International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 8 Issue 01 Ver. II || Jan 2019 || PP 01-13

Hierarhical Forest Custer and Heuristic Poll Classification For Mining Web Access Logs

B.K. Mathan Nagan[1], Dr.C.Chandrasekar[2]
[1]*Research scholar, Department of Computer Science, Dravidian University kuppam, A.P, India and Assistant Professor, Caussanel College of Arts and Science, Ramanathapuram, Tamil Nadu, India*
[2]*Professor, Department of Computer Science, Periyar University, Salem, Tamil Nadu, India*
Corresponding Author: B.K. Mathan Nagan

Abstract: *Modern web clickstream data comprises multi levels of granularity based on the visitor behavior on its websites, making it cumbersome to scrutinize. Following the overarching idea that the event visualization should provide information regarding the click events at multiple levels of granularity and identifying navigation patterns across several applications, three levels of granularity are identified for analyzing the navigation patterns, namely, web access sequences, set of all users along with log entry. This method is called as Hierarchical Forest Cluster and Heuristic Poll Classification (HFC-HPC). With the objective of mining web access logs and improving the scalability of visualization interface, three parts are designed. They are preprocessing using Localized Sequence Level model, clustering using Certainty Hierarchical Forest and classification via Heuristic Poll. Three identifications are considered during preprocessing namely, user, user session and page view identifications to extract information of interest with minimal time. Followed by the extracted information, Certainty Hierarchical Forest Clustering algorithm is applied with which optimal clusters are said to be generated with minimal computation overhead. Finally, heuristic poll is applied to the generated optimal clusters of applications to address scalability in the context of high level patterns. Experimental results based on web log files illustrate the applicability of the proposed method in achieving better performance than existing methods in terms of computational time, overhead and error.*

Keywords*: Hierarchical Forest Cluster, Heuristic Poll, Classification, web access logs, Localized Sequence Level, information of interest,*

Date of Submission: 26-12-2018 Date of acceptance: 11-01-2019

I. Introduction

Even sequence collection and analysis occurs in several domains. For example, e-commerce organizations pursue to discern customer behaviors from clickstream data. Accordingly, decisions are informed to the marketing team. In case of healthcare field, electronic health records acts as the sources of information regarding the health updates of the patients. Here, the absolute volume and complexity of event sequences poses several questions in the visual analysis of those data. Therefore, linear patterns were exploited to mine sequential pattern with the parameters such as frequency and profit.

Mining And Querying User Interface (MAQUI) [1] portrayed the process involved in interwoven querying and mining. This method enabled recursive event sequence exploration. According to the analysts' tasks, the need for integrating querying and mining to obtain event sequences in a recursive manner was also performed. Next, a framework of interwoven querying and mining were combined and designed to determine the atomic user actions in order to refine the analytic context in an iterative manner and concentrate therefore during analysis, therefore improving the speed of frequent events and frequent patterns in an efficient manner. However, one of the limitations concerned in MAQUI was stability, involving large number of event types. This in turn resulted in high latency by potentially hampering user's performance during analysis.

Modern web clickstream data includes high-dimensional sequences involving multivariate events, making it too cumbersome to scrutinize. With intention of afford information at multiple levels and allow users' and effective navigate across these levels, four levels in clickstream analysis were detected. They were referred to as, patterns, segments, sequences and events.

Three stages were involved in the design of understanding common visitors paths, namely, pattern mining, pattern pruning and coordinated exploration among patterns and sequences and was referred as novel visualization and interaction technique. Accordingly, the characteristics of maximal sequential patterns were also investigated to minimize the frequency of patterns and obtain design considerations for extracting the sequential patterns along with the analogous raw sequences. Mutual information between patterns was used to automate the process of selecting interesting patters with minimum time and overhead. However, with the

increase in the pattern size, the quality of final patterns being pruned remains a challenge, therefore compromising scalability.

Product reviews forms as the basis for obtaining individual's opinions or belief about certain product provided by several business establishments. Such product reviews helps the business establishments in planning and monitoring their business profitability in terms of both quality and quantity. However, due to the high volume of spam, data validity remains a major concern. In [4], Gradient Boost and Generalized Boosted Regression Model were designed to enhance spam detection rate. A survey on enabling protocols was presented based on hands-on experience in [5].

Several data mining methods are investigated for mining useful patterns present in text documents. But, the most open research issue to be addressed is the efficient use of update discovered patterns, especially in the domain of text mining. To address these issues, an effective pattern discovery technique including pattern deploying and pattern evolving was designed in [6] with the objective of identifying both relevant and addressing information.

An enhanced document reading method depends on contextual visualization technique to navigate access content present overall the text was investigated in [7]. However, the economic impact of the navigated patterns was not analyzed. To address this issue, Random Forest-based Classifiers were designed in [8] to examine the relative importance of reviewer related features, review subjectivity features and review readability features. However focus was made only towards e-commerce retailer.

Aiming to overcome such limitations, this study proposes using a method that contains hierarchical and heuristic based polls for mining web access logs noted in 'NASA' dataset. The method has been designed and then evaluated through the machine learning approach which includes the Hierarchical Forest Cluster and Heuristic Poll Classification. The aim was to see if they are efficient in mining web access longs noted in 'NASA' dataset. For the purpose of this study, data were then implemented into the dataset to facilitate analysis. It was deduced that this combination in method is able to infer useful interpretation that enhance the precision and lessen the error rate. Accordingly, the contributions of this study are as follows:

- NASA HTTP dataset from [3] was used as the public dataset for accessing several navigation patterns and inferring meaningful results by applying Hierarchical Forest Cluster and Heuristic Poll Classification (HFC-HPC) method.
- This study focuses on the machine learning approach to investigate navigation patterns and obtain most relevant user information by designing HFC-HPC method.
- To select only the information of interest by applying Localized Sequence Level Preprocessing model, therefore reducing the computational time.
- To obtain optimal clusters with minimal overhead by incorporating Certainty Hierarchical Forest Clustering model in HFC-HPC method.
- Finally, to extract user profiles with higher scalability by designing Heuristic Poll Classification model

This paper is structured as follows: Section 2 discusses related work, some definitions about navigating web logs, machine learning techniques and so on. Section 3 proposes the techniques of navigating patterns from web access logs using Forest Cluster and Heuristic Poll Classification (HFC-HPC) method. Section 4 presents experimental setting and results for evaluating the proposed method followed by discussion in Section 5. Finally, Section 6 gives concluding remarks.

II. Related works

With the increasing surge of Web 2.0, a voluminous amount of product reviews are jumping up on the Web. From these review results, users can gain assessments of product information. Besides, the manufacturers also do obtain instant assessment and chances to enhance the product quality. Hence, mining opinions has become an urgent ask and has engaged a great deal of observation from researchers.

In [9], graph-based co-ranking algorithm was applied to evaluate the confidence and extract opinion targets or opinion words in a timely fashion. However, results with psychometric test did not evolve efficient mining. To address this issue, K-Means Clustering [10] was exploited for extracting cognitive style and hence interesting insights were provided according to navigation behavior of users.

Nowadays, the World Wide Web is exponentially growing at a speedy rate than ever and hence the online available resources spread in Internet present a large source of knowledge for several business establishments and research scholars. An annotation process based on Natural Language Processing (NLP) was designed in [11] to provide insight into gathering geographical based location information. Yet another progressive filtering approach was investigated in [12] by applying an iterative method for high quality web images.

The tourism industry has gained a switch from offline to online travelers and thus has made the application of intelligent systems in tourism industry crucial. This information has to provide both consumers and service providers with the information to be of the most relevant. A non-invasive web mining system was presented in [13] combining usage and content information. According to the users involved in web mining and the activities concerned, the results of the navigation pattern retrieved differs. Hence, for different users, different navigation patterns are said to be retrieved. In [14], dominance principle was applied for multi criteria web mining, called, Dominance-based Rough Set Approach (DRSA). By applying DRSA, the retrieved search results were more effective.

Mobile cloud computing (MCC) is received higher height of interest than ever before where the CC is integrated into mobile computing environments. The new MCC model breaks through the resource limitation by moving data processing and storage from mobile device to cloud via wireless networks. A novel Service Selection and Recommendation Model (SSRM) was investigated in [15] by computing user similarity depends on user context information and interest, therefore improving the accuracy of ranking results. In [16], deep neural networks were applied to comprehensive features based on the apps descriptions into different categories.

In [17], an efficient semantic web index was constructed during query processing with the objective of minimizing the time consuming process. Though extracting useful and meaningful navigation patterns present in web log files, these web applications are nowadays becoming prone to several threats. To address this issue, a cross site scripting was designed in [18] based on fuzzy logic, therefore improving the accuracy and reduction in the false positive rate.

Identifying unique features of the enterprises, for example, the adoption of e-commerce is considered to be one of the most imperative and preliminary tasks for various economic activities. This type of information is regularly collected via surveys, which involves higher cost due to the number of persons involved in the task. In [19], four classification algorithms, namely, Support Vector Machines, Random Forest, Logical Analysis of Data and Logistic Classifier was applied, therefore resulting in the improvement of classification accuracy with time. Certain web mining techniques applied in e-commerce was examined in [20].

From the above works, while a great number of researchers focused on the service selection and recommendation in web access logs based on the user behavior, little attention are devoted to multi levels of granularity. Different from these existing approaches, our work focuses on how to afford information regarding the click events at multiple levels of granularity and detect the navigation patterns across several applications. Thus, we propose a Hierarchical Forest Cluster and Heuristic Poll Classification (HFC-HPC) method for mining web access logs, which considers information of interest in order to design a clustering model with the objective of extracting unique subclasses per application and utilize hierarchical poll classification among several applications to address user profiles with higher scalability.

III. Methodology

To address the challenges of mining evolving user profiles data incorporating hierarchical forest cluster and classification in each of the examined applications, our proposed classification method utilizes Certainty Hierarchical Forest Cluster analysis and Hierarchical Poll Classification in tandem. The overview of the web usage mining for mining evolving user profiles classification method is shown in Figure 1 with an explanation of foremost steps as follows.

As shown in the figure, the block diagram of HFC-HPC method for mining web access logs includes three steps. They are preprocessing, clustering and classification. In the preprocessing stage, web sequential patterns are collected from the corresponding web logs over a period of time and marked with application labels accordingly (e.g., Skype, Google, YouTube) using a localized operational sequence-level classifier.

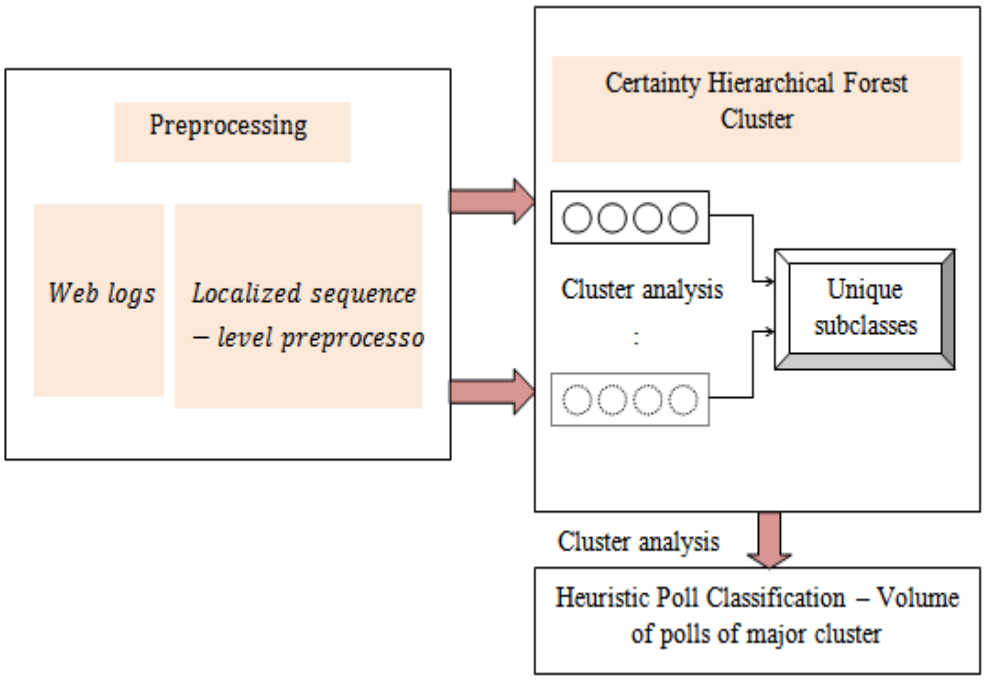

Figure 1 Block diagram of Hierarchical Forest Cluster and Heuristic Poll Classification for mining web access logs

Application labeled sequence is next exported as flows associating to discrete applications for cluster analysis. Next, in the cluster analysis stage, using Certainty Hierarchical Forest Cluster model, flows associating to discrete applications are individually clustered to extract unique subclasses per application, providing a splendid granularity of the classification (e.g., Skype, Google and YouTube flows would be classed as streaming, searching and browsing). Finally, classification is performed by applying Certainty Hierarchical Forest Cluster, indicating the subclass they belong to, are afterwards fed to a Heuristic Poll Classifier for supervised training, leading to a web access log mining with minimal time and overhead addressing scalability.

1.1 Localized Sequence Level Preprocessing model

Classical data preprocessing for event categorization involves three steps: categorize by domain name, categorize by token frequency and categorize by topics [2]. Our solution for Web Usage Mining (WUM) includes what we call advanced data preprocessing. This consists of a data summarization step or localized sequence-level preprocessing, which will allow the analyst to choose only information of interest. Figure 2 shows the block diagram of Localized Sequence Level Preprocessing model.

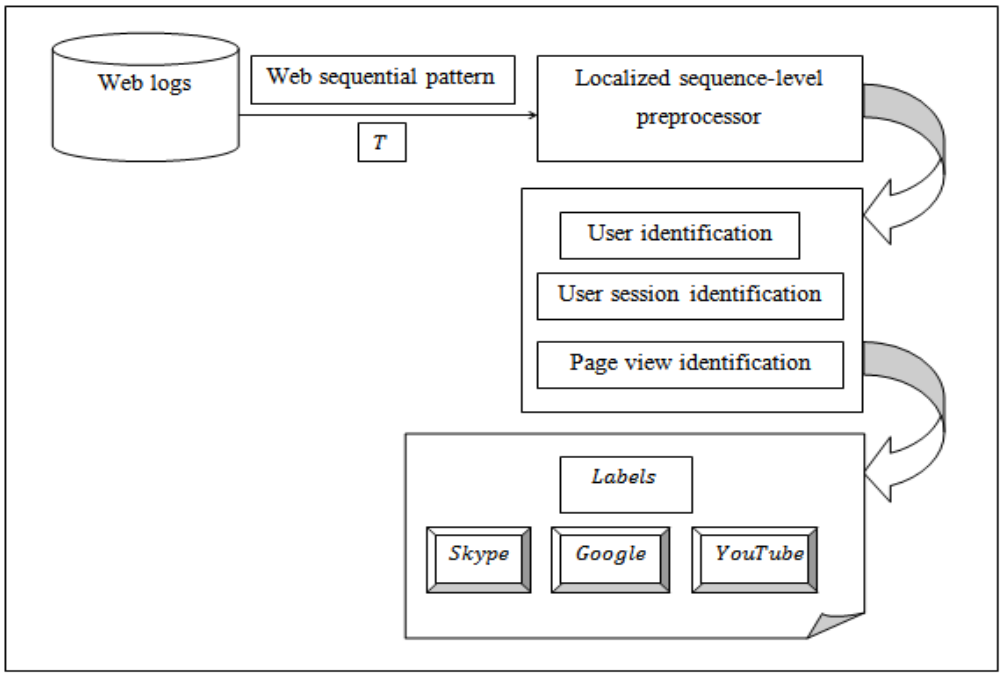

Figure 2 Block diagram of Localized Sequence Level Preprocessing model

As shown in the figure, the web logs are modeled as sequences of events to identify the correlation between web pages on the basis of sequential patterns '$P = p_1, p_2, \dots, p_n$' over a period of time 't', with each sequence represented by set of events 'E'. Each web access sequence '$S = e_1, e_2, \dots, e_n, e_n \in E$' is represented as discrete events corresponding to one of various feasible classes of web pages requested by the user, while 'n' represents the access sequence length. Let us further consider the set of all user '$U = u_1, u_2, \dots, u_n$' that have accessed the web pages '$e_1, e_2, \dots, e_n$', then the log entry is mathematically represented as given below.

$$L_i = \{u_i \in U, t, rs, e_i \in E\} \quad (1)$$

From the above equation (1), each log entry 'L_i' includes, user '$u_i \in U$', access time 't', request status 'rs', whether access provided or denied for the 'i' user and '$e_i \in E$' representing the events recorded from a web site. Besides, the Localized Sequence Level Preprocessing model includes user identification, session identification and page view identification so that it only selects the information of interest.

Here, the user identification refers to the identification of IP address of the corresponding user. Followed by which the session identification is made. Here, the session identification refers to the session address for the corresponding user. The user identification '$User_{ID}$' and session identification '$Session_{ID}$' is mathematically represented as given below.

$$User_{ID} = IP\_Addr(U) \quad (2)$$
$$Session_{ID} = Session\_Addr(U) \quad (3)$$

Finally, the page view identification is made. As far as page view identification is concerned, if the request for the page view, 'p_i' is present in the web log file 'L_i', the log entries corresponding to the embedded sequence of events are removed from 'p_i', and only the request for 'p_i' is kept. On the other hand, if the request for 'p_i' is absent and only some entries for the corresponding sequence of events are present, the entries corresponding to the events are replaced with a request for 'p_i'. Finally, the time of the page view request is mathematically formulated as given below.

$$T(p_i) = MIN\{Time(L_i)\} \quad (4)$$

From the above equation (4), the time of the page view '$T(p_i)$', is obtained from the respective log entry 'L_i' for the sequence 'e_i'. With the above three said factors, the localized sequence-level factors includes the following and is mathematically formulated as given below

$$LS = \{u_i[IPAddr], u_i[SessionAddr], MIN\{Time(L_i)\}\} \quad (5)$$

From the above equation (5), the information of interest are extracted using the three localized sequence-level factors 'LS', namely, IP address 'u_i[IPAddr]', session address 'u_i[SessionAddr]' and the page view 'MIN {Time (L_i)}' respectively. The pseudo code representation of Localized Sequence Preprocessing is given below.

Input: sequential patterns 'P = $p_1, p_2, \dots, p_n$', web access sequence 'S = $e_1, e_2, \dots, e_n$', set of user 'U = $u_1, u_2, \dots, u_n$', access time 't'
Output: Extract information of interest with minimal time
1: Begin 2: **For** each sequential patterns 'P' with each web access sequence 'S' 3: **For** each set of user 'U' 4: Obtain the log entry using equation (1) 5: Obtain user identification using equation (2) 6: Obtain session identification using equation (3) 7: Obtain time of the page view request using equation (4) 8: Obtain localized sequence-level factors using equation (5) 9: **End for** 10: **End for** 11: **End**

Algorithm 1 Localized Sequence Level Preprocessing

As given in the above Localized Sequence Level Preprocessing algorithm, the objective of the algorithm remains in extracting the information of interest with less computation time. With this objective, during the preprocessing stage, the log entries are first grouped upon availability. Besides, with the existing log entries, three factors are derived, namely, user identification, session identification and page view, therefore making the analysts select only the information of interest with minimal time and effort.

1.2 Certainty Hierarchical Forest Cluster model

Followed by the information of interest obtained via Localized Sequence Level Preprocessing algorithm, the second step is to design a clustering model with the objective of extracting unique subclasses per application.

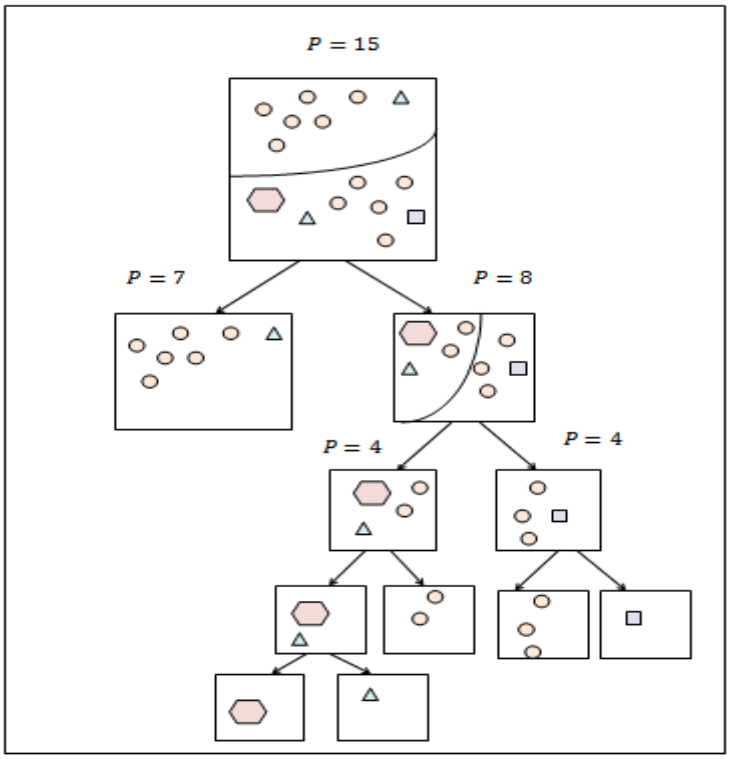

Figure 3 Example of Certainty Hierarchical Forest Cluster with '15' different patterns

With this objective, a forest of cluster hierarchies is designed, where each individual hierarchy can be interpreted as unique subclasses per application. Figure 3 given below, shows an example of Certainty Hierarchical Forest Cluster with '15' dissimilar patterns. As given below, a sample of '15' patterns was cluster

analyzed independently for each application using the computationally efficient Certainty Hierarchical Forest Clustering model (figure 3).

As illustrated in the figure, the distance among two sequence of events is determined by the tree depth at which the sequence of events are divided and the estimated distance of the two sequence of events from their last separating hyperplane. For example, '15' different patterns are extracted at two different time intervals, where, the circle symbol denotes 'YouTube', triangle symbol denotes 'Skype', hexagon symbol denotes 'LinkedIn' and square symbol denotes 'WhatsApp' social networking site respectively. The first tree divided into two hyperplanes. The first hyperplane, represents '6' 'YouTube' applications, '1' 'Skype' application. In a similar manner, the second hyperlane involves '5' 'YouTube' applications, '1' 'LinkedIn' applications, '1' 'WhatsApp' and 'Skype' respectively. In a similar manner, each leaf node is further split to form clusters. Since value of certainty influences the number of clusters or classes per application, hyper parameter is employed to determine certainty factor per application. With the optimal selection of hyper parameter, the maximum within cluster between successive values is chosen according to certainty factors with the objective of arriving optimal cluster number per application. Then, the particular hierarchy or the specific hierarchy for clustering per application is mathematically formulated as given below.

$$H_m(e_1\ e_2) = \begin{cases} 0, & \text{if } H_{mn}(e_1, e_2) \text{is leaf node} \\ C_m(e_1, e_2).\frac{H_{mn}(e_1,e_2)}{N}, & \text{Otherwise} \end{cases} \quad (6)$$

From the above equation (6), '$H_m(e_1\ e_2)$' refers to the particular hierarchy for two sequence of events '$(e_1\ e_2)$', with 'H_{mn}' representing the 'nth' node in 'H_m'. Besides, '$C_m(e_1, e_2)$', represent the certainty factor determined by distance of sequence of events '(e_1, e_2)' and '$H_{mn}(e_1, e_2)$', denoting the lowest node in 'H_m' that contains both 'e_1' and 'e_2' respectively. The certainty factor is evaluated as given below.

$$C_m(e_1, e_2) = \left[\frac{1}{1+\left(\beta\left(E_{mn}(e_1,e_2)(x_{e_1})\right)\right)}\right] - \left[\frac{1}{1+\left(\beta\left(E_{mn}(e_1,e_2)(x_{e_2})\right)\right)}\right] \quad (7)$$

From the above equation (7) the certainty factor 'C_m', for sequence of events '(e_1, e_2)' is derived based on the hyper parameters 'β' that controls sensitivity of 'C'. On the other hand, '$E_{mn}\ (e_1, e_2)$' represents the estimation function at '(e_1, e_2)'. The pseudo code representation of Certainty Hierarchical Forest Cluster is given below.

Input: sequence of events with information of interest '(e_1, e_2)'
Output: Optimal clusters with minimal computational overhead
1: **Begin** 2: **For** each sequence of events with information of interest '(e_1, e_2)' 3: Measure specific hierarchy for clustering per application using equation (6) 4: Measure certainty factor using equation (7) 5: **End for** 6: **End**

Algorithm 2 Certainty Hierarchical Forest Cluster

As given in the above Certainty Hierarchical Forest Clustering algorithm, for each sequence of activities with information of interest, the objective here remains in clustering unique subclasses per application with minimal computational overhead. This is said to be attained by applying hierarchy forest clustering via certainty factor. With this, the hierarchical forest cluster allows us to relax the scalability of the visualization interface [2]. Rather than repeatedly performing hierarchical pattern mining on input pattern, by applying the above algorithm, at each hierarchy node we can optimize visualization interface by using certainty factor. By selecting visualization interface in this way, we can avoid higher latency [1], and thereby better model hierarchical forest clusters with minimal overhead.

1.3 Heuristic Poll Classification model

For a given test datum, MAQUI [1] concerns scalability, i.e., it is highly sensitive to datasets with a large number of event types. In other words, in some scenarios, a small event types may suffice for the classification, whereas in other scenarios, there arises a situation to examine larger event types. Hence, appropriate event types may differ notably. This introduces a trade-off. By posing a comprehensive and appropriately high event types good accuracy may be said to be arrived at. But, the computational cost of determining the event increases for large event type values. Here, higher computational cost minimizes the

scalability for larger pattern types. By keeping a small event types, we get low computational cost, but this may impact accuracy in a negative manner. What is, thus, required is an algorithm that will combine good accuracy and low computational cost, by locally adapting the required value of event types and therefore ensure scalability. In this work, with the optimal clusters generated with minimum computational time and overhead, a Heuristic Classification model based on Poll Volume, called, Heuristic Poll Classification model is presented. Figure 4 given below shows the block diagram of Heuristic Poll Classification model.

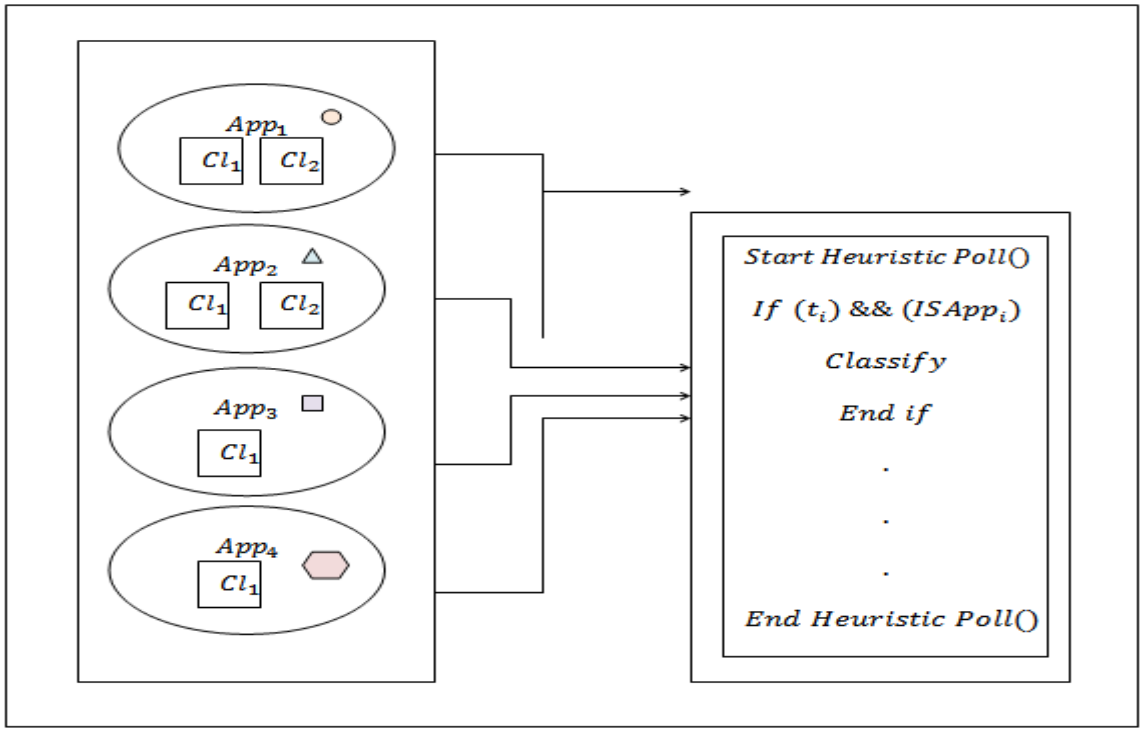

Figure 4 Block diagram of Heuristic Poll Classification

The Heuristic Poll Classification model is depends on the superiority of the major cluster. The superiority is determined by the difference between the volume of polls of major cluster and the volume of polls of all the other clusters. The parameter '$DFactor$' (Distinct Factor) defines the superiority level of the major cluster that must be met to perform a heuristic. The value of '$DFactor$' is arrived at based on the certainty factor. Then, the heuristic poll for classification of event types according to the application and time interval is mathematically formulated as given below.

$$\text{VPPC} > \text{DFactor} \cdot [\sum_{i=1}^{n} \text{VPC}_i - \text{VPPC}] \quad (8)$$

From the above equation (8), the classification is determined according to the volume of polls of major cluster 'VPPC', number of clusters 'n' and 'VPC_i' represents the volume of polls of cluster 'i'.

Input: Application 'App = $\text{App}_1, \text{App}_2, .., \text{App}_i$', Cluster 'Cl = $\text{Cl}_1, \text{Cl}_2, ..., \text{Cl}_i$', number of clusters 'n', volume of polls of major cluster 'VPPC'
Output: User profiles with higher scalability
1: **Begin** 2: **Let** EventCounter = 0 3: **For** each application 'App' with cluster 'Cl' 4: **If** 'VPPC is satisfied' (as given in equation (8)) then 5: Classify the new object 'O' in the cluster where the most event types belong to 6: EventCounter = EventCounter + 1 7: Object 'O' is classified using EventCounter 8: **End if** 9: **End for** 10: **End**

Algorithm 3 Heuristic Poll Classification algorithm

As given in the above algorithm, heuristic poll is used for classification, which does not require a fixed value for the required number of event types. This is achieved by incorporating a heuristic poll into the

classification algorithm. These heuristic poll classification aims at minimizing the computation costs with scalability and maintaining classification accuracy at a high level.

IV. Experimental settings

We choose NASA Kennedy space center's www server in Florida (http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html) [2] that comprises two traces of two month's worth of all HTTP requests to the NASA Kennedy Space Center WWW server in Florida. The web access logs comprises of an ASCII file with one line per request, including the following columns, host making the request, timestamp in the format "DAY MON DD HH:MM:SS YYYY", where DAY is the day of the week, MON is the name of the month, DD is the day of the month, HH:MM:SS is the time of day using a 24-hour clock, and YYYY is the year. It also includes the request given in quotes, HTTP reply code and bytes in the reply. In the experimental data pre-processing stage, we used the approach designed in [3].

The performance of the proposed Hierarchical Forest Cluster and Heuristic Poll Classification (HFC-HPC) method for mining web access logs is compared with the two existing methods, Mining And Querying User Interface (MAQUI) [1] and Novel visualization and interaction [2]. In order to evaluate the performance of web access logs with the objective of addressing scalability, three different parameters are tested with, computational time, computational overhead and error rate.

Computational time here refers to the time taken or time consumed for mining user profiles for the corresponding web access log files.

$$\mathrm{CT} = \sum_{i=1}^{n} U_i * \mathrm{Time}\,[\mathrm{VPPC}] \qquad (9)$$

From the above equation (9), the computational time 'CT' is measured based on the time consumed for obtaining volume of poll of major clusters 'Time [VPPC]' with respect to the users 'U_i'. Computational overhead refers to the memory consumed for mining user profiles with the corresponding web access log files.

$$\mathrm{CO} = \sum_{i=1}^{n} U_i * \mathrm{MEM}\,[\mathrm{VPPC}] \qquad (10)$$

From the above equation (10), the computational overhead 'CO' is measured depends on the memory consumed for obtaining volume of poll of major clusters 'Mem[VPPC]' with respect to the users 'U_i'. Finally, the error rate is measured. The error rate refers to the ratio of similar interests of web users correctly clustered to the overall users accessed.

$$E = \sum_{i=1}^{n} \left[\frac{\mathrm{SICC}}{U_i}\right] * 100 \qquad (11)$$

From the above equation (11), the error rate 'E' is obtained using the number of similar interests of web users correctly clustered 'SICC' to the overall users 'U_i' considered for experimentation.

V. Discussion

To analyze the performance of accessing web logs, three different experiments were conducted using NASA Kennedy space center's (http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html) dataset. Three different experiments conducted were computational time, computational overhead and error rate. Comparison was made with two different event sequence exploration methods, Mining And Querying User Interface (MAQUI) [1] and Novel visualization and interaction [2].

1.4 Scenario 1: Computational time

At first, computational time is measured and comparison made with [1] and [2]. The results of these folds are plotted in figure 5. Figure 5 illustrates the convergence performance of the three different methods. As illustrated in the figure, the general performance of HFC-HPC is the lowest. Since our algorithm can be applied for complex web access log files, improvements on computational time are obvious. Despite parameters being generalized on the NASA dataset, for the proposed and existing methods, results show that the HFC-HPC method still achieves better results than the state-of-the-art method. The sample calculation is provided below.

Sample calculation

- **Proposed HFC-HPC**: With the time consumed for obtaining volume of poll of major clusters being '0.018ms' for single user, the overall computational time for '15' users is given below.

$$\mathrm{CT} = 15 * 0.018\mathrm{ms} = 0.27\mathrm{ms}$$

- **MAQUI**: With the time consumed for obtaining volume of poll of major clusters being '0.025ms' for single user, the overall computational time for '15' users is given below.

$$\mathrm{CT} = 15 * 0.025\mathrm{ms} = 0.375\mathrm{ms}$$

- **Novel visualization and interaction**: With the time consumed for obtaining volume of poll of major clusters being '0.028ms' for single user, the overall computational time for '15' users is given below.

$$\mathrm{CT} = 15 * 0.028\mathrm{ms} = 0.42\mathrm{ms}$$

From the above sample calculations, time consumed for retrieving volume of poll of major clusters using HFC-HPC was found to be '0.018ms', '0.025ms' using MAQUI and '0.028ms' using Novel visualization and interaction respectively. Hence, the overall computational time was found to be '0.27ms', '0.375ms' and '0.42ms' using HFC-HPC, MAQUI and Novel visualization and interaction [2] respectively.

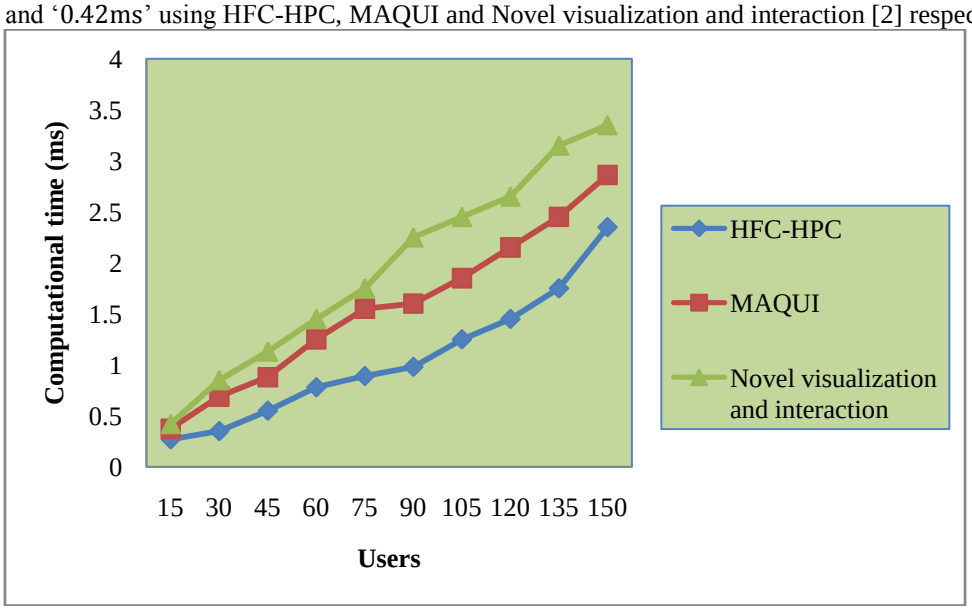

Figure 5 Performance comparison of computational time

Figure 5 given above shows the comparison performance of computational time 150 different users collected at different time intervals. As a result, 150 different users are observed in the x axis and computational time is observed in the y axis. With increase in the number of users the time taken for selecting interesting patterns also increases. As a result, computational time increases with the increase in the number of users. However, it is found to be comparatively better using HFC-HPC method when compared to MAQUI [1] and Novel visualization and interaction [2]. This is because of the inherent extraction of information of interest using the Localized Sequence Level Preprocessing algorithm. By applying the Localized Sequence Level Preprocessing algorithm, three important aspects like, user identification, session identification and page view identification are made in an efficient manner using the respective log entry that in extracts the actual information of interest. As a result, the overall computational time accessing web logs is found to be comparatively lesser using HFC-HPC method by 35% compared to [1] and 47% compared to [2].

1.5 Computational overhead

In the second experiment, computational overhead is measured and comparison is made with the two existing methods [1] and [2]. Figure 6 given below summarizes the results of the experiment to compare the computational overhead performance between the proposed HFC-HPC method and the existing MAQUI [1] and Novel visualization and interaction [2]. The HFC-HPC method performs better than the two existing methods. The sample calculation is provided below with the graphical representation provided in figure 6.

Sample calculation

- **Proposed HFC-HPC**: With the memory consumed for obtaining volume of poll of major clusters being '11KB' for single user, the overall computational overhead for '15' users is given below.

$$\mathrm{CT} = 15 * 11\mathrm{KB} = 165\mathrm{KB}$$

- **MAQUI**: With the time consumed for obtaining volume of poll of major clusters being '13KB' for single user, the overall computational overhead for '15' users is given below.

$$\mathrm{CT} = 15 * 13\mathrm{KB} = 195\mathrm{KB}$$

- **Novel visualization and interaction**: With the time consumed for obtaining volume of poll of major clusters being '16KB' for single user, the overall computational overhead for '15' users is given below.

$$\mathrm{CT} = 15 * 16\mathrm{KB} = 240\mathrm{KB}$$

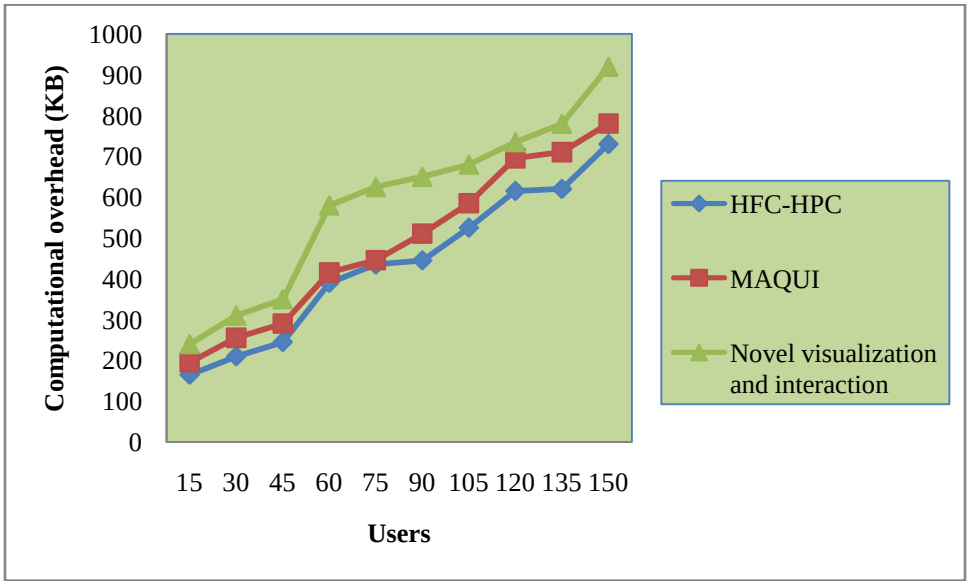

Figure 6 Performance comparison of computational overhead

Figure 6 shows the performance comparison of computational overhead for the NASA datasets as the results of the experiment. In the figure, x axis represents the users and y axis represents the computational overhead. In the case of the NASA dataset, the retrieval of web access logs incurred with '15' users showed the lowest computational overhead. This is because with the increase in the number of users, the web access logs to be access increases and as a result, the computational overhead incurred for the increased number of users also increases. However, the proportionality of increase is found to be lessened by using the HFC-HPC method because of the application of Certainty Hierarchical Forest Clustering algorithm. By applying the Certainty Hierarchical Forest Clustering algorithm, optimal clusters are generated due to the incorporation of certainty factor. With the application of the certainty factor, unique subclasses per application are extracted by considering the estimated distance of the two sequences of events from their last separating hyperplane. Hence, optimal clusters were generated, therefore minimizing the computational overhead incurred during web access logs of higher scalability. The improvement using HFC-HPC method was found to be 11% when compared to [1] and 27% when compared to [2] using NASA dataset.

1.6 Error rate

The experiment results in previous section has HFC-HPC method is more efficient than [1] and [2] in terms of computational overhead. In this section, comparison analysis of error rate is measured with [1] and [2] to illustrate the effectiveness of applying Heuristic Poll Classification algorithm in terms of error rate. The sample calculations are provided below, followed by which the graph convergence is provided.

Sample calculations

- **Proposed HFC-HPC**: With '15' users considered for experimentation and '10' number of dissimilar interests of web users correctly clustered, the overall error rate is given below.

$$\mathrm{E} = \frac{10}{15} * 100 = 66.66\%$$

- **MAQUI**: With '15' users considered for experimentation and '12' number of dissimilar interests of web users correctly clustered, the overall error rate is given below.

$$E = \frac{12}{15} * 100 = 80.33\%$$

- **Novel visualization and interaction**: With '10' users considered for experimentation and '11' number of dissimilar interests of web users correctly clustered, the overall error rate is given below.

$$E = \frac{13}{15} * 100 = 86.66\%$$

From the above sample calculations, when '' users were considered for experimentation, '10' number of dissimilar interests of web users were extracted using the proposed HFC-HPC, '12' number of dissimilar interests of web users were extracted using MAQUI and '13' number of dissimilar interests of web users were extracted using Novel visualization and interaction method. Therefore the error rate was observed to be '66.66%' using HF-HPC, '80.33%' using MAQUI and '86.66%' using Novel visualization and interaction method.

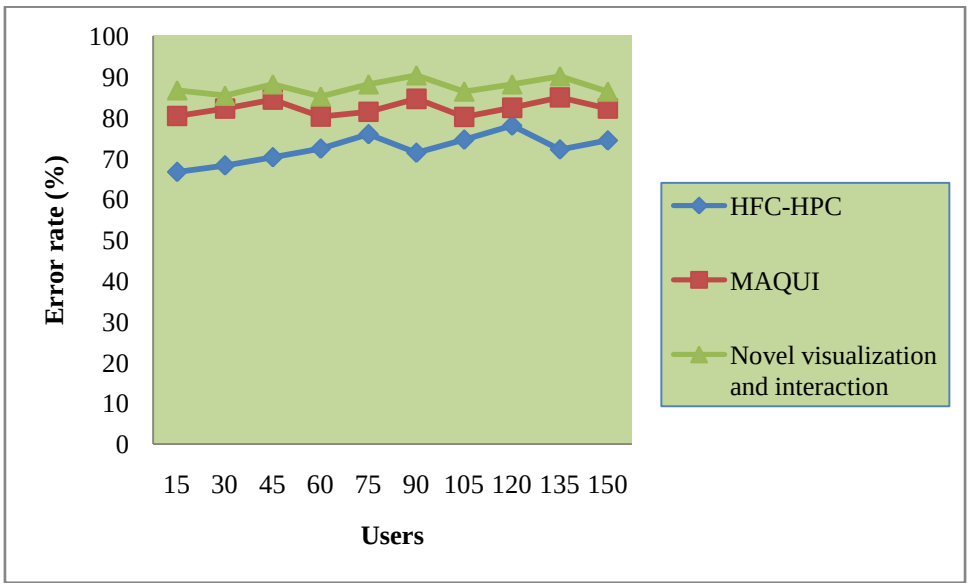

Figure 7 Performance comparison of error rate

The comparison of error rate is provided in the above figure with respect to different users accessing web logs in the range of 15 – 150 at different time intervals. With increase in the size of users, the error rate though not observed in a linear manner. This is because of the different types of web pages requested by different users at different time intervals and of different demographic groups. Different demographical group possess different interests. As a result, the variations are also noted. With the design of Heuristic Poll Classification algorithm, the classification of similar interests with higher scalability of users is addressed or performed via superiority of the major cluster. Here, the value of superiority is observed according to the difference between poll volume major cluster and poll volume of all the other clusters. Accordingly the event counter is also said to be incremented. This in turn reduces the error rate using HFC-HPC method by 12% when compared to [1] and 17% when compared to [2].

VI. Conclusion

HFC-HPC method for mining web access logs has been designed to improve the scalability of visualization interface with minimum computational time and overhead. We adopt localized sequence-level preprocessing model that extracts information of interest with minimal computational time. With maximum information of interest extracted by the analyst with minimal time, the accuracy rate of information extracted is also said to be improved. Then, the Certainty Hierarchical Forest Clustering model is applied to the extracted information to form hierarchical forest, with the objective of generating optimal clusters with minimal computational overhead using certainty factor. Finally, Heuristic Classification model based on Poll Volume is

applied of the hierarchical forest clustered patterns to obtain user profiles with higher scalability rate in a significant manner. Experimental evaluation is conducted to measure the effectiveness of the proposed method and parameter analysis are performed in terms of computational time, computational overhead and error rate with respect to differing user size. Compared to the existing web access log methods, the proposed HFC-HPC method decreases the computational time by 81% and error rate by 29% compared to MAQUI and Novel visualization and interaction.

References

[1]. Po-Ming Law, Zhicheng Liu, Sana Malik, Rahul C. Basole, "MAQUI: Interweaving Queries and Pattern Mining for Recursive Event Sequence Exploration", IEEE Transactions on Visualization and Computer Graphics, Aug 2018 (Mining And Querying User Interface (MAQUI) [1])

[2]. Zhicheng Liu, Yang Wang, Mira Dontcheva, Matthew Hoffman, Seth Walker, Alan Wilson, "Patterns and Sequences: Interactive Exploration of Clickstreams to Understand Common Visitor Paths", IEEE Transactions on Visualization and Computer Graphics (Volume: 23, Issue: 1 , Jan. 2017)

[3]. G. Poornalatha, S. Raghavendra Prakash, "Web sessions clustering using hybrid sequence alignment measure (HSAM)", Social Network Analysis and Mining, Springer, Jun 2013 (data)

[4]. Mohamad Hazim, Nor Badrul Anuar, Mohd Faizal Ab Razak, Nor Aniza Abdullah, "Detecting opinion spams through supervised boosting approach", PLOS ONE | https://doi.org/10.1371/journal.pone.0198884 June 11, 2018

[5]. PavelMasek, Jiri Hosek, Krystof Zeman, Martin Stusek, Dominik Kovac, Petr Cika, JanMasek, Sergey Andreev, and Franz Kröpfl, "Implementation of True IoT Vision: Survey on Enabling Protocols and Hands-On Experience", Hindawi Publishing Corporation, International Journal of Distributed Sensor Networks, Feb 2016

[6]. Ning Zhong, Yuefeng Li, Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining", IEEE Transactions on Knowledge and Data Engineering, VOL. 24, NO. 1, JANUARY 2012

[7]. Sriram Karthik Badam, Zhicheng Liu, Niklas Elmqvist, "Elastic Documents: Coupling Text and Tables through Contextual Visualizations for Enhanced Document Reading", IEEE Transactions on Visualization and Computer Graphics, May 2018

[8]. Anindya Ghose, Panagiotis G. Ipeirotis, "Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics", IEEE Transactions on Knowledge and Data Engineering, VOL. 23, NO. 10, OCTOBER 2011

[9]. Kang Liu, Liheng Xu, Jun Zhao, "Co-extracting Opinion Targets and Opinion Words from Online Reviews Based on the Word Alignment Model", IEEE Transactions on Knowledge and Data Engineering (Volume: 27, Issue: 3 , March 1 2015)

[10]. Marios Belk, Efi Papatheocharous, Panagiotis Germanakos, George Samaras, "Investigating the Relation between Users' Cognitive Style and Web Navigation Behavior with K-means Clustering", International Conference on Conceptual Modeling, Springer, May 2012

[11]. PaoloNesi, GianniPantaleo, MarcoTenti, "Geographical localization of web domains and organization addresses recognition by employing natural language processing, Pattern Matching and clustering", Engineering ApplicationsofArtificial Intelligence, Elsevier, Jun 2016

[12]. Jufeng Yang, Xiaoxiao Sun, Yu-Kun Lai, Liang Zheng and Ming-Ming Cheng, "Recognition from Web Data: A Progressive Filtering Approach", IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE Transactions on Image Processing (Volume: 27, Issue: 11 , Nov. 2018)

[13]. Olatz Arbelaitz, Ibai Gurrutxaga, Aizea Lojo, Javier Muguerza, Jesús Maria Pérez, Iñigo Perona, "Web usage and content mining to extract knowledge for modeling the users of the Bidasoa Turismo website and to adapt it", Expert Systems with Applications, Elsevier, Dec 2013

[14]. Couto, Ayrton Benedito Gaia do, Gomes, Luiz Flavio Autran Monteiro, "Multi-criteria web mining with DRSA", Information Technology and Quantitative Management (ITQM 2016), Elsevier

[15]. Xu Wu, "Context-Aware Cloud Service Selection Model for Mobile Cloud Computing Environments", Hindawi Wireless Communications and Mobile Computing, Mar 2018

[16]. Masoud Reyhani Hamedani, Dongjin Shin, Myeonggeon Lee, Seong-Je Cho, Changha Hwang, "AndroClass: An Effective Method to Classify Android Applications by Applying Deep Neural Networks to Comprehensive Features", Hindawi, Wireless Communications and Mobile Computing, Sep 2018

[17]. Sven Groppe A, Dennis Heinrich A, Christopher Blochwitz B, Thilo Pionteck, "Constructing Large-Scale Semantic Web Indices for the Six RDF Collation Orders", Open Journal of Big Data (OJBD) Volume 2, Issue 1, 2016

[18]. Bakare K. Ayeni , Junaidu B. Sahalu, Kolawole R. Adeyanju, "Detecting Cross-Site Scripting in Web Applications Using Fuzzy Inference System", Hindawi, Journal of Computer Networks and Communications, Aug 2018

[19]. Gianpiero Bianchi, Renato Bruni ,Francesco Scalfati, "Identifying e-Commerce in Enterprises by means of Text Mining and Classification Algorithms", Hindawi, Mathematical Problems in Engineering, Aug 2018

[20]. Ahmad Tasnim Siddiqui, Sultan Aljahdali, "Web Mining Techniques in E-Commerce Applications", International Journal of Computer Applications (0975 – 8887) Volume 69– No.8, May 2013

B.K. Mathan Nagan" Hierarhical Forest Custer and Heuristic Poll Classification For Mining Web Access Logs" International Journal of Engineering Science Invention (IJESI), vol. 08, no. 01, 2019, pp 01-13