Hybrid Approaches for User Navigation in Web Usage Mining

[1]J.Umarani, [2]G.Thangraju and [3]R.Arumugam,
[1]Research Scholar, Department of Computer Science, Bharathiyar University, Coimbatore, India
[2]Research Scholar, Department of Computer Science, Karpagam University, Coimbatore, India
[3]M.Phil-Research Scholar, Department of Computer Science, Thanthai Hans Roever College of Arts and Science, (Autonomous), Elambalur, Perambalu, India

***Abstract*:** With the growth of World Wide Web and large number Hosts are join continuously to the internet, huge number of access events to Web sites pages were recorded by Servers in log files , many users share, send, post and download lot of things from Web Sites, this manner can be difficult to many organization and Agents in order to monitor and control that, the recorded information and type of analysis used to extract useful knowledge and understanding it become a practical challenges to many researchers. Log files can provide many events information regard to Clients activities, server activities and so on. Many organization employee many log files analysis tools to predict, analysis and monitor users behavior towards site contents .In this paper we proposed algorithms to analysis hidden information contents in Log files and discovering patterns by identified users along with them navigation behaviors then clustering similar users based on different interesting log file content for many Web sites that hosted in Web server. Find statistics for every part in log file command lines which are not present in many log files analysis tools are supported here and finally discovering frequent Web sites-Users and user's activities towards those Web sites.

Forming a fine ordered web site to assist users to access the website effectively are the crisis in the modern information age. Web developer must recognize the structure of in viewer's perception. Developer centered website is out of scope now a days. User centered websites have higher rating in the WWW. Navigation is important for novice users because most of the websites are dynamic in nature. Identifying accurate direction for target webpage is the essential commitment of each and every networked user. In this paper, we suggest a most favorable clarification for improved navigation for the users who accessing the dynamic website and using this they can find destination meticulously. Web Usage Mining (WUM) is applied to obtain information that may assist web site reorganization and adaptation. The classification algorithm, Longest Common Subsequence classifies user navigation.

Keywords: *Web Usage Mining, User Navigation, Adaptive Web site.*

I. INTRODUCTION

Internet is a collection of two or more computers networks connected together to share the computer related resources over the web by using standard internet protocols to assist billions of user around the world. It is a *grid of networks* that consists of number of networks of various fields or function that are linked by expansive array of electronic, wireless and optical networking technologies. The Internet carries the wide range of information resources such as hypertext documents and services such as email, transaction services etc.

The arrival of the internet provides a platform for people to acquire knowledge and explore information. Today internet is widely used by 2.27 billion people around the world. In order to satisfy the increasing demands from online customers, firms are heavily investing in the development and maintenance of their websites. The Internet has no centralized dominance in either technological implementation or policies for access and usage; each integral network sets and follows its own policies. The Internet Protocol address space and the Domain Name System are the two fundamental namespaces maintained by the Internet Corporation for Assigned Names and Numbers (ICANN).The Internet Engineering Task Force (IETF) perform the task of standardization of the core protocols.

Regardless of focusing investments in websites, discovery of required information in a website is a difficult task. According to Palmer [1], imperfect website design will lead to website failure. If the user is not able to reach their intended information, then the user can exit the website even containing more quality information find by McKinney et al. [2].

The process of obtaining unknown and useful information from the World Wide Web by using data mining techniques in order to improve the effectiveness of a website is known as web mining. Web mining is used to understand and evaluate user behavior and to evaluate the efficiency of a website .There are three types of web mining.

1. Web content mining:

It is the process of extracting and analyzing the information from the webpage content.

2. Web structure mining:

It is the process of extracting and analyzing the information from the web structure.

3. Web usage mining:

It is the process of extracting and analyzing the information from the web server logs. The information collected by the web mining process can be evaluated by using the data mining parameters such as association, classification, clustering, prediction, analysis of sequential patterns.

A primary cause is that web builders may identify of how a web site should be designed can be different from that of users [3], [4].this cause will lead to the problem for user to locate their intended information. This is difficult to prevent because web developers can develop the webpages and organize it based on their own ideas and prediction. The effectiveness of a website is highly based on the user satisfaction. Hence it is necessary to systematize the website based on the user preferences [5].

II. RELATED WORK

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 pre-dict user future intentions [6].

Web personalization is the process of adapting web pages based on the user navigational behavior information, user preferences and his profile patterns [7].

Web transformation is the process of changing the structure of the web site for large group of users instead of changing the structure to the individual user [8].

Gupta et al. [9] propose a heuristic method based on finding the maximum likely hood of links to improve the navigation by using the user preference data.

Lin [10] develops integer programming models that uses the information about the inter relationship between the pages to restructure the website. it is efficient for small website to reduce information overload, but not for large website because it requires large computation time to obtain optimal solution.

[11] D. S. Hirschberg develops Algorithms for the Longest Common Subsequence Problem which is useful for solving our problem.

III. METHODOLOGY

The author planned to implement her proposed methods in two aspects they are listed here 1). User Navigation with Analysis with Similarity and Dissimilarity Matrix 2). User Navigation through the LCS Patterns. The following sections are deal with the concepts implemented in the mentioned above methods.

A. User Navigation with Analysis with Similarity and Dissimilarity Matrix

A.1. PHASES OF WEB USAGE MINING

In order to extract knowledge from log file, several problems exist when extract useful information from that log file and also there are many outlier records need to be eliminate from it in this case we are applying general phases of Web Usage Mining to analysis and understand the extracted and valid information. The general phases of Web Usage Mining as follow:

PHASE 1: PREPROCESSING

Preprocessing phase include some activities can be applied on log file for cleaning, identifying users, valid URL path and also eliminate outliers from log file, tasks on preprocessing phase as follow [13]:

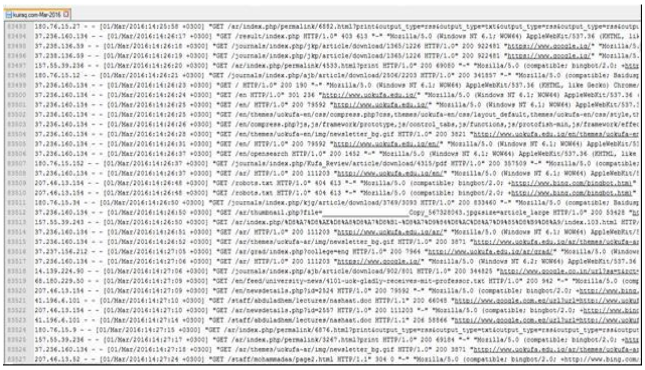
Data Cleaning

log file contain several records are irrelevant to our work like redirect path to other Sites, entries belong to top/bottom frames and records contain server error message. Error message identified through the status code that has been sent by server when user request particular content, server status code can be vary and valid status codes are show in **table 3.1**.

Table 3.1. HTTP server status codes

T	Code	Status	Description
	Syntax	Code	
1	1xx	100	Countinue
		101	Switching Protocol
		102	Processing
	2xx	200	Ok
		201	Created
		202	Accepted
		203	Non- Authoritative Information
3	3xx	301	Moved Permanently
		302	Foud
		303	See Other
		304	Not Modified
4	4xx	400	Bad Request
		401	Authorization Required
		402	Payment Required
		403	Not Found
5	5xx	500	Internal Server Error
		501	Method Not Implemented
		502	Bad Gateway
		503	Service Unavailable
		504	Gateway Time Out

Status code show the success and failures users request, records with status code less than 200 and greater than 299 are considered failure records and eliminated from log file entries. Data cleaning also include eliminated records that browsed irrelevant paths such as CSS content, main site paths, gif, icons and maps etc. by checked suffix part of URL.



```
180.76.15.27 - - [01/Mar/2016:14:25:58 +0300] "GET /ar/index.php/permalink/6852.html?print&output_type=rss&output_type=txt&output_type=rss&output_type=rss&outpu
37.236.160.134 - - [01/Mar/2016:14:26:17 +0300] "GET /result/index.php HTTP/1.0" 403 613 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, lik
37.238.136.59 - - [01/Mar/2016:14:26:18 +0300] "GET /journals/index.php/jkp/article/download/1365/1226 HTTP/1.0" 200 922481 "https://www.google.iq/" "Mozilla/5.
37.238.136.59 - - [01/Mar/2016:14:26:19 +0300] "GET /journals/index.php/jkp/article/download/1365/1226 HTTP/1.0" 200 922481 "https://www.google.iq/" "Mozilla/5.
157.55.39.236 - - [01/Mar/2016:14:26:20 +0300] "GET /ar/index.php/permalink/4533.html?print HTTP/1.0" 200 69080 "-" "Mozilla/5.0 (compatible; bingbot/2.0; +http
180.76.15.12 - - [01/Mar/2016:14:26:21 +0300] "GET /journals/index.php/ajb/article/download/2506/2203 HTTP/1.0" 200 341857 "-" "Mozilla/5.0 (compatible; Baidusp
37.236.160.134 - - [01/Mar/2016:14:26:23 +0300] "GET / HTTP/1.0" 200 190 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/
37.236.160.134 - - [01/Mar/2016:14:26:24 +0300] "GET /en HTTP/1.0" 301 236 "http://www.uokufa.edu.iq/" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (
37.236.160.134 - - [01/Mar/2016:14:26:25 +0300] "GET /en/ HTTP/1.0" 200 79592 "http://www.uokufa.edu.iq/" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.3
37.236.160.134 - - [01/Mar/2016:14:26:25 +0300] "GET /en/themes/uokufa-en/css/compress.php?css,themes/uokufa-en/css/layout_default,themes/uokufa-en/css/style,th
37.236.160.134 - - [01/Mar/2016:14:26:26 +0300] "GET /en/compress.php?js,js/framework/prototype,js/control_tabs,js/functions,js/protofish-min,js/framework/effec
37.236.160.134 - - [01/Mar/2016:14:26:28 +0300] "GET /en/themes/uokufa-en/img/newsletter_bg.gif HTTP/1.0" 200 3821 "http://www.uokufa.edu.iq/en/themes/uokufa-en
37.236.160.134 - - [01/Mar/2016:14:26:31 +0300] "GET /en/ HTTP/1.0" 200 79592 "http://www.uokufa.edu.iq/en/" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/53
37.236.160.134 - - [01/Mar/2016:14:26:37 +0300] "GET /en/opensearch HTTP/1.0" 200 1452 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like
180.76.15.152 - - [01/Mar/2016:14:26:37 +0300] "GET /journals/index.php/Kufa_Review/article/download/4315/pdf HTTP/1.0" 200 357509 "-" "Mozilla/5.0 (compatible;
37.236.160.134 - - [01/Mar/2016:14:26:37 +0300] "GET /ar/ HTTP/1.0" 200 111203 "http://www.uokufa.edu.iq/en/" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/5
207.46.13.154 - - [01/Mar/2016:14:26:48 +0300] "GET /robots.txt HTTP/1.0" 404 613 "-" "Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)"
207.46.13.154 - - [01/Mar/2016:14:26:48 +0300] "GET /robots.txt HTTP/1.0" 404 613 "-" "Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)"
180.76.15.34 - - [01/Mar/2016:14:26:50 +0300] "GET /journals/index.php/kjg/article/download/3769/3093 HTTP/1.0" 200 833460 "-" "Mozilla/5.0 (compatible; Baidusp
37.236.160.134 - - [01/Mar/2016:14:26:50 +0300] "GET /ar/thumbnail.php?file=____________________Copy_567328043.jpg&size=article_large HTTP/1.0" 200 53428 "ht
157.55.39.243 - - [01/Mar/2016:14:26:50 +0300] "GET /ar/index.php/%D8%A7%D8%AE%D8%A8%D8%A7%D8%B1-%D8%A7%D9%84%D8%AC%D8%A7%D9%85%D8%B9%D8%A9/index.103.html HTTP/
37.236.160.134 - - [01/Mar/2016:14:26:51 +0300] "GET /ar/ HTTP/1.0" 200 111203 "http://www.uokufa.edu.iq/ar/" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/5
37.236.160.134 - - [01/Mar/2016:14:26:52 +0300] "GET /ar/themes/uokufa-ar/img/newsletter_bg.gif HTTP/1.0" 200 3871 "http://www.uokufa.edu.iq/ar/themes/uokufa-ar
37.237.156.212 - - [01/Mar/2016:14:27:05 +0300] "GET /ar/grad/index.php?college=eng HTTP/1.0" 200 7964 "http://www.uokufa.edu.iq/ar/grad/" "Mozilla/5.0 (Windows
37.236.160.134 - - [01/Mar/2016:14:27:06 +0300] "GET /ar/ HTTP/1.0" 200 111203 "https://www.google.iq/" "Mozilla/5.0 (Windows NT 6.2; WOW64) AppleWebKit/537.36
14.139.224.90 - - [01/Mar/2016:14:27:06 +0300] "GET /journals/index.php/ajb/article/download/902/801 HTTP/1.0" 200 344825 "http://www.google.co.in/url?sa=t&rct=
68.180.229.50 - - [01/Mar/2016:14:27:09 +0300] "GET /en/feed/university-news/4101-uok-gladly-receives-mit-professor.txt HTTP/1.0" 200 942 "-" "Mozilla/5.0 (comp
207.46.13.154 - - [01/Mar/2016:14:27:09 +0300] "GET /en/newsdetails.php?id=2524 HTTP/1.0" 200 79592 "-" "Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.
41.196.6.101 - - [01/Mar/2016:14:27:10 +0300] "GET /staff/abduladhem/lectures/nashaat.doc HTTP/1.1" 200 66048 "http://www.google.com.eg/url?url=http://www.uokuf
207.46.13.154 - - [01/Mar/2016:14:27:10 +0300] "GET /ar/newsdetails.php?id=2557 HTTP/1.0" 200 111203 "-" "Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing
41.196.6.101 - - [01/Mar/2016:14:27:14 +0300] "GET /staff/abduladhem/lectures/nashaat.doc HTTP/1.1" 206 58846 "http://www.google.com.eg/url?url=http://www.uokuf
180.76.15.9 - - [01/Mar/2016:14:27:15 +0300] "GET /ar/index.php/permalink/6876.html?print&output_type=rss&output_type=txt&output_type=rss&output_type=rss&output
157.55.39.236 - - [01/Mar/2016:14:27:17 +0300] "GET /ar/index.php/permalink/3267.html?print HTTP/1.0" 200 69184 "-" "Mozilla/5.0 (compatible; bingbot/2.0; +http
37.236.160.134 - - [01/Mar/2016:14:27:18 +0300] "GET /ar/themes/uokufa-ar/img/newsletter_bg.gif HTTP/1.0" 200 3871 "http://www.uokufa.edu.iq/ar/themes/uokufa-ar
207.46.13.52 - - [01/Mar/2016:14:27:24 +0300] "GET /staff/mohammadea/page2.html HTTP/1.1" 304 0 "-" "Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/
```

Figure 3.2: Portion of Web Server Log file format

The result of this step produce the valid entries in log file, next step used to identifying unique users and distinguish users that belong to same IP address. The following algorithm in **Figure 3.2** used for eliminated irrelevant entries in log file data.

Data Cleaning Algorithm

Input: Web Server Log file data

Output: Log file data

```
Step 1: Read log file record from (Web Server Log File).
Step 2: IF (log File Record) .URL == (gif, Css, Main.php, index.php )
        AND (Status code < 200 ∧ Status code > 209)
        Remove from log file.
        End IF.
    Step 3:Repeat Step 1 and Step 2 until EOF (Web Server Log
           file).
    Step 4:Stop and Save file in Data base.
   END
```

Figure 3.1: Data Cleaning Algorithm Steps

User Identification

Web Usage Mining does not required knowledge for user's identifying; there is a need to distinguish among different user's behavior. Server logs record of multiple sessions for user may visit Web site frequently. By absent authentication mechanisms in many Web Server some Web site used Cookies in Client-side, Due to privacy content this feature may disable by users, therefore IP address alone not sufficient to identify unique users in general by assigning many sessions to map IP address [15]. In case of absent user authentication and client-side cookies the possible accurate user identifying method by combination IP addresses with User agent and referrer [13].

```
User Identification Algorithm
Input: log file data
Output: Unique Users Table.
Step1: Initialization
Create Table include the following field:
([User ID, IP's address, Date, Time, Request, Site name, User
Agent, Size)].
Step2: Read record from Log file data
Step3: User's IP addresses of tow sequential records are
compared.
Step4: IF ((IP address) is not in Users Table) THEN
Assign User ID to IP address
Add both to Users Table
ELSE
IF ((IP address) is in User's Table) THEN
Check (User Agent if same) then Add it with Same User ID
ELSE Assign (next User ID) to IP address
Add both to Users Table
Step5: Repeat Step2-5 until EOF (log file data)
Step6: STOP, Store Result.
End.
```

Figure.3.3: User-Identification Algorithm Steps

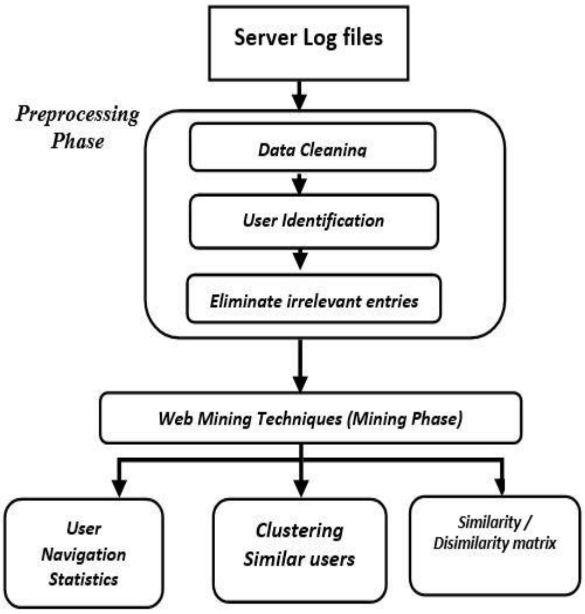

Server Log files
Preprocessing Phase
Data Cleaning
User Identification
Eliminate irrelevant entries
Web Mining Techniques (Mining Phase)
User Navigation Statistics
Clustering Similar users
Similarity / Dismilarity matrix

Figure 3.4.Architecture of Proposed Methods

B. User Navigation Through The Lcs Patterns.

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 = <$w1$, $w2$, $w3$, $w4$, ⋯, wk > is a set of k web pages as a user navigational patterns for each $1 \leq i \leq n$. The sequence ω = <$w1$, $w2$, $w3$, $w4$, ⋯, wm> 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 difference 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.

We utilized the LCS algorithm for website improvement. We have user navigational patterns (U). Let it is <U1, U2, U3, U4, ⋯, Un> The following algorithm finds the LCS between (U1, U2), (U1, U2, U3), (U1, U2, U3, U4) 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 ob-tained 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 = Ui
        for (i = 1; i ≤ n; i++)
            {
            Ai =LCS (Ui+1, B);
            B = Ai
            }
End
```

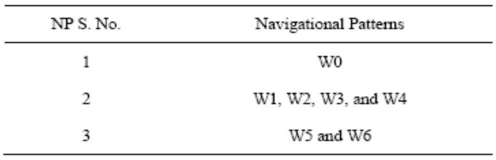
IV. IMPLEMENTATION

The navigational patterns four deferent users are given in **Table 2** as follows: Here W0, W1, W2, ⋯, W10 are the web pages corresponding to web pages 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.

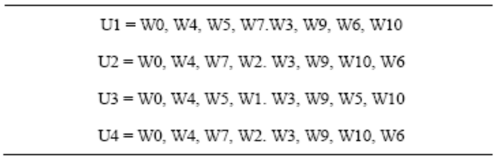
Now we found the LCS L3 that common subsequence of all above given patterns. Now we can reshuffle our website according to this common sub sequence. The pages on the website should be arranged in the following order that is W0, W4, W3, W9, W10. Home page W0 second page should be fourth page W4 then third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

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

CONCLUSION

The first research work focuses on discovering the hidden information from main server general log file, main server contain combination for all Web sites access information that hosted on it in text format, this file include navigation activities for many Web sites in order to understand the behaviors of users towards those sites not for single Web site, the contribution of the paper is to extract information from huge log file and consider novel approaches to deal and analysis users patterns, then extracted useful information for valid sessions after that clustering approach has been applied to grouping similar users navigations behaviors, this can give as indicators frequent users interest towards different Web sites content, monitor users activities for particular Web site, consume bandwidth for each user during selected period, monitor Web sites visits and browsed content and many others activities for future works.

The second work 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 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.

References

[1] J. Palmer, “Web Site Usability, Design, and Performance Metrics,” Systems Research, vol. 13, no. 2, pp. 151-167 2002.

[2] V. McKinney, K. Yoon, and F. Zahedi, “The Measurement of Web-Customer Satisfaction: An Expectation and Disconfirmation Approach,” Information Systems Research, vol. 13, no. 3, pp. 296-315, 2002.

[3] T. Nakayama, H. Kato, and Y. Yamane, “Discovering the Gap between Web Site Designers’ Expectations and Users’ Behavior,”Computer Networks, vol. 33, pp. 811-822, 2000.

[4] M. Perkowitz and O. Etzioni, “Towards Adaptive Web Sites:Conceptual Framework and Case Study,” Artificial Intelligence,vol. 118, pp. 245- 275, 2000.

[5] J. Lazar , User-Centered Web Development . Jones and Bartlett Publishers, 2001.

[6] 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.

[7] M. Eirinaki and M. Vazirgiannis, “Web Mining for Web Personalization,”ACM Trans. Internet Technology, vol. 3, no. 1pp. 1-27, 2003.

[8] C.C. Lin and L. Tseng, “Website Reorganization Using an Ant Colony System,” Expert Systems with Applications, vol. 37, no. 12, pp. 7598- 7605, 2010.

[9] R. Gupta, A. Bagchi, and S. Sarkar, “Improving Linkage of Web Pages,” INFORMS J. Computing, vol. 19, no. 1, pp. 127-136, 2007.

[10] C.C. Lin, “Optimal Web Site Reorganization Considering Information Overload and Search Depth,” European J.Operational Research, vol. 173, no. 3, pp. 839-848, 2006.

[11] 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