
Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2020 SCPE. Volume 21, Issue 2, pp. 189–202, DOI 10.12694:/scpe.v21i2.1645

ENHANCED DBSCAN WITH HIERARCHICAL TREE FOR WEB RULE MINING

NEELIMA GULLIPALLI* AND SIREESHA RODDA†

Abstract. Like other mining, web mining is also necessary to increase the power of web search engine to identify the intended web page and web document. While processing with large datasets, there arises several issues associated with space availability, similarity relationships between different webpage's and running time. Hence, this paper intends to develop an enhanced web mining model based on two contributions. At first, the hierarchical tree is framed, which produces different categories of the searching queries (different web pages). Next, to hierarchical tree model, enhanced Density-Based Spatial Clustering of Applications with Noise (DBSCAN) technique model is developed by modifying the traditional DBSCAN. This technique results in proper session identification from raw data. Moreover, this technique offers the optimal level of clusters necessitated for hierarchical clustering. After hierarchical clustering, the rule mining is adopted. The traditional rule mining technique is generally based on the frequency; however, this paper intends to enhance the traditional rule mining based on utility factor as the second contribution. Hence the proposed model for web rule mining is termed as Enhanced DBSCAN-based Hierarchical Tree (EDBHT). It benefits in providing the search results depending on high level information (e.g., location), so that the ability of search engine in providing the interesting association rules can be improved. Next, to the implementation, the performance of proposed EDBHT is found to be enhanced when compared over several traditional models.

Key words: Rule Mining; Hierarchical Clustering; Searching Behaviour; DBSCAN; A priori Algorithm.

AMS subject classifications. 68M11

1. Introduction. With the increase in usuage of devices [35], weblog analysis software can be used to analyze the server logfile otained from web server and depending on the standards present in the log file,insights into the manner in which pages are accessed , the user accessing the relavent webpages and the duration for which perticular webpage is accessd, are gained [1]. The web server normally generates log files earlier; hence the original data is available in advance. Also, the web server consistently files each deal it makes [2]. Log files include details on visits from search engine spiders. Companies and organizations depend on the corresponding websites to communicate with their clients [3,38]. Maintaining present customers and drawing effective websites thrust such companies, associations,and foundations to come across the striking way to create their websites helpful and capable [4,34,37]. To attain this objective, several reviewing efforts have to be made [5]. These tasks can be done in two random modes. Clients of a particular website could be sought to assess their practice of browsing [5,6]. Subsequently, the performance will be engaged in progressing the construction and/or content depending on the response that is arrived to offer a feedback [2,7,36]. The involuntary navigational account recorded by clients' is also checked up consequently [1,8].

Web Mining [9,10] can be categorized into three diverse categories, based on the types of data to be mined. They include web content mining, web usage mining and web structure mining [11,12]. A lot of web testing equipment subsist however they were restricted,and the effectiveness of such equipments is in a state of excellence [13,14]. Several data mining algorithms have been successfully employed to weblog analysis to better understanding of the user behaviour. Clustering and classification are proving helpful with such issues [3,15,16]. There exist a lot of schemes for producing association rules. A priori algorithm [17] is a well-known and significant approach to discover association rules [5,18]. Improvements obtained from the research that are held over years are integrated to existing web systems for acquiring more successful suggestions.

In addition, data mining approaches have been helpful to deal with sparsity and presentation issues as they were not only dependent on invention assessments but also on various other attributes [19,20]. Hence,

*Associate Professor Vignan's Institute of Information Technology (VIIT), Visakhapatnam (gullipalli.neelima@gmail.com).
†Professor GITAM University, Visakhapatnam

it is essential to discover such algorithms that are considerably perceptive to sparsity for obtaining accurate recommendations [6,21,22]. The algorithm scrutinizes the training dataset once more to construct a frequent pattern tree (FP-tree) [6,23,24]. These trees are well-organized data structures to offer linear time solutions to risky problems in string [2]. A tree for all the data record can be obtained by adding a non-natural parent node in addition to nodes of data record [1,25,26,27]. Decision tree can also provide solution to complex issues in a string [28]. A decision tree is a decision support tool that adopts a tree-like design or graph of decisions and their possible effects, including utility, resource costs,and chance event outcomes [29,30]. Though several methods exist for supporting web mining, still need to be properly addressed.

This paper contributes an improved web mining scheme depending on two contributions. Initially, the hierarchical tree is framed that generates various categories of searching queries. Then, a relevant model is implemented by modifying the conventional clustering method known as DBSCAN that can also be referred as enhanced DBSCAN technique. This model results in appropriate session identification from raw data. From the DBSCAN, the optimal level of clusters can be obtained, which is then provided to hierarchical model. Subsequent to the formation of clusters from hierarchical tree model, the A priori algorithm is adopted from which the interesting association rules can be obtained. The paper is organized as follows. Section II analyses the related works and reviews done on this topic. In addition, section III describes the newly adopted web rule mining model and section IV explains Enhanced DBSCAN-based Hierarchical Tree. Section Vconfirms the results. At last, section VIconcludes the paper.

2. Literature Survey.

2.1. Related works. Sheng sheng Shi et al. [1] have presented a paper based on a novel scheme said to be AutoRM that mines data accounts from a Web page without human intervention. This includes three steps, namely, building the DOM tree of the specified web page, mining the entire groups of neighboring Candidate data Records (C-Records) from the DOM tree, drawing out real data accounts from C-Records. In several pages of web, analogous data accounts are dispersed in larger and neighboring objects that are similar. The results obtained from experiments show that AutoRM is very efficient, and outperforms conventional approaches.

Abdelghani Guerbas et al. [2] has presented a novel algorithm called DBSCAN algorithm to enhance the weblog mining procedure and online navigational pattern assumption. The procedure involves three diverse mechanisms, (1) implementing an advanced time-out dependent heuristic for session detection. (2) Suggestion of using a particular density-dependent technique for navigational pattern discovery. (3) At last, an innovative mechanism for well-organized online as sumption was provided. The performed analysis reveals the significance and usefulness of the presented method.

V. Sujatha et al. [3] presented a novel saliency detection algorithm called Patterns Using Clustering and Classification (PUCC). which offers the assumption from web log data. In the initial step PUCC concerns on partitioning the probable clients in weblog data, and in the subsequent step, clustering is adopted to assemble the effective users with equal attention,and in the third step, the outcomes of categorization and clustering are employed to guess the prospect of requests from users. The analyzed outcomes signified that this mechanism could produce abetter quality of clustering for customer navigation pattern. Rui Liu [4] has presented web-video-mining-supported surgical workflow modeling (webSWM) that offers workflow model with a little cost and labor efficient methods. Also, a testing method for determining the quality of video depending upon study and sentiment investigation methods is produced to choose videos with high-quality from large and noisy videos in the web. Moreover, a numerical learning technique is adopted to construct the workflow representation depending on certain videos. Better out comes confirmed the exactness of the produced SWM and SWM-associated skills.

Mingxing Wu et al. [5] suggested an approach called A priori algorithm that depends on web mining to study the product usability process. This method adopts the enormous online consumer reviews on similar products and characteristics as data basis that is simple to obtain from web and can reproduce the most efficient client opinions on the usage of products. Association rule mining methods are utilized to take out the opinions of consumers about the use of product and its characteristics. Moreover, it is employed for mining organization rules, depending on which a usability valuation technique is offered. Maria N.Morenoet al. [6] presented a paper on MovieLensdatabase that proposes a total structure to agree with certain significant sparsity, scalability, first rater and cold start issues. Even though the structure is represented to movies' suggestion and considered

in the corresponding framework, it can be simply comprehended in various domains. It maintains diverse assumption designs for constructing recommendations based on specific situations. Such designs are induced by data mining approaches that are used as data for input in both product and consumer attributes ordered based on specific domain ontology. Experimental results on a various multispectral data sets and an evaluation with other methods exhibit the strength of the proposed method in identifying objects.

Ziang Li et al. [7] has presented effective Support Vector Regressions (SVR) architecture which is underpinned by a domain ontology that obtains joblessness associated concepts and their dealings to assist the mining of helpful assumption characteristics from related search engine queries. Further, conventional feature selection techniques and data mining designs like neural networks (NN) and SVR are utilized to improve the efficiency of unemployment rate assumption. The results obtained from experiments demonstrate that the presented method performs better than other techniques that were deployed for unemployment rate prediction.

Dirk Thorleuchteret al. [8] presented a paper based on Latent Semantic Indexing (LSI) algorithm that focuses on automated recognition of weak signals. This enhances prevailing knowledge dependent methods, as LSI considers the aspects of meaning and thus, related textual patterns in diverse contexts can be identified. A new weak signal maximization approach is established that replaces the usually adopted prediction modeling in LSI. It measures the numerous related weak signals that are addressed in singular value decomposition (SVD) dimensions. Thus, it is revealed that the proposed method enables organizations to recognize weak signals from the internet for a known suggestion.

2.2. Review. Table 2.1 shows the methods, features, and challenges of conventional techniques based on web block data mining. At first, Auto RM algorithm exhibits certain unique properties that require only one Web page. Further, there is no any necessity for vertically distributed data records,but it requires a area containing two data records and also it could not obtain and align data items between unique data records [1]. Moreover, DBSCAN algorithm is presented where the log file can be accessed without knowing the user's identity. The lacking factor in this algorithm is that it is highly complex and there is no any consideration of combined density-based clustering [2]. PUCC algorithm separates the potential users in weblog data,and it also improves the quality of clustering for user navigation pattern. However, there is no analyzation of performance efficiency,and there are no suggestions in the direction of exploiting association rules for prediction engine [3]. Furthermore, webSWM algorithm solves the knowledge scalability issue in surgical workflowdesign and selects videos from enormous, noisy web videos. Here, there is no consideration of computer-vision-based scheme to fragment the surgical video and moreover, there is no implementation of produced SWM knowledge on the prediction of phase which is said to be a challenge in this paper [4]. On the contrary, A priori algorithm helps modelers to analyze the usability of product features and also it provides decision supports to alleviate FF in product development. However, there are limitations for some unpopular products in the related online reviews,and this method is computationally expensive [5]. In addition, MovieLens database possesses the ability to handle various predictive models for generating suggestions depending on particular circumstances,and it can be extended to various other domains, anyhow, it is more complex due to the application of several procedures [6]. SVR is much suited for obtaining the lowest average RMSE and MAE,and further, it achieves the best prediction performance. However, it should be combined with wrapper-based forward selection for better achievement [7]. LSI algorithm enables an organization to identify weak signals and helps strategic planners to react ahead of time, yet, the occurrence of new weak signals is not visible,and there are no any applications indicating parameter selection procedure [8]. Thus the challenges in various techniques enforce to improve the web mining more effectively in the current work.

3. Newly Contributed Web Rule Mining Model.

3.1. Proposed architecture. The overall architecture of the proposed model is given by Fig. 3.1. for enhancing the ability of search engine in offering the interesting association rules. Initially, the raw data is processed using four processing steps which include data cleaning, user identification, session identification and path completion. Data cleaning is the process of eliminating the irrelevant items from the log file. After cleaning, the cleaned data is forwarded for user identification; there how many users visited the website is identified with the help of IP address. Next to user identification, session identification process takes place. Session is the time between logged in and logged out. This activity is to find the sequence of pages and trace

TABLE 2.1
Review on state-of-the-art of web rule mining techniques

Author [citation]	**Adopted methodology**	**Features**	**Challenges**
Shengsheng Shi et al. [1]	AutoRM algorithm	Requires single Web page as input No need for vertically distributed data records	Requires at least two data records No extraction of data items among unique data records
Abdelghani Guerbas et al. [2]	DBSCAN	Log file can be accessed No necessity to know about user's identity	Highly complex No consideration of combined density-based clustering
V. Sujatha et al. [3]	PUCC	Separates the potential users in weblog data Improves the quality of clustering	No analyzation of effective computation. No investigation in the direction of prediction engine
Rui Liu et al. [4]	webSWM	Resolves the scalability crisis in surgical workflowdesign Chooses high-quality videos	No consideration infragmenting the surgical video No implementation of phase predicted SWM knowledge on
Mingxing Wu et al. [5]	A priori algorithm	Helps researchers to investigate the usability of product characteristics Provides decision supports to improve FF	Limitation for certain unpopular products in the related online reviews Computationally expensive
Maria N.Morenoet al.[6]	MovieLens data base	Extends over various domains easily. Handles various predictive models for on specific situations	Highly complex
Ziang Li et al. [7]	SVR	Achieves the lowest average RMSE and MAE Achieves the best prediction performance	It should be combined with wrapper-based forward selection for better achievement
Dirk Thorleuchteret al. [8]	LSI	Enables an organization to identify weak signals Assists deliberated planners to respond ahead of time	No application of parameter selection procedure. The occurrence of new weak signals is not visible

the user activity. This is because the user may visit many pages at a time. Moreover, the final phase of pre-processing is path completion. The path completion is the process of discovering the users travel pattern. After pre-processing, the processed data is clustered using hierarchical clustering model. In hierarchical clustering, the item sets are first identified and then in order In order to decide which clusters should be combined or where a cluster should be split, a measure of dissimilarity between sets of observations is required. This is achieved by measuring the distance between pairs of observations (Refer: https://en.wikipedia.org/wiki/Hierarchical clustering). The distance matrix is computed followed by the computation of mean and standard deviation. Through this model, the number of clusters has to be determined optimally, for which DBSCAN algorithm is adopted. This is because DBSCAN has the property of grouping together the points that are close to each other based on a distance measurement. In DBSCAN, the cluster computation is performed based on the density reachable points with respect to equilibrium point and minimum points (min points) that results in proving maximum possible clusters. After obtaining the accurate number of clusters from DBSCAN, it is given to hierarchical clustering, and the clustered web pages are obtained with associated level of clustering. Hence, the searching of web pages in each area can became to know. Further, Apriori algorithms help in determining the searching behavior on each web page in each area.

3.2. Notations. The dataset comprises of fields and records, indicated by D_{ij}, where d_{ij}, indicates the data of i^{th}, and j^{th}, elements. The record i can be referred as $i = 1, 2, \ldots, N_r$, in which N_r is the number of records and the fields are denoted by j, that can be referred as $j = 1, 2, \ldots, N_f$, where N_f is the total number of fields. The data of each field belongs to each record i, i.e. $d_j \forall i, I \in A$, includes a set of attributes denoted by A, that can be represented as $a_1, a_2, \ldots, a_n$, where n is the total set of attributes. The spatial information and searching behavior of people in various locations can be obtained as given by the following sections.

4. Enhanced DBSCAN-Based Hierarchical Tree.

4.1. Location Centric Equilibrium Point Estimation. Determining the equilibrium points ϵ is a non-trivial issue even in the nonexistence of uncertainty as it amounts to resolving a system of nonlinear equations.

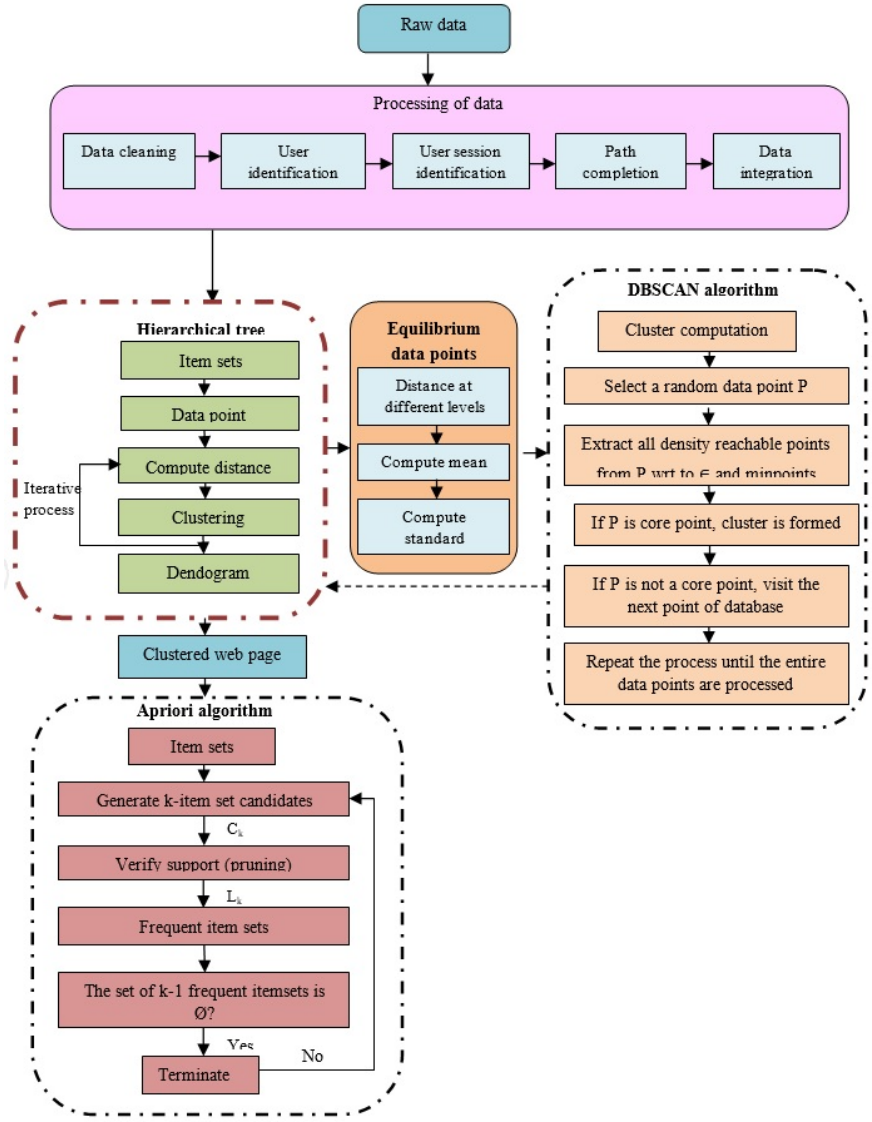

FIG. 3.1. *Overall architecture of the proposed Web rule mining model*

Moreover, the (ϵ) estimation offers the location-based information. In this research work, web transaction, and the IP address is necessary for performing the clustering process. Web transaction is nothing but the web pages from which the corresponding IP address visited.

For example, consider 5 levels with IP address and corresponding web pages as given by Table 4.1. A user from this IP address visited web page 1, web page 2 and so on.

The main steps of hierarchical tree are given below.

Step 1: Distance matrix computation. Let i indicates the records and $j$denotes the fields in the dataset. Here, N_w indicates the total number of web pages. The levels of clustering are denoted as K, where $K = 1, 2, \ldots, N_L$, where N_L indicates the total number of levels.

The distance matrix for K^{th}level is expressed in Eq. (4.1), where d_{ij}^k represents the distance between i^{th} and j^{th} web pages at k^{th} level.

$$DI_k = d_{ij}^{(k)} \tag{4.1}$$

TABLE 4.1
Various Levels Of Web Transactions

Sl.no	IP address	Web transactions
1	199.55.6.7	1 2
2	199.45.4.5	1 2 3
3	199.48.5.8	1 4 5
4	199.36.9.8	1 2
5	199.57.5.6	1 3

TABLE 4.2
Model of Web Transaction for Distance Computation

Sl.no	IP address	Web transactions
1	1	1 2
2	2	1 3
3	3	1 3
4	6	4 2
5	7	3 4

$$d_{ij} = 0 \quad \text{if } i = j \tag{4.2}$$

The number of levels may increase or decrease based on number of web pages, which is given in Eq. (4.3).

$$N_L = N_w - 1 \tag{4.3}$$

Moreover, the number of clusters C_i formed at K^{th} level is also based on number of web pages, as given by Eq. (4.4).

$$N_c^{(k)} = N_w - k \tag{4.4}$$

$$DI_k = \begin{bmatrix} d_{11} & d_{12} & d_{13} & \dots & d_{1N_w} \\ d_{21} & d_{22} & d_{23} & \dots & d_{2N_w} \\ d_{31} & d_{32} & d_{33} & \dots & d_{3N_w} \\ d_{41} & d_{42} & d_{43} & \dots & d_{4N_w} \\ \vdots & \vdots & \vdots & & \vdots \\ d_{N_w1} & d_{N_w2} & d_{N_w3} & \dots & d_{N_wN_w} \end{bmatrix} \tag{4.5}$$

There is a procedure for determining the distance between the IP addresses of various web pages. Let A be the set of web pages, which is searched by different IP addresses. Similarly, B be another set, which is searched by different IP addresses. If the similar web pages are detected in both A and B, i.e. $A \cap B$, then the similar data in both the web transactions have to be discarded, as given by Eq. (4.6) and Eq. (4.7).

$$\overline{A} = A - (A \cap B) \tag{4.6}$$

$$\overline{B} = B - (A \cap B) \tag{4.7}$$

An example for finding the distance between the web pages for various IP addresses is given by Table. 4.1.

Let us assume the web transactions 1 and 2, which can be determined as given in the subsequent steps. For example, let us consider the web address 199.55.10.8 as 1, 199.16.1.0 as 2, 199.55.6.7 as 3, 201.50.60.1 as 6 and 199.55.6.7 as 7. From Table 4.1, the web page transactions from page 1 is (1, 2, 3). Similarly, the web page transaction from web page 2 is (1, 6). The common IP address , found in both the transactions can be eliminated. Thus, the unique address selection for web page 1 can be represented as (2, 3) and the unique

Levels	1	2	3	4	5
1	0	1	0	2.23	2.64
2	1	0	1	3.46	0
3	2	1	0	2.64	3
4	2 .23	3.46	2.64	0	2
5	2.64	0	3	2	0

FIG. 4.1. *Representation of distance matrix*

address selection for web page 2 can be indicated as 6. Consider 2 as 199.16.1.0 and 3 as 199.55.6.7 and 6 as 201.50.60.1. On considering the Euclidean distance between (2, 6), and (3,6) the Eq. (4.8) can be obtained.

$$T = \sqrt{(}199-201)2+(16-50)2+(6-60)2+(0-1)2, \quad T = \sqrt{(}199-201)2+(55-50)2+(6-60)2+(0-1)2 \quad (4.8)$$

Thus the final distance T_d can be evaluated by dividing the distance between (2, 6) and (3, 6) by the average distance as given by Eq. (4.9):

$$T_d = \frac{T(2,6)}{T_A} = \frac{68.13}{61.364}, \quad T_d = \frac{T(3,6)}{T_A} = \frac{54.598}{61.364} \quad (4.9)$$

On assuming 5*5 levels of data points, for the given web pages, the distance matrix can be computed as given by Fig. 4.1.

Step 2: Computation of mean and standard deviation. The mean $\overline{\mu}$ is computed for every field j and is repeated for different levels. Hence at the end, $\mu_1\mu_2\mu_{3NL}$ will be calculated, where μ_1 is the mean of the first level, μ_2 is the mean of the second level andμ_{NL} is the mean of the last level. The overall mean of the entire levels can be computed as given by Eq. (4.10).

$$\overline{\mu} = \frac{1}{N_L}\sum_{k=1}^{N_L}\mu_k \quad (4.10)$$

Similarly, the standard deviation is evaluated for every field j, and then the standard deviation is computed. This overall standard deviation can be determined as given by Eq. (4.11).

$$\overline{\delta} = \sqrt{\frac{\sum_{k=1}^{N_c}(\mu_k - \overline{\mu})^2}{N_L - 1}} \quad (4.11)$$

The computation of mean and standard deviation for the various levels is given by Fig. 4.2. From the Fig. 4.2, the computation of mean is obtained for all the five levels. Fig. 4.2.(a) gives the value of μ_1 as 1.52, similarly, Fig. 4.2.(b) gives the value of μ_2 as 1.41, also, from Fig. 4.2.(c) the mean value of μ_3 can be obtained as 0.33 and from Fig. 4.2.(d), the mean value of μ_4 can be attained as 0.5. On taking the average of all the mean values, $\overline{\mu}$ =0.75 can be achieved. Moreover, the standard deviation can be formulated as given by Eq. (4.11). Accordingly, the standard deviation $\overline{\delta}$ can be obtained as 0.38.

Step 3: Computation of Equilibrium points. The equilibrium point is evaluated further using the overall mean and overall standard deviation. In the proposed architecture, the value of ϵ is based on the hierarchical clustering model, which can be evaluated by summation of $\overline{\mu}$ and $\overline{\delta}$ as shown by Eq. (4.12).

$$\epsilon = \overline{\mu} + \overline{\delta} \quad (4.12)$$

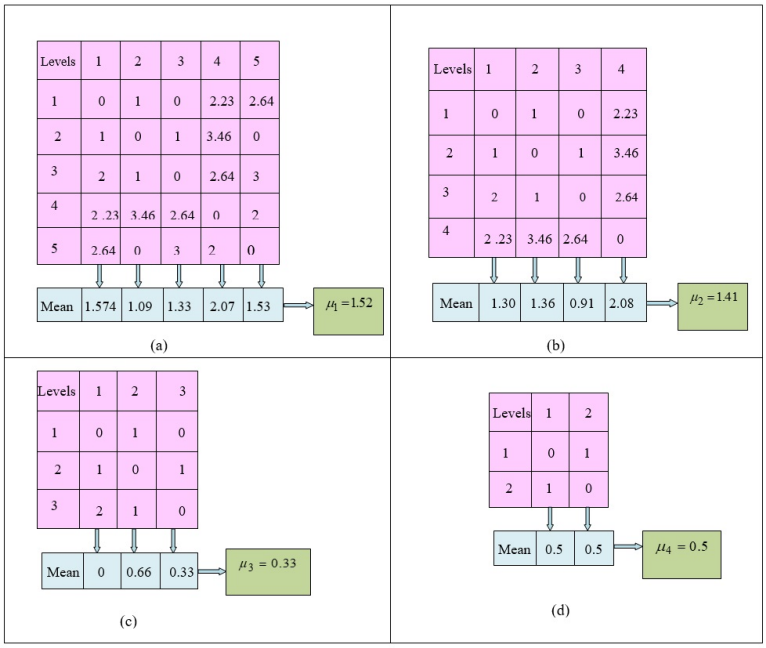

FIG. 4.2. *Representation of mean computation in hierarchical clustering for (a) First low level (b) second low level (c) Third level (d) Fourth level*

Let us consider the field and records as 5×5 where the distance for the entire fields are represented by d_{ij}. Moreover, the mean of each level is given by μ_{C1} to μ_{C5}. Accordingly, the overall mean for all the levels are evaluated and considered as $\overline{\mu}$=0.75. Similarly, the standard deviation of the entire levels are calculated and regarded as $\overline{\delta}$=0.38. Now, the ϵ evaluates to 1.13 based on Eq. (4.12).

4.2. DBSCAN Model. Using DBSCAN, the optimal number of clusters for the entire levels can be obtained, by which hierarchical clustering takes place. DBSCAN algorithm describes the cluster as an area of densely linked points partitioned by areas of non-dense points. If the equality evaluation is considered as Euclidean distance, the region is a hypersphere of radius at the specified point as center denoted by p.

ϵ-neighbourhood: For a point, the ϵ - neighborhood indicates the group of points, where the distance from x $\leq$ ϵ. The cardinality of ϵ-neighbourhood explains the threshold density of x.

ϵ-connected: On considering a pair of points, if x, y ϵ D,if $\|$x-y$\|\leq\epsilon$, then x, y are ϵ-connected points.

In DBSCAN technique, all the points in the dataset would rely on either border point or core point. Moreover, a border point may be density connected point or noise point.

Core point: At this point, the condition threshold dencity $\geq$ min pts is followed.

Border point: At this point, threshold dencity min pts is followed.

Noise point: At this point, p is a noise point if the threshold dencity (p) <min pts and the entire points in the ϵ-neighbourhood of p are border points.

Density-connected point: It is also a border point with a minimum of one core point in itsϵ-neighbourhood. The algorithm for DBSCAN is in Algorithm 1.

4.3. Extracting dendrogram. Based on the number of clusters obtained from the DBSCAN technique, the dendogram is designed from which the desired level can be attained. E.g., consider the number of clusters as 3 that is obtained from DBSCAN. Therefore, the dendogram can be modeled as given by Fig. 4.1, where the level 4 includes one cluster, level 3 includes two clusters, level 2 includes three clusters, and level 1 includes four clusters. From the Fig. 4.1, for = 3, the cluster 1 consists of web pages 1 and 2. Similarly, cluster 2 includes web pages 3 and 4. Also, cluster 3 comprises of web page 5. Therefore, it can be concluded that web page combinations of (1, 2) are visited by users of similar or nearby locations. Also, web page combination of

Algorithm 2: DBSCAN Technique

```
Mark all patterns in D as unvisited
Cluid ← 1
for each unvisted pattern x in D do
    z ← Discover neighbors (x,ϵ, min pts)
if |z|<min pts then
    Point out x as noise
else
    Point out x and every pattern of z with cluid
    Queue-list← every unvisited patterns of z
    repeat
        y← Remove a pattern from Queue-list
        z← Discover neighbors (x,ϵ, min pts)
        if |z |≥ min pts then
            for all pattern do
                if w in z then
                    Point out w with cluid
                if w is unvisited then
                    Queue-list← ∪ Queue-list
        Point out y as visited end until
    until Until Queue-list is empty;
    Point out x as visited cluid←+1
Output: every patterns in D pointed with cluid or noise
```

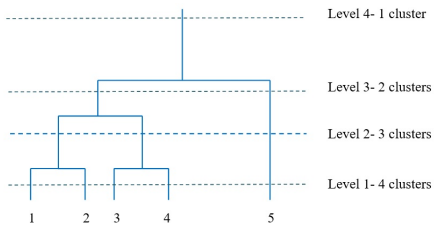

FIG. 4.3. *Representation of dendrograms in Hierarchical clustering*

(3, 4) is visited by users of similar or nearby locations and web page 5 is visited by users of similar or nearby locations. Fig. 4.3 shows the representation of dendogram in hierarchical clustering model.

4.4. A priori algorithm. The spatial distribution of various web pages can be obtained using A priori algorithm. Association rule generation is generally partitioned into two phases [32]. The initial one is the minimum support, which is deployed to discover the entire frequent item sets contained in a database. The next one is the frequent itemsets and the minimum confidence constraint that are exploited to form rules.

The initial phase requires more consideration, but the second phase is straightforward. Discovering the entire frequent itemsets in a database is complex as it entails exploring the entire feasible combinations of items. The set of probable itemsets is the power set over and it includes a size of $2n - 1$. Even though the size of the power setenlarges gradually in the count of items n in i, proficient search is made feasible by means of the downward-closure characteristics of support that assures that, for a frequent itemset, all the corresponding subsets are frequent and similarly, for an infrequent item set, its relative super sets should be infrequent.By deploying this feature, proficient techniques could be able to detect the entire frequent item sets. The pseudo

code of A priori algorithm is shown in Algorithm 2, which can easily determine the searching behavior of clustered web pages.

Algorithm 3: A priori algorithm

```
Data: D is the database and min is the minimum support
Procedure A priori (D, min support)
L_1= frequent items
for k=2; L_k − 1! = Θ; k++ do
    C_k=candidates generated from L_{k−1}
    // that is Cartesian product L_{k−1} * L_{k−1} and k − 1 size item set that is not frequent
    for every transaction t in the database do
        #increment the count of the entire candidates in C_k that are available in t
        L_k= candidates in C_k with min support
Return U_k L_k;
```

4.5. Proposed EDBHT Algorithm. The proposed web rule mining algorithm is shown in Algorithm 3.

Algorithm 4: Proposed EDBHT model

```
Data: Web data D
Result: Web rules R_i, i = 1, 2, ..., N_c
for every level in Hierarchical tree T_H do
    Determine d_ij
    Calculate μ_k
Calculate μ̄ and δ̄ from μ using Eq. (4.10) and Eq. (4.11), respectively
Estimate ε using Eq. (4.12)
Construct T_H
N_L ← DBSCAN(ε, Min point)
Extract cluster from T(K) : k = N_L
for every cluster do
    R_i ← apriori(C_i)
Return R_i
```

5. Results And Discussion.

5.1. Simulation Procedure. The proposed scheme was implemented in JAVA, and the results were obtained. The proposed scheme was executed based on the NASA weblog dataset [33]. The dataset includes two traces that comprises of two month's worth of all HTTP requests to the NASA Kennedy Space Center WWW server in Florida. The proposed EDBHT model was compared with conventional algorithms such as DBHT 1 (ϵ=19), DBHT 2 (ϵ=21), DBHT 1 (ϵ=25) and traditional algorithm and the outcomes were analyzed.

5.2. Spatial Index Analysis. The implemented scheme for extracting the interesting association rules was analyzed based on spatial index analysis. Accordingly, from Table 5.2, for cluster 1, when the support value is 0.05, the proposed model is 2.65% better than DBHT 1, 0.61% better than DBHT 2, and 10.41% better than DBHT 3 techniques. Similarly, from Fig. 5.1.(a), for cluster 1, it can be observed that, when the support value is 0.2, the implemented design is 2.65% superior to DBHT 1, 0.61% superior to DBHT 2, and 10.41% superior to DBHT 3 methods. Also from Fig. 5.1.(b), when the support value is 0.4, the suggested method is 2.65% better than DBHT 1, 0.61% better than DBHT 2 and, 10.41% better than DBHT 3 algorithms.

TABLE 5.1
Spatial Index Computation Of Proposed Over Conventional Web Rule Mining With Respect To Support Value At 0.05

Support value 1=0.05			
Method	**Cluster 1**	**cluster 2**	**cluster 3**
Method	Cluster 1	cluster 2	cluster 3
Proposed	364.658	364.259	365.5082
DBHT 1	374.3258	373.8744	375.1143
DBHT 2	366.9055	370.9248	372.1198
DBHT 3	402.6452	385.5523	381.4983
Traditional	288.9584	-	-

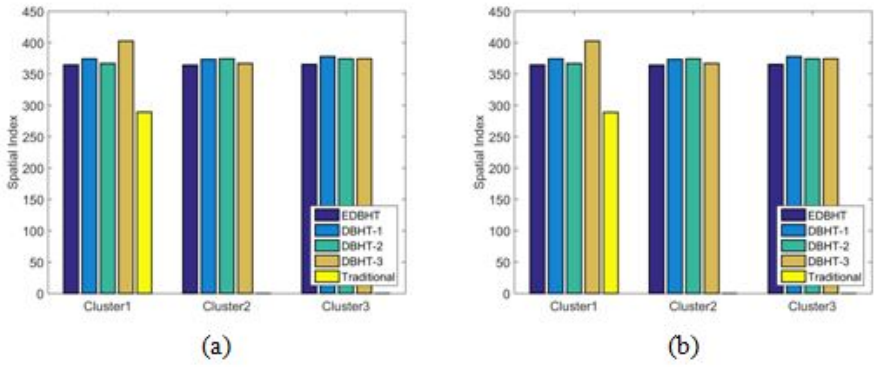

FIG. 5.1. *Spatial index analysis for proposed over conventional rule mining with respect to support value at (a) 0.2 (b) 0.4*

TABLE 5.2
Spatial index computation of proposed over conventional web rule mining with respect to confidence value at 0.5

Confidence value 1=0.5			
Method	**Cluster 1**	**cluster 2**	**cluster 3**
Method	Cluster 1	cluster 2	cluster 3
Proposed	364.658	364.259	365.5082
DBHT 1	374.3258	373.4833	378.0071
DBHT 2	366.9055	374.3258	374.3258
DBHT 3	402.6452	366.9055	374.3258
Traditional	288.7682	-	-

In addition, from Table 5.2, for cluster 1, when the confidence value is at 0.5, the presented model is 2.65% superior to DBHT 1, 0.61% superior to DBHT 2 and, 10.41% superior to DBHT 3 methods. Moreover, from Fig. 5.2.(a), when the confidence value is set at 0.3, the proposed design is 2.65% better than DBHT 1, 0.61% better than DBHT 2, and 10.41% better than DBHT 3 methods. Also, from Fig. 5.2.(b), when the confidence value is fixed at 0.1, the implemented mechanism is 2.65% superior to DBHT 1, 0.61% superior to DBHT 2, and 10.41% superior to DBHT 3 algorithms. The traditional technique was seemed to attain better results in some cases, but the process of clustering was not adopted in the conventional method, and hence the obtained results are not reliable. Thus the superiority of the implemented scheme was verified successfully.

5.3. Frequency rule analysis. The suggested EDBHT method for rule mining was analyzed in terms of the frequency rule analysis for database. From Table 5.3, the when the value of support value is 0.05, the implemented model is 46% superior to DBHT 1, 12% superior to DBHT 2, 56.7% superior to DBHT 3 techniques. Similarly, from Fig. 5.3.(a), it can be observed that, when the support value is 0.2, the implemented design is 46% better than DBHT 1, 12% better than DBHT 2 and, 56.7% better than DBHT 3 methods. Also from Fig. 5.3.(b), when the support value is 0.4, the suggested method is 46% superior to DBHT 1, 12% superior to DBHT 2, 56.7% better than DBHT 3 algorithms.

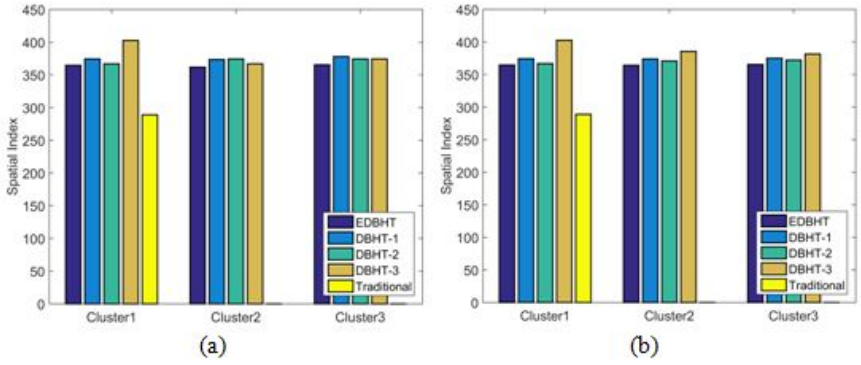

FIG. 5.2. *Spatial index analysis for proposed over conventional rule mining with respect to confidence value at (a) 0.3 (b) 0.1*

TABLE 5.3
Frequency rule computation of proposed over conventional web rule mining with respect to support value at 0.05

Support value 1=0.05			
Method	**Cluster 1**	**cluster 2**	**cluster 3**
Method	Cluster 1	cluster 2	cluster 3
Proposed	0.003	0.001	0.0014
DBHT 1	0.004385	0.004143	0.004525
DBHT 2	0.003364	0.003917	0.004081
DBHT 3	0.004667	0.004043	0.004168
Traditional	0.126708	-	-

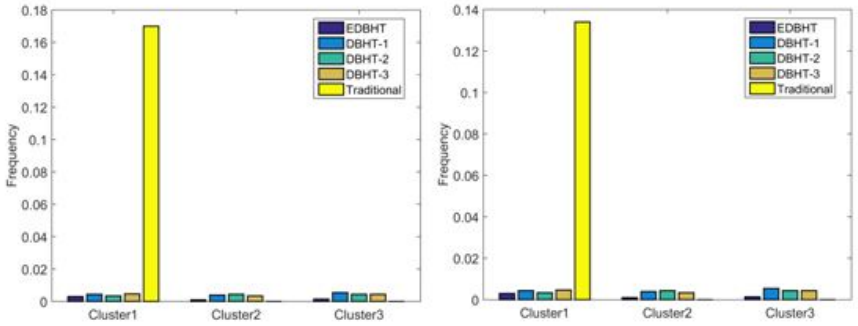

FIG. 5.3. *Frequency rule analysis for proposed over conventional rule mining with respect to confidence value at (a) 0.2 (b) 0.4*

Moreover, from Table 5.4, when the confidence value is at 0.5, the presented model is 46% better than DBHT 1, 12% better than DBHT 2, 56.7% better than DBHT 3 approaches. Moreover, from Fig. 5.4.(a), when the confidence value is set at 0.6, the proposed design is 46% superior to DBHT 1, 012% superior to DBHT 2, 56.7% superior to DBHT 3 techniques. Also, from Fig. 5.4.(b), when the confidence value is fixed at 0.7, the implemented mechanism is 46% better than DBHT 1, and 12% better than DBHT 2, and 56.7% better than DBHT 3 schemes. The compared traditional scheme is seemed to be better, but as clustering technique was not adopted in traditional scheme, the result could not be considered reliable.

6. Conclusions. This paper has presented an enhanced web mining method on the basis of two contributions. Primarily, the hierarchical tree was developed that produced several categories of searching queries. Subsequently, the enhanced DBSCAN technique was adopted from which the required levels of clusters can be attained. Moreover, this scheme results in appropriate session identification from raw data. The level of clusters obtained from DBSCAN was again given to the hierarchical tree model. Following the formation of

TABLE 5.4
Frequency Rule Computation Of Proposed Over Conventional Web Rule Mining With Respect To Confidence Value At 0.5

Confident value 1=0.5			
Method	**Cluster 1**	**cluster 2**	**cluster 3**
Method	Cluster 1	cluster 2	cluster 3
Proposed	0.003	0.001	0.0014
DBHT 1	0.004385	0.003933	0.005417
DBHT 2	0.003364	0.004385	0.004385
DBHT 3	0.004667	0.003364	0.004385
Traditional	0.141667	-	-

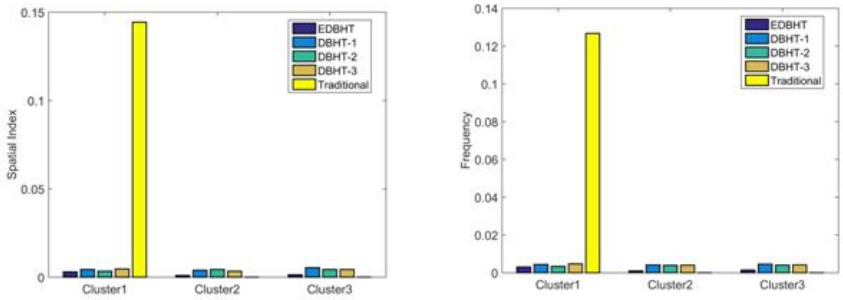

FIG. 5.4. *Frequency rule analysis for proposed over conventional rule mining with respect to confidence value at (a) 0.6 (b) 0.7*

hierarchical tree structure, the enhanced rule mining model was exploited from which the interesting association rules can be attained. Moreover, the proposed technique was compared with the conventional algorithms like DBHT 1, DBHT 2, DBHT 3 and traditional schemes and the results were obtained. From the spatial index analysis, when the support value is 0.05, the suggested model was 2.65% better than DBHT 1, 0.61% better than DBHT 2 and 10.41% better than DBHT 3 techniques. Thus the superiority of the implemented scheme was proved successfully over traditional web rule mining models.

REFERENCES

[1] S. SHI, C.LIU, Y. SHEN, C.YUAN, Y. HUANG,*AutoRM: An effective approach for automatic Web data record mining*, Knowledge-Based Systems, vol. 89, pp. 314-331, November 2015

[2] A. GUERBAS, O. ADDAM, O. ZAAROUR, M. NAGI, R.ALHAJJ,*Effective web log mining and online navigational pattern prediction*, Knowledge-Based Systems,vol.49, pp. 50-62, September 2013.

[3] V. SUJATHA, PUNITHAVALLI, *Improved user Navigation Pattern Prediction Technique from Web Log Data*, Procedia Engineering vol.30, pp. 92-99, 2012

[4] R. LIU, X. ZHANG, H. ZHANG,*Web-video-mining-supported workflow modeling for laparoscopic surgeries*, Artificial Intelligence in Medicine vol. 74, pp. 9-20, November 2016.

[5] M. WU, L. WANG, M. LI, H. LONG , *An approach of product usability evaluation based on Web mining in feature fatigue analysis*, Computers & Industrial Engineering, vol. 75, pp. 230-238, September 2014

[6] M.N. MORENO, S. SEGRERA, V. F. LÓPEZ, M. D. MUÑOZ, Á.L.SÁNCHEZ,, *Web mining based framework for solving usual problems in recommender systems. A case study for movies recommendation*, Neurocomputing, vol. 176, pp. 72-80, 2 February 2016

[7] Z. LI, W. XU, L. ZHANG, R.Y.K. LAU, , *An ontology-based Web mining method for unemployment rate prediction,* Decision Support Systems, vol. 66, pp. 114-122, October 2014.

[8] D.THORLEUCHTER, D. V. DEN POEL, *Weak signal identification with semantic web mining,* Expert Systems with Applications, vol. 40, no. 12, pp. 4978-4985, 15 September 2013.

[9] C. ZAÍN, M. PRATAMA, E. LUGHOFER, S. G. ANAVATTI , *Evolving type-2 web news mining,* Applied Soft Computing, Vol. 54, pp. 200-220, May 2017.

[10] A.B. GAIA DO COUTO, L. F. A. MONTEIRO GOMES, *Multi-criteria Web Mining with DRSA*, Procedia Computer Science, vol. 91, pp. 131-140, 2016.

[11] V. MEDVEDEV, O. KURASOVA, J. BERNATAVIČIENĖ, P.TREIGYS, G.DZEMYDA, *A new web-based solution for modelling data mining processes*, Simulation Modelling Practice and Theory, vol.76, pp. 34-46, August 2017.

[12] J. A. Iglesias, A. Tiemblo, A.Ledezma, A. Sanchis, *Web news mining in an evolving framework*, Information Fusion, vol. 28, pp. 90-98, March 2016.
[13] E. Habibi, S.H. M. Hosseinabadi, *Event-driven web application testing based on model-based mutation testing*, Information and Software Technology, vol. 67, pp. 159-179, November 2015.
[14] F. Simeonov, N. Palov, D. Ivanova, D. Kostova-Lefterova, J. Vassileva , *Web-based platform for patient dose surveys in diagnostic and interventional radiology in Bulgaria: Functionality testing and optimisation,* Physica Medica, 4 May 2017.
[15] H.Tongal, B. Sivakuma , *Cross-entropy clustering framework for catchment classification*, Journal of Hydrology, Vol. 552, pp. 433-446, September 2017.
[16] J. Luo, L. Jiao, R. Shang, F. Liu, , *Learning simultaneous adaptive clustering and classification via MOEA*, Pattern Recognition, vol. 60, pp. 37-50, December 2016.
[17] L. Hanguang, N. Yu , *Intrusion Detection Technology Research Based on Apriori Algorithm*, Physics Procedia, vol. 24, Part C, pp. 1615-1620, 2012.
[18] A. Bhandari, A.Gupta, D. Das, , *Improvised Apriori Algorithm Using Frequent Pattern Tree for Real Time Applications in Data Mining*, Procedia Computer Science, vol. 46, pp. 644-651, 2015.
[19] M.R.F. Coelho, J. M. Sena-Cruz, L.A.C. Neves, M. Pereira, T. Miranda, *Using data mining algorithms to predict the bond strength of NSM FRP systems in concrete*, Construction and Building Materials, vol. 126, pp. 484-495, 15 November 2016.
[20] S. García, J. Luengo, F. Herrera , *Tutorial on practical tips of the most influential data preprocessing algorithms in data mining*, Knowledge-Based Systems, vol. 98, pp. 1-29, 15 April 2016.
[21] D. Apiletti, E. Baralis, T.Cerquitelli, P. Garza, L. Venturini, *Frequent Itemsets Mining for Big Data: A Comparative Analysis*, Big Data Research, 24 August 2017.
[22] F. Alam, R.Mehmood, I. Katib, A.Albeshri, *Analysis of Eight Data Mining Algorithms for Smarter Internet of Things (IoT)*, Procedia Computer Science, vol. 98, pp. 437-442, 2016.
[23] F. Alavi, S. Hashemi, *DFP-SEPSF: A dynamic frequent pattern tree to mine strong emerging patterns in streamwise features*, Engineering Applications of Artificial Intelligence, vol. 37, pp. 54-70, January 2015.
[24] J. Zhang, X. Zhao, S. Zhang, S. Yin, *Interrelation analysis of celestial spectra data using constrained frequent pattern trees*, Knowledge-Based Systems, vol. 41, pp. 77-88, March 2013.
[25] F. M.Bianchi, A. Rizzi, A. Sadeghian, C. Moiso, *Identifying user habits through data mining on call data records*, Engineering Applications of Artificial Intelligence, vol. 54, pp. 49-61, September 2016.
[26] M. Kool, E. Bastiaannet, C.J.H. Van de Velde, P.J. Marang-van de Mheen, *Reliability of self-reported treatment data by breast cancer patients compared with medical record data*, Clinical Breast Cancer, 18 August 2017.
[27] C. Bingen, C. E. Robert, K.Stebel, C. Brühl, S. Pinnock, , *Stratospheric aerosol data records for the climate change initiative: Development, validation and application to chemistry-climate modelling*, Remote Sensing of Environment, 8 July 2017.
[28] K. Kim, J.Hong, *A hybrid decision tree algorithm for mixed numeric and categorical data in regression analysis*, Pattern Recognition Letters, vol. 98, pp. 39-45, 15 October 2017.
[29] A. Saettler, E. Laber, F. d. A. Mello Pereira, *Decision tree classification with bounded number of errors*, Information Processing Letters, vol. 127, pp. 27-31, November 2017.
[30] R. Yan, Z. Ma, Y. Zhao, G. Kokogiannakis, *A decision tree based data-driven diagnostic strategy for air handling units*, Energy and Buildings, vol. 133, pp. 37-45, 1 December 2016.
[31] K. Mahesh Kumar, A. Rama Mohan Reddy, *A fast DBSCAN clustering algorithm by accelerating neighbor searching using Groups method*, Pattern Recognition, vol. 58, pp. 39-48, October 2016.
[32] Y. Djenouri, M. Comuzzi , *Combining Apriori heuristic and bio-inspired algorithms for solving the frequent itemsets mining problem*, Information Sciences, vol. 420, pp. 1-15, December 2017.
[33] *http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html*
[34] G. Singh, V.K. Jain, A. Singh , *Adaptive network architecture and firefly algorithm for biogas heating model aided by photovoltaic thermal greenhouse system,* Journal of Energy Environment, pp.1–25, 2018.
[35] Sherifi and Senja , *Internet Usage On Mobile Devices And Their Impact On Evolution Of Informative Websites In Albania* Vol. 3, no. 6, pp. 37-43, 2018
[36] A. H. Sable and K. C. Jondhale, *Modified Double Bilateral Filter for Sharpness Enhancement and Noise Removal,* 2010 International Conference on Advances in Computer Engineering, Bangalore, 2010, pp. 295-297, doi: 10.1109/ACE.2010.76
[37] R Remmiya and C Abisha , *Artifacts Removal in EEG Signal Using a NARX Model Based CS Learning Algorithm,* Multimedia Research, Vol. 1, no. 1, pp. 1-8.
[38] Q. N. L. Hoang Thuy To, T.N. Phong, DBH Vy , *A hybrid multi criteria decision analysis for engineering project manager evaluation*, International Journal of Advanced and Applied Sciences, 4 (4), 49-52, 2017.

Edited by: P. Vijaya
Received: Nov 10, 2019
Accepted: Apr 1, 2020