# Hilbert Space Clustering based Chronological Backward Search for Effective Web Sequential Pattern Mining

A. P. Selva Prabhu
PhD Research Scholar
Department of Computer Science,
Research and Development centre
Bharathiar University Coimbatore and
Assistant Professor, Bharathiar University
Arts and Science College Sivagiri, Erode, Tamil
Nadu, India

T. Ravi Chandran, PhD
Dean and HoD
Electronics and Communication Engineering
SNS College of Technology, Coimbatore, Tamil
Nadu, India

## ABSTRACT

Web data mining is an important research topic because it attains a significant amount of interest from both academic and industrial environments. Web sequential pattern mining is an imperative for analyzing the access behavior of web users. Recently, few research works have been designed for mining the web sequential patterns. However, performance of existing techniques was not effectual. In order to overcome such limitation, Hilbert Space clustering based Chronological Backward Search (HSC-CBS) Technique is proposed. HSC-CBS Technique is designed in order to improve the performance of web sequential patterns mining. The HSC-CBS Technique at first used Hilbert space clustering in order to group the similar user's interest web patterns in web log database which resulting in improved clustering accuracy. The clustering of frequent web patterns in web log database helps for minimizing the space and time complexity of web sequential pattern mining. After clustering, HSC-CBS Technique applied chronological backward search algorithm in order to efficiently mine the web sequential patterns and improving true positive rate of web sequential pattern mining. The HSC-CBS Technique conducts the experimental works on the parameters such as execution time, space complexity, clustering accuracy, true positive rate of mining and scalability. The experimental results show that the HSC-CBS Technique is able to improve the true positive rate of pattern mining and also reduces the execution time as compared to state of the art works.

## Keywords

Chronological backward search, Hilbert Space clustering, mining, web user**,** web log database, web pattern.

## 1. INTRODUCTION

As the Internet increases rapidly, more and more people go to visit diverse types of websites for obtaining the information they want. Their behavior has an effect on the website. Data Mining is the process of mining practical information from a large repository of data. Web mining is one of the kinds of data mining and specific as the process of discovery and analysis of practical information from the data corresponding to World Wide Web. There are three categories of web mining such as web structure mining, web content mining and web usage mining.

Web sequential pattern mining is a significant way to recognize the access behavior of web users. Web sequential patterns are the frequent access subsequences for providing user-specified minimum support. They describe the most frequent access sequential relationships of the web pages that people visit. By using web sequential patterns, the topology of website is improved, so that users can acquire more information with fewer operations. Besides, web sequential patterns can also help to provide personalized service for the users, which will make the users served better. So web sequential pattern mining has a higher quality view in data mining. An IncWTP algorithm was designed in [1] to discover a fast and efficient incremental mining algorithm based on dynamic characteristics of Web access information. IncWTP algorithm lessens the data storage space and easy access to the required information. But, time taken web sequential pattern mining was more. An automatic annotation approach was presented in [2] for grouping and extracting the same set of web pages from web databases. However, true positive rate of mining efficiency was lower.

A PLWAP tree structure was introduced in [3] to revise the web sequential patterns without scanning the whole database even when previous items become frequent. However, the performance of web sequential pattern mining was not effectual. A PrefixSpan algorithm was developed in [4] that divide the human user sessions into transactions to identify users' meaningful sequential patterns in transaction database with higher mining efficiency. However, space complexity of web sequential pattern mining was remained unsolved. A novel time-constrained sequential pattern mining method was intended in [5] to mine the sequential patterns in which we are interested with minimum execution time. But, efficiency of time-constrained sequential pattern mining was not at required level. An improved approach of Gap-BIDE algorithm was introduced in [6] to mine the user access patterns from web log data. However, time complexity for mining web sequential patterns was higher.

A sequence tree algorithm was presented in [7] for discovering and extracting more number of frequent web sequential patterns through formation of a tree with lower time. A Web Log Mining method was developed in [8] by using multi item sequential pattern based on Plwap for extracting useful patterns from the web log data and web recommendation and personalization. However, the web log mining performance was not effectual. An advanced graph based techniques was presented in [9] to mine the frequent sequential access pattern and predict the navigation behavior of user which will help modifications of web sites or web pages. But, this technique consumes more time to finding the frequent pattern. A Web Log Frequent Sequential Pattern Mining Algorithm was intended in [10] with application of WAP-Tree in order to reduce the execution time and memory consumption of web sequential pattern mining. But, memory utilization was more. In order to solve the above mentioned existing issues, Hilbert Space clustering based Chronological Backward Search (HSC-CBS) Technique is designed. The major contribution of HSC-CBS Technique is formulated as,To improve the performance of web sequential pattern mining, HSC-CBS Technique is developed with application of Hilbert space clustering and chronological backward search algorithm.To group the frequent web pages browsed by user's stored in web log database and thereby reducing the time and space complexity of web sequential pattern mining, Hilbert space clustering algorithm is applied in HSC-CBS Technique.To efficiently mine the clustered web sequential patterns with higher true positive rate, chronological backward search algorithm is employed in HSC-CBS Technique. The chronological backward search algorithm extracts the grouped similar user's interest web patterns in a reverse chronological order of time. The rest of paper is organized as follows: In Section 2, the proposed Hilbert Space clustering based Chronological Backward Search (HSC-CBS) Technique is described with aid of architecture diagram. In Section 3, Experimental settings are presented and the analysis of results is explained in Section 4. Section 5 introduces the background and reviews the related works. Section 6 provides the conclusion of the paper.

# 2. HILBERT SPACE CLUSTERING BASED CHRONOLOGICAL BACKWARD SEARCH (KC-CBS) TECHNIQUE

Web sequential pattern mining is based on web access log. Web access log records the access information of users such as IP address, access time, request URL, referrer, and user-agent and so on. Besides, web log database is typically dynamic. Web log records are generated continuously and user access patterns will change according to the time. For handling such huge amounts of dynamic data, efficient web sequential pattern mining technique is required in order to reducing the time and space complexity. In order to overcome such limitations, Hilbert Space clustering based Chronological Backward Search (HSC-CBS) Technique is introduced. The HSC-CBS Technique is designed to find a fast and efficient web sequential pattern mining based on dynamic characteristics of web access information. The overall architecture diagram of Hilbert Space clustering based Chronological Backward Search (HSC-CBS) Technique for effectual web sequential pattern mining is shown in below Figure 1.

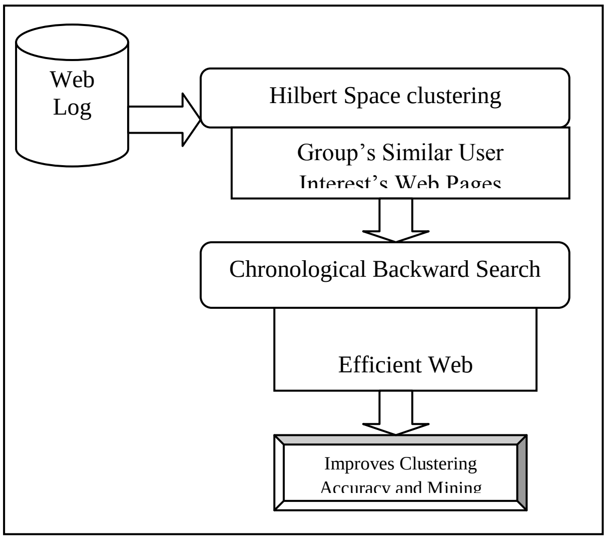
Web Log
Hilbert Space clustering
Group's Similar User Interest's Web Pages
Chronological Backward Search
Efficient Web
Improves Clustering Accuracy and Mining

**Figure 1 Clustering based Chronological Backward Search for Web Sequential Pattern Mining**

As shown in Figure 1, HSC-CBS Technique initially takes web log database as input. Then, HSC-CBS Technique applied Hilbert Space clustering for grouping the similar user's interest web patterns in web log database. The Hilbert Space clustering efficiently cluster the frequent web pages browsed by users in web log database with higher clustering accuracy. This clustering process lessen the time and space of mining the web sequential pattern in a significant manner. After that, HSC-CBS Technique applied Chronological Backward Search algorithm in order to mine the web sequential patterns with higher true positive rate of mining. The elaborate explanation about HSC-CBS Technique is described in forthcoming sub sections.

## 2.1 Hilbert Space based Web Pattern Clustering

The HSC-CBS Technique used Hilbert Space Clustering (HSC) algorithm in order to group the similar user interest's web pages in web log database based on the user query. Web log database is a typically dynamic database. It updates continually all the time. The user access patterns also change with time and therefore changes of user access patterns are very significant for efficient web sequential pattern mining. Generally clustering is the process of grouping which user visits the similar web pages of a web site depends on their relationship. The key objective of clustering is that the webpage's within a group is similar to one another and dissimilar from the web pages in other groups. In HSC-CBS Technique, the user sessions are grouped depends on the order in which web users visit diverse pages of a web site. The user sessions are derived from the host name and time fields. These user sessions are stored in a web log database. The session includes number of web pages visited by a web user in the sequence within a specified time. For instance, a user browsed web pages $wp_1$,$wp_4$, $wp_5$ of a web site in a series, then the session is mathematically expressed as below,

$$s = (wp_1, wp_4, wp_5) \quad (1)$$

From equation (1), $s$ represents the sessions. In order to efficiently group the web access pattern according to session (i.e. time) and thereby improving the performance of web sequential pattern mining, HSC algorithm is employed in HSC-CBS Technique. The HSC algorithm clusters the frequent web pages visited by the web users using the session.

The results of HSC algorithm generate who are interested to use similar type of web pages according to session. The process involved in HSC algorithm for web access pattern clustering is shown in below Figure 2.

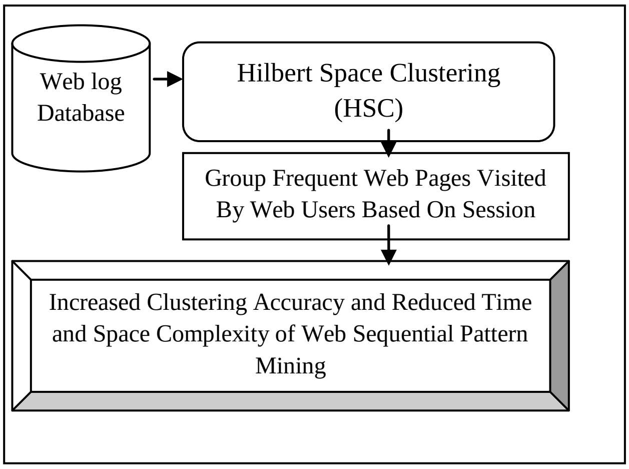
Web log Database
Hilbert Space Clustering (HSC)
Group Frequent Web Pages Visited By Web Users Based On Session
Increased Clustering Accuracy and Reduced Time and Space Complexity of Web Sequential Pattern Mining

**Figure 2 Process of Hilbert Space Clustering For Web Pattern Clustering**

As shown in Figure 2, HSC algorithm significantly clusters the similar user interest's web pages in web log database with respect to session with higher clustering accuracy. Clustering of web access pattern according to session is also helps for minimizing the space and time complexity of web sequential pattern mining in an effective manner. The description of web access pattern is given as, if $I = \{WP_1, WP_2, WP_3, .. WP_n\}$ is the collection of web pages, the access pattern or sequences $S =< wp_1, wp_2, wp_3, ... wp_m >$ ($wp_i \in I$, $1 \le i \le m$) is a series of web pages with respect to the access time in which each page can be visited frequently. The length of the access pattern $S$ is the number of browsing web pages. The following Table 1 shows the user access pattern in web log database.

**Table 1 User Pattern in Web Log Database**

| $TID$ | User Pattern |
|---|---|
| 01 | $wp_1, wp_2, wp_3, wp_5, wp_4$ |
| 02 | $wp_1, wp_2, wp_3, wp_4$ |
| 03 | $wp_3, wp_4, wp_5, wp_1, wp_5$ |
| 04 | $wp_3, wp_4, wp_5, wp_1, wp_2$ |
| 05 | $wp_3, wp_4, wp_1, wp_2$ |
| 06 | $wp_1, wp_2, wp_4, wp_3$ |

As shown in Table 1, an access pattern is the collection of a series of web users' record. Each record comprises $TID$ and user pattern. A user pattern is the sequence of access or visiting web pages and represents the complete browsing behavior. During the Hilbert space clustering process, the similar user interest's web patterns in web log database are grouped according to access time. In Hilbert space clustering, the web access pattern of users is mapped into a high-dimensional Hilbert space. Let us assume the smallest sphere in the Hilbert space which encloses the access patterns of web users along with a session ($s$).

Let us consider the set of web access pattern $wp_1, wp_2, \dots, wp_m \in \{y_i\}$ in high dimensional data space. Mapping function ($\varphi$) is employed in HSC for mapping the access pattern into Hilbert space (H) which is mathematically formulated as,

$$\varphi: y_i \rightarrow H \qquad (2)$$

In HSC, Hilbert curve is used to discover imilar web access patterns in web log database and to form a cluster with high accuracy. Hilbert curve is a continuous path which goes across each data objects (i.e. web access patterns) in a Hilbert space and also presents direct link among the coordinates of the data objects. A Hilbert curve is also termed as Hilbert space-filling curve. The Hilbert space-filling curve of web access patterns for clustering is shown in below Figure 3.

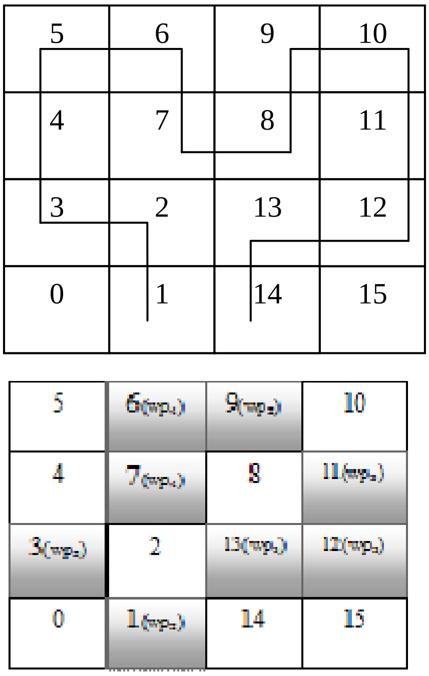

**Figure 3 Hilbert Curve of Order 2 for Web Patterns Clustering**

Figure 3 illustrates the Hilbert curve of second order for clustering similar user interest's web patterns in web log database. In HSC, the web patterns are arranged in rectangular block. The paths of a space filling curve build a linear ordering in the grid points which is determined by beginning at one end of the curve and subsequent path to the other end. As exposed in figure, if two numbers (i.e. web patterns) are continuous in the two dimensional Hilbert space, then web patterns are grouped in one cluster as a similar user interests. If it is not continuous, they are grouped in different cluster. This process is continued for all the rectangular blocks in Hilbert curve. The algorithmic process of Hilbert Space Clustering (HSC) algorithm for grouping the frequent web patterns of web users in web log database along with session is shown in below,

```
// Hilbert Space based Web Pattern Clustering Algorithm
Input: Collection of web access patterns from Amazon Commerce Website, data block B[i] is varied from 0 to n-1.
Output: Improved clustering accuracy with minimum time and space complexity
Step 1:Begin
Step 2:    For each web access patterns from Amazon Commerce Website
Step 3:      Mapping the web patterns into Hilbert space using (2)
Step 4:      For each rectangle box in Hilbert space
Step 5:        if block status B[i]$\neq$ 0 then
Step 6:          if block '$i$' objects and their previous block '$i-1$' objects are continuous then
Step 7:            Group the web patterns in same cluster
Step 8:          else
Step 9:            Group the web patterns in different cluster
Step 10:         End if
Step 11:       End if
Step 12:     End for
Step 13: End for
Step 14: End
```

**Algorithm 1 Hilbert Space based Web Patterns Clustering**

Algorithm 1 demonstrates the step by step process of Hilbert Space based Web Patterns Clustering Algorithm for grouping the similar user interest's web patterns with respect to access time (i.e. session). For each patterns in web log database, initially the mapping of web patterns into Hilbert space is carried out with help of mapping function. After that, the each rectangular box in Hilbert space is ensured if it is empty or not. If the block status is not equal to zero (i.e. contain the web pattern), then it verifies each block and their previous block are continuous or not. If it is continuous, then it is clustered in similar cluster. Otherwise, it clustered into diverse cluster. As a result, the HSC-CBS Technique significantly enhances the clustering accuracy of web sequential pattern mining. The grouping of similar user interest's web patterns according to the session is also supports for HSC-CBS Technique to lessen the space and time complexity of web sequential pattern mining in an effectual manner. After grouping the frequent web pages visited by web users in web log database, HSC-CBS Technique used Chronological Backward Search algorithm for mining the clustered web patterns in a sequential order of their time (i.e. accessing time).

## 2.2 Chronological Backward Search based Web Sequential Pattern Mining Algorithm

In data mining, a search algorithm is employed to mine the information stored within some data structure. Backward search algorithm is to mine the information within some data structure from a goal state (i.e. latest) to initial state (i.e. old). Therefore, HSC-CBS Technique applied Chronological Backward Search algorithm with objective of mining the web sequential patterns in a reverse chorological order of time (i.e. from current to previously visited similar user interest's web patterns). The Chronological Backward Search algorithm initially arranges the clustered web patterns in a chronological order of time. After, Chronological Backward Search algorithm efficiently mines the web sequential patterns with aid of backward search algorithm.

The process involved in Chronological Backward Search Algorithm for web sequential pattern mining is shown in Figure 4. As shown in Figure 4, the chronological backward search algorithm initially takes clustered similar user interest's web patterns as input. Then, chronological backward search algorithm arranges the similar user interest's web patterns in a chronological order of time. After that, chronological backward search algorithm extracts the web sequential patterns by using the concepts of backward search. This helps for HSC-CBS Technique to improve the performance of web sequential mining in an effective manner. Thus, HSC-CBS Technique enhanced the true positive rate of web sequential mining. Let consider the grouped similar user interest's web patterns as $wp_1, wp_2, \ldots, wp_n$ .Thus, the arrangement of clustered web patterns in a chorological order of time is mathematically represented as,

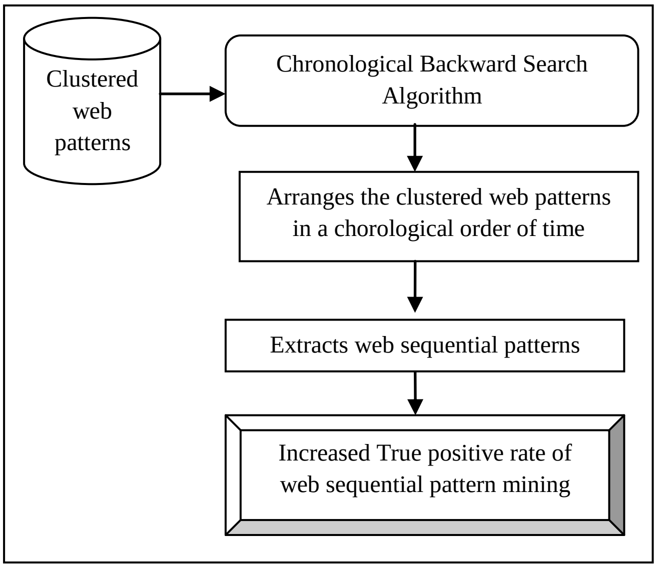
Clustered web patterns
Chronological Backward Search Algorithm
Arranges the clustered web patterns in a chorological order of time
Extracts web sequential patterns
Increased True positive rate of web sequential pattern mining

**Figure 4 process of Chronological Backward Search for Web Sequential Pattern Mining**

$$Y = wp_{t_i}, wp_{t_2}, wp_{t_3}, \ldots wp_{t_n} \quad (3)$$

From equation (3), $t_i$ represents the time at which the web pattern is accessed by user. Next, chronological backward search algorithm employs the concepts of backward search for efficient web sequential mining. The backward search algorithm starts the search at a goal state and work backward until the initial state $y_{initial}$ is attained. In backward search algorithm, considered that there is a single goal state $y_{goal}$. In backward search, a new state $y'$ is determined by using the preceding state $y \in Y$ and action $u \in U(y)$such that,

$$y' = f(y, u) \quad (4)$$

For many problems, it may be preferable to pre-determine a representation of the state transition function $f$ that is "backward" to be consistent with the search algorithm. Some convenient notation used for the backward version of $f$. Let $U^{-1} = \{(y, u) \in Y \times U | y \in Y, u \in U(y)\}$ which denotes the set of all state action pairs and also which considered as the domain$f$. Let consider from a given state $y' \epsilon X$, the set of all $(y, u) \in U^{-1}$ which map to $y'$ using $f$. This is called as backward action space. The backward action space for any $y' \epsilon Y$ is mathematically formulated as,

$$U^{-1}(\ y') = \{(y, u) \epsilon U^{-1} |\ y' = f(y, u)\} \quad (5)$$

Besides for convenience, let us consider $u^{-1}$ represents a state action pair $(y, u)$ that belongs to some $U^{-1}(\ y')$. From any $u^{-1} \in U^{-1}(\ y')$, there is a unique $y \in Y$. Thus, $f^{-1}$ indicates a backward state transition function which yields $x$ from $y'$ and $u^{-1} \in U^{-1}(\ y')$. The backward state transition equation is mathematically expressed as,

$$y = f^{-1}(y', u^{-1}) \qquad (6)$$

From equation (6), the interpretation of $f^{-1}$ is easy to capture in terms of the state transition graph i.e. reverse the direction of every edge. The backward state transition function is the variant of $f$ that is acquired after reversing all of the edges. Each $u^{-1}$ is a reversed edge (i.e. represents a backward search). The state transition graph of backward search for web sequential pattern mining is shown in below Figure 5.

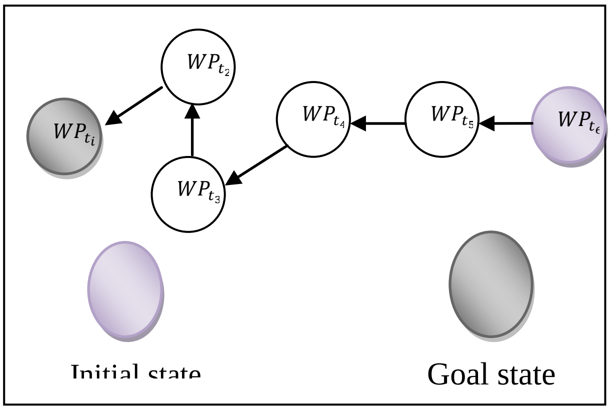

**Figure 5 State Transition Graph of Backward Search for Web Sequential Pattern Mining**

The algorithmic process of Chronological Backward Search for mining the web sequential patterns in shown in below,

```
// Chronological Backward Search based Web Sequential Pattern Mining Algorithm
Input: Clustered web patterns
Output: Improved True Positive rate of mining
Step 1: Begin
Step 3:     Arrange the grouped similar user interest's web patterns in a chorological order of time
            using (3)
Step 4:   If $Q$ is empty, then
Step 5:       $y' \leftarrow Y.GetFirst()$
Step 6:       if $y \neq y_{Initial}$ then
Step 8:         for all $u^{-1} \in U^{-1}(y)$
Step 9:           $y \leftarrow f^{-1}(y', u^{-1})$

Step 10:          if the web pattern in current state ($y$) is extracted, then
Step 12:          Q.Insert($y$)// the web pattern in current state is stored in Q
Step 11:          mark $y$ as visited
Step 13:          end if
Step 14:      end if
Step 15:  end if
Step 16:  Display $Q$
Step 17:end
```

**Algorithm 2 Chronological Backward Search based Web Sequential Pattern Mining**

As shown in above algorithm 2, initially Chronological Backward Search algorithm arranges the clustered web patterns in a chronological order of time. If the queue Q is empty, then Chronological Backward Search algorithm get the web pattern from a goal state (i.e. new state). After obtaining the web pattern, the new state is considered as current state. If the current state is not equal to initial state, then the backward state transition is performed to extract the web patterns. These mined web patterns are stored in queue $Q$. This process is continued until all the state in a graph is visited. Finally, Chronological Backward Search algorithm display Q which comprises of web sequential patterns of similar users interest. As a result, HSC-CBS Technique increases the performance of web sequential mining with higher true positive rate.

# 3. EXPERIMENTAL SETTING

In order to analyze the performance of proposed, Hilbert Space clustering based Chronological Backward Search (HSC-CBS) Technique is implemented in Java language using Amazon Commerce reviews set dataset [21] extracted from UCI repository. The Amazon Commerce reviews set dataset is obtained from the customer's reviews in Amazon Commerce Website. For conducting the experimental work, HSC-CB Technique takes 50 active users (represented by a unique ID and username) who frequently posted reviews in these newsgroups. The number of reviews gathered for each author is 30.The performance of HSC-CBS Technique is compared against with existing automatic annotation approach [1] and IncWTP algorithm [2]. The effectiveness of HSC-CBS Technique is measured in terms of execution time, space complexity, clustering accuracy, true positive rate of mining and scalability.

# 4. RESULT AND DISCUSSION

In this section, the result analysis of HSC-CBS Technique is presented. The efficiency of HSC-CBS Technique is compared against with automatic annotation approach [1] and IncWTP algorithm [2] respectively. The performance of HSC-CBS Technique is analyzed along with the following metrics with the help of tables and graphs.

## 4.1 Measure of Execution Time

In HSC-CBS Technique, Execution Time ($ET$) measures the amount of time taken for extracting the web sequential patterns from web log database. The execution time is measured in terms of milliseconds (ms) and mathematically represented as,

$$ET = N* \qquad (7)$$

*Time (extracting one web sequential pattern)*

From equation (7), time complexity of HSC-CBS Technique is determined. While an Execution Time is lower, the method is said to be more efficient. Table 2 depicts the comparative result analysis of execution time for web sequential pattern mining based on different number of patterns.

**Table 2 Tabulation of Execution Time**

| Number of patterns | Execution Time (ms) | | |
|---|---|---|---|
| | Automatic Annotation Approach | IncWTP algorithm | HSC-CBS Technique |
| 25 | 0.14 | 0.11 | 0.08 |
| 50 | 0.20 | 0.15 | 0.12 |
| 75 | 0.28 | 0.22 | 0.16 |
| 100 | 0.31 | 0.25 | 0.18 |

| | | | |
|---|---|---|---|
| 125 | 0.38 | 0.29 | 0.23 |
| 150 | 0.45 | 0.31 | 0.27 |
| 175 | 0.51 | 0.36 | 0.30 |
| 200 | 0.55 | 0.39 | 0.35 |
| 225 | 0.56 | 0.42 | 0.37 |
| 250 | 0.63 | 0.50 | 0.41 |

The HSC-CBS Technique considers the framework with diverse numbers of web patterns in the range of 25-250 using Java language for conducting experimental works. While considering the 100 web patterns for experimental process, proposed HSC-CBS Technique takes 0.18 ms execution time for mining web sequential patterns whereas automatic annotation approach [1] and IncWTP algorithm [2] takes 0.31 ms and 0.25 ms respectively. Thus, the execution time for web sequential pattern mining using proposed HSC-CBS Technique is lower when compared to other existing works [1], [2].

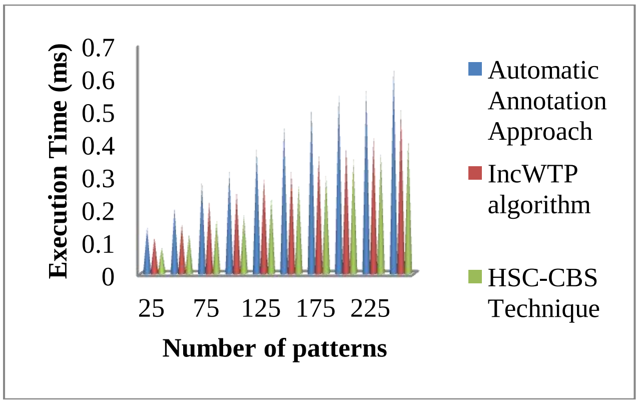

**Figure 6 Measure of Execution Time versus Number of Web Patterns**

Figure 6 explains the impact of execution time for web sequential pattern mining versus diverse number of patterns in the range of 25-250. As shown in figure, the proposed HSC-CBS Technique provides better execution time for mining the sequential patterns from web log database when compared to automatic annotation approach [1] and IncWTP algorithm [2]. Besides, while increasing the number of web patterns for conducting the experiments, the execution time is also gets increased for all three methods. But, comparatively the execution time using proposed HSC-CBS Technique is lower. This is due to application of Hilbert space clustering algorithm and chronological backward search in HSC-CBS Technique. The Hilbert space clustering algorithm groups the frequent web pages visited by user's stored in web log database in order to increase the speed of web sequential pattern mining which resulting in reduced time. Further, chronological backward search efficiently mines the web sequential patterns with minimum time. This helps for HSC-CBS Technique to lessen the execution time in a significant manner. As a result, proposed HSC-CBS Technique minimizes the time of web sequential pattern mining by 39 % and 19% when compared to automatic annotation approach [1] and IncWTP algorithm [2] respectively.

## 4.2 Measurement of Space Complexity

In HSC-CBS Technique, Space Complexity $(SC)$ measures the amount of memory taken for storing the extracted web sequential patterns from web log database The space complexity is measured in terms of kilo bytes (KB) and mathematically expressed as,

$$\mathrm{SC} = \mathrm{N} * memory(storing\ one\ web\ sequential\ pattern) \qquad (8)$$

From equation (8), space complexity of HSC-CBS Technique is computed. While a space complexity is lower, the method is said to be more effectual.

**Table 3 Tabulation of Space Complexity**

| Number of patterns | Space Complexity (KB) | | |
|---|---|---|---|
| | Automatic Annotation Approach | IncWTP algorithm | HSC-CBS Technique |
| 25 | 260 | 215 | 142 |
| 50 | 274 | 238 | 148 |
| 75 | 288 | 241 | 156 |
| 100 | 299 | 256 | 163 |
| 125 | 320 | 268 | 170 |
| 150 | 325 | 279 | 175 |
| 175 | 336 | 284 | 183 |
| 200 | 345 | 305 | 201 |
| 225 | 351 | 320 | 219 |
| 250 | 360 | 331 | 235 |

The comparative result analysis of space complexity for web sequential pattern mining using three methods with respect to different number of patterns in the range of 25-250 is presented in Table 3. While considering the 150 web patterns for experimental process, proposed HSC-CBS Technique acquires 175 KB space complexity whereas automatic annotation approach [1] and IncWTP algorithm [2] acquires 325 KB and 279 KB respectively. Therefore, the space complexity of web sequential pattern mining using proposed HSC-CBS Technique is lower when compared to other existing works [1], [2].

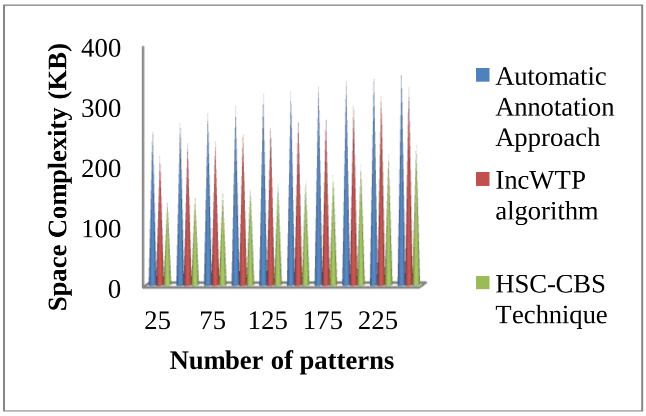

**Figure 7 Measure of Space Complexity versus Number of Web Patterns**

Figure 7 describes the impact of space complexity for web sequential pattern mining versus different number of patterns in the range of 25-250. As demonstrated in figure, the proposed HSC-CBS Technique provides better space complexity for mining the sequential patterns from web log database as compared to automatic annotation approach [1] and IncWTP algorithm [2]. In addition, while increasing the number of web patterns for conducting the experiments, space complexity is also gets increased for all three methods. But, comparatively the space complexity using proposed HSC-CBS Technique is lower. This is owing to application of Hilbert space clustering algorithm in HSC-CBS Technique. Hilbert space clustering algorithm clusters only a similar user interest's web pages in web log database. This helps for reducing the memory space for storing the web patterns in an effectual manner. Hence, proposed HSC-CBS Technique lessens the space complexity of web sequential pattern mining by 44 % and 35 % when compared to automatic annotation approach [1] and IncWTP algorithm [2] respectively.

## 4.3 Measurement of Clustering Accuracy

In HSC-CBS Technique, the clustering accuracy (CA) is defined as ratio of the number of correctly clustered web patterns to the total number of web patterns taken as input. The clustering accuracy is measured in terms of percentages (%) and formulated as,

$$CA = \frac{number\ of\ clustered\ web\ patterns\ as\ similar}{total\ number\ of\ web\ patterns} * 100 \tag{9}$$

From equation (9), clustering accuracy of HSC-CBS Technique is determined with respect to different number of web patterns. While a clustering accuracy is higher, the method is said to be more efficient.

**Table 4 Tabulation of Clustering Accuracy**

| Number of patterns | Clustering Accuracy (%) | | |
|---|---|---|---|
| | Automatic Annotation Approach | IncWTP algorithm | HSC-CBS Technique |
| 25 | 64.46 | 72.53 | 84.21 |
| 50 | 66.14 | 74.16 | 85.68 |
| 75 | 68.47 | 77.15 | 86.98 |
| 100 | 71.02 | 79.65 | 88.12 |
| 125 | 73.56 | 81.02 | 90.12 |
| 150 | 74.15 | 82.98 | 90.92 |
| 175 | 76.98 | 84.02 | 92.01 |
| 200 | 77.16 | 85.36 | 92.65 |
| 225 | 78.93 | 86.74 | 93.56 |
| 250 | 81.06 | 88.56 | 94.86 |

The result analysis of clustering accuracy for web sequential pattern mining using three methods depends on diverse number of patterns in the range of 25-250 is portray in Table 4. While considering the 175 web patterns for carried outing experimental process, proposed HSC-CBS Technique acquires 92.01 % clustering accuracy whereas automatic annotation approach [1] and IncWTP algorithm [2] acquires 76.98 % and 84.02 % respectively. For that reason, the clustering accuracy of web sequential pattern mining using proposed HSC-CBS Technique is higher when compared to other existing works [1], [2].

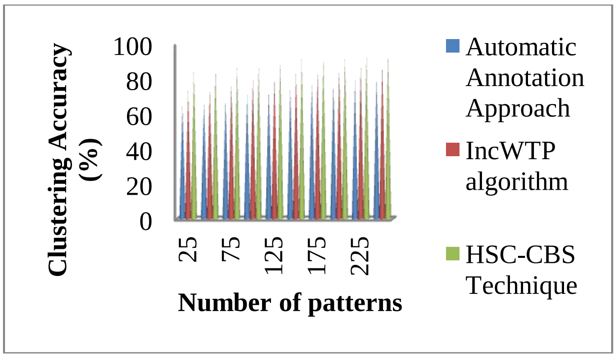

**Figure 8 Measure of Clustering Accuracy versus Number of Web Patterns**

Figure 8 presents the impact of clustering accuracy for web sequential pattern mining versus varied number of patterns in the range of 25-250. As exposed in figure, the proposed HSC-CBS Technique provides better clustering accuracy for mining the sequential patterns from web log database as compared to automatic annotation approach [1] and IncWTP algorithm [2]. As well, while increasing the number of web pattern for conducting the experiments, clustering accuracy is also gets increased for all three methods. But, comparatively the clustering accuracy using proposed HSC-CBS Technique is higher. This is because of application of Hilbert space clustering algorithm in HSC-CBS Technique where the mapping of web patterns into Hilbert space is performed initially with assist of mapping function for each patterns in web log database. After that, then Hilbert space clustering algorithm ensures each block and their previous block are continuous or not. If it is continuous, then it is grouped in similar cluster. Otherwise, it clustered into diverse cluster. This helps for enhancing the clustering accuracy in an effective manner. Thus, proposed HSC-CBS Technique increases the clustering accuracy of web sequential pattern mining by 23 % and 11 % when compared to automatic annotation approach [1] and IncWTP algorithm [2] respectively.

## 4.4 Measurement of True Positive Rate of Mining

In HSC-CBS Technique, True Positive Rate (TPR) computes the ratio of number of correctly mined web access patterns from web log to the total number of web access patterns taken as input. The true positive rate of web sequential pattern mining is measured in terms of percentages (%) and mathematically expressed as,

$$TPR = \frac{correctly\ mined\ web\ patterns\ form\ web\ log}{total\ number\ of\ web\ patterns} * 100 \tag{10}$$

From equation (10), True Positive Rate of HSC-CBS Technique is determined. While a True Positive Rate is higher, the method is said to be more efficient.

**Table 5 Tabulation of True Positive Rate of timing**

| Number of patterns | True Positive Rate of Mining (%) | | |
|---|---|---|---|
| | Automatic Annotation Approach | IncWTP algorithm | HSC-CBS Technique |
| 25 | 56.13 | 64.25 | 76.32 |
| 50 | 57.16 | 66.01 | 77.85 |
| 75 | 59.32 | 67.89 | 79.20 |
| 100 | 62.05 | 70.12 | 81.56 |
| 125 | 63.97 | 71.65 | 83.65 |
| 150 | 66.12 | 73.96 | 84.98 |
| 175 | 69.01 | 77.14 | 87.12 |
| 200 | 70.15 | 79.22 | 88.37 |
| 225 | 72.06 | 81.79 | 91.72 |
| 250 | 75.69 | 84.23 | 93.02 |

The true positive rate of web sequential pattern mining is obtained with respect to various number of patterns in the range of 25-250 using three methods is shown in Table 5. While considering the 200 web patterns for performing experimental work, proposed HSC-CBS Technique obtains 88.37 % clustering accuracy whereas automatic annotation approach [1] and IncWTP algorithm [2] acquires 70.15 % and 79.22 % respectively. Hence, the true positive arte of web sequential pattern mining using proposed HSC-CBS Technique is higher when compared to other existing works [1], [2].

Figure 9 illustrates the impact of true positive rate of web sequential pattern mining versus different number of patterns in the range of 25-250. As revealed in figure, the proposed HSC-CBS Technique provides better true positive rate for mining the sequential patterns from web log database as compared to automatic annotation approach [1] and IncWTP algorithm [2]. Further, while increasing the number of web pattern for conducting the experiments, true positive rate is also gets increased for all three methods. But, comparatively the true positive rate using proposed HSC-CBS Technique is higher.

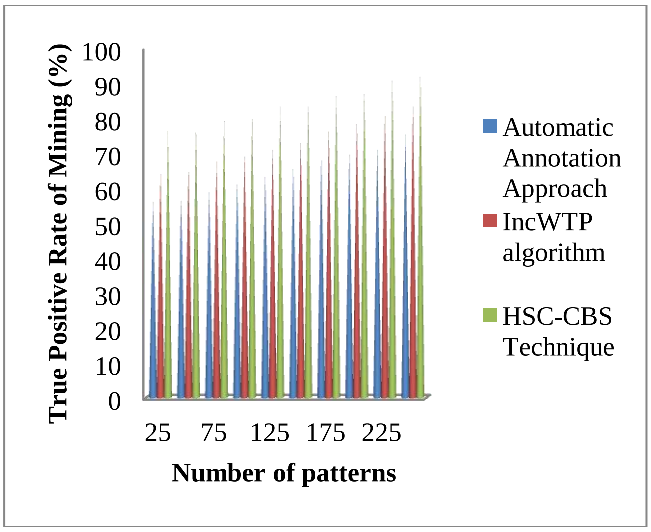

**Figure 9 Measure of True Positive Rate of Mining versus Number of Web Patterns**

This is owing to application of chronological backward search in HSC-CBS Technique. The chronological backward search considerably mines web patterns from a web log database in a reverse chorological order of time. This helps for increasing the true positive rate of web sequential pattern mining in an efficient manner. Therefore, proposed HSC-CBS Technique increases the true positive rate of web sequential pattern mining by 30 % and 15 % when compared to automatic annotation approach [1] and IncWTP algorithm [2] respectively.

## 4.5 Measurement of Scalability

Scalability measure the efficiency of HSC-CBS Technique to handle large number of web patterns in a simultaneous manner for effectual mining the web sequential patterns. While the scalability is higher, the method is said to be more efficient.

**Table 6 Tabulation for Scalability**

| Methods | Scalability (%) |
|---|---|
| **Automatic Annotation Approach** | 69.15 |
| **IncWTP algorithm** | 73.65 |
| **HSC-CBS Technique** | 85.69 |

The scalability of web sequential pattern mining using three methods is illustrated in Table 6. From the table value, it is clear that the scalability rate of web sequential pattern mining using proposed HSC-CBS Technique is higher when compared to other existing automatic annotation approach [1] and IncWTP algorithm [2].

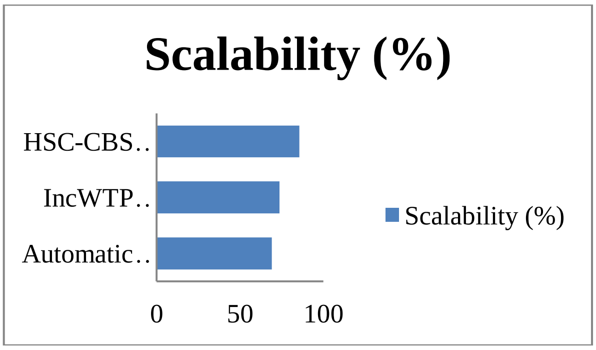

**Figure 10 Measure of Scalability versus Number of Web Patterns**

Figure 10 depicts the impact of scalability for web sequential pattern mining while increasing the input web patterns. As demonstrated in figure, the proposed HSC-CBS Technique provides better scalability for web sequential pattern mining as compared to automatic annotation approach [1] and IncWTP algorithm [2]. Moreover, while increasing the number of web pattern for experimental work, scalability is also gets increased for all three methods. But, comparatively the scalability using proposed HSC-CBS Technique is higher. This is owing to application of Hilbert space clustering algorithm and chronological backward search in HSC-CBS Technique. The Hilbert space clustering algorithm applied in HSC-CBS Technique is provides better scalability while increasing the number of web patterns for sequential mining. Further, chronological backward search is presents better mining efficiency as increasing the number of web sequential patterns mining. This assists for HSC-CBS Technique to increase the scalability of web sequential pattern mining in an effective manner. As a result, proposed HSC-CBS Technique increases the scalability of web sequential pattern mining by 24 % and 16 % when compared to automatic annotation approach [1] and IncWTP algorithm [2] respectively.

## 5. RELATED WORKS

In [11], weighted sequential web access patterns are mined from weighted sequential database to build the recommendation model. An efficient approach was designed in [12] for recognizing the frequent web pages through clustering web users. Though, clustering accuracy of sequential pattern mining was not considered. MapReduce-based web mining was developed in [13] to predict navigation patterns of web users. With the aid of frequent-sequence-pattern-mining algorithms, web users are designed using the programming model of MapReduce. But, space and time complexity was remained unsolved. A novel Sequential Pattern Mining algorithm was introduced in [14] for capturing the list of users to discover the patterns which can be recommended to the users with higher precision. However, due to presence of various user patterns time complexity is higher.

An agglomerative clustering technique was used in [15] to discover users with comparable significance and to establish the inspiration for visiting a website with higher clustering accuracy. But, time taken for pattern mining was more. A new model was designed in [16] that integrate clustering with a new pattern-extraction algorithm for extracting web access patterns under user-defined criterion. However, the performance of web sequential pattern mining was not effectual. A novel web classification algorithm was introduced in [17] with application of fuzzy association rule mining. Here, individual web pages are categorized into various web user categories by using fuzzy association rule and it uses the user sessions for mining. But, execution time was high. A Constraint Programming Approach was developed in [18] for solving the problem of web sequential pattern mining and to handle diverse constraints.

A novel web sequential pattern mining system was designed in [19] that extract most users interesting usage access pattern from the statistical web log data through assigning weights on the activities, interest period and visiting item counts of the user. However, the pattern mining efficiency was not effectual. Pattern-Based Web Mining was intended in [20] with aid of data mining techniques in order to improve the performance of web sequential mining. But, scalability of web sequential mining was remained addressed.

## 6. CONCLUSION

An efficient Hilbert Space clustering based Chronological Backward Search (HSC-CBS) Technique is developed with objective of enhancing the performance of web sequential mining with higher true positive rate. The objective of HSC-CBS Technique is achieved by using the Hilbert Space clustering and Chronological Backward Search algorithm. At first, The HSC-CBS Technique employed Hilbert space clustering with aim of grouping the similar user's interest web patterns in web log database. This process resulting in increased clustering accuracy for mining the web sequential patterns. The grouping of frequent web patterns is also supports for minimizing the space and time complexity of web sequential pattern mining. Finally, HSC-CBS Technique used chronological backward search algorithm for efficiently mining the web sequential patterns and thereby enhancing true positive rate of web sequential pattern mining. The efficiency of HSC-CBS Technique is test with the metrics such as execution time, space complexity, clustering accuracy, true positive rate and scalability. With the experiments carried out for HSC-CBS Technique, it is clear that the true positive rate provides more accurate results for web sequential pattern mining as compared to state-of-the-art works. The experimental results illustrates that HSC-CBS Technique is presents better performance with an enhancement of true positive rate and the reduction of execution time when compared to the state-of-the-art works.

## 7. REFERENCES

[1] Dawei Liu, Saifeng Cai, Xiaohong Guo, "Incremental sequential pattern mining algorithms of Web site access in grid structure database", Neural Computing and Applications, Springer, Volume 28, Issue 3, Pages 575–583, March 2017

[2] Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng, Clement Yu, "Annotating Search Results from Web Databases", IEEE Transactions On Knowledge And Data Engineering, Volume 25, Issue 3, Pages 514 – 527, March 2013

[3] C. I. Ezeife, Yi Liu, "Fast incremental mining of web sequential patterns with PLWAP tree", Data Mining and Knowledge Discovery, Springer, Volume 19, Issue 3, Pages 376–416, December 2009

[4] Jingjun Zhu, Haiyan Wu and Guozhu Gao, "An Efficient Method of Web Sequential Pattern Mining Based on Session Filter and Transaction Identification", Journal Of Networks, Volume 5, Issue 9, Pages 1018-1024, September 2010

[5] Kuo-Wei Hsu, Efficiently and Effectively Mining Time-Constrained Sequential Patterns of Smartphone Application Usage", Hindawi Mobile Information Systems, Volume 2017, Article ID 3689309, Pages 1-18, 2017

[6] Xiuming Yu,Meijing Li, Taewook Kim, Seon-phil Jeong and Keun Ho Ryu, "An Application of Improved Gap-BIDE Algorithm for Discovering Access Patterns", Hindawi Publishing Corporation, Applied Computational Intelligence and Soft Computing, Volume 2012, Article ID 593147, Pages 1-7, 2012

[7] Rajashree Shettar, "Sequential Pattern Mining from Web Log Data", International Journal of Engineering Science & Advanced Technology, Volume 2, Issue 2, Pages 204 – 208, 2012

[8] Jaymin Desai, Risha Tiwari, "Web Log Mining Using Multiitem Sequntial Pattern Based On PLWAP", International Journal For Technological Research In Engineering, Volume 4, Issue 7, Pages 1015-1016, March-2017

[9] Hemraj K. Varachhia, Ankur N. Shah, "A Systematic Review on Mining Web Navigation Pattern Using Graph Based Techniques", International Journal of Scientific & Technology Research Volume 3, Issue 7, Pages 76-79, July 2014

[10] Zhengyu Zhu, Meiyu Zheng, Yihan Wu, "A Web Log Frequent Sequential Pattern Mining Algorithm Linked WAP-Tree", Journal of Software, Volume 10, Number 10, Pages 1228-1234, October 2015

[11] K. Suneetha, M. Usha Rani, "Finding of Weighted Sequential Web Access Patterns for Effective Web Page Recommendations", International Journal of Computer

Science and Technology, Volume 3, Issue 3, Pages 884-888, 2012

[12] Xiuming Yu, Meijing Li, Kyung Ah Kim, Jimoon Chung and Keun Ho Ryu, “Emerging Pattern-Based Clustering of Web Users Utilizing a Simple Page-Linked Graph”, Sustainability, Volume 8, Pages 1-18, 2016

[13] Meijing Li, Xiuming Yu, Keun Ho Ryu, “MapReduce-based web mining for prediction of web-user navigation”, Journal of Information Science, Volume 40, Issue 5, Pages 557-567, 2014

[14] Kuldeep Singh Rathore, Sanjiv Sharma, “Web Personalization Based on Enhanced Web Access Pattern using Sequential Pattern Mining”, International Journal Of Engineering And Computer Science, Volume 5, Issues 7, Pages 17152-17159, 2016

[15] A. Anitha, “An Efficient Agglomerative Clustering Algorithm for Web Navigation Pattern Identification”, Circuits and Systems, Volume 7, Pages 2349-2356, 2016

[16] Oznur Kirmemis Alkan, Pinar Karagoz, “WaPUPS:Web access pattern extraction under user-defined pattern scoring”, Journal of Information Science, Volume 42, Issue 2, Pages 261 – 273, 2015

[22]

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

[18] Amina Kemmar, Yahia Lebbah, Samir Loudni, “A Constraint Programming Approach for Web Log Mining”, International Journal of Information Technology and Web Engineering, Volume 11 Issue 4, Pages 24-42,October 2016

[19] Nu Yin Kyaw, “Creating User Interesting Usage Access Pattern using Statistical Data”, International Journal of Scientific Engineering and Technology Research, Volume 03, Issue 18, Pages: 3695-3700, August-2014

[20] Sheng-Tang Wu and Yuefeng Li, “Pattern-Based Web Mining Using Data Mining Techniques”, International Journal of e-Education, e-Business, e-Management and e-Learning, Volume 3, Issue 2, Pages 163-167, April 2013

[21] Amazon Commerce reviews set Data Set: https://archive.ics.uci.edu/ml/datasets/Amazon+Commerce+reviews+set