

ENHANCING WEB NAVIGATION USABILITY USING WEB USAGE MINING TECHNIQUES

Swapnil S. Patil¹, Hridaynath P. Khandagale²

¹M.Tech. Scholar, Swapnil S. Patil, Dept. Of Technology, Shivaji University, Kolhapur, Maharashtra, India

²Asst. Professor H.P. Khandagale, Dept. Of Technology, Shivaji University, Kolhapur, Maharashtra, India

Abstract - World Wide Web consists of huge source of information resources and services. That's why there is an explosive growth in web traffic. Each and every website has some form of navigation which is decided by the web developer. And most of the time we knew that these navigational paths are decided without looking into users web interest, which results into the navigational problems have to face by user. This paper "Enhancing web navigation usability by using web usage mining techniques" discusses in detail and provides a standard way for web developers to recognize actual usage behavior and anticipated usage behavior with the use of server side access log record file. Along with this it discusses and provides facility for updating web links in an automated manner based on sequential pattern mining and thereafter pattern analysis results. Anticipated Usage behavior helpful for user to provide better effectiveness and efficiency for their tasks as well as updating links in an automated manner reduces the time expended by developer. Overall this system is more useful from web developers' as well as users' point of view.

Key Words: Web Usage Mining, Web Usability, Sequential Pattern Discovery, Meta Mining, Web Personalization in automated manner.

1. INTRODUCTION

Web navigation refers to the process of navigating a network of information resources in the World Wide Web, which is organized as hypertext or hypermedia. Just about every website has some form of navigation. Unfortunately, not every website's navigation is good. Most of the time, a website's navigation is put together by Web designers who know a lot about making pretty websites, but very little about marketing a website or creating a website built from the users point of view.

Therefore, it is necessary to identify navigation-related Web usability problems which will be helpful to users to provide better effectiveness (higher task completion rate) and efficiency (less time for given tasks) [1]. One of the greatest advantages of designing web-based user interfaces over traditional user interfaces is the ability to keep track of user interactions with the site. This information is stored on most web servers by default.

Web Usage mining [2] applies data mining techniques to extract knowledge from these web log files. This paper elaborate research work done from data preparation and preprocessing phase to personalized website in automated manner.

1.1 Data Preparation and Pre-processing

Data preparation and preprocessing is one of the important steps to generate and analyze the result of research work. Pre-processing [11] the log files and converting back to sequential data. The data pre-processing phase includes data cleaning, user identification, session identification, and site structure and link details formation.

1.2 Sequential Pattern Extraction Component

"Sequential Pattern mining" [3], [4] includes a set of sequences and support threshold and we have to find the complete set of frequent subsequences. To find users' navigational patterns from the sequential data using different pattern mining algorithms can be used. The raw log files from the web server on which the Sheet Exchange website resides are first simplified and converted into sequential data. Then a number of pattern finding algorithms are applied. Sequential pattern mining algorithms include: Generalized Sequential Pattern (GSP) Mining Algorithm and Prefix-Span like algorithms.

1.3 Pattern Analysis (Meta Mining)

Pattern analysis phase makes use of cognitive user model [14], [15] which will discover anticipated usage behavior. Cognitive user model is used to simulate or predict human behavior or performance based on discovered patterns.

1.4 Updating Links

Result of Meta mining phase helps web developer to find out most common interest of users' through graphical representation. To reduce the work of web developer we can implement the automated system for Updating web-navigation links [8].

2. RELATED WORK

Web Usage Mining[2], [6], [7] used to discover interesting usage patterns from Web log data, in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered. They are web server data, application server data and application level data. Web server data [13] correspond to the user logs that are collected at Web server. Some of the typical data collected at a Web server include IP addresses, page references, and access time of the users and is the main input to the present Research. Assessing Web Site Usability from Server Log Files is a difficult task". [19] Presents various assessment methods and techniques in assessing web-usability. This Research work concentrates applying web usage mining techniques for improving web navigation and in particular focuses on discovering the anticipated usage patterns for websites from the server log files and reducing the time of web developers by updating web links to website in automated manner.

Delivery of efficient service through a web site makes it compulsory in the redesigning stage to take into account the behavior of the users, which can be studied by means of a web log file that partially records information about user visits. The reconstruction of all of the sequences of pages that are visited by users who browse a web site is known as the web sessionization problem, and it has been formulated by means of an integer programming model; however, because a web log can accumulate a large amount of information, it is necessary to reconstruct the sessions over a period of weeks or months, thus the solution to this problem requires a long computational processing time. A heuristic approach based on simulated annealing is useful for the sessionization problem. Using [11] and [12] this approach, it has been possible to reduce the processing time up to 166 times compared to the time that is required for the integer programming model. Furthermore, the meta-heuristic solution finds new optimum values, which achieve increases on the order of 17% in the best cases.

The data stored in the log files do not present an accurate picture of the users' accesses to the Web site. Hence, preprocessing of the Web log data is an essential and pre-requisite phase before it can be used for knowledge-discovery or mining tasks. [10] The preprocessed Web data can then be suitable for the discovery and analysis of useful information referred to as Web mining. Web usage mining, a classification of Web mining, is the application of data mining techniques to discover usage patterns from click stream and associated data stored in one or more Web servers [2] this paper presents an overview of the various steps involved in the preprocessing stage like: Data Cleaning, User Identification, and Session Identification.

Link prediction and path analysis is an important part in web usage mining [16] presents the notion of probabilistic link prediction and path analysis using Markov chains is proposed and evaluated. Markov chains allow the system to dynamically model the URL access patterns that are observed in navigation logs based on the previous state. Furthermore, the Markov chain model can also be used in a generative mode to automatically obtain tours. The Markov transition matrix can be analyzed further using eigenvector decomposition to obtain 'personalized hubs=authorities'. The utility of the Markov chain approach is demonstrated in many domains: HTTP request prediction, system-driven adaptive Web navigation.

Usability testing [5], [17] is a technique that is used to elicit the quality of systems and sites by regular tests carried with the potential users. [18] Testing Web sites for usability requires an understanding of the sites' audience, category, content, usability goal and how to measure to achieve these goals. There is a whole variety of usability testing methods available for Web based testing. But our main objective was to make a pick of those right methods that could be implemented to test usability of any Web application given a short time, small budget and fewer resources. GSP is the Apriori based Horizontal formatting method and Prefix-SPAN is Projection-based pattern growth method.

3. IMPLEMENTATION

3.1 DATA PREPARATION AND PREPROCESSING

Data preparation [9] phase includes the collection of data from various sources for research work.

We have collected data source (Web server log data) of the <http://www.southtexasshooting.org> website. Win-track website copier Tool is used for downloading of whole website along with their web server log data. Total 1, 98,938 log records in year 2016 are selected for research work.

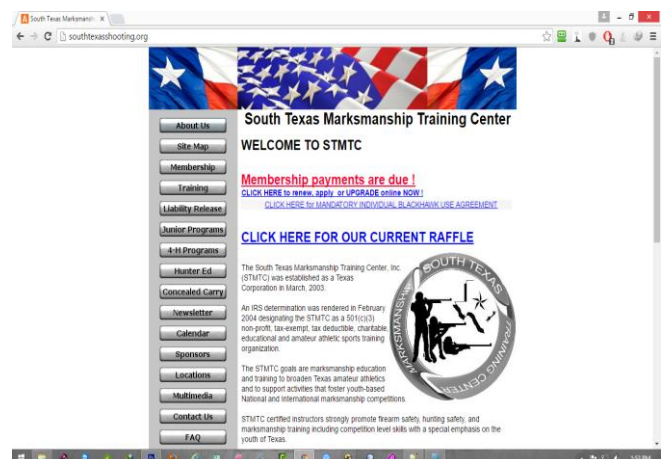


Fig -1: Original <http://www.southtexasshooting.org> Website

Selected web log data is in Combined Log Format.

Syntax Host IP Address, Proprietor, Username, date: time, request method, status code, byte size, referrer, User agent, Cookie.

Example 174.46.233.4 - - [30/Dec/2015:16:10:21-0800]"GET/sitebuilder/images/Jason_NRA_Record_Certificate-750x630.png HTTP/1.1" 200 719769 "https://www.google.com/" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_5) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/47.0.2526.106Safari/537.36" "southtexasshooting.org". Data pre-processing phase involves data cleaning, user identification, session identification.

3.1.1 DATA CLEANING

All log entries with file name suffixes such as .jpg, .jpeg, .mpg, .mpeg, .gif, .ico, .css, .js, .class, .lst, .properties, .log, .bat, .inf, .dat, .exe, .vcd, .wav, etc are removed. The data_cleaning_Algorithm

Input: Raw web log file

Output: Cleaned log file

- 1: Read record in file
- 2: For each record in file
- 3: If field {*.jpg, *.jpeg, *.mpg, *.mpeg, *.gif, *.ico, *.css, *.js, *.class, *.lst, *.properties, *.log, *.bat, *.inf, *.dat, *.exe, etc }then
- 4: Skip Record
- 5: Next Record
- 6: End of the file

3.1.2 USER IDENTIFICATION

Cleaned log file records applied for the user identification. User identification is done through the combination of IP address and user agent. The algorithm for User_Identification

Input: Cleaned web log records

Output: Users

- 1: For each record in the table

- 2: For each IP+ Agent
- 3: If IP + Agent=current.IP + Agent then
- 4: Insert record in existing user
- 5: Else
- 6: Make new user entry
- 7: End if
- 8: Next IP + Agent
- 9: Next Record

3.1.3 SESSIONIZATION

Sessions are created for each user using session duration and page stay time heuristics. Algorithm used to identify sessions.

Input: Identified Users

Output: Sessions

- 1: For each user
- 2: For each record of the user
- 3: If(current_request_time-first_request.time<1800) sec then
- 4: Add current request to existing session
- 5: Else
- 6: Start new session
- 7: End if
- 8: Next Record
- 9: Next User

3.2. SEQUENTIAL PATTERN EXTRACTION

Here, we have modified GSP and prefix span sequential pattern mining algorithms which is chosen from SPMF [20]. There is major problem with this technique is to set minimum support threshold. Usually is done by using trial and error. This problem is solved using auto adjusting the minimum support threshold according to the data size.

3.2.1 Algorithm: Modified GSP algorithm

Is used for extracting frequently occurring sequences and is proposed by Agrawal and Srikant.

Input: A Sequence database

Output: The complete set of sequential patterns.

Procedure:

Calculating $\text{minsup}(x) = (e^{-ax-b}) + c$

F1 = the set of frequent 1-sequences $k=2$, do while $F(k-1) \neq \text{Null}$;

Generate candidate sets C_k (set of candidate k -sequences);

For all input sequences s in the database D

Do

Increment count of all a in C_k if s supports a

$F_k = \{a \in C_k \text{ such that its frequency exceeds the Threshold}\}$

$k = k+1$;

Result = Set of all frequent sequences is the union of all F_k s

End Do

End do

3.2.2 Algorithm: Modified Prefix Span algorithm

Is used for discovery of sequential pattern and performs better in mining large sequences. It is proposed by Jian Pei et.al

Input: A Sequence database S

Output: The complete set of sequential patterns.

Procedure:

Calculating $\text{minsup}(x) = (e^{-ax-b}) + c$

Method:

Call Prefix Span $(\langle \rangle, 0, S)$

Subroutine: Prefix Span $(\alpha, l, S | \alpha)$

Parameters: α : A sequential pattern; l : The length of α ; $S | \alpha$: The α projected database,

If $\alpha \neq \langle \rangle$; otherwise the sequence database S

3.3. PATTERN ANALYSIS (Meta Mining)

Meta mining is not the new concept but, still researchers are not familiar with it. The result of discovered patterns in sequential pattern extraction is not meaningful and needs to analyze and visualize once again as per requirement of the research work.

3.4. UPDATING LINKS IN AUTOMATED MANNER

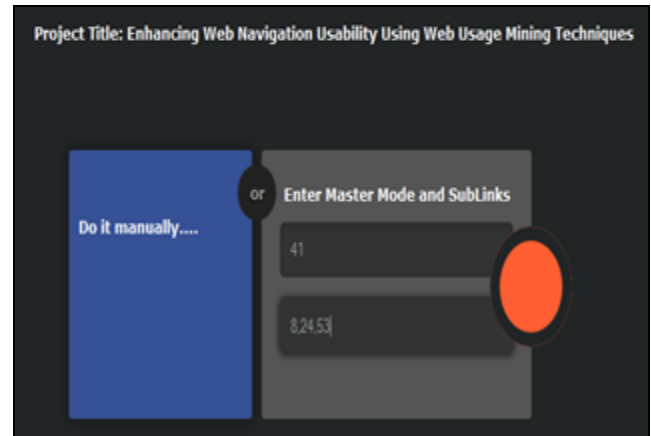
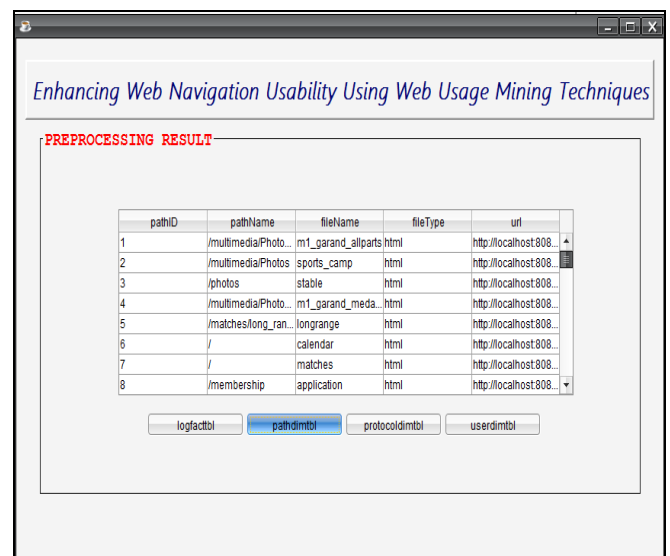


Fig. -2 Web Application for updating links

At the start of paper we talk about predicting useful patterns using our modified sequential pattern mining algorithms and also reducing the valuable time of web developers. So to reduce the valuable time of web developers our web application helps to update navigational links within few seconds.

4. EXPERIMENTAL RESULT AND ANALYSIS

Following figure is the result of pre-processed log data.



pathID	pathName	fileName	fileType	url
1	/multimedia/Photo...	m1_garand_allparts	html	http://localhost:808...
2	/multimedia/Photos	sports_camp	html	http://localhost:808...
3	/photos	stable	html	http://localhost:808...
4	/multimedia/Photo...	m1_garand_meda...	html	http://localhost:808...
5	/matches/long_ran...	longrange	html	http://localhost:808...
6	/	calendar	html	http://localhost:808...
7	/	matches	html	http://localhost:808...
8	/membership	application	html	http://localhost:808...

Fig -3: Result of Pre-Processing

Following table shows the result of preprocessed log data in terms of amount of memory and number of transactions used. The efficiency of pre-processing phase result is measured in terms of memory required to store these results. Following chart-1 represents the total number of transactions to that of preprocessed result.

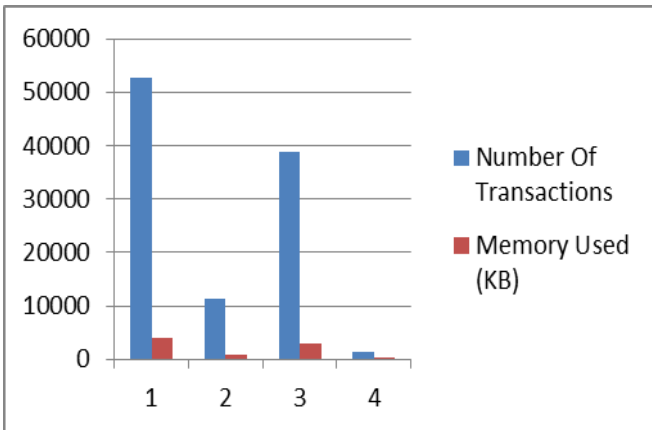


Chart -1: No of Transaction v/s Pre-processed result

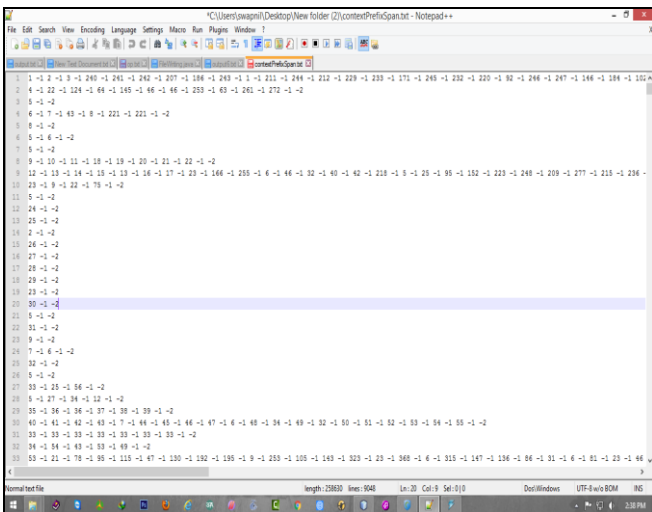


Fig -4: Result of Pre-Data discovery phase

The above figure is the result of pre-Data discovery phase where sequence logs should be delaminated by -1 and user sessions are ended with -2.

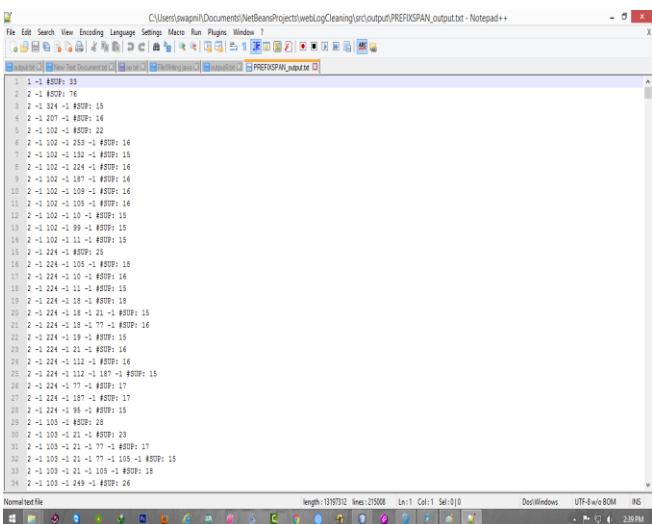


Fig -5: Result of sequential Pattern mining Algorithm

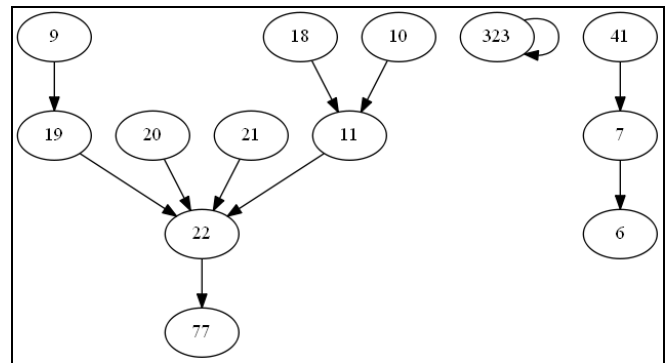


Fig -6: Anticipated Usage Behavior Using GSP Algorithm

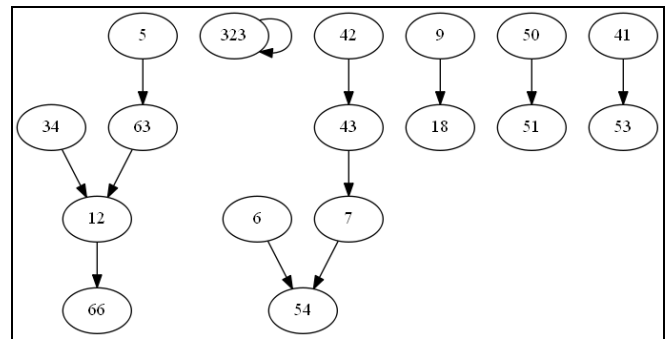


Fig -5: Anticipated Usage Behavior Using Prefix-Span Algorithm

The above two figures are the result of Meta Mining phase for predicting anticipated usage behavior using standard GSP-algorithm and prefix-span algorithm.

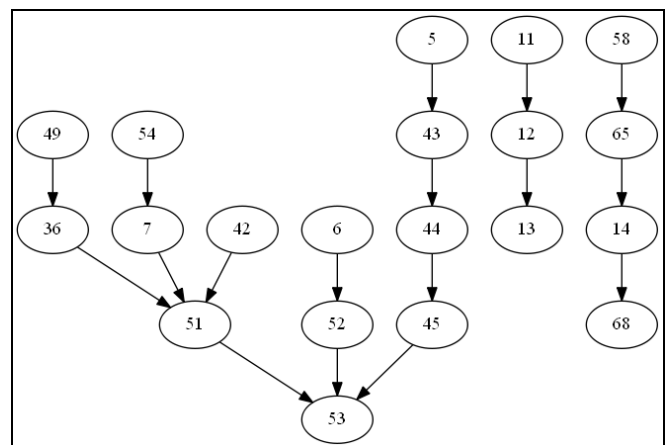


Fig. 6- Anticipated Usage Behavior Using Modified GSP Algorithm

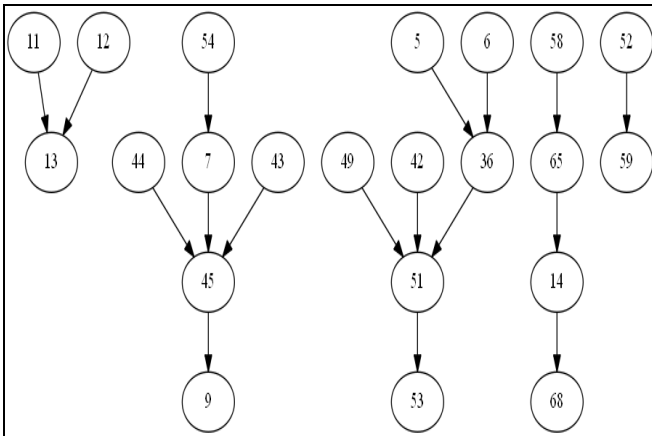


Fig. 7- Anticipated Usage Behavior Using Modified prefix-span Algorithm

The above two figures are the result of Meta Mining phase for predicting anticipated usage behavior using our own modified GSP-algorithm and modified prefix-span algorithm.

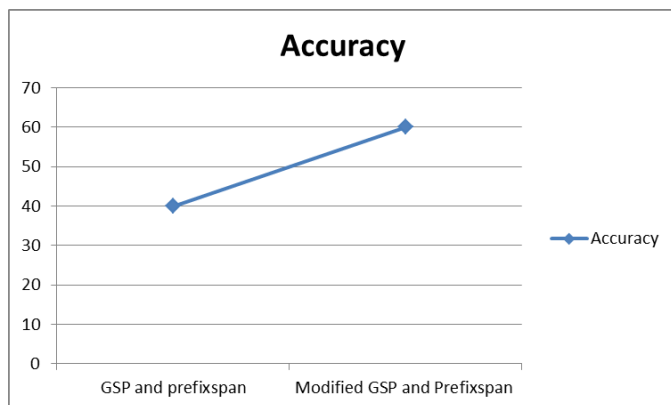


Chart -2: Accuracy of Modified GSP and modified prefix-span Algorithms for Anticipated Usage Pattern Discovery.

Using these results web developers can easily get look into it and predict the anticipated usage behavior of the users. The following figure shows the result of our web-application for updating links in automated manner.

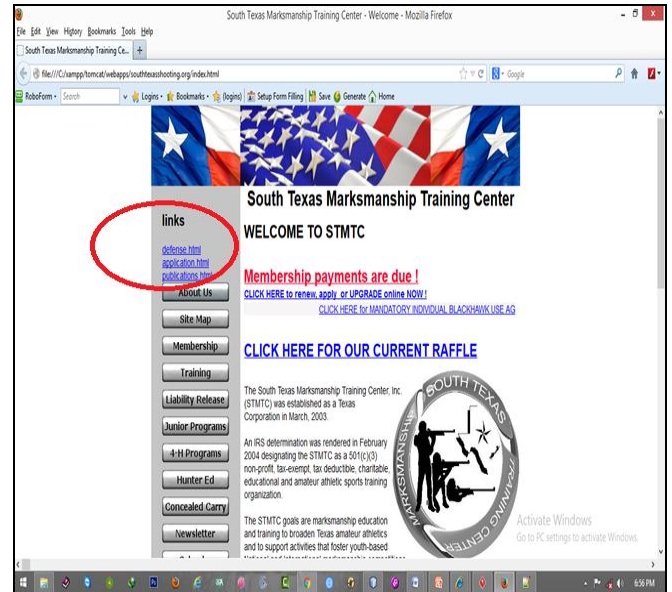


Fig. 8- Personalized Website

5. CONCLUSIONS

Drawback of the massive availability of information on web is that, users do not get the specific information within a precise period of time from the internet. This huge availability of data on the web has made it necessary to find the ways to retrieve the information needed in collecting the requested data. Web mining is the mining technique applied on the various web resources. Web resources are the web log files which contain web navigation information. Our method can contribute significantly to continuous usability improvement over these prolonged maintenance cycles performed over web navigations of the website. The usability improvement in successive iterations can be quantified by the progressively better effectiveness (higher task completion rate) and efficiency (less time for given tasks).

ACKNOWLEDGEMENT

I express my sincere thanks to guide Mr. H. P. Khandagale for their valuable guidance. I also express my sincere thanks to the department of technology, Shivaji University, Kolhapur for providing the necessary resources to carry out this work.

REFERENCES

[1]Ruili Geng, Jeff Tian, "Improving Web Navigation Usability by Comparing Actual and Anticipated Usage", IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, VOL. 45, NO. 1, February 2015.

[2]DeMin Dong, "Exploration on Web Usage Mining and Its Application" IEEE 2009, 978-1-4244-3894-5/09, 1-4.

[3]Cooley R., Mobasher. Mobasher B. And Srivastava J, "Web mining: Information and pattern discovery on the World Wide Web" in proceeding of the 9th IEEE international Conference on tools with Artificial intelligence, 1997.

[4]Baoyao Zhou, Siu Cheung Hui, Kuiyu Chang "An Intelligent Recommender System using Sequential Web Access Patterns", Proceedings of the 2004 IEEE Conference on Cybernetics and Intelligent Systems Singapore, 1-3 December, 2004

[5]T. Carta, F. Patern, and V. F. D. Santana, "Web usability probe: A tool for supporting remote usability evaluation of websites", in Human-Computer Interaction INTERACT 2011. New York, NY, USA: Springer, pp. 349-357, 2011.

[6]Magdalini Eirinaki and Michalis Vazirgiannis, "Web mining for web personalization" communications of the ACM, vol.3, No.1, Feb. 2003 pp. 2-21.

[7]J. Srivastava, R. Cooley, M. Deshpande, and P. -N. Tan "Web Usage Mining: Discovery and Applications of Usage patterns from Web Data". ACM SIGKDD Explorations, 1(2):12 -23, 2000.

[8]M. Y. Ivory and M. A. Hearst, "The state of the art in automating usability evaluation of user interfaces", ACM Computer. Surveys, vol. 33, no. 4, pp. 470-516, 2001.

[9]R. Cooley, B. Mobasher, and J. Srivastava,, "Data preparation for mining world wide web browsing patterns" Journal of Knowledge Information System, vol. 1, no. 1, pp. 532, 1999.

[10]Om Kumar C. U. and P. Bhargavi, "Analysis of web server log by web usage mining for extracting usage patterns", Vol. 3, Issue 2, ISSN 2249-6831, 123-136, (IJCSEITR), June 2013.

[11]C.P.Sumathy, R. Padmaja Valli, T. Santhanam, "An overview of pre-processing of web log files for web usage mining", JATIT, vol. 34 No. 1, ISSN: 1992-8645, 15th Dec.2011.

[12]T. Arce, P. E. Romn, J. D. Velsquez, and V. Parada, "Identifying web sessions with simulated annealing", Expert Syst. Appl., vol. 41, no. 4, pp. 15931600, 2014.

[13]Mr. Akshay Upadhyay, Mr. Balram Purswani, "Web usage mining has pattern discovery", International Journal of Scientific and Research Publications, Volume 3, Issue 2, February 2013.

[14]V. SUJATHA, PUNITHAVALLI "Improved user navigation pattern prediction technique from web log data" Sci-Verse Science Direct, International conference on communication technology system design 2011, procedia Engineering 30 (2012) 92-99.

[15]M. Heinath, J. Dzaack, and A. Wiesner, "Simplifying the development and analysis of cognitive models", in Proc. Eur. Cognitive Sci. Conf., Delphi, Greece, 2007, pp. 446-451.

[16]Ramesh R. Sarukkai, "Link prediction and path analysis using Markov chains" Elsevier Computer Networks 33 (2000) 377-386.

[17]T. Tullis and B. Albert, "Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics (Interactive Technologies)". San Mateo, CA, USA: Morgan Kaufmann, 2008.

[18]Chris Menezes, Blair Nonnecke, "UX-Log: Understanding Website Usability through Recreating Users' Experiences in Logfiles", International Journal of Virtual Worlds and Human-Computer Interaction Volume 2:47-55, Year 2014.

[19]Tec-Ed, Inc. "Assessing Web Site Usability from Server Log Files". December 1999.

[20]SPMF: A Sequential Pattern Mining FrameWork <http://www.philippe-fournier-viger.com/spmf/>

BIOGRAPHIES



Mr. Swapnil S. Patil had completed B.E. in Information Technology in 2013 from PVPIT Budhgaon. Currently he is pursuing M. Tech degree in computer science and Technology from Department of Technology, Kolhapur.



Mr. H. P. Khandagale perceived B.E. and M. Tech in computer science and Technology. Presently he is working as Asst. Professor in Department of Technology, Shivaji University, Kolhapur. His interest lies in Application Programming, Data Mining and Web Technology. He is member of ISTE.