ORIGINAL RESEARCH PAPER
Volume-6 | Issue-12 | December-2017 | ISSN No 2277 - 8179 | IF : 4.176 | IC Value : 93.98
INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

LINEAR DISCRIMINANT CLASSIFIER BASED PEARSON CORRELATION FOR WEB TRAFFIC PATTERN MINING

Computer Science

Ulaganathan.N. HOD Assistant Professor, NIFT-TEA College of Knitwear Fashion SIDCO, Mudalipalayam, Tirupur Affiliated to Bharathiar University East of TEKIC, Coimbatore-641 046, Tamilnadu, India

ABSTRACT

With the fast growing popularity of the WWW, Websites plays an important role to communicate knowledge and information to the users. The task of mining web traffic patterns is very difficult when the weblog database is enormous. Few research works has been designed for predicting the traffic web patterns to analyze the user's behaviours. But, performance of existing traffic web patterns mining was not efficient. In order to overcome such limitations, MapReduce Pearson Correlation Fisher's Linear Discriminant Classifier (MPC-FLDC) technique is proposed. The MPC-FLDC technique efficiently extracts web traffic patterns from weblogs through pre-processing, classification and correlation analysis. The MPC-FLDC technique initially carried outs pre-processing in which MapReduce framework is used to group the web patterns according to different sessions. After preprocessing, MPC-FLDC technique applied Fisher's Linear Discriminant (FLD) Classifier to classify the web patterns at a different sessions as frequent or non-frequent based on hit ratio. This process resulting in improved classification accuracy of web patterns. Finally, MPC-FLDC technique used Pearson correlation analysis that evaluates the web patterns correlation between a different user sessions in order to efficiently predict the traffic web patterns with higher accuracy and minimum time. This process assists for MPC-FLDC technique to improve the traffic patterns prediction rate and reducing the prediction time. The performance of MPC-FLDC technique is evaluated with parameters such as classification accuracy, traffic patterns prediction accuracy, prediction time, and true positive rate with respect to different web patterns. The experimental result shows that MPC-FLDC technique is able to increases the traffic patterns prediction accuracy and also lessens the prediction time when compared to state-of-the-art-works.

KEYWORDS

Fisher's Linear Discriminant Classifier, Mapreduce Framework, Pearson Correlation Analysis, Pre-Processing, Traffic Web Patterns, Web Users

1. Introduction

The World Wide Web (WWW) attains greater attention with the increasing information transaction from web servers and number of requests from web users. Discovering information regarding web user's usage patterns is essential for marketing strategies to increase the future growth. Most of the existing technique presents only statistical information without real useful knowledge for Web managers. Therefore, mining web traffic and user access patterns is significant for marketing and management of E-business, E-services, E-searching, and E-education.

Following are the several research works designed for web usage mining to analyze the user behaviour. A web usage mining approach was presented in [1] for predicting online navigational behavior. But, prediction performance of web user behaviour was not efficient. Besides, predicting the web traffic patterns was remained unsolved. A fuzzy clustering was used in [2] to predict the maximum visited pages by user from website. However, prediction time was higher.

A novel method was designed in [3] for identifying web usage patterns through client-side logging. However, time complexity for identification of usage patterns was more. Sequence- based clustering was developed in [4] for web usage mining to evaluate elder self-care behavior patterns. But, performance of web usage mining was not efficient.

A novel web usage mining approach was presented in [5] to predict sequences of navigation patterns. However, prediction accuracy was remained unsolved. In [6], web mining integrated with the electronic commerce application in order to improve the performance of web user behaviour analysis.

A review of different clustering techniques designed for web usage mining was analyzed in [7] to improving the prediction accuracy. K-Nearest Neighbor (KNN) classification method was intended in [8] for automatic web usage data mining based on user behaviors. But, a web traffic pattern mining was remained unaddressed.

A survey of diverse techniques designed for mining users' navigational behavior from weblog was presented in [9]. A novel algorithm was presented in [10] for discovering frequent web sites-users. However, prediction accuracy and time was not at required level.

In order to solve the above mentioned existing issues, The MapReduce Pearson Correlation Fisher's Linear Discriminant Classifier (MPC-FLDC) technique is introduced. The major contribution of MPC-FLDC technique is organized as follows,

- To improve the prediction performance of web traffic patterns from a weblog database, MPC-FLDC technique is designed. The MPC-FLDC technique is developed with application of Fisher's Linear Discriminant Classifier and Pearson Correlation Analysis for efficient daily/hourly traffics prediction.
- To group the web patterns in a weblog database according to different sessions, MapReduce framework is applied in MPC-FLDC technique. The MapReduce framework used two phases such as Map and Reduce for efficiently grouping web patterns at different sessions.
- To classify the web patterns as frequent or non-frequent at different sessions, Fisher's Linear Discriminant (FLD) Classifier is applied in MPC-FLDC technique. The FLD Classifier finds optimal projection direction for improving the classification accuracy of web patterns based on hit ratio.
- To predict the web traffic patterns among different sessions with minimum time, Pearson Correlation Analysis is applied in MPC-FLDC technique. The Pearson Correlation Analysis determines the web patterns correlation between a different user sessions for efficiently mining the traffic web patterns.

The remaining structure of the paper is organized as follows: In Section 2, MapReduce Pearson Correlation Fisher's Linear Discriminant Classifier (MPC-FLDC) technique is described with help of architecture diagram. In Section 3, Simulation settings are presented and the result discussion is explained in Section 4. Section 5 introduces the background and reviews the related works. Section 6 provides the conclusion of the paper.

2. MapReduce Pearson Correlation Fisher's Linear Discriminant Classifier framework

The MapReduce Pearson Correlation Fisher's Linear Discriminant Classifier (MPC-FLDC) technique is developed with objective of improving the performance of web traffic patterns mining with higher accuracy. The MPC-FLDC technique used Fisher's Linear Discriminant (FLD) Classifier for categorizing the web patterns as frequent or non-frequent web patterns in weblog database according to different sessions. This helps for MPC-FLDC technique to improve the classification accuracy for finding the frequent web pages (i.e. web patterns) browsed by different web users. After classification, MPC-

FLDC technique employed Pearson correlation analysis to evaluate the web patterns correlation among different sessions for efficient prediction of daily/hourly traffics with minimum time complexity. The overall architecture diagram of MapReduce Pearson Correlation Fisher's Linear Discriminant Classifier (MPC-FLDC) technique is shown in below Figure 1.

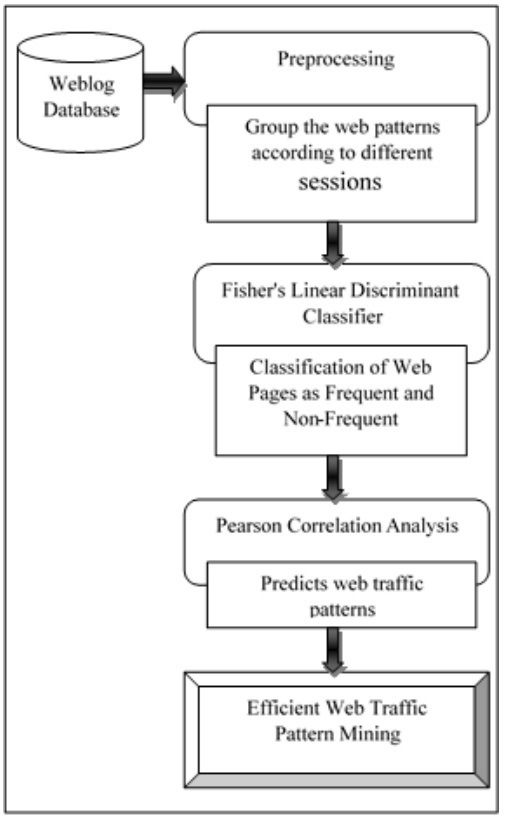


Figure 1Architecture of MapReduce Pearson Correlation Fisher's Linear Discriminant Classifier for Web Traffic Pattern Mining

Figure 1 shows the process of MPC-FLDC technique for efficiently mining web traffic patterns. As shown in Figure 1, MPC-FLDC technique at first takes weblog database (i.e. Apache log samples dataset) as input. Then, MPC-FLDC technique performs preprocessing to group the web patterns in a weblog database based on a sessions. After that, MPC-FLDC technique applies Fisher's Linear Discriminant Classifier to classify the web patterns as frequent or non-frequent with higher classification accuracy. Finally, MPC-FLDC technique used Pearson Correlation Analysis in order to predict the web traffic patterns efficiently with minimum time. The elaborate description about MPC-FLDC technique is explained in next sub sections.

2.1 Preprocessing

In MPC-FLDC technique, preprocessing is carried out in order to group web patterns in a weblog database based on different sessions (i.e. Time Interval) using MapReduce framework. MapReduce framework partitions the web user's activities into a sequences (sessions) based on access time. The MapReduce is a divide-and-conquer program model which comprises of "Map" and "Reduce" phase. The input to the MapReduce is divided into a list of key/value pairs. The map and reduce task processes jobs of data on all nodes stored in a local machine. The MapReduce framework used in MPC-FLDC technique segments the original web logs into a number of sessions based on access time. Generally Web server stores the access activities of web users in Weblog database. The weblog database consists of client IP address, time, requested URL, HTTP status code, referrer, etc. The separation of web patterns according to diverse sessions helps for MPC-FLDC technique to find the web pages visited by web users in a particular period of time for effective web traffic pattern mining.

The MPC-FLDC technique applies MapReduce framework on weblog database to divide the weblog database into a number of sessions '

From equation (1), *wp1,wp3,..wp7* are web pages visited by users in a specified time period (i.e. session *S.*

The MapReduce framework used in MPC-FLDC process huge volumes of weblog database in a parallel manner and efficiently group web patterns. In MPC-FLDC, the technique MapReduce framework takes weblog database an input and then map task transforms it into a Key-Value pairs in which the key includes classes (different sessions) and their related web patterns which is formulated as,

$S = (wp1, wp3, wp5, .. wp7)$ (1)

$Map(Y) \rightarrow (Key, Value)$ (2)

$(Key) \rightarrow (Class, web\ patterns)$ (3)

From equation (2) and (3), *key* denotes the different time intervals (i.e. sessions). Thus, map phase in MapReduce framework helps for grouping the web patterns in a weblog database according to diverse sessions. The process of MapReduce framework for separating web patterns in a weblog database is shown in below,

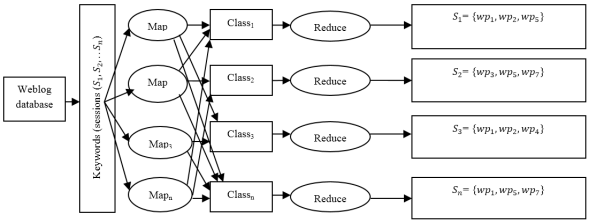
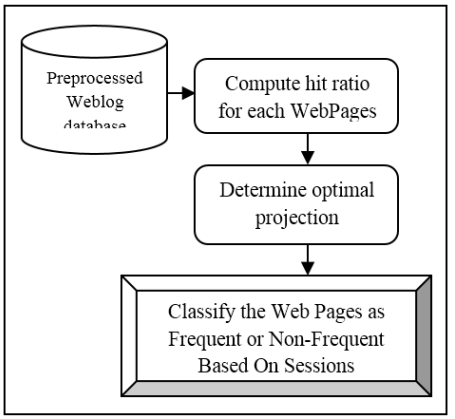


Figure 2 Process of Mapreduce Framework for Grouping Web Patterns

As shown in Figure 2, A MapReduce Frameworks initially takes weblog database as input and then divides the input into number of sessions which are processed by the map tasks in an entirely parallel manner. Subsequently sorts the outputs of the maps and input to the reduce tasks. Finally, the output of reduce tasks includes of different sessions and their associated web patterns from a weblogs. This assists for MPC-FLDC to reduce the time complexity for efficient web traffic pattern mining.**2.2 Fisher's Linear Discriminant Classifier** After completing the preprocessing task, MPC-FLDC technique used Fisher's Linear Discriminant (FLD) Classifier in order to classify the web pages (i.e. web patterns) visited by web users stored in weblog database as frequent or non-frequent based on their sessions. Session represents that the web user's actions performed within a period of time. The FLD Classifier designed in MPC-FLDC technique categorizes the web pages which are most frequently browsed by web users within a period of time as a frequent web patterns and the remaining web pages as non-frequent based on hit ratio. Here the hit ratio is defined as the number of times that the particular web page found in weblog database to the total number of web pages. Besides, the hit ratio also represents that the probability of web pages being in the weblog database. The basic idea of FLD Classifier is that the web pages are projected onto a line and the classification is performed in this one-dimensional space. The projection increases the distance between means of two classes as reducing the variance in each class. The process of FLD Classifier is shown in below Figure 3.



As shown in Figure 3, initially FLD Classifier takes the preprocessed weblog database as input and then evaluates the hit ratio for each web pages stored in weblog database. Based on measured hit ratio of web pages, then FLD Classifier determines optimal projection direction for efficient web patterns classification. Finally, FLD Classifier separates the frequent web patterns and non-frequent web patterns through identified optimal projection direction according to their different sessions 'S1,S2,..Sn'.

Let us consider the weblog database consists of numerous web pages like {wp1,wp2,..wpN } where wp(i)∈R^n in which n1samples are in class 1 denoted C1 (frequent web pages) and n2 samples in class 2 denoted C2(non-frequent web pages). The FLD Classifier only performs the classification between two classes with respect to different sessions of web users. The hit ratio of each web page in weblog database is evaluated using following mathematical formula,

$$HR_{wp_i} = \frac{\text{nt}\,(wp_i)}{\text{N}} \qquad (4)$$

From equation (4), N denotes the total number of web pages in weblog database whereas nt (*wpi*) represents the number of times that the particular web page found in weblog database. After evaluating the hit ratio, the class separability function of FLD Classifier is determined. The class separability function in a direction $v \in R^n$ at a particular session *Si* defined as,

$$J(v) = \frac{v^T SC_B v}{v^T SC_W v} \qquad (5)$$

From equation (5), *SCB* and *SCW* indicates the scatter matrixes between-class and within class which are measured as,

$$SC_B = (m_1 - m_2)(m_1 - m_2)^T \qquad (6)$$

$$SCw = \sum_{i=1,2} \sum_{wp \in C_i} (x - m_i)(x - m_i)^T \qquad (7)$$

Then, subsequently sample mean of the respective classes m_i is evaluated as,

$$m_i = \frac{1}{n_i} \sum_{wp \in C_1} wp \qquad (8)$$

Thus, The Fisher linear discriminant is given by the vector v that maximizes the class separability function *J(v)*. The equation (5) is a particular case of the generalized Rayleigh quotient and therefore assuming that *SCW* is a non-singular matrix. From that, it is potential to determine an analytic expression for v using below mathematical expression which maximizes *J(v)*,

$$w = S_w^{-1}(m_1 - m_2) \qquad (9)$$

By using the equation (9), the optimal projection direction v is determined that helps for classifying the web pages into an each one of the two classes. The discovered optimal projection direction 'v' efficiently categories the web pages in weblog database as frequent or non-frequent web patterns according to different sessions 'S1,S2,..Sn'. The following diagram shows the FLD Classifier results for a particular session *Si* .

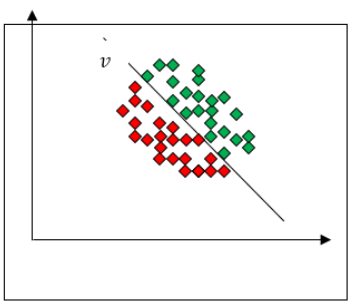


Figure 4 demonstrates the output of FLD Classifier for separating frequent or non-frequent web patterns in web log database to predict the traffic web patterns. As shown in Figure 4, green color diamond represents the frequent web patterns browsed by web users in a weblog database at a particular session Si whereas red color diamond indicates the non frequent web patterns. The algorithmic process of Fisher's Linear Discriminant Classifier is shown in below,

// Fisher's Linear Discriminant Classifier Algorithm
Input: Preprocessed Weblog Database
Output: Classify the Web Pages as Frequent or Non-Frequent patterns according to Sessions *Si*

Step 1: Begin
Step 2: For each web page in web log database
Step 3: Measure hit ratio using (4)
Step 4: Define class separability function using (5)
Step 5: Compute scatter matrixes between-class and within class using (6) and (7)
Step 6: Sample mean of the respective classes is determined using (8)
Step 7: Find optimal projection direction to classify the web pages as frequent or non frequent at a particular session Si using (9)
Step 8: End for
Step 9: End

Algorithm 1 Fisher's Linear Discriminant Classifier

Algorithm 1 shows the step by step process of FLD Classifier for categorizing the frequent or non-frequent patterns. As depicted in algorithm, FLD Classifier initially determines the hit ratio for all web pages in a weblog database. Then, FLD Classifier defines class separability function through measuring the scatter matrixes between-class and within class. After that, Sample mean of the respective classes is estimated. At last, FLD Classifier discovers optimal projection direction to classify the web pages as frequent or non-frequent patterns based on a sessionSi. This process is continued until all the web pages in weblog database are classified with respect to different number of sessions 'S1,S2,..Sn'. This helps for MPC-FLDC technique to increase the classification accuracy in an effective manner. After classification, the classified web traffic patterns are employed for efficient web traffic pattern prediction which is detailed explained in next section.

2.3 Pearson Correlation Analysis based Web Traffic Pattern Mining

The MPC-FLDC technique used Pearson Correlation Analysis for efficient web traffic patterns prediction (i.e. daily/hourly traffic) from a weblog database. The web traffic patterns represents a web pages that are visited more number of times by a number of web users. The web traffic patterns predicted from a classified frequent web patterns through Pearson correlation. The MPC-FLDC technique employed Pearson Correlation analysis to determine degree of web pages correlation among different sessions for web traffic predictions. In MPC-FLDC technique, Pearson Correlation measures the web pages similarity between the user sessions. With the aid of computed degree of correlation i.e. similarity, then MPC-FLDC technique efficiently forecast the daily and hourly traffic volume. This helps for MPC-FLDC technique to achieve higher prediction rate for mining web traffic patterns from a weblog database. The process involved in Pearson Correlation analysis for effective web traffic patterns prediction is shown in below Figure 5.

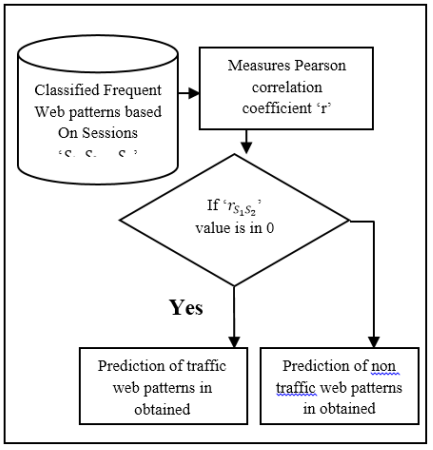


Figure 5 Process of Pearson Correlation Analysis for Web Traffic Patterns Prediction

Figure 5 shows the process of Pearson correlation analysis for web traffic patterns predictions and mining. As shown in figure, Pearson correlation analysis initially evaluates the Pearson correlation coefficient *'r S1 S2 '* between the web patterns in different sessions *'S1,S2,..S*n'. If Pearson correlation coefficient '*r S1 S2'* values is lies between the 0 to +1.00, then the prediction of traffic web patterns in obtained. Otherwise, prediction of non traffic web patterns in obtained. As a result, MPC-FLDC technique increases the prediction rate of web traffic patterns mining with minimum time. This process results in minimum prediction time for web traffic patterns analysis.

The proposed MPC-FLDC technique used Pearson correlation analysis for computing the degree of the correlation between web patterns in diverse '*S1,S2*,..S n' to forecast the web traffic patterns on an hourly and daily basis. The following mathematical expression is used to compute the Pearson correlation coefficient to predict the web traffic patterns between the two user session S_1 and S_2 ,

$$r_{S_1 S_2} = \frac{\sum S_1 S_2 - \frac{(\sum S_1)(\sum S_2)}{n}}{\sqrt{\left(\sum {S_1}^2 - \frac{(\sum S_1)^2}{n}\right)\left(\sum {S_2}^2 - \frac{(\sum S_2)^2}{n}\right)}} \qquad (10)$$

From equation (10), *n* represents the number of user sessions 'S$_1$,S$_2$,..Sn', S$_i$and S$_1$ are ith sessions. Here, $\sum S_1 S_2$ refers the sum of cross product of S$_1$ andS$_2$, $\sum S_1$ is the sum of frequent web patterns in sessionS$_1$, $\sum y$ is the frequent web patterns in sessionS$_2$, $\sum x^2$ is the sum of squared of session S$_1$ and $\sum y^2$ is the sum of squared of frequent web patterns of session S$_1$. With help of equation (7), web patterns correlation among the different user sessions 'S$_1$,S$_2$,..S$_n$' is determined in order to predicts the traffic web patterns. The results of Pearson correlation coefficient r value is lies between the -1.00 to +1.00. If the measured r value between two user sessions is lies between the 0 to +1.00, then the prediction of traffic web patterns in acquired. Otherwise, prediction of non traffic web patterns in acquired. The algorithmic process of Pearson Correlation analysis for Web Traffic Pattern Mining is shown in below.

// Pearson Correlation Analysis based Web Traffic Pattern Mining Algorithm
Input: Classified frequent web pages according to diverse sessions "
Output: Improved traffic pattern prediction rate and reduced prediction time
Step 1: Begin
Step 2: **For** each session of classified frequent web pages
Step 3: Compute the Pearson correlation coefficient " using (10)
Step 4: **If** (" value is lies between 0 to +1.00)
Step 5: The web patterns in sessions are predicted as traffic web patterns
Step 6: **else If** (" value lies between -1.00 to 0)
Step 7: The web patterns in sessions are predicted as non traffic web patterns
Step 8: **End if**
Step 9: Mine the predicted traffic web patterns
Step 10: **End for**
Step 11: End

Algorithm 2 Pearson Correlation Analysis based Web Traffic Pattern Mining As shown in Algorithm 2, at first Pearson correlation analysis measures the Pearson correlation coefficient value for each user sessions of classified frequent web patterns in a weblog database. After that, Pearson correlation analysis algorithm ensures if r *S1 S2* value is ranges from 0 to +1.00 which tells that there is a perfect positive web patterns similarity between the user sessions. Therefore web patterns in sessions are predicted as traffic web patterns. Otherwise, Pearson correlation analysis checks if r*S1 S2* value is ranges from-1.00 to 0 which tells that there is a perfect negative web patterns similarity between the user sessions. Therefore, web patterns in sessions are predicted as non traffic web patterns. At last, Pearson correlation analysis mines the predicted web traffic patterns efficiently. The measure of Pearson correlation coefficient among diverse sessions 'S_1,S_2,..S_n' is also helps for effectively predicting the hourly and daily web traffics from a weblog database with minimum time. As a result, MPC-FLDC technique attains higher prediction accuracy and minimum prediction time for web traffic pattern analysis.

3. Experimental Settings

In order to evaluate the performance of proposed, MapReduce Pearson Correlation Fisher's Linear Discriminant Classifier (MPC-FLDC) technique is implemented in Java language using Apache log samples dataset [21]. The Apache log samples dataset comprises of access activities of numerous web users such as IP address, Date, Time of Access, Port Number, accessed Webpage's. The performance of MPC-FLDC technique is compared against with existing web usage mining approach [1] and fuzzy clustering [2]. The effectiveness of MPC-FLDC technique is measured in terms of classification accuracy, traffic patterns prediction rate, prediction time and true positive rate. The experimental evaluation of MPC-FLDC technique is conducted for several instances with respect to diverse number of web patterns and averagely ten results is depicted in table and graph for performance analysis.

4. Results and Discussions

In this section, the result of MPC-FLDC technique is presented. The performance of MPC-FLDC technique is compared against with existing web usage mining approach [1] and fuzzy clustering [2] respectively. The efficiency of MPC-FLDC technique is analyzed along with the following metrics with the aid of tables and graphs.

4.1 Measure of Classification Accuracy In MPC-FLDC technique, Classification Accuracy (CA) is defined as the ratio of the number of web patterns that are correctly classified as frequent to the total number of web patterns. The Classification Accuracy is evaluated in terms of percentage (%) and formulated as,

$$CA = \frac{number\ of\ web\ patterns\ that\ are\ correctly\ classified\ as\ frequent}{total\ number\ of\ web\ patterns} * 100 \qquad (11)$$

From equation (11), the classification accuracy for web traffic pattern analysis is determined with respect to different number of web patterns. While classification accuracy is higher, the method is said to be more efficient.

Table 1 Tabulation for Classification Accuracy

Number of Patterns	Classification Accuracy (%)		
	Web Usage Mining Approach	Fuzzy Clustering	MPC-FLDC technique
30	66	72	83
60	69	73	85
90	70	76	86
120	71	77	88
150	73	78	89
180	74	80	90
210	76	81	92
240	77	83	93
270	79	84	96
300	80	86	97

Table 1 depicts the tabulation results of classification accuracy for mining web traffic patterns for weblogs database based on various number of web patterns using three methods. The MPC-FLDC technique considers the framework with different number of web patterns in the range of 30-300 using java language. While considering the 150 web patterns for conducting experimental works, the proposed MPC-FLDC technique achieves 89 % classification accuracy for finding the frequent web patterns whereas web usage mining approach [1] and fuzzy clustering [2] obtains 73 % and 78 % respectively. Thus, it is illustrative that the classification accuracy using proposed MPC-FLDC technique is higher as compared to other existing [1], [2].

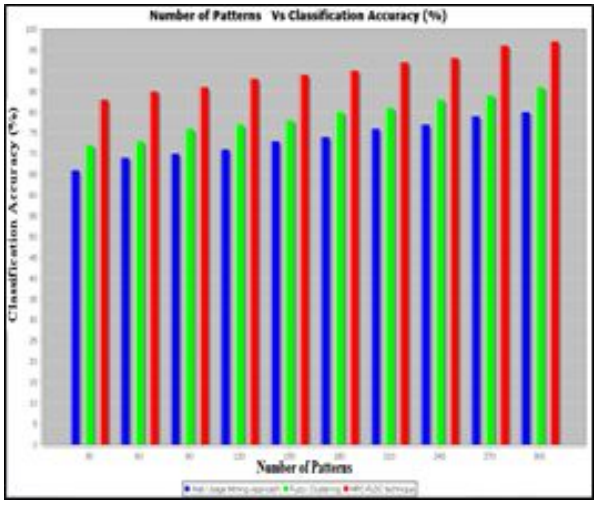


Figure 6 Measurement of Classification Accuracy versus Different Number of Patterns

Figure 6 portrays the impact of classification accuracy for discovering frequent web patterns visited by web users versus diverse number of patterns in the range of 30-300 using three methods. As shown in figure, the proposed MPC-FLDC technique provides better classification accuracy for finding frequent web patterns from web log database when compared to web usage mining approach [1] and fuzzy clustering [2]. Besides, while increasing the number of web patterns, the classification accuracy is also gets increased for all three methods. But, comparatively the classification accuracy using proposed MPC-FLDC technique is higher. This is due to application of FLD Classifier in MPC-FLDC technique. FLD Classifier designed in MPC-FLDC technique determines optimal projection direction to efficiently categorize the web pages as frequent or non-frequent patterns based on a session Si. This in turn assists for MPC-FLDC technique to enhance the classification accuracy in a significant manner. As a result, proposed MPC-FLDC technique improves the classification accuracy by 22 % and 14 % when compared to web usage mining approach [1] and fuzzy clustering [2] respectively.

4.2 Measure of Traffic Patterns Prediction Rate

In MPC-FLDC technique, Traffic Patterns Prediction Rate (TPPR) is defined as the ratio of the number of web patterns that are predicted as traffic patterns to the total number of web patterns. The traffic patterns prediction rate is measured in terms of percentage (%) and mathematically expressed as,

$$TPPR = \frac{number\ of\ web\ patterns\ that\ are\ predicted\ as\ traffic\ pattern}{total\ number\ of\ web\ patterns} \times 100 \quad (12)$$

From equation (12), the prediction rate for web traffic patterns is estimated with respect to dissimilar number of web patterns. While traffic patterns prediction rate is higher, the method is said to be more effectual.

Table 2 Tabulation for Traffic Patterns Prediction Rate

Traffic Patterns Prediction Rate (%)		
Web Usage Mining Approach	**Fuzzy Clustering**	**MPC-FLDC technique**
62	70	86
63	71	88
65	74	89
66	75	90
67	77	92
70	78	93
71	80	94
72	81	95
75	82	97
76	84	98

Table 2 demonstrates the results of traffic patterns prediction rate for mining web user's behaviours with respect to diverse number of web patterns in the range of 30-300 using three methods. While considering the 210 web patterns for performing experimental evaluation, the proposed MPC-FLDC technique attains 94 % traffic patterns prediction rate whereas web usage mining approach [1] and fuzzy clustering [2] acquires 71 % and 80 % respectively. From that, it is expressive that the traffic patterns prediction rate using proposed MPC-FLDC technique is higher as compared to other existing [1], [2].

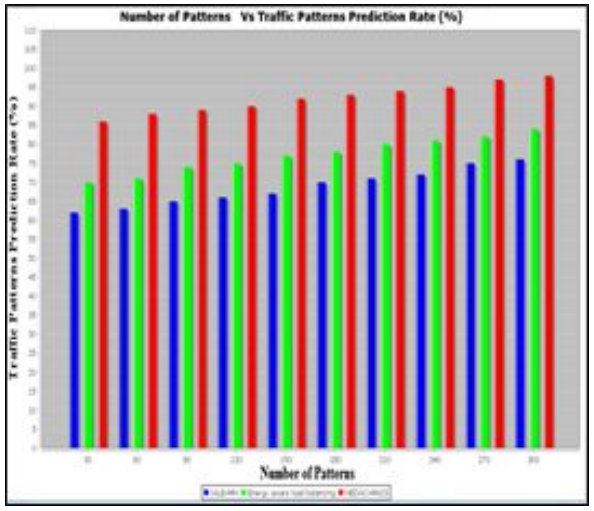


Figure 7 Measurement of Traffic Patterns Prediction Rate versus Different Number of Patterns

Figure 7 explains the impact of traffic patterns prediction rate versus different number of patterns in the range of 30-300 using three methods. As exposed in figure, the proposed MPC-FLDC technique provides better web traffic patterns prediction rate for analyzing web user's behaviours when compared to web usage mining approach [1] and fuzzy clustering [2]. In addition, while increasing the number of web pattern, traffic patterns prediction rate is also gets increased for all three methods. But, comparatively the traffic patterns prediction rate using proposed MPC-FLDC technique is higher. This is owing to application of Pearson Correlation Analysis in MPC-FLDC technique. Pearson correlation analysis used in MPC-FLDC technique determines the correlation between the web pages among different user sessions in order to efficiently predict the web traffic patterns with higher accuracy. This process helps for MPC-FLDC technique to improve the traffic patterns prediction rate in an effectual manner. Hence, proposed MPC-FLDC technique increase the traffic patterns prediction rate by 34 % and 20 % when compared to web usage mining approach [1] and fuzzy clustering [2] respectively.

4.3 Measure of Prediction Time

In MPC-FLDC technique, Prediction Time *(PT)* measures the amount of time required for finding the web traffic patterns in a weblog database. The prediction time is measured in terms of milliseconds (ms) and mathematically represented as,

PT=N time(predicting one web traffic pattern) (13)*

From equation (13), the time for predicting web traffic patterns is evaluated with respect to diverse number of web pages. While prediction time is lower, the method is said to be more effective.

Table 3 Tabulation for Prediction Time

Number of Patterns	Prediction Time (ms)		
	Web Usage Mining Approach	**Fuzzy Clustering**	**MPC-FLDC technique**
30	21.9	18.1	13.5
60	26.1	25.6	16.2
90	31.2	29.8	22.6
120	36.7	34.2	29.7
150	42.5	39.7	35.2
180	48.6	45.3	40.7
210	54.8	51.2	44.9
240	60.3	58.9	49.8

270	68.7	65.1	55.3
300	75.9	71.8	62.5

The performance analysis of prediction time for mining web traffic patterns based on various number of web patterns in the range of 30-300 using three methods is presented in Table 3. While considering the 240 web patterns for carried outing experimental process, the proposed MPC-FLDC technique takes 49.8 ms time for predicting web traffic patterns whereas web usage mining approach [1] and fuzzy clustering [2] takes 60.3 ms and 58.9 ms respectively. Therefore, the prediction time using proposed MPC-FLDC technique is lower as compared to other existing [1], [2].

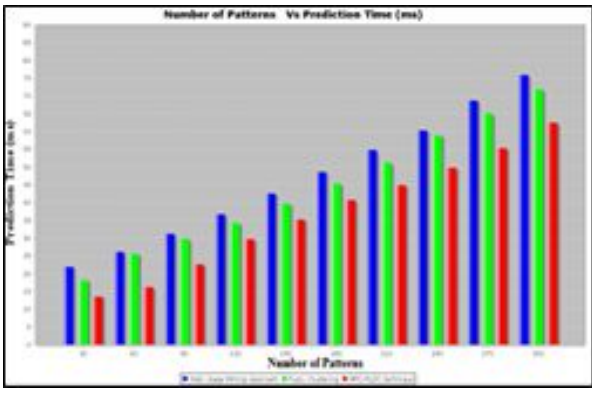


Figure 8 Measurement of Prediction Time versus Different Number of Patterns

Figure 8 shows the impact of prediction time versus dissimilar number of patterns in the range of 30-300 using three methods. As demonstrated in figure, the proposed MPC-FLDC technique provides better prediction time for mining web traffic patterns when compared to web usage mining approach [1] and fuzzy clustering [2]. Further, while increasing the number of web pattern, prediction time is also gets increased for all three methods. But, comparatively the prediction time using proposed MPC-FLDC technique is lower. This is because of application of Pearson Correlation Analysis in MPC-FLDC technique where it finds the relationship between the web pages among diverse user sessions to significantly predict the web traffic patterns with minimum time. This process assists for MPC-FLDC technique to reduce the prediction time in an effective manner. As a result, proposed MPC-FLDC technique minimizes the prediction time by 23 % and 18 % when compared to web usage mining approach [1] and fuzzy clustering [2] respectively.

4.4 Measure of True Positive Rate

In MPC-FLDC technique, True Positive Rate *(TPR)* determines the ratio of the number of web patterns correctly predicted as traffic patterns to the total number of web patterns. The true positive rate is measured in terms of percentage (%) and mathematically formulated as,

$$TPR = \frac{number\ of\ web\ patterns\ \ correctly\ predicted\ as\ traffic\ patterns}{total\ number\ of\ web\ patterns} * 100 \qquad (14)$$

From equation (14), the true positive rate for web traffic patterns mining is evaluated with respect to various number of web patterns. While true positive rate for web traffic patterns mining is higher, the method is said to be more efficient.

Table 4 Tabulation for True Positive Rate

Number of Patterns	True Positive Rate (%)		
	Web Usage Mining Approach	Fuzzy Clustering	MPC-FLDC technique
30	56	63	80
60	58	66	81
90	59	67	83
120	62	69	84
150	63	70	87
180	65	71	88
210	66	73	90
240	69	77	91
270	70	78	92
300	72	79	94

The tabulation result of true positive rate for predicting web traffic patterns with respect to diverse number of web patterns in the range of 30-300 using three methods is illustrated in Table 4. While considering the 270 web patterns for conducting experimental process, the proposed MPC-FLDC technique obtains 92 % true positive rate for predicting web traffic patterns whereas web usage mining approach [1] and fuzzy clustering [2] attains 70 % and 78 % respectively. As a result, the true positive rate using proposed MPC-FLDC technique is higher as compared to other existing [1], [2].

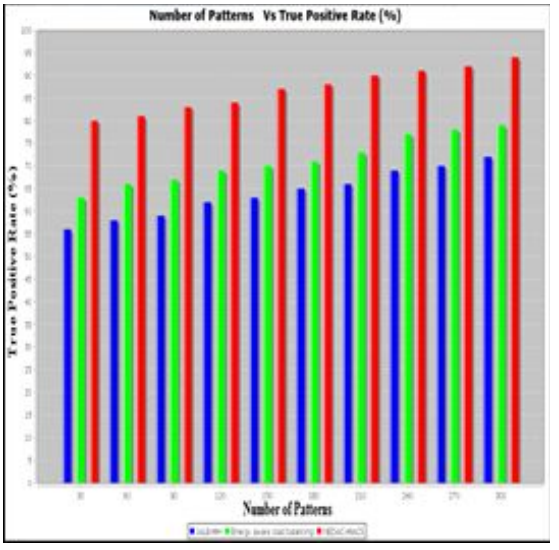


Figure 9 Measurement of True Positive Rate versus Different Number of Patterns

Figure 9 describes the impact of true positive rate versus diverse number of patterns in the range of 30-300 using three methods. As illustrated in figure, the proposed MPC-FLDC technique provides better true positive rate for predicting web traffic patterns when compared to web usage mining approach [1] and fuzzy clustering [2]. Moreover, while increasing the number of web pattern, true positive rate is also gets increased for all three methods. But, comparatively the true positive rate using proposed MPC-FLDC technique is higher. This is due to application of Pearson Correlation Analysis based Web Traffic Pattern Mining algorithm in MPC-FLDC technique. With aid of Pearson Correlation Analysis, MPC-FLDC technique effectually finds the web traffic patterns among the different user sessions for analyzing the web user behaviours to enhance the quality of Web information service performances. This in turn assists for MPC-FLDC technique to achieve higher true positive rate in an efficient manner. Thus, proposed MPC-FLDC technique increases the true positive rate by 36 % and 22 % when compared to web usage mining approach [1] and fuzzy clustering [2] respectively.

5. Related works

A linear-temporal logic model checking approach was intended in [11] for analysis of users' behaviors. However, traffic pattern prediction was not considered. An Efficient Hybrid Successive Markov Prediction Model was presented in [12] to identify web user usage behavior with higher accuracy. But, true positive rate of prediction was lower.

A novel algorithm for web personalization was designed in [13] by combining the web user profiles and behavioral patterns. However, time for predicting behavioral patterns was more. Hybrid Data Aggregation Technique was developed in [14] to classify the web users to discover knowledge about the web users with reduced time complexity. However, classification accuracy was poor.

A novel method was intended in [15] for analyzing user behavior pattern through evaluating web users in websites. But, prediction

performance of user behavior pattern was not effectual. A k-means clustering algorithm was presented in [16] to classify the daily traffic patterns and to predict the current daily traffic patterns. However, prediction accuracy of web traffic patterns was lower.

Apriori prefix tree (PT) algorithm was designed in [17] for predicting and mining the patterns of user's visit web pages. But, the efficiency of mining was not efficient. A four-gram unified event model was intended in [18] to present high quality data sources for web mining algorithms and increase the quality of intelligent services.

A survey of different techniques developed for web mining was presented in [19] to effectively predict the web user behaviors using classification, clustering, statistical analysis and association rule. A web classification algorithm was designed in [20] by using fuzzy association rule mining to find out the correlation among the web pages visited by users. However, accuracy of finding frequent web pages visited by web users was not at required level.

6.Conclusion

An effective MapReduce Pearson Correlation Fisher's Linear Discriminant Classifier (MPC-FLDC) technique is designed in order to efficiently mine web traffic patterns from weblogs with higher true positive rate and minimum time. The MPC-FLDC technique includes of three processes such as pre-processing, classification and traffic web patterns prediction. During the pre-processing, MPC-FLDC technique groups the web patterns in a weblog database based on diverse user sessions. Afterward, MPC-FLDC technique categorizes the web patterns at a different user sessions as frequent or non-frequent based on hit ratio with application of (FLD) Classifier which resulting in enhanced classification accuracy of web patterns. At last, MPC-FLDC techniques efficiently predict the traffic web patterns with aid of Pearson correlation analysis. This in turn supports for MPC-FLDC technique to achieve higher the traffic patterns prediction rate with minimum prediction time. The efficacy of MPC-FLDC technique is test with the parameters such as classification accuracy, traffic patterns prediction rate, prediction time and true positive rate. With the experiments conducted for MPC-FLDC technique, it is illustrative that the true positive rate presents more precise results for mining the web traffic patterns as compared to state-of-the-art works. The experimental result demonstrates that MPC-FLDC technique is provides better performance with an improvement of true positive rate and the reduction of prediction time when compared to the state-of-the-art works.

REFERENCES

[1] 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, Pages 50–62, 2013

[2] D. Anandhi, M. S. Irfan Ahmed, "Prediction of user's type and navigation pattern using clustering and classification algorithms", Cluster Computing, Springer, Pages 1–10, 2017

[3] Vagner Figueredo deSantana, Maria Cecília CalaniBaranauskas, "WELFIT: A remote evaluation tool for identifying Web usage patterns through client-side logging", International Journal of Human-Computer Studies, Elsevier, Volume 76, Pages 40-49, April 2015

[4] Yu-ShiangHung, Kuei-Ling B.Chen,Chi-TaYang, Guang-FengDeng, "Web usage mining for analyzing elder self-care behavior patterns", Expert Systems with Applications, Elsevier, Volume 40, Issue 2, Pages 775-783, February 2013

[5] OritRaphaeli, AnatGoldstein, LiorFink, "Analyzing online consumer behavior in mobile and PC devices: A novel web usage mining approach", Electronic Commerce Research and Applications, Elsevier, Volume 26, Pages 1-12, November–December 2017

[6] Ahmad Tasnim Siddiqui, Sultan Aljahdali, "Web Mining Techniques in E-Commerce Applications", International Journal of Computer Applications, Volume 69, Issue 8, Pages 39-43, May 2013

[7] Mohammed Asad, Girish P. Potdar, "A Survey on Different Clustering Techniques for Web Usage Mining", International Journal of Computer Science and Information Technology & Security, Volume 6, Issue 2, Pages 200-204, 2016

[8] D.A.Adeniyi, Z.Wei, Y.Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method", Applied Computing and Informatics, Applied Computing and Informatics, Elsevier, Volume 12, Issue 1, January 2016, Pages 90-108

[9] Maryam Jafari, Farzad SoleymaniSabzchi, Shahram Jamali, "Extracting Users'Navigational Behavior from Web Log Data: a Survey", Journal of Computer Sciences and Applications, Volume 1, Issue 3, Pages 39-45, 2013

[10] 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, Pages 8797-8805, 2016

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

[12] V.V.R.Maheswara Rao, Valli Kumari, "An Efficient Hybrid Successive Markov Model for Predicting Web User Usage Behavior using Web Usage Mining", International Journal of Data Engineering, Volume 1, Issue 5, Pages 43-62, 2011

[13] Doddegowda B J, Sunil Kumar S Manvi, G T Raju, "A Novel Algorithm for Web Personalization through Integration of Web User Profiles and Behavioral Patterns", International Journal of Computer Science and Information Technology & Security, Volume 7, Issue 2, Pages 4-13, Mar-April 2017

[14] E. Manohar, D. Shalini Punithavathani, "Hybrid Data Aggregation Technique to Categorize the Web Users to Discover Knowledge about the Web Users", Wireless Personal Communications, Springer, Pages 1–15, 2017

[15] Shilpa Mahajan, Shilpa Yadav, "Analyzing HTTP Traffic Patterns for Monitoring and Analyzing User Behavior", Indian Journal of Science and Technology, Volume 9, Issue 48, Pages 1-7, December 2016

[16] Ricardo Garcia-R ´odenas, Maria L. Lopez-Garcia & Maria Teresa Sanchez-Rico, "An Approach to Dynamical Classification of Daily Traffic Patterns", Computer-Aided Civil and Infrastructure Engineering, Volume 32, Issue 3, Pages 191–212, March 2017

[17] R Geetharamani, P Revathy And Shomona G Jacob, "Prediction of users webpage access behaviour using association rule mining", Indian Academy of Sciences, Springer, Volume 40, Issue 8, Pages 2353–2365, December 2015

[18] Xinyao Zou, "A four-gram unified event model for web mining", Cluster Computing, Springer, Pages 1–9, 2017

[19] Tawfiq A. Al-asadi, Ahmed J. Obaid, Rahmat Hidayat, Azizul Azhar Ramli, "A Survey on Web Mining: Techniques and Applications", International journal on advanced science engineering information technology, Volume 7, Issue 4, Pages 1178-1184, 2017

[20] Binu Thomas and G. Raju, "A Novel Web Classification Algorithm Using Fuzzy Weighted Association Rules", Hindawi Publishing Corporation, ISRN Artificial Intelligence, Volume 2013, Article ID 316913, Pages 1-10, 2013

[21] Apache log samples dataset: http://www.monitorware.com/en/logsamples/apache.php