technique	fuzzy clustering algorithm
100	88	83	78	72	68
200	92	85	80	76	71
300	94	86	82	78	73
400	93	84	79	76	71
500	89	82	78	74	70
600	92	85	82	78	72
700	91	86	83	80	75
800	93	88	85	82	79
900	95	89	86	84	80
1000	93	87	84	76	73

The experimental result analysis of pattern mining accuracy during the processes of web user behaviour identification using five methods namely proposed MIP-BSLR technique and conventional works Sequential search pattern analysis and clustering [1], FPM-CSNUIA [2], LTL-based model checking technique [3] and fuzzy clustering algorithm [4] is presented in Table 1.

Graphical result analysis of pattern mining accuracy of five methods i.e. proposed MIP-BSLR technique and conventional works Sequential search pattern analysis and clustering [1], FPM-CSNUIA [2], LTL-based model checking technique [3] and fuzzy clustering algorithm [4] are shown in figure 4 along with varied numbers of web patterns ranging from 25-250 and 100 to 1000 from Apache web log dataset and NASA dataset.

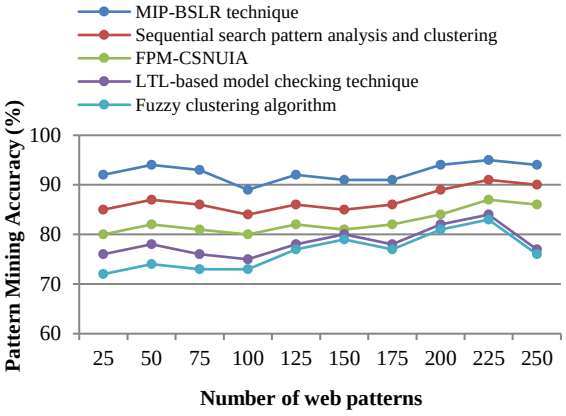
Fig. 4 show cases the suggested MIP-BSLR technique achieves enhanced accuracy while varying the web patterns count as input to exactly mine frequent web patterns and thereby discovering web usage behaviours when compared to existing [1-4]. This is owing to application of Mutual Information Based Pre-processing and Broken-Stick Linear Regression Analysis in proposed MIP-BSLR technique on the divergent to conventional works. Hence, proposed MIP-BSLR technique increases the proportion of the web patterns count frequently accessed by the user is properly extracted when compared to other conventional methods.

Accordingly, proposed MIP-BSLR technique improves web patterns mining accuracy using Apache web log dataset by 6% and 12%, 18 % and 21 % compared to existing Sequential search pattern analysis and clustering [1], FPM-CSNUIA [2], LTL-based model checking technique [3] and fuzzy clustering algorithm [4] correspondingly. Similarly, the pattern mining accuracy of NASA dataset is improved by 8%, 13%,19% and 26% as compared to existing Sequential search pattern analysis and clustering [1], FPM-CSNUIA [2], LTL-based model checking technique [3] and fuzzy clustering algorithm [4] respectively.

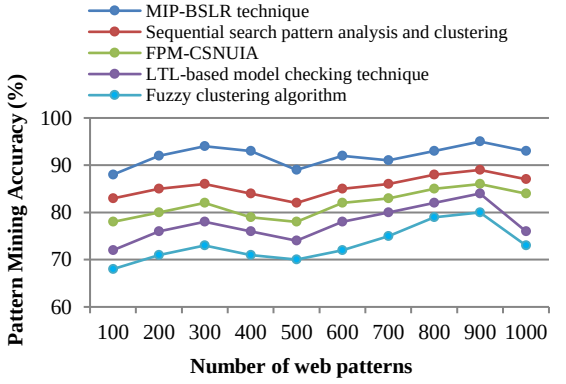
5.2 Case 2: Performance measure of false positive rate

In MIP-BSLR technique, false positives\ rate (FPR) is calculated as the proportion of the count of user accessed web patterns incorrectly mined as frequent(FWPIM) to the overall web patterns count (NWP). The mathematical estimation of false positive rate is as follows,

$$FPR = \frac{FWPIM}{NWP} \times 100 \qquad (5)$$



(a)



(b)

Figure. 4: (a) Results of pattern mining accuracy using apache web log dataset and (b) results of pattern mining accuracy using NASA dataset

The false positive rate is computed in percentage (%) using mathematical Eq. 5.

Table 2 (a) and (b) shows the comparative result analysis of false positive rate involved during the processes of web user behaviour mining using five methods namely proposed MIP-BSLR technique and conventional Sequential search pattern analysis and clustering [1], FPM-CSNUIA [2], LTL-based model checking technique [3] and fuzzy clustering algorithm [4].

Graphical result shown in figure 5 presents the false positive rate with respect to diverse number of web patterns using five methods i.e. MIP-BSLR technique and conventional methods [1-4]. From Fig. 5, proposed MIP-BSLR technique gets marginal false positive rate with increasing web pattern access count as input to precisely extract frequent web pattterns and thereby predicting web usage behaviours when compared to existing works of Sequential search pattern analysis and clustering [1], FPM-CSNUIA [2], LTL-based model checking technique [3] and fuzzy clustering algorithm [4].

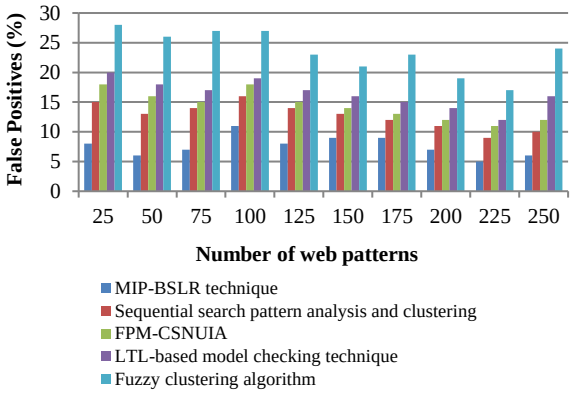
Table 2. (a) False positive rate results using apache web log dataset

Number of web patterns	False Positives (%)				
	MIP-BSLR technique	Sequential search pattern analysis and clustering	FPM-CSNUIA	LTL-based model checking technique	fuzzy clustering algorithm
25	8	15	18	20	28
50	6	13	16	18	26
75	7	14	15	17	27
100	11	16	18	19	27
125	8	14	15	17	23
150	9	13	14	16	21
175	9	12	13	15	23
200	7	11	12	14	19
225	5	9	11	12	17
250	6	10	12	16	24

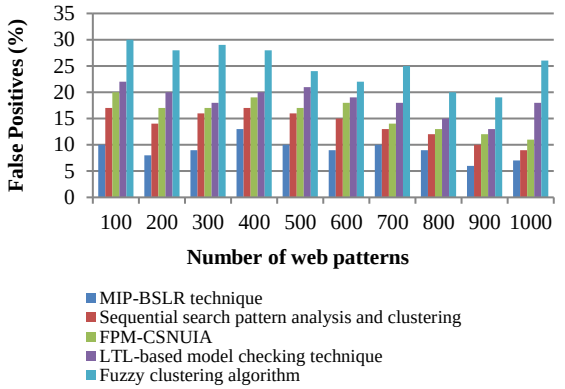
Table 2. (b) False positives results using NASA dataset

Number of web patterns	False Positives (%)				
	MIP-BSLR technique	Sequential search pattern analysis and clustering	FPM-CSNUIA	LTL-based model checking technique	fuzzy clustering algorithm
100	10	17	20	22	30
200	8	14	17	20	28
300	9	16	17	18	29
400	13	17	19	20	28
500	10	16	17	21	24
600	9	15	18	19	22
700	10	13	14	18	25
800	9	12	13	15	20
900	6	10	12	13	19
1000	7	9	11	18	26

This is owing to application of Mutual Information Based Pre-processing and Broken-Stick Linear Regression Analysis in proposed MIP-BSLR technique. Therefore, proposed MIP-BSLR technique reduces the count of web patterns frequently accessed by the user is wrongly extracted when matched to other conventional methods. For that reason, proposed MIP-BSLR technique minimizes false positive rate of web pattern mining through the Apache web log dataset by 40% ,47%, 54 % and 67 % and NASA web log dataset by 34%, 41%, 49% and 63% as compared Sequential search pattern analysis and clustering [1], FPM-CSNUIA [2], LTL-based model checking technique [3] and fuzzy clustering algorithm [4] correspondingly.



(a)



(b)

Figure. 5: (a) Outcome of false positive rate using apache web log dataset and (b) outcome of false positive rate using NASA dataset

5.3 Case 3: Performance measure of time requirements

In MIP-BSLR technique, Time Requirements (TR) estimates amount of time utilized to extract the frequent web patterns and thereby identifying web user behaviours. Mathematical representation of Time Requirements (TR) is as follows,

$$TR = NWP \times (miningonewebpattern) \quad (6)$$

From the mathematical expression (6), time requirements are determined in terms of milliseconds (ms).

Table 3 (a) and 3 (b) shows the tabulation result analysis of time complexity involved during the processes of web user behavior discovery using five methods namely proposed MIP-BSLR technique and existing Sequential search pattern analysis and clustering [1], FPM-CSNUIA [2], LTL-based model checking technique [3] and fuzzy clustering algorithm [4].

Table 3. (a) Time requirements results using apache web log dataset

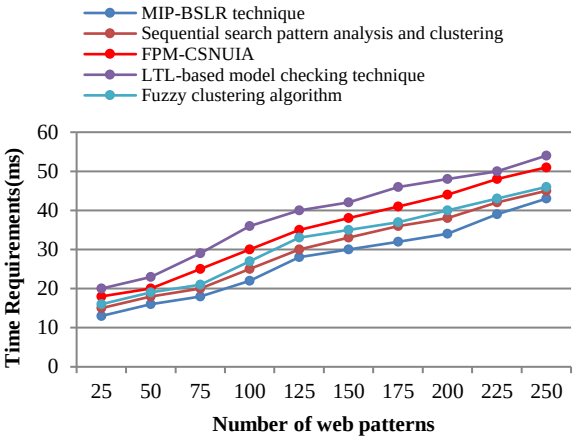
Number of web patterns (NWP)	Time Requirements(ms)				
	MIP-BSLR technique	Sequential search pattern analysis and clustering	FPM-CSNUIA	LTL-based model checking technique	fuzzy clustering algorithm
25	13	15	18	20	16
50	16	18	20	23	19
75	18	20	25	29	21
100	22	25	30	36	27
125	28	30	35	40	33
150	30	33	38	42	35
175	32	36	41	46	37
200	34	38	44	48	40
225	39	42	48	50	43
250	43	45	51	54	46

Table 3. (b) Time requirements results using NASA dataset

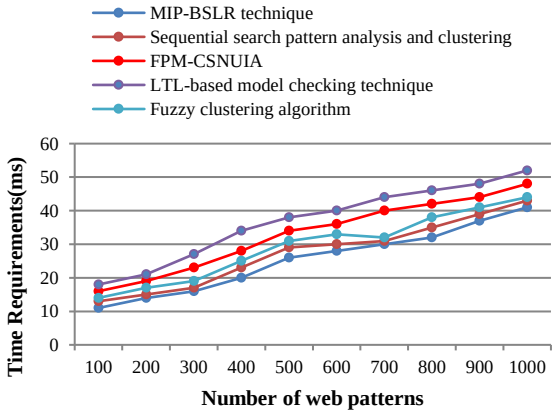
Number of web patterns (NWP)	Time Requirements(ms)				
	MIP-BSLR technique	Sequential search pattern analysis and clustering	FPM-CSNUIA	LTL-based model checking technique	fuzzy clustering algorithm
100	11	13	16	18	14
200	14	15	19	21	17
300	16	17	23	27	19
400	20	23	28	34	25
500	26	29	34	38	31
600	28	30	36	40	33
700	30	31	40	44	32
800	32	35	42	46	38
900	37	39	44	48	41
1000	41	43	48	52	44

Figure 6 (a) and 6 (b) portrays experimental result of time requirements according to different numbers of web patterns using five methods i.e. MIP-BSLR technique and existing Sequential search pattern analysis and clustering [1], FPM-CSNUIA [2], LTL-based model checking technique [3] and fuzzy clustering algorithm [4].

From the figure 6(a) and 6(b) suggested MIP-BSLR technique attains lower amount of time complexity with increasing number of web patterns as input to correctly take out frequent web patterns and thereby detecting web usage behaviours when matched with conventional Sequential search pattern analysis and clustering [1], FPM-CSNUIA [2], LTL-based model checking technique [3] and fuzzy clustering algorithm [4]. This is due to application of



(a)



(b)

Figure. 6: (a) Comparative result of time requirements using apache web log dataset and (b) comparative result of time requirements using NASA dataset

Mutual Information Based Pre-processing and Broken-Stick Linear Regression Analysis in proposed MIP-BSLR.

Thus, proposed MIP-BSLR technique minimizes the amount of time used to extract the frequent web patterns and thereby determining web user behaviours when matched with existing methods. Time requirements of web pattern mining using Apache web log dataset is reduced in the proposed MIP-BSLR technique by 10% and 23% 30% and 14% when matched Sequential search pattern analysis and clustering [1], FPM-CSNUIA [2], LTL-based model checking technique [3] and fuzzy clustering algorithm [4] respectively. In addition, Time requirements of MIP-BSLR is minimized through NASA dataset by 8%, 24%, 32% and 14% as compared to existing methods Sequential search pattern analysis and clustering [1], FPM-CSNUIA [2], LTL-based model checking technique [3] and fuzzy clustering algorithm [4].

Table 4. (a) Tabulation result of space complexity using apache web log dataset

Number of web patterns (NWP)	Space Requirements (KB)				
	MIP-BSLR technique	Sequential search pattern analysis and clustering	FPM-CSNUIA	LTL-based model checking technique	fuzzy clustering algorithm
25	10	11	14	16	12
50	13	14	17	19	15
75	15	17	20	23	18
100	18	19	23	28	21
125	19	21	25	30	23
150	20	22	28	33	24
175	24	26	30	37	27
200	26	28	32	40	30
225	27	30	35	43	33
250	29	33	38	46	36

Table 4. (b) Tabulation result of space complexity using NASA dataset

Number of web patterns (NWP)	Space Requirements (KB)				
	MIP-BSLR technique	Sequential search pattern analysis and clustering	FPM-CSNUIA	LTL-based model checking technique	fuzzy clustering algorithm
100	8	9	11	14	10
200	11	12	14	17	13
300	13	15	18	21	16
400	18	19	23	26	20
500	17	19	24	28	21
600	18	20	26	31	22
700	22	23	28	35	25
800	24	26	30	38	28
900	25	29	33	41	31
1000	27	30	36	44	34

5.4 Case 4: Performance measure of space requirements

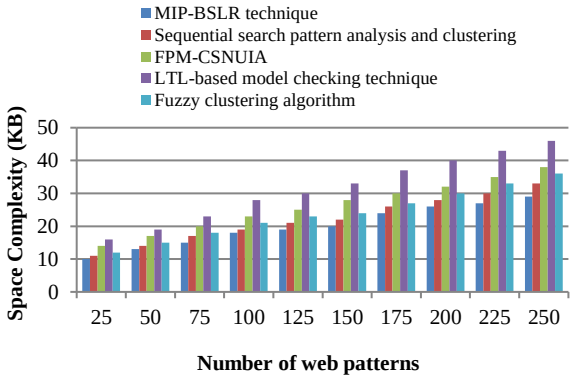
Space Requirements ' SR ' in MIP-BSLR technique evaluates an amount of memory space used for storing the extracted frequent web patterns. The space requirements are mathematically measured as follows,

$$SR = NWP \times Memory\ (\ onewebpattern) \quad (7)$$

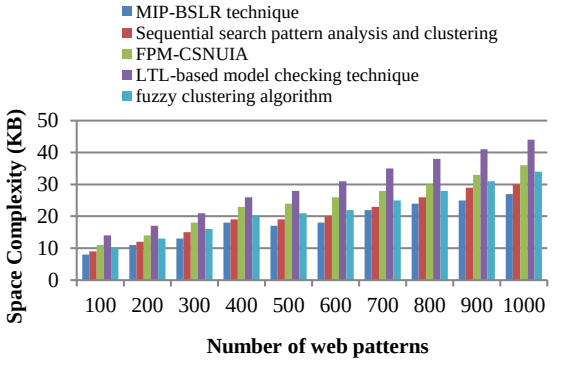
From mathematical formula (7), in terms of kilobytes (KB) the space requirements is calculated.

The performance result analysis of space requirements involved during the processes of web user behavior detection using five methods namely proposed MIP-BSLR technique and state-of-the-art Sequential search pattern analysis and clustering [1], FPM-CSNUIA [2], LTL-based model checking technique [3] and fuzzy clustering algorithm [4] is depicted in Table 4.

Figs. 7 (a) and 7 (b) shows space requirements result of web pattern mining based on dissimilar numbers of web patterns using five methods i. e. MIP-BSLR technique and existing Sequential search pattern analysis and clustering [1], FPM-CSNUIA [2], LTL-based model checking technique [3] and fuzzy clustering algorithm [4]. As per graphical represenation in Figs. 7 (a) and 7 (b), proposed MIP-BSLR technique obtains minimal space requirements with increasing number of web patterns usage and behaviour mining when matched with conventional works in [1-4]. This is because of application of Mutual Information Based Pre-processing and Broken-Stick Linear Regression Analysis in proposed MIP-BSLR technique. Extracted frequent



(a)



(b)

Figure. 7: (a) Comparative result of space requirements using apache web log dataset and (b) comparative result of space requirements using NASA dataset

web patterns storage space utilization is reduced in the proposed MIP-BSLR technique when compared to other conventional methods. Therefore, proposed MIP-BSLR technique reduces web pattern mining space requirements using Apache web log dataset by 9%, 24%, 36% and 16% when compared [1], [2], [3] and [4] respectively. Also, the space complexity of MIP-BSLR technique through the NASA dataset is reduced by 9%, 25%, 38% and 17% as compared to existing [1-4] methods.

6. Conclusion

The MIP-BSLR technique is designed with the intention of enhancing the web usage behaviour mining performance. The MIP-BSLR technique is desigend with the contribution of Mutual Information based Pre-processing (MI-P) method and Broken-Stick Linear Regression analysis (BLRA). At first, MI-P is applied in MIP-BSLR technique to identify the mutual dependance between web patterns for selecting relevant patterns. This in turns, highly reduces the time requirements and space requirements. Then the BLRA is applied on the preprocessed web patterns for identifying the frequent web patters through detecting the relationship between patterns. This aids to increases the accuracy of patterm identification with less error rate. Experimental evaluation is carried out using Apache web log dataset and NASA dataset with the parameters such as pattern mining accuracy, time requirement, space requirements and false postive rate. The results and discussion shows that MIP-BSLR technique improves the pattern mining accuracy by 14% with minimum false positive rate by 52%, time requirement by 19% and space requirements by 21% using Apache web log dataset than the state-of-the-art methods. Also, the pattern mining accuracy of NASA dataset is enahnced by 16% with less time requirement by 20%, false positives by 47% and space requirements by 22% as compared to existing [1-4] methods. In future, we perform the web usage behaviour mining in the large scale applications with more metrics.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

The contributions by the authors for this research article are as follows: "conceptualization, methodology Gokulapriya Raman and Ganesh Kumar Raj; Formal analysis, data curation, visualization and writing—original draft preparation Gokulapriya Raman; Result validation, data curation, resources, formal analysis writing—review and editing and supervision Ganesh Kumar Raj;"

Acknowledgments

Authors wishes to acknowledge the technical and infrastructural help rendered by the faculty members from Department of CSE CHRIST (Deemed to be University), Bangalore, India.

References

[1] I. Wu and H. Yu, "Sequential Analysis and Clustering to Investigate Users' Online Shopping Behaviours based on Need-States", *Information Processing and Management,* Vol. 57, No. 6, pp. 1-18, 2020.

[2] K. Deng, L. Xing, L. Zheng, H. Wu, P. Xie, and F. Gao, "A User Identification Algorithm based on User Behaviour Analysis in Social Networks", *Applications of Big Data in Social Sciences, IEEE Access*, Vol. 7, No. 1 pp. 47114 – 47123, 2019.

[3] S. Hernández, P. Álvarez, J. Fabra, and J. Ezpeleta, "Analysis of Users' Behaviour in Structured e- Commerce Websites", *IEEE Access*, Vol. 5, No. 1, pp. 11941–11958, 2017.

[4] D. Anandhi and M. S. Irfan Ahmed, "Prediction of User's Type and Navigation Pattern using Clustering and Classification Algorithms", *Cluster Computing*, Vol. 22, No. 1, pp. 10481–10490, 2019.

[5] R. Hinami and S. Satoh, "Audience Behaviour Mining: Integrating TV Ratings with Multimedia Content*", IEEE Multimedia,* Vol. 24, No. 2, pp. 44–54, 2017.

[6] S. Maiti and R. B. V. Subramanyam, "Mining Behavioural Patterns from Spatial Data", *Engineering Science and Technology*, Vol. 22, No. 2, pp. 618-628, 2019.

[7] J. Ya, T. Liu, Q. Li, J. Shi, H. Zhang, P. Lv, and L. Guo, "Mining Host Behaviour Patterns from Massive Network and Security Logs", *Procedia Computer Science*, Vol. 108, No. 1, pp. 38–47, 2017.

[8] A. Apaolaza and M. Vigo, "Assisted Pattern Mining for Discovering Interactive Behaviours on the Web", *International Journal of Human-Computer Studies*, Vol. 130, No. 1, pp. 196-208, 2019.

[9] O. Raphaeli, A. Goldstein, and L. Fink, "Analyzing Online Consumer Behaviour in Mobile and PC Devices: A Novel Web Usage Mining Approach", *Electronic Commerce*

Research and Applications, Vol. 26, No. 1, pp. 1-12, 2017.

[10] T. Chovanak, O. Kassak, M. Kompan, and M. Bielikova, "Fast Streaming Behavioural Pattern Mining", *New Generation Computing*, Vol. 36, No. 4, pp. 365–391, 2018.

[11] Q. Yi, S. Xiong, B. Wang, and S. Yi, "Identification of Trusted Interactive Behaviour Based on Mouse Behaviour Considering Web User's Emotions", *International Journal of Industrial Ergonomics*, Vol. 76, No. 1, pp. 1-10, 2020.

[12] N. Gal-Oz, Y. Gonen, and E. Gudes, "Mining Meaningful and Rare Roles from Web Application Usage Patterns", *Computers & Security*, Vol. 82, No. 1, pp. 296-313, 2019.

[13] C. Shen, Y. Chen, X. Guan, and R. A. Maxion, "Pattern-Growth based Mining Mouse Interaction Behaviour for an Active User Authentication System", *IEEE Transactions on Dependable and Secure Computing*, Vol. 17, No. 2, pp. 335-349, 2020.

[14] M. Bilal, M. Marjani, M. I. Lali, and N. Malik, "Profiling Users' Behaviour, and Identifying Important Features of Review "Helpfulness", *IEEE Access*, Vol. 8, No. 1, pp. 77227 – 77244, 2020.

[15] DS. Sisodia, V. Khandal, and R. Singhal, "Fast Prediction of Web User Browsing Behaviours using Most Interesting Patterns", *Journal of Information Science*, Vol. 44, No. 1, pp. 74-90, 2018.

[16] G. Bello-Orgaz, RM. Mesas, C. Zarco, V. Rodriguez, O. Cordón, and D. Camacho, "Marketing Analysis of Wineries using Social Collective Behaviour from Users' Temporal Activity on Twitter", *Information Processing and Management*, Vol. 57, No. 5, pp. 1-20, 2020.

[17] V. F. de Santana and M. C. C. Baranauskas, "WELFIT: A Remote Evaluation Tool for Identifying Web Usage Patterns through Client-Side Logging", *International Journal of Human-Computer Studies*, Vol. 76, No. 1, pp. 40-49, 2015.

[18] A. D. Kasliwal and G. S. Katkar, "Web Usage mining for Predicting User Access Behaviour", *International Journal of Computer Science and Information Technologies*, Vol. 6, No. 1, pp. 201-204, 2015.

[19] Q. F. Ying, D. M. Chiu, S. Venkatramanan, and X. Zhang, "User Modelling and Usage Profiling based on Temporal Posting Behaviour in OSNs", *Online Social Networks and Media*, Vol. 8, No. 1, pp. 32-41, 2018.

[20] Apache web log dataset: http://www.almhuette-raith.at/apache-log/access.log

[21] NASA dataset: https://opensource.indeedeng.io/imhotep/docs/sample-data/#nasa-apache-web-logs