ENHANCING WEB NAVIGATION USABILITY BY MINING WEB LOG FILE

Payal Chintawar[1], Dr. Kishor Wagh[2]
[1] PG Scholar, [2]Assistant Professor
[1]Department of Computer Science and Engineering.
[2]Department of Information Technology.
[1,2]Government College of Engineering, Amravati, India.

Abstract : In last few decades, there is a huge increase in the number of users of Internet resulting in the huge growth of Web traffic and websites. Every website has some form of navigation decided by Web developer. But not every website has a good navigation as per users' point of view and comfort. Web navigation usability can be improved significantly by using Web usage mining techniques. First of all data preparation and preprocessing is discussed which is very important in web mining. In this paper, a new approach is proposed which uses time window approach for transaction identification and GSP (Generalized Sequential Pattern) algorithm which is a sequential pattern mining algorithm used for pattern extraction. The last part covers implementation details of the proposed approach along with the results. Overall this system is more useful from web developers' as well as users' point of view.

Index Terms **- Web usage mining, Web navigation usability, sequential pattern mining, GSP.**

I.INTRODUCTION

World Wide Web is a very large network of hypermedia or hypertext. Due to the ease of availability of website editors and supplementary tools used for website development, publishing documents on the Web has become very easy. Thus there is a large number of websites developed for different purposes are available on the Internet. But not every website is functionally convenient by users' point of view. While developing a website Web developer have to consider the factors like purpose of the website, the potential users using the website, arrangement of content of the website and the navigation paths to be followed by the users while using the website. Thus there is need of identification of navigation related issues of website in order to improve effectiveness and efficiency of the website.

Log data called actual data is the user activity recorded at Web server when user visits the website. Usage data is used to understand the user behavior and pattern of user Web navigation. It is also used to guide the graphical user interface design of the website. Data mining techniques are applied on Web usage data to extract knowledge called Web mining. Sequential mining is one of the classes of Web usage mining techniques used to find the frequent navigation patterns and to understand the user's behavior. This method is used to analyze how the website is used and to evaluate website's effectiveness and efficiency.

II.RELATED WORK

Preprocessing and transaction identification are part of data preparation phase and are very important in Web usage mining. Clustering, sequential pattern generation and association rule generation are commonly used for Web usage mining. Browser behavior model are constructed using server logs. This work states the limitations of maximum forward reference approach and auxiliary content transactions. A method to create semantically meaningful transactions from user sessions is tested successfully against other two methods. WEB-MINER system is developed for discovering association rules from the real world data. Task analysis is very important. This method helps to improve design of website and user satisfaction by analyzing several university websites, their departmental hardcopies, search engine queries and interviews of some of the website users[2].

Application of data mining techniques to huge Web data is called as Web mining. Server logs called actual data can be automatically captured at the Web server. Usage data captures the information about the users visiting the website. This information includes page access time , IP aggress and page references made by users. Apriori algorithm and Frequent Pattern Growth algorithms are captured and compared based on memory usage and time usage[8].

A comparison between three kinds of algorithm named as GSP (Generalized Sequential Pattern), Span (Prefix-projected Sequential Pattern Mining) and SPADE (An efficient Algorithm for mining Frequent Sequences) is given. GSP is the horizontal formatting method which is Apriori based and Prefix-SPAN is projection based pattern growth method. This paper demonstrates number of iterations required in each algorithm and elaborates step wise explanation of each algorithm [4].

Identification of Web usability problems by extracting usage patterns is done using cognitive user model and comparing it with actual usage pattern. Anticipated user behavior is captured with the help of cognitive user models whereas actual user behavior is captured from Web server logs. Proposed IUIP model which captures human behavior cognition represents part of cognitive experts' work. The method corrects usability problems leading to better functional convenience as characterized by better effectiveness and efficiency. Continuous usability improvement of the Web system is the key feature of this method. The processing time of the tasks is reduced upto 166 times using this method[1].

III.PROPOSED APPROACH

Following figure 1 shows the architecture of proposed approach followed by the details of modules used in the approach.

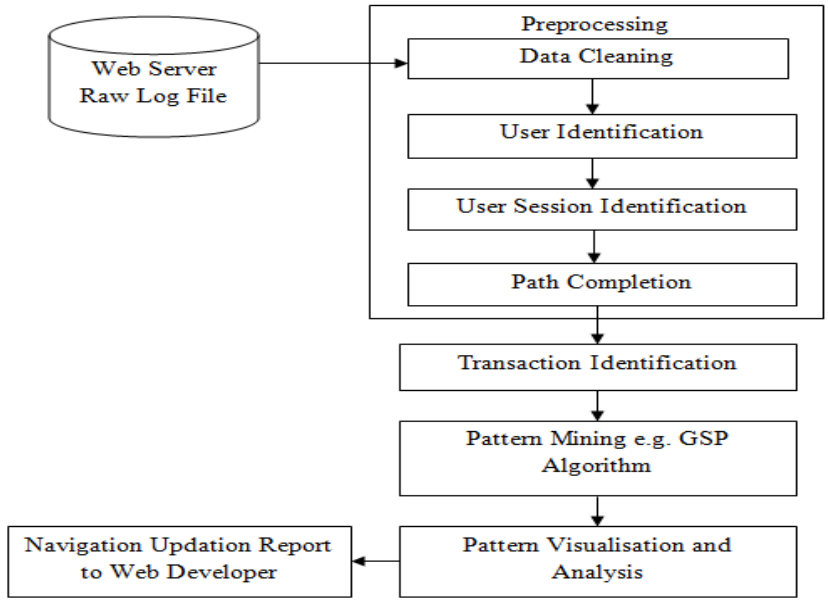

Figure 1. Architecture of Proposed Approach

3.1 Data Preparation and Preprocessing

Data preparation involves preprocessing of data, integration of data from different sources and transformation of data into a suitable form for particular Web mining operations to be applied on it. Server log file is the input to the preprocessing phase. Following figure 2 shows sample entries of log file of the website http://www.southtexasshooting.org.

```
202.203.132.250 - - [13/Nov/2014:10:01:38 +0800] "GET
/pluginfile.php/1/theme_clean/logo/1415779681/log3.jpg HTTP/1.1" 200 46151 "http://202.203.132.250/"
"Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.2; Trident/4.0; .NET CLR 1.1.4322; .NET CLR
2.0.50727; .NET CLR 3.0.4506.2152; .NET CLR 3.5.30729)"

202.203.132.250 - - [13/Nov/2014:10:01:38 +0800] "GET
/theme/image.php/_s/clean/core/1415779681/t/collapsed HTTP/1.1" 200 187 "http://202.203.132.250/"
"Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.2; Trident/4.0; .NET CLR 1.1.4322; .NET CLR
2.0.50727; .NET CLR 3.0.4506.2152; .NET CLR 3.5.30729)"

202.203.132.250 - - [13/Nov/2014:10:01:39 +0800] "GET
/theme/image.php/_s/clean/core/1415779681/t/block_to_dock HTTP/1.1" 200 233 "http://202.203.132.250/"
"Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.2; Trident/4.0; .NET CLR 1.1.4322; .NET CLR
2.0.50727; .NET CLR 3.0.4506.2152; .NET CLR 3.5.30729)"
```

Figure 2. Sample Entries from Web Log File

Data preprocessing involves following tasks:

- Data cleaning: Data cleaning involves task of removing irrelevant items from the log file. For data cleaning checking of suffix of URL name in log file is done. If the filename suffix is html, htm or php then they are kept else removed.
- User identification: IP of the user, referrer and user agent are combined to identify the unique users. All entries of every unique user are grouped together.
- User session identification: Each user data grouped together is partitioned into sessions. Every session represents single visit to the website. For this purpose heuristic method is used partition into session is done by setting a 30 min threshold time elapse between two successive log entries.
- Path completion: In this step, missing page references inferred from the site topology and temporal information taken from server logs are added.

3.2 Transaction Identification

Transaction is the meaningful cluster of references for every user. Some of the data mining algorithms are unable to handle large sessions or fine grained sessions. Thus as per the requirement of the algorithm, divide or merge approach is followed to create the transactions. Time window approach is used to create the meaningful transactions.

3.3 Pattern Extraction

The goal of pattern extraction phase is to create and analyze user behavioral model. Frequently used navigational paths of the users inside the website are extracted using sequential data mining techniques. Discovered patterns are filtered and converted to

aggregate user model to give as a input to the applications such as visualization tools, recommendation engine, Web analytics and report generation tools. In this phase, GSP algorithm is applied on preprocessed log file.

Algorithm: GSP algorithm proposed by Agrawal and Srikant is used for extracting frequently occurring sequences.
Input: Sequence database obtained by output of data preparation and preprocessing phase.
Output: The complete set of frequent sequential patterns.
Procedure: F1 = the set of frequent 1-sequence k=2,
do while F (k-1)! = Null;
Generate candidate sets Ck (set of candidate k-sequences);
For all input sequences s in the database D
Do
Increment count of all a in Ck if s supports a
Fk = {a ∈ Ck such that its frequency exceeds the Threshold}
k = k+1;
Result = Set of all frequent sequences is the union Of all Fks
End Do
End do

3.4 Pattern Visualization and Analysis

Visualization of extracted pattern is also called as meta mining. The result of discovered patterns in sequential pattern extraction is not in human understandable form and needs to be analyzed and visualized once again as per requirement of the research work. Thus visualization is used.

Sometimes the pages which are used lesser number of times has important contents. Sometimes unnecessarily long convoluted traversal paths are given in the website. From the visualization output such usability issues are checked. This tells that the site structure is not in intuitive manner and needs to be modified.

IV. EXPERIMENTAL ANALYSIS

Figure 3 given below shows Screenshot of output of transaction identification using time window approach followed by Figure 4 which shows the output of GSP algorithm.

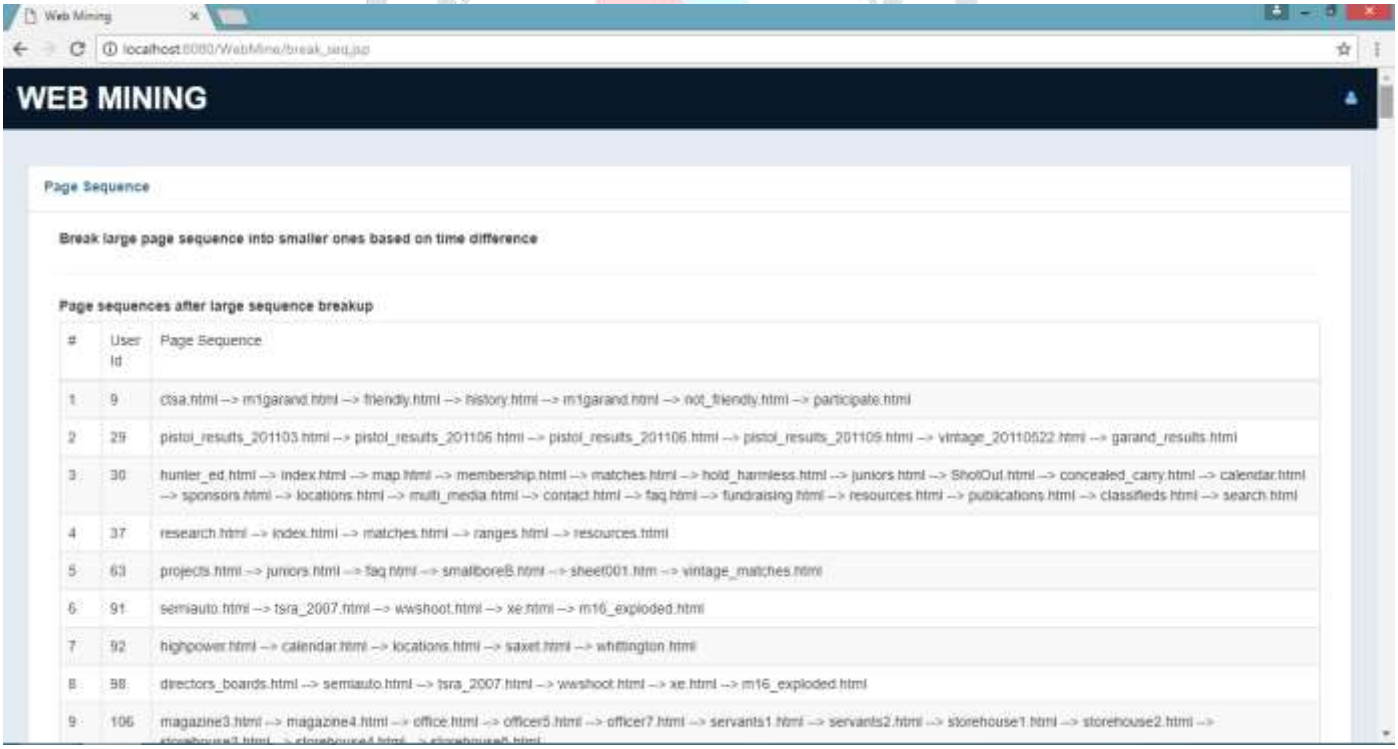

Figure 3. Screenshot of Output of Transaction Identification using Time Window Approach

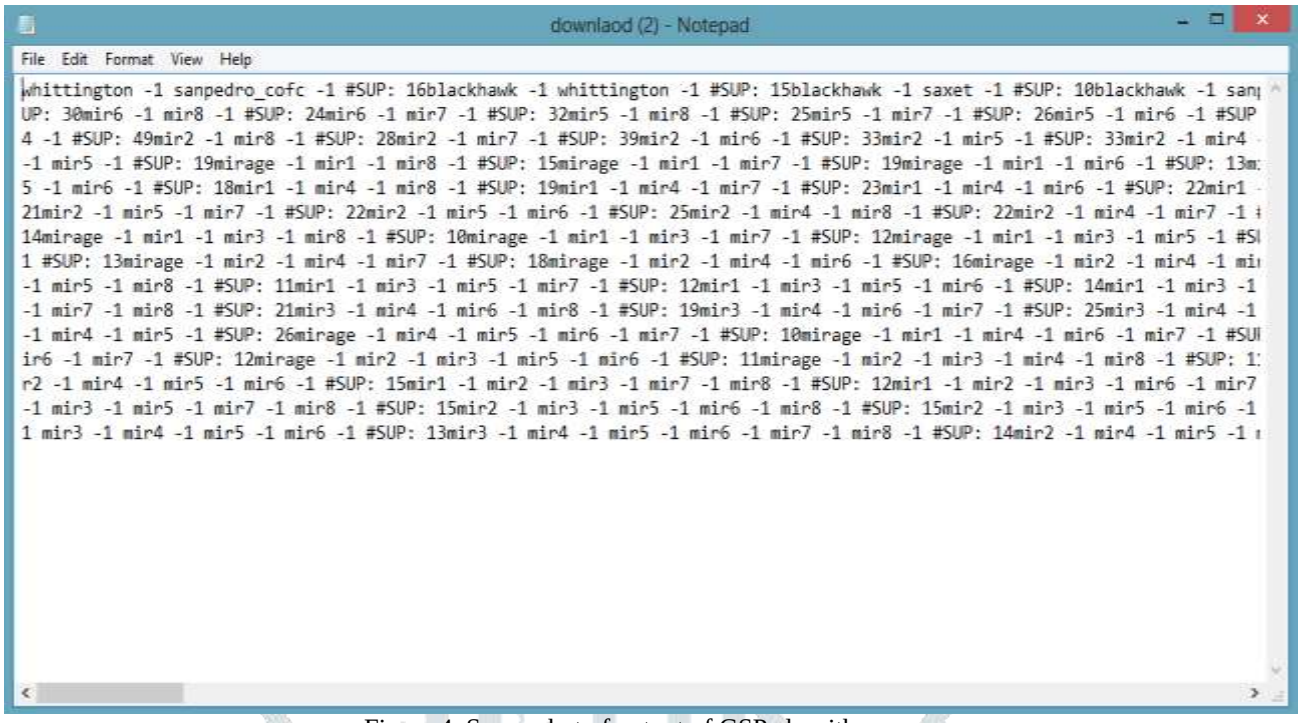

Figure 4. Screenshot of output of GSP algorithm

Figure 5 given below shows results of pattern analysis phase. The results of visualization helps Web analyst to Web developer how the entire Web application is being used by the users. The Web developer can bring resources which are more frequently visited by users to higher level of tree. Thus efforts of users to access these resources are reduced resulting in improved website usability. Also Web analyst can focus on least visited links, which can be removed to free up Web space which ultimately improves Web application performance. Nodes in the visualization output represents the pages of the website whereas arrows connecting them represents the navigation paths followed by the users.

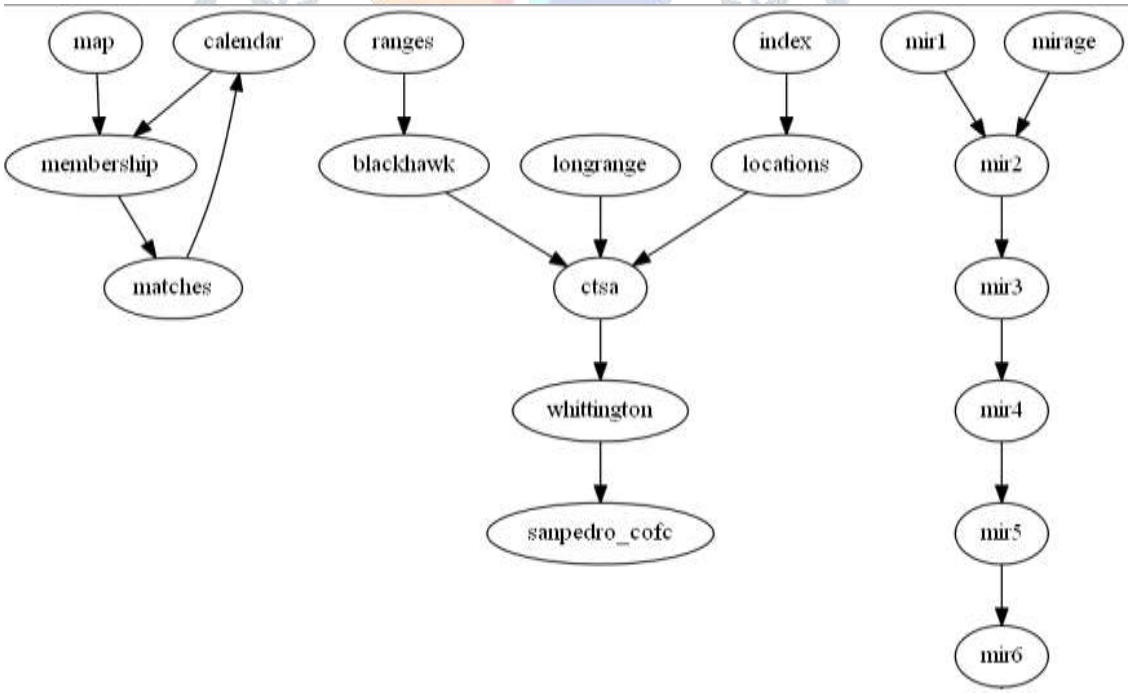

Figure 5. Results of Pattern Analysis Phase

V. CONCLUSIONS

Proposed approach in this paper is used to design and restructure the physical layout of Web server. This approach is also used to improve the site's static structure within the underlying hypertext system.

In future work usage information can be used for providing the list of popular destinations from the particular Web page. Usage information can also be used to customize and adapt the site's interface for the individual user. Additional pattern extraction methods can be explored.

REFERENCES

[1] RuiliGeng, and Jeff Tian, Member, IEEE, "Improving Web Navigation Usability by Comparing Actual and Anticipated Usage", IEEE Transactions on Human–Machine Systems, Volume No. 45, 2015.

[2] R. Cooley, B. Mobasher, and J. Srivastava,, "Data preparation for mining World Wide Web browsing patterns," Knowl. Inf. Syst., vol. 1, no. 1, pp. 5–32, 1999.

[3] F. E. Ritter, A. R. Freed, and O. L. Haskett, "Discovering user information needs: The case of university department Web sites," ACM Interactions, vol. 12, no. 5, pp. 19–27, 2005.

[4] 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.

[5] 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.

[6] 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. J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin, "An integrated theory of the mind," Psychol. Rev., vol. 111, pp. 1036–1060, 2004.

[7] Guosheng Kang, Mingdong Tang, Jianxun Liu Xiaoqing and Buqing Cao, "Diversifying Web Service Recommendation Results via Exploring Service Usage Histor" IEEE Transactions on Services Computing Vol. 9, Issue 4, July-Aug. 2016.

[8] C. M. Nadeem Faisal, Martin Gonzalez-Rodriguez, Daniel Fernandez-Lanvin, and Javier de Andres-Suarez, "Web Design Attributes in Building User Trust, Satisfaction, and Loyalty for a High Uncertainty Avoidance Culture", IEEE Transactions on Human–Machine Systems, Volume No. 47, Issue 6, Dec 2017.

[9] Sneha V. Dehankar, K. P. Wagh and P. N. Chatur, "Web Page Classification Using Apriori Algorithm and Naive Bayes Classifier" IJARCSMS volume 3, issue 4, April 2015, pg-527- 533.

[10] Hetal C. Chaudhari, K. P. Wagh and P. N. Chatur, "Search Engine Results Clustering Using TF-IDF Based Apriori Approach" IJECS vol. 4, Issue 5, May 2015, Pg-11956- 1196.

[11] Sanjay D. Sawaitul, Prof. K. P. Wagh, Dr. P. N. Chatur," Classification and Prediction of Future Weather by usingBack Propagation Algorithm-An Approach",IJETAE,ISSN 2250-2459, Volume 2, Issue 1, January 2012.

[12] 10. P. H. Govardhan, K. P. Wagh, P. N. Chatur," Web Document Clustering using Proposed Similarity Measure", International Journal of Computer Applications (0975 – 8887) National Conference on Emerging Trends in Computer Technology (NCETCT-2014).

[13] M. F. Arlitt and C. L. Williamson, "Internet Web servers: Workload characterization and performance implications," IEEE/ACM Trans. Netw., vol. 5, no. 5, pp. 631–645, Oct. 1997.

[14] C. Kallepalli and J. Tian, "Measuring and modeling usage and reliability for statistical Web testing," IEEE Trans. Softw. Engin., vol. 27, no. 11, pp. 1023–1036, Nov. 2001.

[15] Tec-Ed, "Assessing Web site usability from server log files," White Paper, Tec-Ed, 1999.