International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August 2019

Research Aligned Analysis on Web Access Behavioral Pattern Mining for User Identification

Gokulapriya R, Ganesh Kumar R.

Abstract*— Human activity understanding includes activity recognition and activity pattern discovery. Monitoring human activity and finding abnormality in their activities used by many field like medical applications, security systems etc. Basically it helps and support in decision making systems. Mining user activity from web logs can helps in finding hidden information about the user access pattern which reveals the web access behaviour of the users. Clustering and Classification techniques are used for web user identification. Clustering is the task of grouping similar patterns for web user identification. Classification is the process of classifying web patterns for user identification. In this paper we have implemented the existing works and discussed the results here to find the limitations. In existing methods, many data mining techniques were introduced for web user behaviour identification. But, the user identification accuracy was not improved and time consumption was not reduced. Our objective is to study the existing work and explore the possibility to improve the identification accuracy and reduce the time consumption using machine learning and deep learning techniques.

Keywords*— Human activity understanding,web usage mining, web user behavior, data mining, web patterns,User Idenifiction

I. INTRODUCTION

Web mining is a subset of data mining, web mining helps in extracting useful information from the web. The data available in websites can be divided into three parts majorly

1. Images , videos, audios, text
2. Hyperlinks
3. Data available from browsing

Based on these three types of mining can be carried out. Content mining is carried out on the first data source. Structure mining is carried out on the second data source as in all the hyperlinks of the pages are analyzed so that all the relevant information can be extracted. The last type of mining done is web usage mining. This mining is done on the third source of data which is mentioned above and these data is available in the form of log files from the server. The log files are generated each time the user access the website. These log file include unstructured data like the IP address, status code, number of bytes transmitted, class and time stamp. In order to utilize this unstructured data various preprocessing methods are applied and data is converted to structured data. Base on the type of information to be retrieved we select the parameters from the preprocessed structured data and then the prediction is done.

This paper discuss the web user behavior mining as follows: Section II discusses the review on different user behavior mining techniques, Section III analysis the existing web user behavior pattern mining techniques, Section IV showcase the comparison between them. In Section V, the discussion and limitations on the existing techniques are portrayed and Section VI concludes the paper. The focus of this work is to find the research gap to improve the performance of web user identification based on web user behavior patterns.

Revised Manuscript Received on August 30, 2019.

* Correspondence Author

Gokulapriya R.*, Research Scholar, Computer Science and Engineering, CHRIST (Deemed to be University), EMail id: r.gokulapriya @ res.christuniversity.in

Ganesh Kumar R., Associate Professor, Computer Science and Engineering, CHRIST (Deemed to be University), EMail id: ganesh.kumar@christuniversity.in

II. LITERATURE REVIEW

A special model known as Linear-temporal logic model was proposed in [1]. This model maps the web logs with the e-commerce website structure, changing the normal web logs to event transactions from which the user patterns are derived. However this model fails to automate the discovery of behavioural patterns. MiND discussed in [2] is a cluster based method, focusing on collecting the internet measurements at end user location. The identification happens here on two fronts, i.e. users based on internet access behaviour and users based on anomalous services. The user measurement over time is represented through histogram and observed through two level clustering. However MiND did not improve the accuracy of clustering. Identifying daily human behaviour plays a vital role. [3] Proposes a Scalable approach which mines the pattern from multivariate temporal data collected from smartphones. The proposed algorithm identified the frequent behavioural patterns with temporal fineness based on the division of the time created by the users. Personality based product recommender system proposed in [4] identifies the personality of the user and gives the recommendations accordingly. The method worked fine with the social media but cannot be extended to applicant personality and their fitness with a certain organization in e-recruiting process in crowd sourcing market.

Focusing on various online materials a process oriented approach proposed in [5] dealt with student behavioral patterns through the mining of their interaction with different quizzes from different course. The drawback of this model was that the time required to complete the mining process was not minimized.

For detecting large scale events on the basis of their frequency and time duration an evaluation method was proposed in [6]. The aim of the work was to identify the event importance and segregate them into big uniform events. The method was not able to minimize the processing time. The analysis based on spam pages improved the search engine results [7] by minimizing the energy consumption. It used Normal and gamma distribution for computing probability but failed to identify the web access control.

The major issues identified from the above-said literature are lack of investigating additional behavioural patterns, less user behaviour pattern identification accuracy, more execution time for mining the user behaviour patterns from the weblog and so on. In order to solve the issues from the existing literature, the research work can be carried out using machine learning and deep learning techniques for increasing the accuracy of the web behaviour pattern mining with minimum time consumption

III. WEB ACCESS BEHAVIORAL PATTERN MINING

Web user behavioral mining is the activity of user behavior from the previous navigation history. The detection of the behavioral patterns has a range of application in variety of fields. User behavioral pattern mining finds its usage in health care to transportation systems. The steps involved in web user behavioral mining are collecting the data from different sources, cluster the data and then analyze those clusters for the usage identification.

A.User Behavioral Analysis in e-Commerce Websites

Understanding and identifying user behavioral pattern play a major role for e-commerce website to flourish. The information to be mined is obtained from the server logs. Linear temporal logic model is used in [1] for examination of structured e-commerce logs. This model maps each page involved with the e-commerce site with the web log data and thus converts the normal web log into event log. This simplifies the process of extracting the behavioral patterns as certain predefined queries are used to detect different user patterns based on various user actions performed by user during their session.

Temporal logic model were proposed keeping in mind applicability with open systems. A new methodology was proposed for analyzing the e-commerce websites for the customer behavioral patterns. The web logs were prepared for applying the temporal logic model by removing the traces. The user generates events by interaction with the site page, for ex. if the user adds a certain product to wish list, searching for a particular product etc. The queries help us examine the probable relationship existing between various events obtained from the client session. The technique provides casual relations between the events of user trace than providing global view of session. The end goal in mind is to predict and modify the website according to global session and thus improve the overall usability of the site by changing the structure as per user needs.

B. Performing Cluster Analysis to group the user having Similar Internet Access

MiND standing for Mining Neubot Data is cluster based methodology used to examine periodic internet calculation gathered at end user location[2]. MiND groups the data on the basis of users having similar internet access and users with anomalous services. The measurements collected over time are modelled through the utilization of histogram and examined through two-way clustering strategy. The data evaluated from MiND is obtained from Neubot which is voluntary installed by users. The grouping of majority users is done into homogeneous and cohesive clusters which are in line with Internet access service and users collecting the anomalous service are grouped as outliers. The anomalous behaviors in the services are identified by monitoring the services offered by ISP.

The statistical distribution obtained from user location is used by MiND to group the users into groups which are consistent with their access service. There are no instances of the users having lesser or no occurrences of download speed being lesser than the threshold value in users gaining access through regular access service. The distribution of download speed when similar to other group indicates a service coherent to the subscribed ones. This provides the users with the options of analysing various other options present and the ISP providers to find the reason for the anomalous behaviour and fixing it.

MiND has fourfold objective. The measurements which are collected are represented through histograms which emphasize the relationship of internet access in terms of bandwidth. The bandwidths are spitted into intervals. Histogram is generated for each bin.This is used to model the measurements generated by similar user over times. Two level clustering utilizes DBSCAN and K-means algorithm to identify homogeneous users i.e. users consistent with download speed actually experienced and users with anomalous patterns. This method is applicable on internet measurements with both noise and outlier data. The proposed method also groups the users into separated clusters. DBSCAN algorithm uses distance measure for detecting noise and outliers. The evaluation of this method was done using real datasets from single ISP and different geographical location during varied time intervals.

C.User Behavioral Pattern Mining from Multivariate Temporal Data

The scalable algorithms were introduced in [3] to detect the human access behavior. These patterns prove to be of significant value to the end users and third parties who provide service depending on the information. The algorithm performs the data analysis the sensor data available in mobile device is converted into machine readable data from heterogeneous data. The algorithm is designed in such a way that it can be integrated on smaller devices which has less processing power compared to desktop computers. The algorithm is lightweight, having measurable execution time in terms of scalability. Daily behavioral patterns are mined using scalable approach on the data obtained from sensors.

This approach was introduced to mine the variations on time stamps to show human perception of time. Frequent behavioral pattern approach is not dependent on the single information source and is generic. Algorithm reduces uncertainty by using combination of information obtained from various sources.

D.Personality-mining in social networks to detect User behavioral patterns

A Personality based product recommender framework was introduced in [4] which identifies the user based on the personality and gives suggestion accordingly. This framework was proposed utilizing social network dataset. PBPR consists of three engines which are an engine to retrieve the user personality through social network features. Figure 1 shows pictorial representation of the composition of PBPR. The frameworks uses personality traits obtained from the social network to predict the user personality and recommend products accordingly.

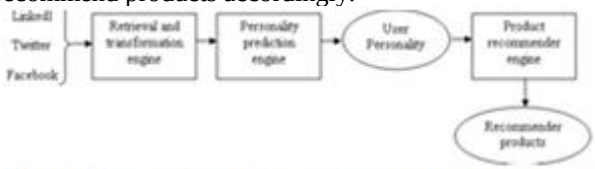

Figure 1 Personality-based product recommender framework

Retrieval & transformation engine:

The first engine which is retrieval & transformation engine retrieves the social media parameters which are obtained from different social media platforms and stores them as standardized vectors. Popular social media platforms provide programming interfaces for easy retrieval of the required traits.

Personality prediction engine:

Human personality is determined through various personality traits. They are categorized as endogenous, constant, hierarchically structured dispositions. Mostly these traits are managed or influenced through biological factors such as genes and brains structure. Traits observed for a particular person remain constant over lifetime and influences people's thoughts, feel and behavior.

Product recommender engine:

Product recommender engine recommends the products based on the predicted user personality. Recommender engine finds a relationship between the consumer's personality and product features. Every product has a functional utility and symbolic meaning. The product itself is denoted by the symbolic meaning and it is explained with the consumer's personality which in turn forms the product personality. Self-congruence considered as one of the important factors in directing consumer preferences. The selection of product and symbolic meaning is dependent on consumer's self concept.

General framework discussed in [8] for mining audience behavior used TV rating data providing the active user intentions. However using extensive video content highlighted the interests of viewer in specific manner. Computing a similarity measure was proposed in [9] for mining different entity sets. This was done to explore the interaction between objects of different entity set with the end goal to determine the behavioral similarity between the objects in the region of interest. Space complexity was not minimized in the proposed method. Log mining for behavior pattern, a statistical model is highlighted in [10]. This model gathers every host networks behavior patterns with non-negative matrix factorization algorithm. The algorithm improved compatibility and enhanced interpretations, thus reducing the complexity involved but failed to reduce the false positive rates.

IV. COMPARISON OF WEB USER BEHAVIORAL PATTERN MINING & SUGGESTIONS

Table 1 summarizes the various performance metrics of existing behavior pattern mining techniques on web.

Table 1 Tabulation for Parameters in Existing Methods

S. No.	Existing Method Name	Execution Time	Purchase intention rate	Sensitivity	Specificity	Scalability	Robustness	Error rate	Precision	Accuracy
		Parameter Name								
1	Linear-Temporal Logic Model Checking Approach	✓	✓	NA	NA	NA	NA	NA	NA	NA
2	MiND method	NA	NA	✓	NA	NA	✓	NA	NA	NA
3	Scalable approach	✓	NA	NA	NA	✓	NA	NA	NA	✓
4	PBPR framework	NA	NA	✓	✓	NA	NA	NA	✓	✓

A. *Impact of Web User Identification Accuracy*

Web User Identification Accuracy (WUIA) is measured as,

WUIA=(Number of web patterns are grouped)/(Total number of web patterns)*100 (1)

From (1), web user identification accuracy is calculated in terms of percentage (%). When the web user identification accuracy is higher, the method is said to be more efficient.

Table 2 describes the web user identification accuracy with respect to number of web patterns. Web user identification accuracy comparison takes place on existing methods.

Table 2 Tabulation for Web User Identification Accuracy

Number of web patterns (number)	Web User Identification Accuracy (%)			
	Linear-Temporal Logic Model Checking Approach	MiND method	Scalable approach	PBPR framework
10	54	78	59	66
20	55	79	61	68
30	57	80	63	70
40	59	82	65	72
50	61	84	67	75
60	63	86	69	77
70	65	88	71	80
80	67	90	73	82
90	69	91	75	85
100	71	92	77	87

The graphical representation of web user identification accuracy is illustrated in figure 2.

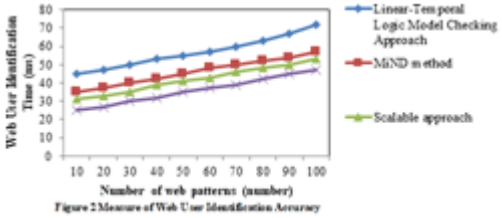

Figure 2 Measure of Web User Identification Accuracy

From the figure 2, it is clear that the web user identification accuracy using Mining Neubot Data (MiND) method is higher when compared to Linear-Temporal Logic Model Checking Approach,

Scalable approach, and personality-based product recommender (PBPR) framework. MiND evaluated on real data collected by Neubot which gathers the Internet measurements. MiND examined statistical distribution of download speed at user locations to group the Internet users into homogeneous and cohesive groups consistent with broadband access service. This in turn helps to increase the web user identification accuracy. Research in Mining Neubot Data (MiND) method increases the web user identification accuracy by 37% when compared to Linear-Temporal Logic Model Checking Approach, by 25% when compared to scalable approach and by 12% when compared to PBPR framework.

B. IMPACT OF WEB USER IDENTIFICATION TIME

Web user identification time (WUIT) is defined as the amount of time required for classifying the user based on access behavioral patterns. Web user identification time is calculated as,

WUIT=Ending time-Starting time of web user identification (2)

From (2), web user identification time is calculated in terms of milliseconds (ms). When the web user identification time is lesser, the method is said to be more efficient.

Table 3 Tabulation for Web User Identification Time

Number of web patterns	Web User Identification Time (ms)			
	Linear-Temporal Logic Model Checking Approach	MiND method	Scalable approach	PBPR framework
10	45	35	31	25
20	47	37	33	27
30	50	40	35	30
40	53	42	39	32
50	55	45	41	35
60	57	48	43	37
70	60	50	46	39
80	63	52	48	42
90	67	54	50	45
100	72	57	53	47

Table 3 explains the web user identification time with number of web patterns ranging from 10 to 100. Web user identification time is compared with existing methods. From table 3, it is observed that the web user identification time using PBPR framework is lesser when compared to Linear-Temporal Logic Model Checking Approach, Mining Neubot Data, Mining Neubot Data (MiND) method and Scalable approach. This is because of examining the social media data to forecast the user personality and deriving the personality-based product preferences. PBRS framework computed the IT-artifact with distinctive online social network using XING dataset. This in turn helps to minimize the web user identification time. Research in PBPR framework reduces the web user identification time by 37% when compared to Linear-Temporal Logic Model Checking Approach, by 18% when compared to Mining Neubot Data (MiND) method and by 11% when compared to scalable approach.

The graphical representation of web user identification time is described in figure 3.

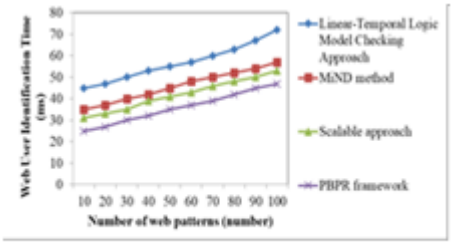

Figure 3 Measure of Web User Identification Time

C. IMPACT OF SPACE COMPLEXITY

Space complexity (SC) is defined as the amount of memory space utilized for storing the web user behavior patterns for user identification. It is measured in terms of kilobytes (KB). The space complexity mathematically given,

SC=Number of web patterns*space (storing the web patterns) (3)

When the space complexity is lesser, the method is said to be more efficient. Table 4 explains the space complexity with respect to number of web patterns.

Table 4 Tabulation for Space Complexity

Number of web patterns (number)	Space Complexity (KB)			
	Linear-Temporal Logic Model Checking Approach	MiND	Scalable approach	PBPR framework
10	6	8	12	15
20	8	10	15	18
30	10	11	17	21
40	12	14	19	24
50	15	17	24	27
60	17	20	27	31
70	20	22	30	35
80	24	26	34	39
90	26	30	37	43
100	29	33	38	45

Space complexity comparison takes place on existing Linear-Temporal Logic Model Checking Approach, Mining Neubot Data (MiND) method, scalable approach and PBPR framework. The graphical representation of space complexity is shown in figure 4.

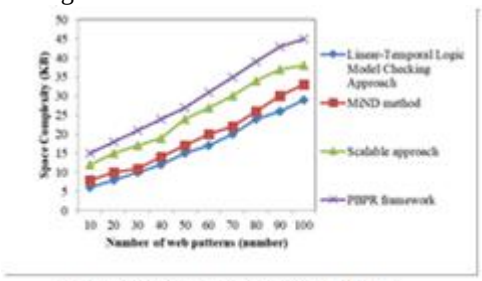

Figure 4 Measure of Space Complexity

From the figure 4, it is clear that the space complexity using Linear-Temporal Logic Model Checking Approach is lesser when compared to PBPR framework, Mining Neubot Data, Mining Neubot Data (MiND) method, and Scalable approach. This is because of providing the causal relations among the events of user trace than providing global view of whole session. The designed approach predicts the possible events and allows having the global view of sessions, global analysis of user behavior and assisting the re-design of website for adaptation to the user necessities. The temporal granularity transformation algorithm in designed approach changes the timestamps to the mirror human perception of time.

The designed approach minimized the problem of uncertainty through combining the various information sources to identify the frequent behavioral patterns. This in turn helps to minimize the space complexity. Research in Linear-Temporal Logic Model Checking Approach reduces the space complexity by 14% when compared to Mining Neubot Data (MiND) method, by 37% when compared to scalable approach and by 46% when compared to personality-based product recommender (PBPR) framework.

D. IMPACT OF FALSE POSITIVE RATE

False positive rate (FPR) is described as the ratio of number of web patterns are incorrectly grouped to the total number of web patterns. FPR is measured in terms of percentage (%). The false positive rate is formulated as,

FPR= (Number of web patterns incorrectly grouped)/(Total number of web patterns)*100 (4)

Table 5 portrays the false positive rate with respect to number of web patterns. False positive rate comparison takes place on existing Linear-Temporal Logic Model Checking Approach, Mining Neubot Data (MiND) method, scalable approach and PBPR framework. The graphical representation of false positive rate is illustrated in figure 5.

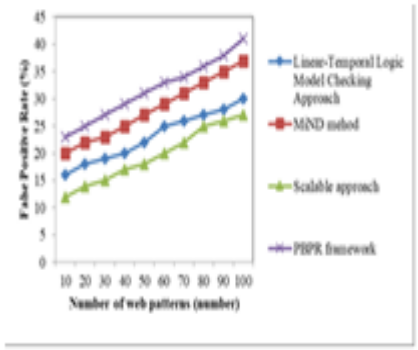

Figure 5 Measure of False Positive Rate

From the figure 4, it is observed that the false positive rate using scalable approach is lesser when compared to PBPR framework, Mining Neubot Data, Mining Neubot Data (MiND) method, and Linear-Temporal Logic Model Checking Approach. This is because of using temporal granularity transformation algorithm with changes on timestamps to mirror human perception of time. The frequent behavioral pattern detection approach used in the approach is generic and not depending on the single information source. The designed approach minimized the problem of uncertainty through combination of information sources to identify the frequent behavioral patterns. This in turn helps to minimize the false positive rate.

Table 5 Tabulation for False Positive Rate

Number of web patterns (number)	False Positive Rate (%)			
	Linear-Temporal Logic Model Checking Approach	MiND	Scalable approach	PBPR framework
10	16	20	12	23
20	18	22	14	25
30	19	23	15	27
40	20	25	17	29
50	22	27	18	31
60	25	29	20	33
70	26	31	22	34
80	27	33	25	36
90	28	35	26	38
100	30	37	27	41

Research in scalable approach reduces the false positive rate by 16% when compared to Linear-Temporal Logic Model Checking Approach, by 31% when compared to Mining Neubot Data (MiND) method and by 39% when compared to personality-based product recommender (PBPR) framework. All parameters are given with the results with respect to the existing methods in table 6.

Table 6 Tabulation for Parameters and Accuracy of results

Parameter Name	Existing Method Name	Linear-Temporal Logic Model Checking Approach	MiND method	Scalable approach	PBPR framework
Web User Identification Accuracy	MiND method	37% ↑	NA	25% ↑	12% ↑
Web User Identification Time	PBPR framework	37% ↓	18% ↓	11% ↓	NA
Space Complexity	Linear-Temporal Logic Model Checking Approach	NA	14% ↓	37% ↓	46% ↓
False positive rate	Scalable approach	16% ↓	31% ↓	NA	39% ↓

V. DISCUSSION ON LIMITATION OF WEB USER BEHAVIORAL PATTERN MINING TECHNIQUES

A linear-temporal logic model checking approach was designed for examination of structured e-commerce Weblogs. Predefined queries were performed to detect the behavioral patterns with different actions carried out by user during session. However, the designed approach failed to examine the additional behavioral patterns and to assist their automatic discovery. MiND was introduced to detect the characteristics of periodic Internet measurements gathered at the end user location. MiND determined the group of users with similar Internet access behavior user service. MiND computed on the real data by Neubot voluntary that gathers the Internet measurements. User measurements was modeled through the histograms and observed through the two-level clustering strategy. But, the clustering accuracy was not improved using MiND approach.

Scalable approach was designed for daily behavioral pattern mining from multiple sensor information. The designed framework detected the frequent behavioral patterns with the temporal granularity inspired by individuals. The patterns were supportive to both end-users and third parties who present the services depending on information. However, the concept drift model was not employed for churn based on the human behavior. PBPR framework was introduced to examine the social media data to predict the user personality and to derive the personality-based product preferences. Personality Prediction Engine and the Product Recommender Engine avoided the interference between the evaluation performances of the engines. But, personality prediction engine was not used for applicant personality – organizational culture congruency fit during the e-recruiting activities in crowdsourcing market.

A general framework was implemented in [8] for mining the audience behaviors. The designed framework used on the change points in TV ratings data that provide the active users intentions. But, the combination of extensive video content analysis exposed the interests of viewers in specific manner. A similarity measure was computed in [9] for mining the behavioral patterns in different entity set. The designed measure was employed to know the interaction between objects from different entity set to determine the behavioral similarity between objects inside region of interest. But, the space complexity was not minimized by similarity measure.

Log Mining for Behavior Pattern (LogM4BP) model was statistical model in [10] for gathering the each host network behavior patterns with the non-negative matrix factorization algorithm. The designed algorithm improved the interpretation and comparability of behavior patterns to reduce the complexity. But, LogM4BP model failed to minimize the false positive rate.

VI. CONCLUSION

A comparison of different existing web user behavioral pattern mining techniques for user identification is studied. From the study, it is observed that the linear-temporal logic model checking approach failed to analyze the additional behavioral patterns and to assist their automatic discovery. The survival review shows that the existing MiND approach failed to improve the clustering accuracy. In addition, personality prediction engine was not employed for the personality–organizational culture congruency fit during the e-recruiting activities in crowdsourcing market. The wide range of experiments on existing methods computes the performance of the many web user behavioral pattern mining techniques with its limitations. Finally, from the result, the research work can be carried out using machine learning and deep learning techniques for enhancing the web user identification accuracy and minimizing the time consumption measure based on web user behavior pattern mining.

REFERENCES

1. Sergio Hernández , Pedro Álvarez , Javier Fabra ,Joaquín Ezpeleta, “Analysis of Users’ Behavior in Structured e-Commerce Websites”, IEEE Access, Volume 5, 2017, Pages 11941 – 11958
2. Tania Cerquitelli, Antonio Servetti, Enrico Masala, “Discovering users with similar internet access performance through cluster analysis”, Expert Systems With Applications, Elsevier, Volume 64, 2016, Pages 536–548
3. Reza Rawassizadeh, Elaheh Momeni, Chelsea Dobbins, Joobin Gharibshah and Michael Pazzani, “Scalable Daily Human Behavioral Pattern Mining from Multivariate Temporal Data”, IEEE Transactions on Knowledge and Data Engineering, Volume 28, Issue 11, November 2016, Pages 3098 – 3112
4. Ricardo Buettner, “Predicting user behavior in electronic markets based on personality-mining in large online social networks”, Electronic Markets, Springer, 2017, Volume 27, Pages 247–265
5. Libor Juhanak, Jirí Zounek and Lucie Rohlíkov, “Using process mining to analyze students' quiz-taking behavior patterns in a learning management system”, Computers in Human Behavior, Elsevier, December 2017, Pages 1-11
6. Tao Xie, Qinghua Zheng and Weizhan Zhang, “Mining Temporal Characteristics of Behaviors from Interval Events in E-learning”, Information Sciences, Elsevier, Volume 447, June 2018, Pages 169-185
7. Aaisha Makkara and Neeraj Kumar “User behavior analysis -based Smart Energy Management for Webpage Ranking: Learning Automata-based Solution”, IEEE Access, Volume 5, July 2017, Pages 1941-11958
8. Ryota Hinami and Shin'ichi Satoh, “Audience Behavior Mining: Integrating TV Ratings with Multimedia Content”, IEEE Multimedia, Volume 24, Issue 2, April-June 2017, Pages 44 - 54
9. Sandipan Maiti and R.B.V. Subramanyam, “Mining behavioural patterns from spatial data”, an International Journal Engineering Science and Technology, Elsevier, 2018, Pages 1-11
10. Jing Ya, Tingwen Liu, Quangang Li, Jinqiao Shi, Haoliang Zhang, Pin Lv, and Li Guo, “Mining Host Behavior Patterns from Massive Network and Security Logs”, Procedia Computer Science, Elsevier, Volume 108, 2017, Pages 38–47.
11. Yadav D, Yadav J, Vashistha R, Goyal DP, Chhabra D. Modeling and simulation of an open channel PEHF system for efficient PVDF energy harvesting. Mechanics of Advanced Materials and Structures. 2019 Apr 10:1-5.