Scientific
Research

A Classification Algorithm to Improve the Design of Websites

Hemant Kumar Singh[1], Brijendra Singh[2]

[1]Department of Computer Applications, Azad Institute of Engineering & Technology, Lucknow, India; [2]Department of Computer Science, University of Lucknow, Lucknow, India.
Email: hemantbib@gmail.com, drbri_singh@hotmail.com

ABSTRACT

In very short time today web has become an enormously important tool for communicating ideas, conducting business and entertainment. At the time of navigation, web users leave various records of their action. This vast amount of data can be a useful source of knowledge for predicting user behavior. A refined method is required to carry out this task. Web usages mining (WUM) is the tool designed to do this task. WUM system is used to extract the knowledge based on user behavior during the web navigation. The extracted knowledge can be used for predicting the users' future request when user is browsing the web. In this paper we advanced the online recommender system by using a Longest Common Subsequence (LCS) classification algorithm to classify users' navigation pattern. Classification using the proposed method can improve the accuracy of recommendation and also proposed an algorithm that uses LCS method to know the user behavior for improvement of design of a website.

Keywords: Web Usage Mining; Web Personalization; Recommender Systems; Classification Algorithms

1. Introduction

WUM deals with understanding user behavior in interacting with the web or with a website. One of the aims is to obtain information that may assist web site reorganization or assist site adaptation to better suit the user. Web usage mining model is a kind of mining to server logs and its aim is getting useful users' access information in logs to make sites can perfect themselves with appropriate users' requirements, serve users better and get more economy benefit [1].

Due to the rapid growth rate of the Web, rise of e-commerce, Web services, and Web-based information systems, the volumes of click streams and user data collected by Web-based organizations in their daily operations have reached huge proportions. Substantial increase in the number of Web sites presents a challenging task for Webmasters to organize the content of the Web sites to cater user needs. Modeling and analyzing Web navigation behavior are helpful in understanding the kind of information in demand by online users. The analyzed results from Web navigation behaviors are indispensable knowledge to intelligent online applications and Web based personalization system in improving searching accuracy during information seeking. in traditional personalization techniques [2].

Web site personalization can be defined as the process

of customizing the content and structure of a Web site to the specific and individual needs of each user taking advantage of the user's navigational behaviour. The steps of a Web personalization process include: 1) the collection of Web data; 2) the modelling and categorization of these data (pre-processing phase); 3) the analysis of the collected data and 4) the determination of the actions that should be performed. The ways that are employed in order to analyze the collected data include content-based filtering, collaborative filtering, rule-based filtering and Web usage mining. The site is personalized through the highlighting of existing hyperlinks, the dynamic insertion of new hyperlinks that seem to be of interest for the current user, or even the creation of new index pages [3].

A general web personalization process follows the high level steps shown in the following **Figure 1**.

The WUM prediction process is structured according to two components performed online and off-line with respect
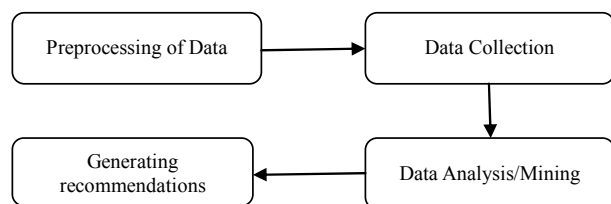

Figure 1. High level taxonomy of web personalization.

to the Web server activity. The off-line component is aimed at building the knowledge base by analyzing historical data, such as server access log files, that is then used in the online component. The main functions carried out by this component are Preprocessing, *i.e.* data cleaning and session identification, and Pattern Discovery, *i.e.* the application of DM techniques, like association rules, sequential patterns, clustering or classification. The online component is devoted to the generation of personalized content. On the basis of the knowledge extracted in the off-line component, it processes a request to the Web server by adding personalized content which can be expressed in several forms, such as links to pages, advertisements, and information relating to products or service estimated to be of interest for the current user [4].

In this paper Section 2 describes the work related with the web personalization. Section 3 illustrates the components of recommender system that is offline and online and work performed by these components. Section 4 explains the longest common sub sequence (LCS) algorithm to web navigation pattern classification. Section 5 explains the proposed website improvement algorithm that used LCS to find the sequence of navigation for website improvement. analyzed navigational patterns of azad Institute of engineering and technology (AzadIET), Lucknow and Section 7 shows the practical execution of the proposed algorithm on azad degree college (www.adc. ac.in) and azad institute of engineering and technology (www.aiet.ac.in) and performance of the algorithm is measured by the finding accuracy and coverage.

2. Related Work

Web usage mining (WUM) refers to automatic discovery and pattern analysis in click streams and associated data collected or generated that from user interactions with Web resources on one or more Web sites. Web usage mining has been used effectively as an approach for automatic personalization and as a way to overcome deficiencies in traditional approaches such as collaborative filtering. The goal of personalization based on Web usage mining is to recommend a set of objects to the active user, possibly consisting of links, ads, text, or products, tailored to user perceived preferences. This task is accomplished by matching the active user session with usage patterns discovered through Web usage mining. M. Jalali, N. Mustapha *et al.* developed a Web-based recommendation system known as Web-ORS for online prediction through Web usage mining system. They also proposed a novel approach for classifying user navigation patterns to predict user future intentions [1,2].

In recent times, the majority of research has been proposed to capture the intuition of web users' preference and their navigation behaviour. A Web personalizer system which provides dynamic recommendation, as a list of hyperlinks to users. In the Web personalizer system analysis is based on user data combined with the structure formed by the hyperlinks of site [5]. Analog is one of the first Web usage mining systems. It consist of online and offline component. The offline component builds session clusters by Intuition Deductive Engine analyzing user navigation pattern recorded in the log file. In the online component part, active user (Intelligent Agent) session is classified according to the generated model. The classification allows identifying the pages matches with the active session and returning Captured List for the requested page with list of suggestion. Users Clustering approach of system affected by several limitation especially scalability and accuracy [6].

Another WUM system called SUGGEST provides useful information to optimize web server performance and make easier the web user navigation. SUGGEST adopts a two level architecture composed by an offline creation of historical knowledge and online component that understands the users' behavior [7].

In summary, All of the work done till date attempt to find the architecture and algorithm to improve accuracy of personalized recommendation, but accuracy still does not meet satisfaction. In our work we proposed a classification algorithm to advance the architecture of online recommender system for improving accuracy of recommendation for users.

3. Different Components of Online Recommender System Architecture

According the functionality of the system we have divided the architecture of online recommender system in two phases one online phase and another is offline phase. Bothe the phases are non-overlapping and work collectively. The first one is clustering the filtered sessionized page views into clusters of similar page views. The second step is learning the profiles based on the preformed clusters. The online component is responsible for matching the new user request (current active session) to the profile shares common interests to the user [8].

In the Back End Phase shown in **Figure 2** there are two main module, Data pretreatment and user navigation mining. The main module of the Front End phase shown in **Figure 3** is Prediction Engine.

3.1. Offline Phase

3.1.1. Data Pretreatment
The data pretreatment is to reduce the redundancy of data and eliminate the data imprecise that is of no use for analysis purpose like audio, vedio files and images etc and reformat the original Web user log files to identify all user sessions. The typical web server log consist of following
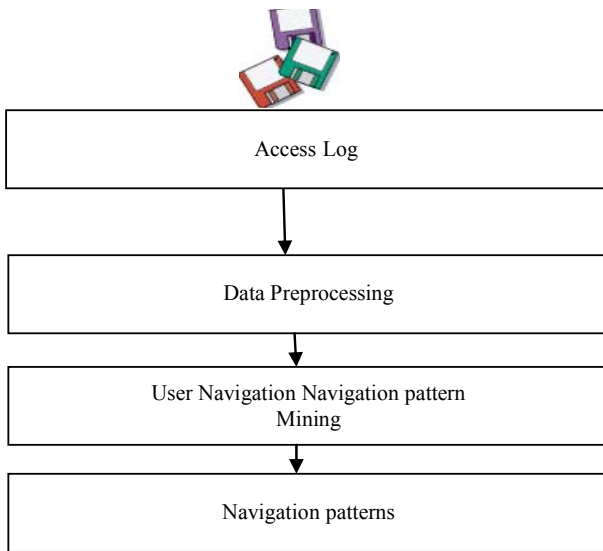
Access Log

↓

Data Preprocessing

↓

User Navigation Navigation pattern Mining

↓

Navigation patterns

Figure 2. Offline phase.

Navigation patterns

↓

Recommendation engine (using LCS Algorithm)
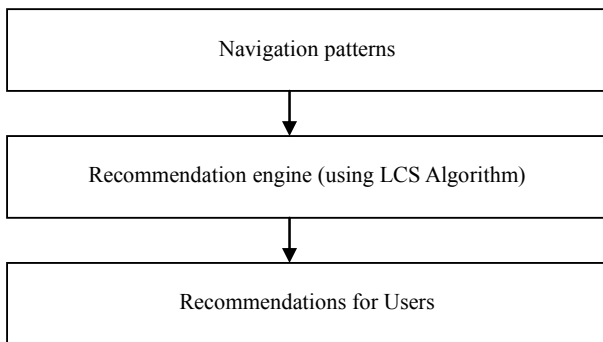
↓

Recommendations for Users

Figure 3. Online phase.

fields: date/time/c-ip/cs-username/s-sitename/s-computername/s-ip/s-port/cs-method/cs-uri-stem/csuri-query/sc-status/time-taken/cs-version/cs-host/cs(User-Agent)/cs(Referer).

Furthermore, numerous pretreatment tasks need to be done before applying navigation pattern mining algorithms on the Web user log files. The tasks include data collection, data cleaning, user identification, and session identification and path completion. The phases of data pretreatment are shown in following **Figure 4**.

3.1.2. Navigation Pattern Mining
After performing the data pretreatment step on the access log, navigation pattern mining is performed on the resultant user access sessions. Now by applying efficient clustering algorithms the clusters of sessions are formed. Clustering of user navigation pattern aims to group sessions into clusters based on their common properties. Access sessions that are obtained by the clustering process are actual patterns of Web user activities. These patterns will be used to further classify current user activities in online phase of the WebORS. In this study, user navigation
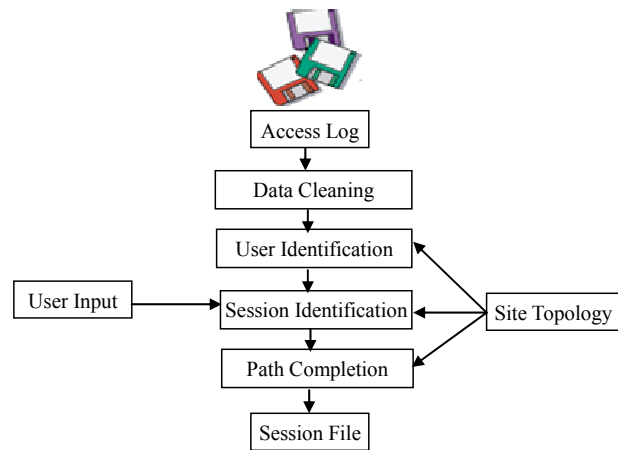

Access Log

↓

Data Cleaning

↓

User Identification

↓

User Input → Session Identification ← Site Topology

↓

Path Completion

↓

Session File

Figure 4. Phases of data preprocessing.

patterns are defined as follows:

We used following clustering algorithm to model navigational patterns in which undirected graph is generated. Web pages are used as nodes of the graph and link between the web pages is used as edges of the graph. The degree of connectivity in each pair of pages depends on two main factors: the time position of two pages in a session *i.e.* time connectivity and the occurrence of two pages in a session is the frequency. *Time Connectivity* measures the degree of visit ordering for each two pages in a session that is calculated by the following formula:

$$T_{ab} = \frac{\sum_{i=1}^{n} \frac{T_i}{T_{ab}} * \frac{f_a(k)}{f_b(k)}}{\sum_{i=1}^{n} \frac{T_i}{T_{ab}}}$$

where T_i is time duration in i-th session, that containing both pages a and b, T_{ab} is the difference between requested time of page a and page b in the session. We consider $f(k) = k$ if web page appears in position k.

Frequency measures the occurrence of two pages in each session. That is calculated by the following formula

$$F_{ab} = \frac{N_{ab}}{\max\{N_a, N_b\}}$$

where N_{ab} is the number of sessions containing both page a and page b. N_a and N_b are the number of session containing only page a and page b. and this formula also has the values between 0 and 1.

In our approach, *Time Connectivity* and *Frequency* plays a major role for finding the degree of connectivity for each pair of web pages and having same importance. We use the harmonic mean of *Time Connectivity* and *Frequency* to represent the connectivity of each two web pages, shown as below. We take this formula for weight of each edge in the undirected graph by the following formula.

$$W_{ab} = \frac{2 * TC_{ab} * FC_{ab}}{TC_{ab} + FC_{ab}}$$

The data structure can be used to store the weights is an adjacency matrix M where each entry M_{ab} contains the value W_{ab} computed according to above given formula. To limit the number of edge in such graph ,element of M_{ab} whose value is less than a threshold are to little correlated and thus removed.

This threshold is named as Minimum Frequency in this contribution [2].

3.1.3. Clustering

Clustering aims to group sessions into clusters based on their common properties. Standard clustering algorithm can partition user access sessions. The result of session clustering is used to represent the set of user navigation patterns.

Traditional collaborative filtering techniques are often based on matching the current user's profile against clusters of similar profiles obtained by the system over time from other users. A similar technique can be used in the context of Web personalization by first clustering user transactions identified in the preprocessing stage. However, in contrast to collaborative filtering, clustering user transactions based on mined information from access logs does not require explicit ratings or interaction with users. In our case, user transactions are mapped into a multidimensional space as vectors of URL references. Standard clustering algorithms generally partition this space into groups of items that are close to each other based on a measure of distance. In the case of Web transactions, each cluster represents a group of transactions that are similar based on co-occurrence patterns of URL references. We apply graph partitioning algorithm to finds groups of strongly correlated pages by partitioning the graph according to its connected components. It will be used for classification based on proposed algorithm. The following **Figure 5** shows the User Navigation pattern Mining.

3.2. Online Phase

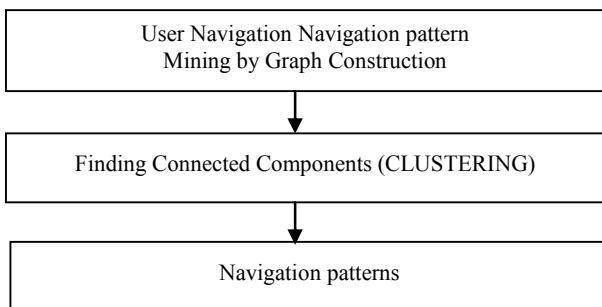Examining **Figure 6**, notice that in the end, there are three
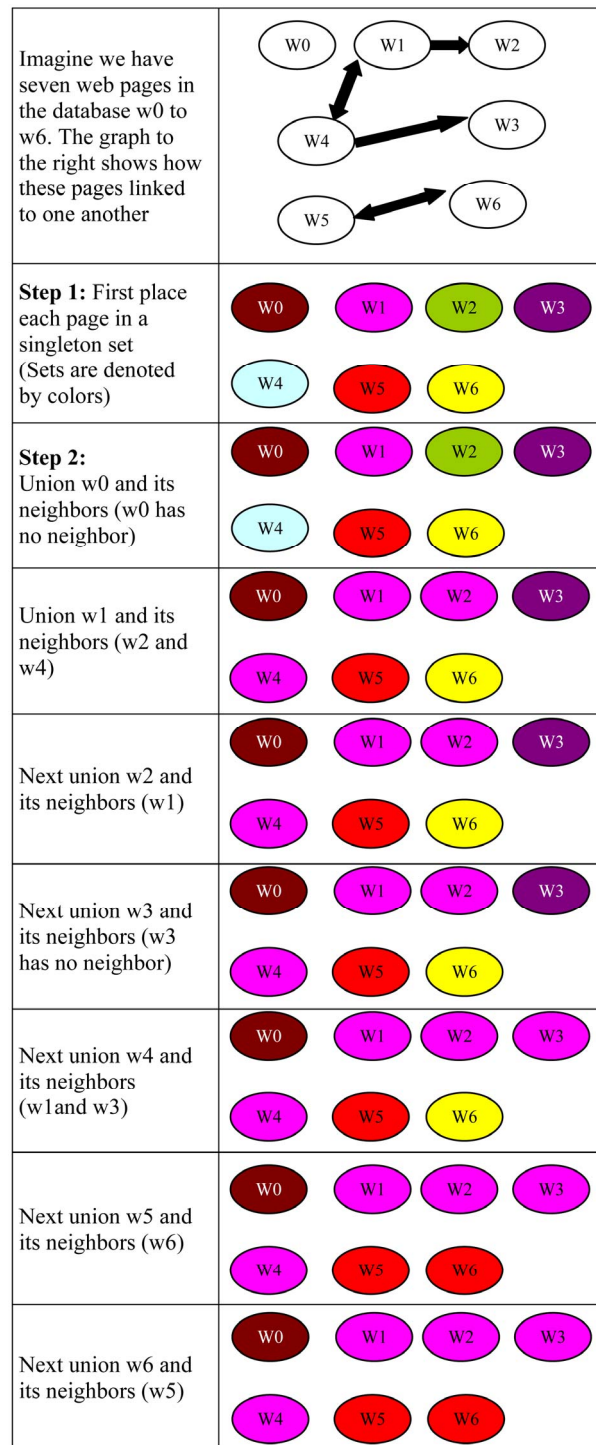

Figure 5. User navigation pattern mining.

Figure 6. Clustering example.

clusters:
- W0
- W1, W2, W3, and W4
- W5 and W6

Recommendation Engine

The recommendation engine is the online component of

the Web personalization system based on usage mining. The task of the recommendation engine is to compute a recommendation set for the current session, consisting of links to pages that the user may want to visit based on similar usage patterns. The recommendation set essentially represents a "short-term" view of potentially useful links based on the user's navigational activity through the site. These recommended links are then added to the last page in the session accessed by the user before that page is sent to the user browser. The over all Architecture for a Usage-Based Web recommendation System is shown in following **Figure 7**.

4. Maximum Length Subsequence Algorithm for Computing Recommendations

In the pattern matching [5], comparing the similarity between the two sequences A and B are the fundamental problem. One of the fundamental problems is to find the maximum length subsequence between two given patterns A and B.

In the proposed classification Algorithm, Our objective is to find the maximum length subsequence between two given patterns $A = \{X_1, X_2, X_3, X_4, \cdots, X_m\}$ and $B = \{Y_1, Y_2, Y_3, Y_4, \cdots, Y_n\}$. Where $X_1, X_2, X_3, X_4, \cdots, X_m$ and $Y_1, Y_2, Y_3, Y_4, \cdots, Y_n$ are web pages. A well known algorithm used by us for classification is given as follows:

Algorithm 1: Longest Common Subsequence (LCS):
Let $X = \{x_1, x_2, x_3, x_4, \cdots, x_m\}$, $Y = \{y_1, y_2, y_3, y_4, \cdots, y_n\}$
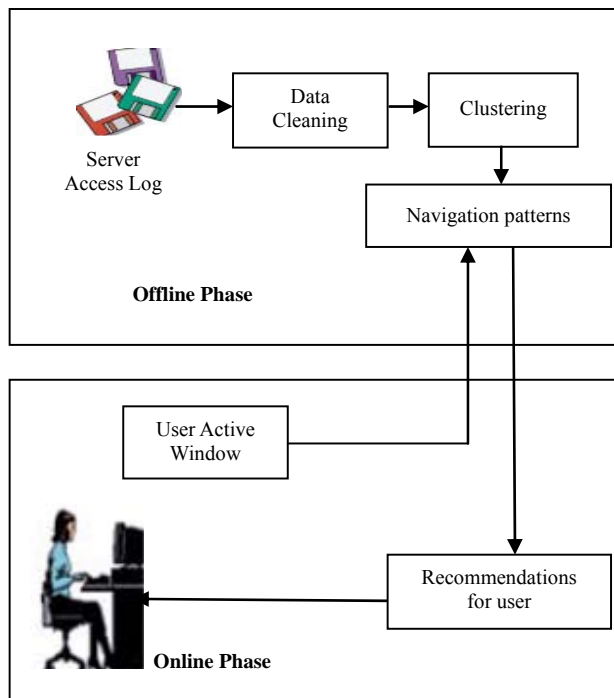

Figure 7. Over all architecture for a usage-based web recommendation system.

be sequences and let $Z = \{z_1, z_2, z_3, z_4, \cdots, z_k\}$ be any LCS of X and Y.

1) If $x_m = y_n$ then $z_k = x_m = y_n$ and Z_{k-1} is an LCS of X_{m-1} and Y_{n-1}

2) If $x_m \neq y_n$ then $z_k \neq x_m$ implies that Z is an LCS of X_{m-1} and Y

3) If $x_m \neq y_n$ then $z_k \neq y_n$ implies that Z is an LCS of X and Y_{n-1}

Let us define $c[i, j]$ to be the length of an LCS of the sequence X_i and Y_j. If either $i = 0$ or $j = 0$, one of the sequence has length 0 so the LCS has length 0. The optimal substructure of the LCS gives the fomula

$$C\lfloor i, j \rfloor$$
$$= \begin{cases} 0 & \text{if } i = 0 \text{ or } j = 0 \\ c[i-1, j-1] + 1 & \text{if } i, j > 0 \text{ and } x_i = y_j \\ \max\left(c[i, j-1], c[i-1, j]\right) & \text{if } i, j > 0 \text{ and } x_i \neq y_j \end{cases}$$

For example, if $X = $ <ABCBDAB> and $Y = $ <BDCABA> then the maximum length subsequence turns out to be <BCBA> of length 4 and similarity $Sim(X, Y) = l/n$ that is 4/7. The more details and proof of the algorithm is available in [9]. We shall not discuss the algorithm in more details since a lot of text already exists on the topic [8-10].

4.1. Classification of Navigation Patterns through LCS

After partitioning the graph in to clusters using the partitioning algorithm for clustering the graph formed by web pages as a node of graph and link between them taken as edges of the graph detailed in Section 3.1.3 of Back End Phase .Now we have a set of cluster $np = $ <$w_1, w_2, w_3, w_4, \cdots, w_k$> is a set of k web pages as a user navigational patterns for each $1 \leq i \leq n$. The sequence $\omega = $ <$w_1, w_2, w_3, w_4, \cdots, w_m$> is a live session window (LSW) where m is the size of live session window.

Step 1: Sorting LSW and Rank the cluster

Before applying the classification we need to order the LSW sequence based on their adjacency weight matrix (WM) constructed in the navigation pattern modeling. Also, we rank all the clusters based on their weight values. Each cluster weight is computed as sum of weight of all its edges.

Step 2: Building the Recommendation List

Now LCS Algorithm is applied on ranked clusters and the web pages in Live Session Window. On applying LCS algorithm, the system finds a cluster with highest degree of LCS in respect to sequence in LSW.

If we get more than one cluster based on LCS Algorithm then that is the job of recommendation engine to select the right cluster among various options. In general recommendation engine chooses that cluster, if the dif-

ference between positions of last elements of longest common subsequence founded in the cluster and the position of first element of this sequence is minimized.

Step 3: Recommendation for User

Finally, Recommendation Engine provides a best cluster of maximum valid matches. Suppose, if the next user activity in live session window is different from the suggested captured list then system has to restart once again.

4.2. Illustration through Example

To illustrate the process, consider the example navigational patterns set given in the following **Table 1**.

In this example we consider the LSW Size = 3, Let LSW is <W2, W3, W4>. np_2 is the cluster that finds with the recommendation engine based on LCS algorithm. In this example the next recommended page will be W1. If there are other pages with W1 (excluding LSW pages) in the np_2 then a recommendation list will be provided to the user.

5. Website Improvement

In this section we utilized the LCS algorithm for website improvement. We have user navigational patterns (U). Let it is <U_1, U_2, U_3, U_4, ⋯, U_n> The following algorithm finds the LCS between (U_1, U_2), (U_1, U_2, U_3), (U_1, U_2, U_3, U_4) and on up to n. It will generate the longest common subsequence among all the navigations U. The obtained sequence may be utilized for arranging the web pages of existing website (site improvement) according to the obtained sequence. This can make a website more useful as well as user friendly to the clients. The Pseudo code for the algorithm is given as follows

```
Algorithm 1, Website Improvement

Begin
i = 1
B = U_i
for (i = 1; i ≤ n; i++)
{
A_i =LCS (U_{i+1}, B);
B = A_i
}
end
```

Algorithm Evolution for four Users' Navigational Patterns

I.
i = 1
A_1 = LCS (U_2, U_1);
B = A_1

II.
i = 2
A_2 = LCS (U_3, A_1);
B = A_1
III.
i = 3
A_3 = LCS (U_4, A_2);

Now A_3 is the LCS of user navigational patterns (U). This means that this is the longest sequence accessed by the users and common in all the navigations. It means all the users access these pages. Therefore we can arrange web pages on our website according this sequence.

We have analyzed the whole algorithm through taking example navigational pattern in the next Section 6 and experimental evaluation in Section 7.

6. Analysis

In this section we have analyzed navigational patterns of azad Institute of engineering and technology (AzadIET), Lucknow (www.aiet.ac.in). The navigational patterns four deferent users at www.aiet.ac.in that are given in **Table 2** as follows:

Here W0, W1, W2, ⋯, W10 are the web pages corresponding to web pages of azadIET like W0 is corresponding the home page, W1 for next page and so on.

In the above table if we analyze the navigation patterns U1 and U2. The LCS of U1 and U2 that is L1 = W0, W4, W7, W3, W9, W10.Now LCS L1 and U3 is L2 = W0, W4, W3, W9, W10. Then LCS of L2 and U4 is L3 = W0, W4, W3, W9, W10.

Now we found the LCS L3 that common subsequence of all above given patterns. Now we can reshuffle our website according to this common sub sequence. The pages on the website should be arranged in the following order that is W0, W4, W3, W9, W10. Home page W0 second page should be fourth page W4 then third place is

Table 1. Navigational patterns generated by graph partitioning algorithm.

NP S. No.	Navigational Patterns
1	W0
2	W1, W2, W3, and W4
3	W5 and W6

Table 2. Example navigational patterns.

U1 = W0, W4, W5, W7.W3, W9, W6, W10
U2 = W0, W4, W7, W2. W3, W9, W10, W6
U3 = W0, W4, W5, W1. W3, W9, W5, W10
U4 = W0, W4, W7, W2. W3, W9, W10, W6

at right position fourth and fifth page should be ninth and tenth pages. Since Most of the users are accessing these pages at least once in their session.

7. System Evaluation

Measuring the quality of the classification algorithm in navigation patterns mining systems needs to characterize the quality of the results obtained. The experimental evaluation was conduced using Azad Degree College (www.adclucknow.org) and Azad Institute of Engineering and Technology (aiet.ac.in) Log files. The only cleaning step performed on this data was the removal of references to auxiliary files (e.g., image files). No other cleaning or preprocessing has been performed in the first phase. The data is in the original log format used by Microsoft IIS.

All evaluation tests were run on a dual processor Intel CPU 1.8 GHz Pentium 4 with 1GBytes of RAM, operating system Windows XP. We implemented the algorithms in C language. We used following metrics for measuring the performance of the system.

7.1. Accuracy

Accuracy measure is defined as degree to which the prediction engine produces accurate recommendations [2].

$$Accuracy = \frac{\left| P\left(LS_{np}, LSW\right) \cap Original_{np}\right|}{\left| P\left(LS_{np}, LSW\right)\right|}$$

LSW: Live Session Window

$P(LS_{np}, LSW)$: Navigation Pattern in predicted future browsing pattern of user.

$Original_{np}$: Original Navigational pattern of user.

7.2. Coverage

Coverage measures the ability of the prediction engine to produce all of the page views that are likely to be visited by the user [2].

$$Coverage = \frac{\left| P\left(LS_{np}, LSW\right) \cap Original_{np}\right|}{\left| Original_{np}\right|}$$

The practical execution of the proposed algorithm on azad degree college (www.adc.ac.in) and azad institute of engineering and technology (www.aiet.ac.in). Our results shown in the following **Figure 8** proves that increase in the size of LSW increases the accuracy and results shown in **Figure 9** proves that increase in the size of LSW coverage decreases and **Figure 10** shows that as the LCS Length (from Algorithm 1) increases more than minimum Threshold *i.e.* equal to 3 that means website requires change in sequence of their arrangement of web pages it means site requires to reshuffle the arrangement of web pages.
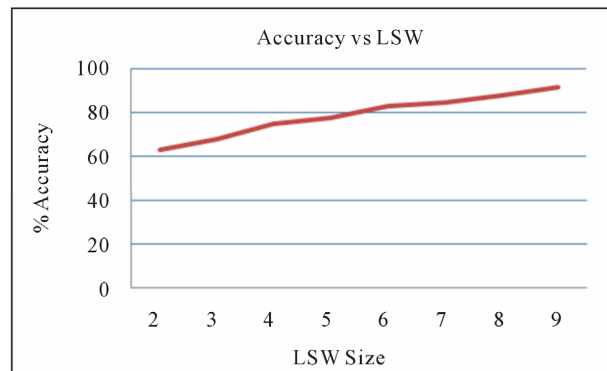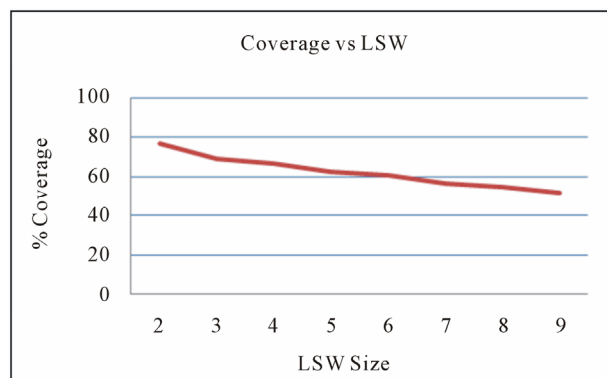

Figure 8. Accuracy.
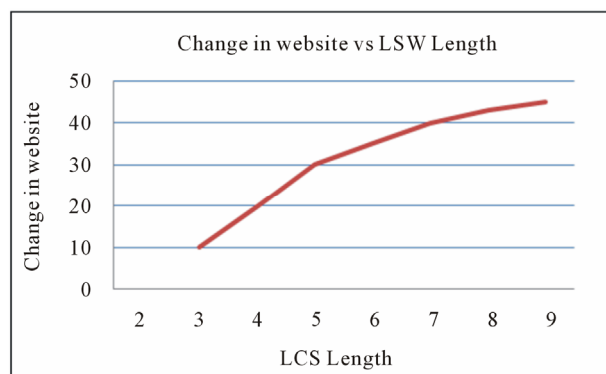

Figure 9. Coverage.

Figure 10. Website improvement.

8. Conclusions and Future Work

This paper describes longest common subsequence algorithm used to classify the user navigation pattern for predicting the recommendation set for the online users and the proposed algorithm provides an efficient way for website improvement to the organizations according the sequence found by the proposed algorithm. That can make the website of any organization more efficient (for site improvement) and user friendly. The quality of the recommendations is measured by the two parameters that are accuracy, coverage and the **Figure 10** shows that as the

length of subsequence increases, the requirement of re-arrangement of pages (site improvement) also increases.

In future it is possible to improve our system by taking the semantic knowledge, time spent by user on particular page, back link etc for quality prediction as well as improvement for the structure of website.

Integrating semantic knowledge and web usage mining can achieve best recommendations in the dynamic large web sites and considering the time spent by user on particular page, back link can improve the design of website to a greater extent.

9. Acknowledgements

REFERENCES

[1] H. K. Singh and B. Singh, "Web Data Mining Research: A Survey," *IEEE International Conference on Computational Intelligence and Computing Research*, Coimbatore, 28-29 December 2010, pp. 1-10. doi:10.1109/ICCIC.2010.5705758

[2] M. Jalali, N. Mustapha, *et al.*, "WebPUM: A Web-Based Recommendation System to Predict User Future Movements," *International Journal Expert Systems with Applications*, Vol. 37, No. 9, 2010, pp. 6201-6212. doi:10.1016/j.eswa.2010.02.105

[3] M. Eirinaki and M. Vazirgiannis, "Web Mining for Web Personalization," *ACM Transactions on Internet Technology*, Vol. 3, No. 1, 2003, pp. 1-27. doi:10.1145/643477.643478

[4] S. K. Shinde and U. V. Kulkarni, "A New Approach for on Line Recommender System in Web Usage Mining," *IEEE International Conference on Advanced Computer Theory and Engineering*, Phuket Island, 20-22 December 2008. doi:10.1109/ICACTE.2008.72

[5] R. Lakshmipathy, V. Mohanraj, *et al.*, "Capturing Intuition of Online Users using a Web Usage Mining," *IEEE International Advance Computing Conference*, Patiala, 6-7 March 2009. doi:10.1109/IADCC.2009.4809222

[6] W. T. Yan, M. Jacobsen and H. Garcia-molina, "From User Access Patterns to Dynamic Hypertext Linking," *Computer Networks and ISDN Systems*, Vol. 28, No. 7-11, 1996, pp. 1007-1014.

[7] R. Baraglia and F. Silvestri, "An Online Recommender System for Large Web Sites," *Proceedings of the* 2004 *IEEE/WIC/ACM International Conference on Web Intelligenc*e, Beijing, 20-24 September 2004.

[8] Y. M. AlMurtadha, M. N. B. Sulaiman, N. Mustapha and N. I. Udzir, "Mining Web Navigation Profiles for Recommendation System," *Information Technology Journal*, Vol. 9, 2010, pp. 790-796. doi:10.3923/itj.2010.790.796

[9] T. H. Cormen, C. E. Leiserson and R. L. Rivest, "Introduction to Algorithms," MIT Press, Cambridge, 1990.

[10] D. S. Hirschberg, "Algorithms for the Longest Common Subsequence Problem," *Journal of the ACM*, Vol. 24, No. 4, 1977, pp. 664-675. doi:10.1145/322033.322044