Original Research Paper

Volume-8 | Issue-8 | August-2018 | PRINT ISSN No 2249-555X

Computer Science

POISSON FREQUENT PATTERNS CLUSTERING AND TEMPORAL SIMILARITY TRAFFIC MINING FOR WEB USER TRACKING

Ulaganathan.n	HOD,Assistant Professor, Department of Computer Science, NIFT-TEA College of Knitwear Fashion, East of TEKIC,SIDCO,Mudalipalayam,Tirupur,Affiliated to Bharathiar University, Coimbatore-641 046,Tamilnadu,India
Prasath.s*	Assistant Professor, Department Of Computer Science, Nandha Arts And Science College,erode-638 052, Affiliated To Bharathiar University, Coimbatore-641 046,tamilnadu,india *corresponding Author

ABSTRACT Web traffic pattern analysis is significant to find the web user behaviors. Few research works has been developed for web traffic pattern mining and web user tracking. However, computational complexity taken for tracking user location of web traffic patterns was higher. In order to overcome such limitation, Poisson Fragment Frequency based Web Pattern Clustering (PFF-WPC) technique is proposed. The PFF-WPC technique is designed with objective of reducing the computational complexity of identifying the location of user interest web pages from a weblog database. Initially PFF-WPC technique performs Poisson Fragment Process with aiming at grouping the web pages in a weblog database according to number of sessions. After session identification, PFF-WPC technique performs Frequency based web patterns clustering with objective of grouping the web pages in an each session as frequent or non frequent web pages with higher clustering accuracy. Finally, PFF-WPC technique carried outs the Temporal Similarity Based Web User Tracking process in which traffic web patterns are detected based on the measurement of temporal similarity among the sessions. Based on identified traffic web patterns, the location of corresponding web users is tracked with help of public IP address stored in weblog database with minimum computational complexity. The PFF-WPC technique conducts the experimental evaluations on factors such as clustering efficiency, computational complexity, true positive rate and space complexity. The experimental result reveals that PFF-WPC technique is able to improve the true positive rate and also reduces computational complexity of web user tracking when compared to state-of-the-art-works.

KEYWORDS : Clustering, Poisson Fragment Process, Sessions, Temporal Similarity, Tracking Web Traffic Pattern

INTRODUCTION

With the development of web technology application, user activity plays an important role on the operation of websites. One of the main tasks in web log mining is to discover the hidden information, for example user browsing behaviour and provide it to the website managers for improving content arrangement. The information regarding web users' behaviour is stored in the weblogs. The analysis of such information using data mining techniques helps to identify web users' behaviour. In addition to that, web user tracking plays a significant role in order to identify the location of traffic web pages and analyzing of visitor behaviour on a website. Therefore, this research work focus on the tracking web user location of traffic web patterns.

A web usage mining approach was designed in [1] with objective of identifying online navigational behavior. However, computational complexity involved in web user behaviour analysis was more. A linear-temporal logic (LTL) model checking approach was developed in [2] in order to analysis the users' behaviors in e-commerce websites. But, traffic pattern mining and tracking was remained unaddressed.

The impact of visitors on page views with google analytics was presented in [3]. But, tracking web user location was not solved. Automated Log Analyzer Tool was used in [4] in order to identify user's behavior from web access logs. However, space complexity taken for behaviour analysis was not considered.

A review of different techniques developed for prediction of user behavior based on web server log files in web usage mining was presented in [5]. Behavior-based tracking technique was designed in [6] with application of Naive Bayes classifier with aiming at tracking web user activities. However, the amount of time needed for tracking web user activities was very higher.

An alternative approach was presented in [7] in order to mine information from a given e-commerce website and predicting users' future behaviour. But, true positive rate was not considered. Fine-Grained HTTP Web Traffic Analysis was intended in [8] to analyze subscriber behaviors with minimum time. However, performance of web traffic analysis was not enough.

Task-level search log analysis was introduced in [9] for discovering user search interests with higher prediction accuracy. But, identification of web traffic patterns was remained open issue. A modified Markov model was developed in [10] to identify the next set of web pages that a user may visit depends on the information of the previously browsed web pages with lower time complexity. However, finding location of user interest web pages was remained unsolved.

In order to solve the above mentioned existing issues, Poisson Fragment Frequency based Web Pattern Clustering (PFF-WPC) technique is introduced. The contribution of proposed PFF-WPC framework is formulated as follows.

To improve the performance of web user tracking, Poisson Fragment Frequency based Web Pattern Clustering (PFF-WPC) technique is developed. The PFF-WPC technique is designed with helps of Poisson Fragment Process, Frequency Based Web Patterns Clustering and Temporal Similarity Measurement.

To efficiently perform session identification process from a weblog database, Poisson Fragment Process is used in PFF-WPC technique. During the session identification process, web pages in a weblog database are grouped into a number of sessions based on access time.

To improve the clustering efficiency of web patterns, Frequency Based Web Patterns Clustering algorithm is designed in PFF-WPC technique. During the clustering process, web pages in each session are grouped as frequent or non frequent based on frequency measurement of each web page.

To reduce the computational and space computation involved during the web user tracking process, Temporal Similarity Measurement is employed in PFF-WPC technique. With the help of computed temporal similarity of each session, PFF-WPC technique predicts the traffic web patterns and locations of corresponding web users with aid of their public IP address.

The rest of paper structure is formulated as follows: In Section 2, Poisson Fragment Frequency based Web Pattern Clustering (PFF-WPC) technique is explained with assists of architecture diagram. In Section 3, Simulation settings are described and the result discussion is presented in Section 4. Section 5 reviews the related works. Section 6 presents the conclusion of the paper.

2.POISSON FRAGMENT FREQUENCY BASED WEB PATTERN CLUSTERING TECHNIQUE

The Poisson Fragment Frequency based Web Pattern Clustering (PFF-WPC) technique is designed with objective of identifying the user locations of web traffic patterns with minimum computational complexity. The process of PFF-WPC technique includes of three processes such as session identification, clustering and location tracking. The PFF-WPC technique initially performs session identification with application of Poisson fragment process in which the web pages in weblog database is grouped based on different sessions. After completing the session identification process, PFF-WPC technique clusters the web pages in each session as frequent or non-frequent with help of frequency based web pattern clustering algorithm. Finally, PFF-WPC technique carried out location tracking process with assists of temporal similarity where the user location of web traffic patterns is identified. In PFF-WPC technique, temporal similarity is determined for each session of frequent web patterns to detect traffic web patterns in a weblog database. For a detected traffic web patterns, a locations of corresponding web user is identified with assist of public IP address stored in weblog database. The overall architecture diagram of Poisson Fragment Frequency based Web Pattern

Clustering (PFF-WPC) technique for tracking web user's location of web traffic pattern is shown in below,

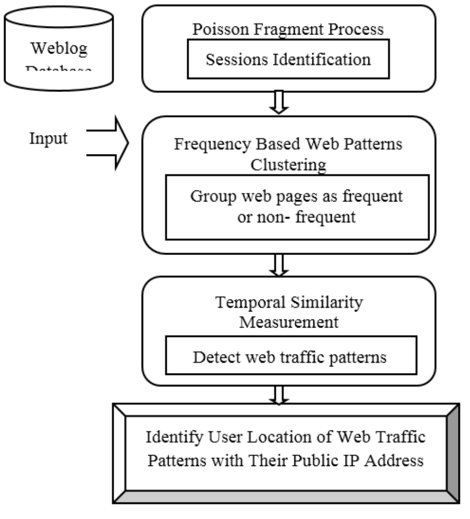

Figure 1 architecture diagram of Poisson Fragment Frequency based Web Pattern Clustering (PFF-WPC) technique for Web User Tracking

Figure 1 shows the process of PFF-WPC technique to detect the user location of web traffic patterns for behaviour analysis of web users. As depicted in Figure 1, PFF-WPC technique initially takes weblog database (i.e. Apache log samples dataset) as input. Next, PFF-WPC technique carried outs sessions identification. Subsequently, PFF-WPC technique employs frequency based web patterns clustering with aiming at clustering the web pages as frequent or non-frequent with higher clustering efficiency. Then, PFF-WPC technique used temporal similarity in order to detect the web traffic patterns. Finally, user locations of web traffic Patterns are efficiently tracked with assist of their public IP address. As a result, PFF-WPC technique improves the clustering efficiency and true positive rate of web traffic mining with lower computational complexity. The detailed explanation about PFF-WPC technique is discussed in forthcoming sub sections.

2.1 Poisson Fragment Process

The PFF-WPC technique used Poisson Fragment Process in order to break the weblog database into sessions. A sequence of web pages viewed by web users during the particular time interval is known as the Session. A Poisson Fragment Process applied in PFF-WPC technique divides the weblog database into a number of sessions. The sessions are generated based on the access time (time field) of web pages stored in weblog database. The Poisson Fragment Process measures the arrival of web user's activities with each session corresponding to a unique user to group the web pages. This helps for PFF-WPC technique to finds out the number of web pages visited by web users during every session. The following diagram shows the process involved in Poisson Fragment Process to identify the sessions.

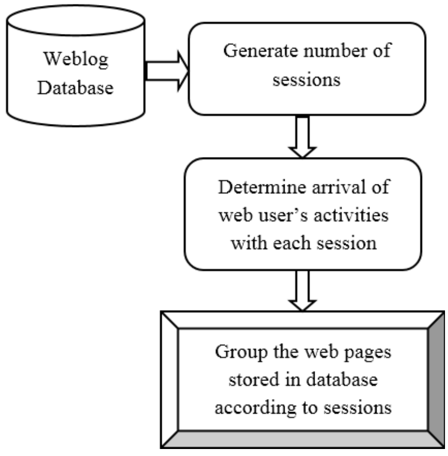

Figure 2 Poisson Fragment Process for Sessions Identification

Figure 2 demonstrates the Poisson Fragment Process for session identification in web user tracking. As depicted in Figure 2, at first Poisson Fragment Process generates the diverse sessions based on access time of web users stored in weblog database. After that, Poisson Fragment Process evaluates the arrival of web user's activities with each session in order to group the web pages stored in database based on generated sessions. The grouping of web pages according to sessions helps for PFF-WPC technique to effectively perform the web user tracking.

Let us consider the arrival of web user activities (i.e. web pages accessed by web users) at random rate λ per unit time. At time t=0, we have no arrivals yet. Therefore, N(0) is considered as 0. Weblog database stored the numerous numbers of web pages visited by web users at different time. Therefore, weblog database is splits into a number of sessions S_ibased on time field as shown in below Figure 3.

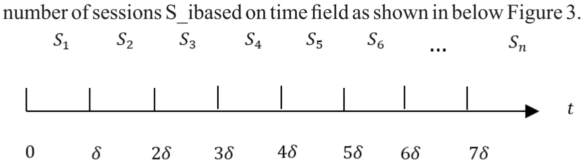

Figure 3 Sessions Identification

As shown in Figure 3, each session corresponds to a time slot of length δ. Thus, weblog database is separated into a number of sessions. In order to group the web pages stored in weblog database based on sessions, consider $N(S_i)$ is the number of arrivals of web user's activities from time 0 to time t. There are $n=\frac{t}{\delta}$ time slots in the interval (0 $_i$). Thus, $N(S_i)$ denotes that the number of web pages browsed by web users during session S_i. The Poisson fragment process for finding web user activities with rate λ>0 according to each session has the following three properties,

- $N(0)=0$
- Poisson fragment process has independent increments.
- The number of arrivals of web user's activities in any interval of length
- The number of arrivals of web user's activities in any interval of length t is a Poisson random variable with parameter λt.

From these properties, the Poisson Fragment Process for expected arrivals of web user's activities is mathematically expressed as,

$E[N(S_i)]=\lambda t$

The Poisson fragment process is interrelated to the Poisson distribution which implies that the probability of a web user activity being equal to n. Thus, the arrivals of web user's activities $N(S_i)$ at the session S is obtained using below mathematical representation,

$N(S_i) = \frac{(\lambda t^n)}{n!} e^{-\lambda t}$

From equation (2), n! indicates n factorial and λ is the single Poisson parameter that is used to define the Poisson distribution. By using equation (2), the number of web pages accessed by web users during each session S_i is obtained. Accordingly, weblog database is splits into a number of sessions S_i as follows,

$$N(S_1) = \{WP_1, WP_3, WP_5, WP_7\} \quad (3)$$

$$N(S_2) = \{WP_2, WP_4, WP_6, WP_8\} \quad (4)$$

$$N(S_3) = \{WP_5, WP_1, WP_3, WP_8\} \quad (5)$$

....

$$N(S_n) = \{WP_1, WP_2, WP_3, WP_7\}$$

From equations (3), (4), (5), (6), PFF-WPC technique obtains the sequence of web pages viewed by web users according to different sessions for efficient web user tracking. The algorithmic process of Poisson Fragment Process for session's identification is shown in below,

//**Poisson Fragment Process based**
Input: Weblog Database
Output: Sessions Identification
Step 1: Begin
Step 2: Generate the number of sessions S_i based on access time
Step 3: For each Session S_i
Step 4: Compute arrival of web user's activities with each session using (2)
Step 5: Group the web pages stored in database according to sessions S_i
Step 6: End for
Step 7: End

Algorithm 1 Poisson Fragment Process Based Session Identification

Algorithm 1 illustrates the step by step process of Poisson Fragment Process to find out the user sessions for web user tracking. With the help of above algorithmic process, PFF-WPC technique effectively group the web pages accessed by web users during each sessions for efficient web user tracking.

2.2 Frequency Based Web Patterns Clustering

The PFF-WPC technique used Frequency Based Web Patterns Clustering technique in order to cluster the web pages in each session as frequent or non-frequent web pages with higher accuracy. Web users spend different amount of time in different web pages. Identifying the web page frequently visited by web users in each session help for PFF-WPC technique to predict user behaviour about the higher traffic web patterns. The web pattern clustering in PFF-WPC technique based on frequency measurement. The frequency in PFF-WPC technique is determined as the ratio of the number of occurrences of particular web page to the total number of web pages in weblog database. With assists of measured frequency, then Frequency Based Web Patterns Clustering technique groups web pages in each session as frequent or non-frequent web pages. The process involved in Frequency Based Web Patterns Clustering technique is depicted in following Figure 4.

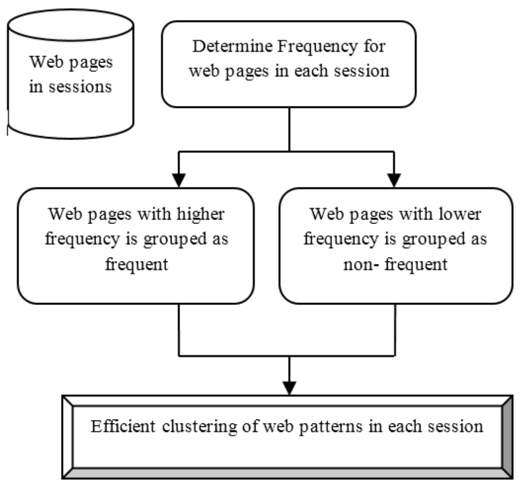

Figure 4 Frequency Based Web Patterns Clustering Process

Figure 4 demonstrates the block diagram of frequency based web patterns clustering process for web user tracking. As shown in figure, initially frequency is computed for each web page in weblog database. Then, web pages with higher frequency are clustered as frequent whereas pages with lower frequency are clustered as non frequent web pages in each session. This process helps for PFF-WPC technique to grouping the web patterns in weblog database as frequent or non frequent web pages with higher clustering efficiency.

Let us consider the weblog database includes of different web pages visited by web users like $\{WP_1, WP_2, .. WP_N\}$. The frequency of each web page in session S_i is evaluated as, $F_{WP_i} = \frac{O_{WP_i}}{N}$ From equation (7), $O_i(WP_i)$ denotes number of occurrences of particular web page 〖 WP_i whereas N indicates the total number of web pages. By using equation (7), the frequency is determined for each web page in weblog database. With aid of frequency measurement, the web pages in each session are clustered as frequent or non frequent web pages with higher accuracy. The algorithmic process of Frequency Based Web Patterns Clustering is shown in below,

// Frequency Based Web Patterns Clustering Algorithm
Input: web pages in different session
Output: Improved clustering efficiency
Step 1: Begin
Step 2: For each session
Step 3: For each web pages in session
Step 4: Compute the frequency of web page using (7)
Step 5: Web pages which have the higher frequency are grouped as frequent
Step 6: Web pages which have the lower frequency are grouped as non-frequent
Step 7: End for
Step 8: End for
Step 9: End for

Algorithm 2 Frequency Based Web Patterns Clustering

Algorithm 2 explains the step by step process involved in Frequency Based Web Patterns Clustering. With the assists of above algorithmic process, PFF-WPC technique clusters the web pages in each session as frequent or non frequent web pages. Hence, PFF-WPC technique significantly improves the clustering efficiency of web pages.

2.3 Temporal Similarity Based Web User Tracking

Temporal Similarity is used in proposed PFF-WPC technique in order to find out traffic web patterns. Based on identified traffic web patterns, a location of corresponding web user is identified. Here, traffic indicates the many users that visit a particular web page more number of times at a specific instance. Temporal similarity is determined for each session of frequent web pages in order to identify traffic web patterns with their location. The following diagram demonstrates the process involved in Temporal Similarity Based Web User Tracking.

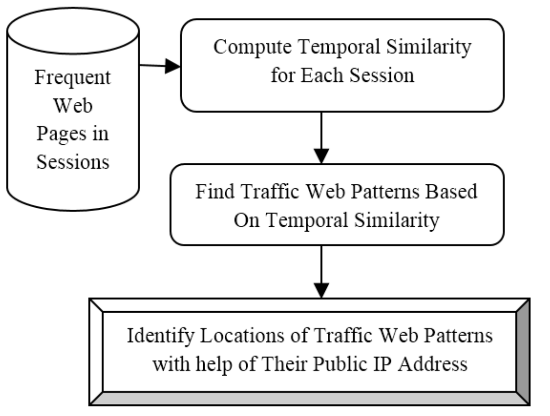

Figure 5 Process of Temporal Similarity Based Web User Tracking

Figure 5 depicts the process of temporal similarity based web user tracking. As demonstrated in figure, temporal similarity based web user tracking technique at first computes the temporal similarity for each sessions. Then estimated temporal similarity of each session is compared with all sessions to discover the web traffic patterns in weblog database. At last, user locations of these web traffic patterns is tracked with aid of their public IP address in weblog.

The temporal similarity among the sessions is determined by its frequency measurement which is defined as $F = F_{t1}, F_{t2}, \ldots F_{tn}$). Here F_{ti}) is represents frequency of web pages for time unit t_i over the whole series of T. Temporal Similarity evaluates a temporally grounded variation that identifies frequency of web patterns between two sessions. Given two session S_i and S_j and their frequency function F_i and F_j, the temporal similarity TS_{SiSj} is mathematically determined as,

$$TS_{S_iS_j} = \frac{1}{n}\sum_i \left(\frac{(ft_i - \mu(F_i)}{\sigma(F_i)}\right)\left(\frac{(ft_j - \mu(F_j)}{\sigma(F_j)}\right)$$

From equation (8), $\mu(F_i)$, $\mu(F_j)$ represents the mean frequencies whereas $\sigma(F_i)$ and $\sigma(F_j)$ indicates the standard deviations between the traffic patterns in two sessions whereas n refers the number of sessions. The temporal similarity of two sessions is a standard measure of how strongly frequents web patterns in these sessions are linearly related. The temporal similarity value always lies between −1 and +1. The result of temporal similarity value + 1 point outs that an exact positive linear relationship between them whereas −1 denotes an exact negative linear relationship. By using the equation (8), the temporal similarity is computed for each session of frequent web patterns. In order to identify the traffic web patterns, the temporal similarity value of every session is compared with all the sessions of frequent patterns. The sessions of frequent web pages with higher temporal similarity is predicted as traffic web patterns among the other sessions. Finally, user location of these traffic web patterns is effectively tracked with assists of their public IP address in weblog database. The algorithmic process of Temporal Similarity for is shown in below,

// Frequency Based Web Patterns Clustering Algorithm
Input: web pages in different session
Output: Improved clustering efficiency
Step 1: Begin
Step 2: For each session
Step 3: For each web pages in session
Step 4: Compute the frequency of web page using (7)
Step 5: Web pages which have the higher frequency are grouped as frequent
Step 6: Web pages which have the lower frequency are grouped as non-frequent
Step 7: End for
Step 8: End for
Step 9: End for

2.3 Temporal Similarity Based Web User Tracking

Temporal Similarity is used in proposed PFF-WPC technique in order to find out traffic web patterns. Based on identified traffic web patterns, a location of corresponding web user is identified. Here, traffic indicates the many users that visit a particular web page more number of times at a specific instance. Temporal similarity is determined for each session of frequent web pages in order to identify traffic web patterns with their location. The following diagram demonstrates the process involved in Temporal Similarity Based Web User Tracking.

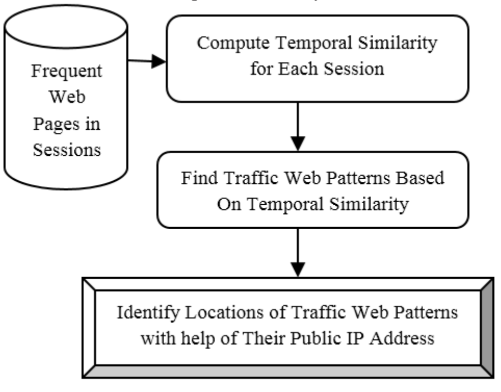

Figure 5 Process of Temporal Similarity Based Web User Tracking

Figure 5 depicts the process of temporal similarity based web user tracking. As demonstrated in figure, temporal similarity based web user tracking technique at first computes the temporal similarity for each sessions. Then estimated temporal similarity of each session is compared with all sessions to discover the web traffic patterns in weblog database. At last, user locations of these web traffic patterns is tracked with aid of their public IP address in weblog.

The temporal similarity among the sessions is determined by its frequency measurement which is defined as $F = F_{t1}, F_{t2}, \ldots F_{tn}$). Here F_{ti} is represents frequency of web pages for time unit t_i over the whole series of T. Temporal Similarity evaluates a temporally grounded variation that identifies frequency of web patterns between two sessions. Given two session S_i and S_j and their frequency function F_i and F_j, the temporal similarity TS_{SiSj} is mathematically determined as,

$$TS_{S_iS_j} = \frac{1}{n}\sum_i \left(\frac{(ft_i - \mu(F_i)}{\sigma(F_i)}\right)\left(\frac{(ft_j - \mu(F_j)}{\sigma(F_j)}\right)$$

From equation (8), $\mu(F_i)$, $\mu(F_j)$ represents the mean frequencies whereas $\sigma(F_i)$ and $\sigma(F_j)$ indicates the standard deviations between the traffic patterns in two sessions whereas n refers the number of sessions. The temporal similarity of two sessions is a standard measure of how strongly frequents web patterns in these sessions are linearly related. The temporal similarity value always lies between −1 and +1. The result of temporal similarity value + 1 point outs that an exact positive linear relationship between them whereas −1 denotes an exact negative linear relationship. By using the equation (8), the temporal similarity is computed for each session of frequent web patterns. In order to identify the traffic web patterns, the temporal similarity value of every session is compared with all the sessions of frequent patterns. The sessions of frequent web pages with higher temporal similarity is predicted as traffic web patterns among the other sessions. Finally, user location of these traffic web patterns is effectively tracked with assists of their public IP address in weblog database. The algorithmic process of Temporal Similarity for is shown in below,

// Frequency Temporal Similarity Based Web User Tracking
Input: Frequent Web Pages in Sessions
Output: Reduced Computational Complexity
Step 1: Begin
Step 2: For each sessions of frequent web patterns
Step 3: Evaluate the temporal similarity using (8)
Step 4: If (" value of session is lies between 0 to +1) then
Step 5: traffic web patterns is identified
Step 6: else If (" value is lies between -1.00 to 0)
Step 7: non traffic web patterns is identified
Step 8: End if
Step 9: Identify web user locations of traffic patterns with their public IP address
Step 10: End for
Step 11: End for

Algorithm 3 Temporal Similarity Based Web User Tracking

Algorithm 3 shows the step by step process of Temporal Similarity Based Web User Tracking. As shown in algorithm 3, Temporal Similarity Based Web User Tracking algorithm initially estimates the temporal similarity for frequent web patterns in each session. Afterward, Temporal Similarity Based Web User Tracking algorithm verifies if temporal similarity value is lies between the 0 to +1 which process helps to predict the traffic web patterns in weblog database. Finally, the location of these traffic web patterns are efficiently detected with assists of their public IP address stored in weblog database with minimum time consumption. Therefore, PFF-WPC technique enhances the true positive rate of web traffic pattern analysis and also reduces the space and computational complexity of web user tracking in an effectual manner.

EXPERIMENTAL SETTINGS

In order to analyze the performance of proposed, Poisson Fragment Frequency based Web Pattern Clustering (PFF-WPC) technique is implemented in Java language with help of Apache log samples dataset. The Apache log samples dataset [21] includes the information about web pages access activities of many web users for example Public IP address, Date, Time of Access, Port Number, web pages accessed at a particular time. For conducting the experimental evaluation, PFF-WPC technique considers the 300 web patterns i.e. web pages.

The effectiveness of PFF-WPC technique is compared against with existing web usage mining approach [1] and Linear-Temporal Logic (LTL) based model [2]. The performance of PFF-WPC technique is evaluated in terms of clustering efficiency, computational complexity, true positive rate and space complexity. The experimental of PFF-WPC technique is performed for numerous instances with respect to varied number of web patterns and averagely ten results is illustrated in table and graph for comparative result analysis.

4. RESULTS AND DISCUSSIONS

In this section, the comparative result analysis of proposed PFF-WPC technique is presented. The efficiency of PFF-WPC technique is compared against with existing Web Usage Mining Approach [1] and Linear-Temporal Logic (LTL) based model [2] respectively. The performance of PFF-WPC technique is estimated along with the following parameters with the aid of tables and graphs.

4.1 Impact of Clustering Efficiency

In PFF-WPC technique, Clustering Efficiency (*CE*) determines the ratio of number of web patterns that are correctly grouped as frequent or non-frequent to the total number of web patterns. The clustering efficiency is measured in terms of percentage (%) and mathematically obtained as follows,

$$CE = \frac{\textit{Number of web patterns correctly clustered as frequent or non-frequent}}{\textit{total number of web patterns}} * 100 \quad (9)$$

From equation (9), clustering efficiency is evaluated with respect to the different number of web patterns taken as input. When clustering efficiency is higher, the method is said to be more effective.

Table 1 Tabulation for Clustering Efficiency

Number of Web Patterns	Clustering Efficiency (%)		
	Web Usage Mining Approach	LTL based model	PFF-WPC technique
30	68	75	89
60	71	77	90
90	72	78	91
120	74	79	92
150	75	82	93
180	78	83	94
210	79	84	95
240	80	86	96
270	82	87	97
300	83	88	98

The comparative result analysis of clustering efficiency based on various number of web patterns in the range of 30-300 using three methods is shown Table 1. PFF-WPC technique considers the framework with different number of web patterns for performing experimental evaluation using java language. When considering the 180 web patterns for conducting experimental process, PFF-WPC technique obtains 94 % clustering efficiency whereas Web Usage Mining Approach [1] and LTL based model [2] acquires 78 % and 83 % respectively. From these results, it is clear that the clustering efficiency using proposed PFF-WPC technique higher than the existing works [1] and [2].

Figure 6 Measurement of Clustering Efficiency versus Number of Web Patterns

Figure 6 describes the impact of clustering efficiency versus diverse number of patterns in the range of 30-300 using three methods namely Web Usage Mining Approach [1], LTL based model [2] and PFF-WPC technique. As shown in figure, the proposed PFF-WPC technique offers higher clustering efficiency for predicting the frequent web pages accessed by web users in a weblog database as compared to web usage mining approach [1] and LTL based model [2]. Further, while increasing the number of web patterns, the clustering efficiency is also gets increased for all three methods. But, comparatively the clustering efficiency using proposed PFF-WPC technique is higher. This is because of the application of frequency based web pattern clustering algorithm designed in proposed PFF-WPC technique.

During the web pattern clustering process, the frequency is estimated for each web page based on number of occurrences of specific web page to the total number of web pages in weblog database. Based on frequency measurement, then web pages in each session are efficiently grouped as frequent or non-frequent web pages. This process helps for PFF-WPC technique to attain higher clustering efficiency in a significant manner. Hence, proposed PFF-WPC technique increases the clustering efficiency by 23 % as compared to web usage mining approach [1] and 14 % as compared to LTL based model [2] respectively.

4.2 Impact of Computational Complexity

In PFF-WPC technique, Computational Complexity (*CC*) determines the amount of time taken for identifying the users location of web traffic patterns. The computational complexity is evaluated in terms of milliseconds (ms) and mathematically expressed as,

$$CC = N * time(\textit{idenifying users location of web traffic pattern}) \quad (10)$$

From equation (10), the computational complexity is computed with respect to various numbers of web patterns (N) as input. While computational complexity is lower, the method is said to be more effectual.

Table 2 Tabulation for Computational Complexity

Number of Web Patterns	Computational Complexity (ms)		
	Web Usage Mining Approach	LTL based model	PFF-WPC technique
30	21.9	15.6	10.2
60	26.1	18.3	14.7
90	31.2	22.1	18.2
120	36.7	28.4	24.3
150	42.5	31.7	29.1
180	48.6	35.8	33.4
210	54.8	40.2	36.7
240	60.3	46.8	41.2
270	68.7	51.3	45.8
300	75.9	60.1	49.6

The comparative result analysis of computational complexity based on different number of web patterns in the range of 30-300 using three methods is shown in Table 2. When considering the 210 web patterns for carried outing experimental work, PFF-WPC technique takes 36.7 ms computational complexity whereas Web Usage Mining Approach [1] and LTL based model [2] acquires 54.8 ms and 40.2 ms respectively. From these results, it is descriptive that the computational complexity using proposed PFF-WPC technique lowers than the existing works [1] and [2].

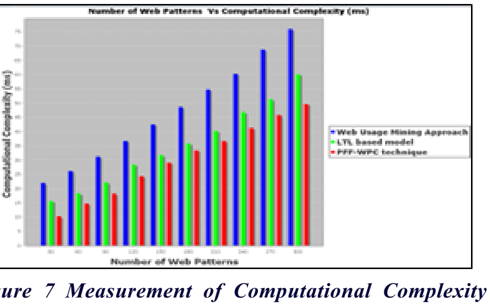

Figure 7 Measurement of Computational Complexity versus Number of Web Patterns

Figure 7 presents the impact of computational complexity versus diverse number of patterns in the range of 30-300 using three methods namely Web Usage Mining Approach [1], LTL based model [2] and

PFF-WPC technique. As described in figure, the proposed PFF-WPC technique offers minimum computational complexity for predicting web traffic patterns and identifying their user's location as compared to web usage mining approach [1] and LTL based model [2]. Moreover, while increasing the number of web pattern, computational complexity is also gets increased for all three methods. But comparatively the computational complexity using proposed PFF-WPC technique is lower. This is owing to the usage of frequency based web pattern clustering and temporal similarity based web user tracking algorithm developed in proposed PFF-WPC technique.

With application of frequency based web pattern clustering algorithm, PFF-WPC technique groups the web pages in each sessions as frequent or non frequent web patterns with higher clustering efficiency. After clustering process, PFF-WPC technique measures the temporal similarity between two sessions of frequent web pages in order to find out the traffic web patterns in a weblog database. Based on determined traffic web patterns, then PFF-WPC technique tracks the web user's location with aid of their public IP address stored in weblog database with minimum time. Therefore, PFF-WPC technique takes minimum amount of time for tracking the user's location of web traffic patterns. As a result, proposed PFF-WPC technique reduces computational complexity by 37 % as compared to web usage mining approach [1] and 15 % as compared to LTL based model [2] respectively.

4.3 Impact of True Positive Rate

In PFF-WPC technique, True Positive Rate (*TPR*) computes the ratio of the number of web patterns correctly detected as traffic patterns to the total number of web patterns. The true positive rate is evaluated in terms of percentage (%) and mathematically represented as,

$$TPR = \frac{number\ of\ web\ patterns\ correctly\ predicted\ as\ traffic\ patterns}{total\ number\ of\ web\ patterns} * 100 \quad (11)$$

From equation (11), the true positive rate for web traffic patterns mining is evaluated with respect to dissimilar number of web patterns. While true positive rate for web traffic patterns mining is higher, the method is said to be more efficient.

Table 3 Tabulation for True Positive Rate

Number of Web Patterns	True Positive Rate (%)		
	Web Usage Mining Approach	LTL based model	PFF-WPC technique
30	56	71	85
60	58	74	86
90	59	75	88
120	62	78	89
150	63	79	91
180	65	81	92
210	66	83	93
240	69	84	95
270	70	85	96
300	72	88	97

Table 3 presents comparative result analysis of true positive rate based on varied number of web patterns in the range of 30-300 using three methods. When considering the 240 web patterns for accomplishing experimental work, PFF-WPC technique attains 95 % true positive rate whereas Web Usage Mining Approach [1] and LTL based model [2] attains 69 % and 84 % respectively. From these results, it is expressive that the true positive rate using proposed PFF-WPC technique is higher than the existing works [1] and [2].

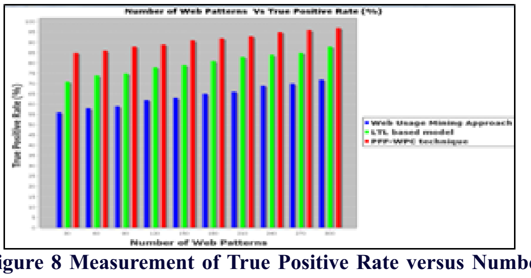

Figure 8 Measurement of True Positive Rate versus Number of Web Patterns

Figure 8 depicts the impact of true positive rate versus various number of patterns in the range of 30-300 using three methods namely Web Usage Mining Approach [1], LTL based model [2] and PFF-WPC technique. As exposed in figure, the proposed PFF-WPC technique offers higher true positive rate for identifying traffic web patterns as compared to web usage mining approach [1] and LTL based model [2]. Furthermore, 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 PFF-WPC technique is higher. This is due to application of Poisson fragment process and frequency based web patterns clustering and temporal similarity in proposed PFF-WPC technique With the aid of poisson fragment process, PFF-WPC technique separates the input weblog database into a number of sessions based on access time. Afterward, web pages in each session are clustered as frequent or non frequent based on frequent measurement by using frequency based web patterns clustering algorithm. Then, the temporal similarity is determined among all sessions of frequent web patterns to find the traffic web patterns. This supports for PFF-WPC technique to identify the traffic web patterns in a given weblog database which results in enhanced true positive rate. Therefore, proposed PFF-WPC technique improves true positive rate by 43 % as compared to web usage mining approach [1] and 14 % as compared to LTL based model [2] respectively.

4.4 Impact of Space Complexity

In PFF-WPC technique, Space Complexity *SC* measures the amount of memory space required for storing the traffic web patterns. The space complexity is estimated in terms of Mega bytes (MB) and mathematically formulated as,

$$SC = N * memory\ (storing\ traffic\ web\ pages) \quad (12)$$

From equation (12), the space complexity is determined with respect to dissimilar number of web patterns. While space complexity for mining the traffic web patterns is lower, the method is said to be more efficient.

Table 4 Tabulation for Space Complexity

Number of Web Patterns	Space Complexity (MB)		
	Web Usage Mining Approach	LTL based model	PFF-WPC technique
30	19	15	9
60	24	18	11
90	27	19	13
120	29	21	16
150	30	22	17
180	32	26	19
210	33	28	22
240	34	29	23
270	36	32	25
300	39	35	28

Table 4 explains comparative result analysis of space complexity based on different number of web patterns in the range of 30-300 using three methods. When considering the 270 web patterns for experimental work, PFF-WPC technique acquires 25 MB space complexity whereas Web Usage Mining Approach [1] and LTL based model [2] attains 36 MB and 32 MB respectively. From these results, it is illustrative that the space complexity using proposed PFF-WPC technique is lower than existing works [1] and [2].

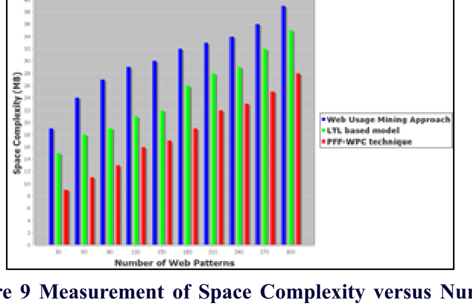

Figure 9 Measurement of Space Complexity versus Number of Web Patterns

Figure 9 illustrates the impact of space complexity versus diverse number of patterns in the range of 30-300 using three methods namely Web Usage Mining Approach [1], LTL based model [2] and PFF-WPC technique. As portrayed in figure, the proposed PFF-WPC technique offers minimum space complexity for web behaviour analysis as compared to web usage mining approach [1] and LTL based model [2]. In addition, while increasing the number of web pattern, space complexity is also gets increased for all three methods. But comparatively the space complexity using proposed PFF-WPC technique is lower. This is because of usage of frequency based web patterns clustering and temporal similarity in proposed PFF-WPC technique.

By using the frequency based web patterns clustering algorithm, PFF-WPC technique efficiently identifies the frequent web pages visited by web users in each session. With the help of detected frequent web pages, then PFF-WPC technique computes the temporal similarity among the all session in order to detect the higher traffic web patterns for web user tracking. The predicted web traffic patterns consume minimum amount of memory space for storage as compared to other existing techniques. Thus, proposed PFF-WPC technique reduces space complexity by 41 % as compared to web usage mining approach [1] and 27 % as compared to LTL based model [2] respectively.

RELATED WORKS

A survey of different techniques designed to track the web user's behaviors and their implications and possible user's defenses was analyzed in [11]. An integration framework was intended in [12] for the analysis of user activity based on user interest. However, web user tracking was remained unsolved.

Hybrid Approach was designed in [13] to minimize the complexity of identifying user's web navigation behavior. But, space complexity of user's web navigation behavior was not solved. A novel Approach was presented in [14] with assists of Apriori algorithm for customer behavior analysis. However, the amount of time taken for behavior analysis was higher.

An ant colony optimization-based algorithm was used in [15] to determine web usage patterns. But, tracking user location was not considered. A concurrent neuro-fuzzy model was intended in [16] to find out useful knowledge from the weblog data for web traffic mining. However, true positive rate of web traffic mining was poor.

K-Nearest Neighbor (KNN) classification method was presented in [17] for web usage data mining. But, a time complexity was more. A clustering with a heuristic-based pattern extraction algorithm was designed in [18] with aiming at mining the web access patterns. However, clustering efficiency of this algorithm was poor.

Association rule mining was employed in [19] to find out the users webpage access behaviour. But, tracking location of traffic web patterns was not performed. In [20], the effect of temporal attribute in relational rule mining was presented to find out interesting patterns.

CONCLUSION

An efficient Poisson Fragment Frequency based Web Pattern Clustering (PFF-WPC) technique is developed with aim of minimizing the computational complexity of tracking the location of user interest web pages from a weblog database. The aim of PFF-WPC technique is achieved with application of poisson fragment process, frequency based web patterns clustering and temporal similarity measurement. At first PFF-WPC technique applied Poisson Fragment Process to group the web pages in a database based on sessions. Then, PFF-WPC technique utilized Frequent based web patterns clustering algorithm to cluster web pages as frequent or non frequent which resulting in enhanced clustering efficiency. Subsequently, PFF-WPC technique employed the temporal similarity measurement to find traffic web patterns among all sessions which in turn assists for increasing the true positive rate in a significant manner. With help of detected traffic web patterns, the location of corresponding web users is identified with public IP address stored in weblog database which resulting in minimum computational complexity. The performance of PFF-WPC technique is test with the metrics such as clustering efficiency, computational complexity, true positive rate and space complexity. The experimental result shows that PFF-WPC technique is presents better performance with an enhancement of true positive rate and the reduction of computational complexity 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. Sergio Hernandez, Pedro Alvarez, Javier Fabra, Joaquín Ezpeleta, "Analysis of Users' Behavior in Structured e-Commerce Websites", IEEE Access, Volume 5, Pages 11941 – 11958, 2017
3. Mohammad Amin Omidvar, Vahid Reza Mirabi and Narjes Shokry, "Analyzing the Impact of Visitors on Page Views With Google Analytics", International Journal of Web & Semantic Technology, Volume 2, Issue1, Pages 14-32, January 2011
4. Neha Goel, C.K. Jha, "Analyzing Users Behavior from Web Access Logs using Automated Log Analyzer Tool", International Journal of Computer Applications, Volume 62, Issue 2, Pages 29-33, January 2013
5. Anurag kumar, Vaishali Ahirwar, Ravi Kumar Singh, "A Study on Prediction of User Behavior Based on Web Server Log Files in Web Usage Mining", International Journal Of Engineering And Computer Science, Volume 6, Issue 2, Pages 20233-20236, 2017
6. Christian Banse, Dominik Herrmann, and Hannes Federrath, "Tracking Users on the Internet with Behavioral Patterns: Evaluation of Its Practical Feasibility", Information Security and Privacy Research, Springer, Pages 235-248, 2012
7. Joao Pedro Dias, Hugo SerenoFerreira, "Automating the Extraction of Static Content and Dynamic Behaviour from e-Commerce Websites", Procedia Computer Science, Elsevier, Volume 109, Pages 297-304, 2017
8. Cheng Fang, Jun Liu, Zhenming Lei, "Fine-Grained HTTP Web Traffic Analysis Based on Large-Scale Mobile Datasets", IEEE Access, Volume 4, Pages 4364 – 4373, 2016
9. Zhen Liao, Yang Song, Yalou Huang, Li-wei He, Qi He, "Task Trail: An Effective Segmentation of User Search Behavior", Transaction on Knowledge and Data Engineering, Volume 26, Issue 12, Pages 3090 – 3102, 2014
10. Mamoun A. Awad, Issa Khalil, "Prediction of User's Web-Browsing Behavior: Application of Markov Model", IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), Volume 42, Issue 4, Pages 1131 – 1142, 2012
11. Tomasz Bujlow, Valentín Carela-Español, Josep Solé-Pareta, Pere Barlet-Ros, "A Survey on Web Tracking: Mechanisms, Implications, and Defenses", Proceedings of the IEEE, Volume 105, Issue 8, Pages 1476 – 1510, 2017
12. Jianping Zeng, Shiyong Zhang, Chengrong Wu, "A framework for WWW user activity analysis based on user interest", Knowledge-Based Systems, Elsevier, Volume 21, Pages 905–910, 2008
13. MeeraNarvekar, Shaikh SakinaBan, "Predicting User's Web Navigation Behavior Using Hybrid Approach", Procedia Computer Science, Elsevier, Volume 45, Pages 3-12, 2015
14. Preeti Sharma, Sanjay Kumar, "An Approach for Customer Behavior Analysis Using Web Mining", International Journal of Internet Computing (IJIC), Volume 1, Issue 2, Pages 1-6, 2011