ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

A Web-Based Recommendation System to Predict User Future Movements

R. Padmapriya[1], D. Maheswari[2]

Research Scholar, Department of Computer Science, Rathnavel Subramaniam College of Arts & Science, Sulur, Coimbatore, TN, India

Assistant Professor & Head research, School of Computer Studies, Rathnavel Subramaniam College of Arts & Science, Sulur, Coimbatore, TN, India

ABSTRACT: Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of web-based applications. Web usage mining consists of three phases, preprocessing, pattern discovery, and pattern analysis. Data preprocessing is the process to convert the raw data into the data abstraction necessary for the further applying the data mining algorithm. The preprocessed web log file can then be suitable for the discovery and analysis of useful information referred to as web mining. To fulfill this requirement the navigations are recorded in web log file as well as the IP address of the website, session of usage & visited web link. This paper presents several data preparation techniques that can be used to improve the performance of data preprocessing in order to identify unique users and user sessions. These techniques and algorithms have been proved valid and efficient by experiments.

KEYWORDS: Web Usage Mining, Data Preprocessing, Longest Common Subsequence, LCS Problem with Fixed Gap

I. INTRODUCTION

Web Usage Mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. The results of Web Usage Mining can be used in personalization, system improvement, site modification, business intelligence, usage characterization and so forth. Generally, Web Usage Mining consists of three processes: data preprocessing, patterns discovery and patterns analysis. As the data sources of patterns discovery, the results' quality of data preprocessing influences the results of patterns discovery directly. Better data sources can not only discover high quality patterns but also improve the algorithm of Web Usage Mining. So, data preprocessing is particularly important for the whole Web Usage Mining processes and the key of the Web Usage Mining's quality. However, Each type of data collection used in data preprocessing differs not only in the terms of the location of the data source, but also the kinds of data available, the segment of population from which the data are collected, and it's method of implementation.

One or several preprocessing techniques individually, cannot guarantee the reliability of overall results of WUM process. Preprocessing phase is a set of inter-connected, coherent, and integrated techniques, applied in a sequence to produce clear and well-defined results. The input for the web usage mining process is a user session file that gives an exact account of who accessed the Web site, what pages were requested and in what order, and how long each page was viewed. A user session is the set of the page accesses that occur during a single visit to a Web site. However, because of the reasons will discuss in the following, the information contained in a raw Web server log does not reliably represent a user session file before data preprocessing. Generally, data preprocessing consists of data cleaning, user identification, session identification and path completion, as shown in Figure 1.

DOI: 10.15680/IJIRCCE.2017. 0504032

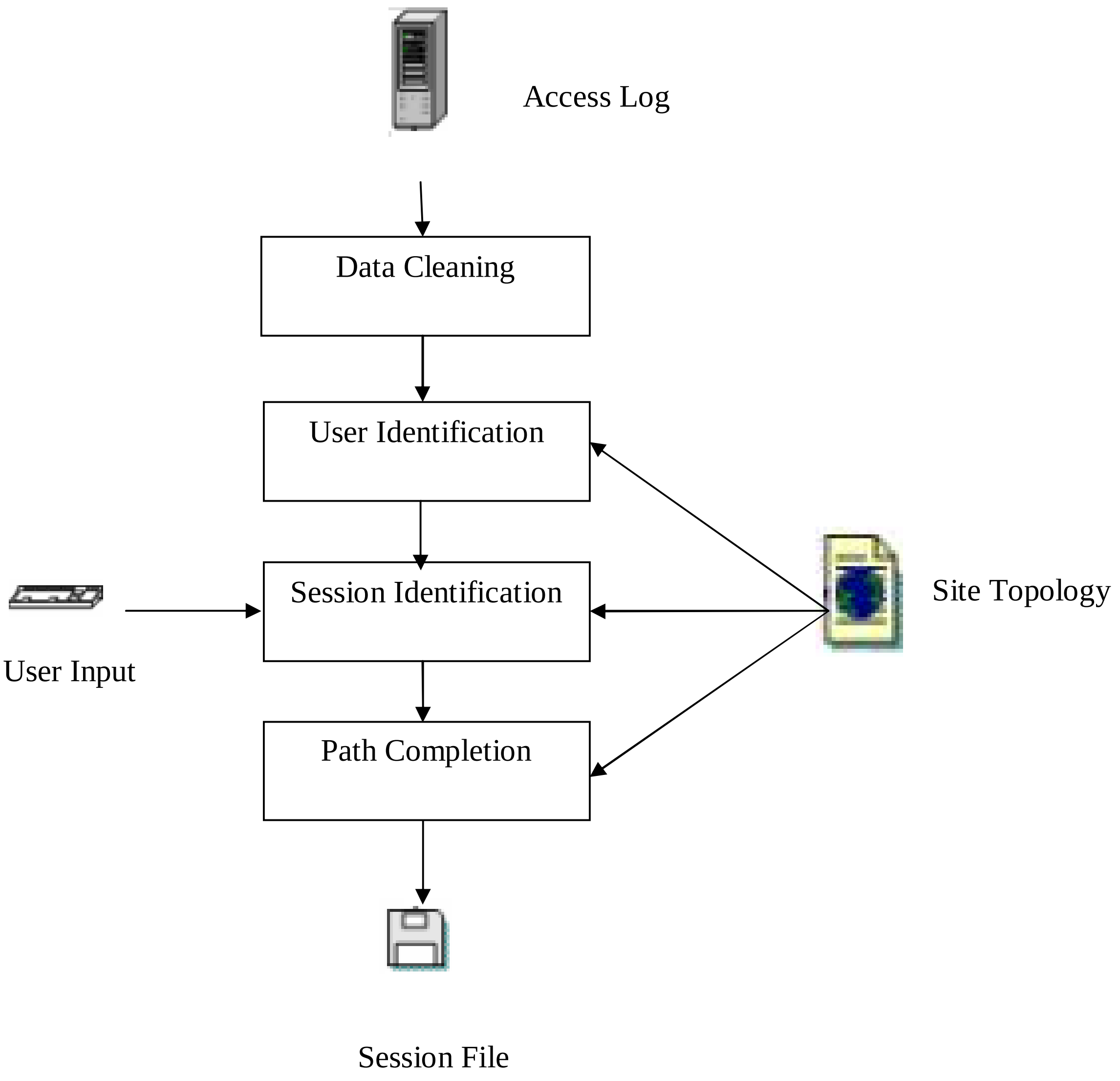

Figure 1: Phases of Data Preprocessing in Web Usage Mining

The research on data preprocessing of Web Usage Mining is a focus field nowadays. This paper attempts to present the process of data preprocess in data preprocessing of Web Usage Mining. The organization of our paper is as follows. In Section 2, we review the related researches in Web Usage Mining. We analyse the processes of data preprocessing in Web Usage Mining in detail and propose the algorithms of each step of data preprocessing in Section 3. In Section 4, an experiment is given to verify the effectiveness and efficiency of our algorithms. Finally, Section 5 concludes this paper with suggestions for future research.

II. LITERATURE SURVEY

At present, the study on Web Usage Mining mainly focuses on pattern discovery (including Association Rules, sequence pattern, etc) and pattern analysis. Given that high-quality data helps a lot in improving Pattern mining precision, Li (2013) studies from this aspects, and proposes the high-effective data preprocessing method. Web mining is an emerging field of data mining used to provide personalization on the web. Jagan and Rajagopalan (2015) focus on web usage mining and algorithms used for providing personalization on the web. Web usage mining (WUM) is a type of Web mining, which exploits data mining techniques to extract valuable information from navigation behavior of World Wide Web users. Aye (2011) mainly focus on data preprocessing stage of the first phase of Web usage mining with activities like field extraction and data cleaning algorithms. Data preprocessing is considered as an important phase of Web usage mining due to unstructured, heterogeneous and noisy nature of log data. Srivastava et al (2015) focused on data fusion, data extraction and data cleaning steps of preprocessing and proposed an algorithm for data extraction which extracts log data according to analysis of time duration. Web Mining plays a vital role in research area in the field of data mining. Hence, Parmar (2015) study some algorithms are presented which can be used according to one's requirement.

Shamsi and Rahul (2012) presents, how web server log data is preprocesses, which includes data cleaning, user identification and Sessionization, path completion. Once the data is preprocessed it is used for discovering some useful patterns. Web usage mining is the process of extracting useful information from users history databases associated to an e-commerce website. Carmona et al (2012) presents the methodology used in an e-commerce website of extra virgin olive oil sale called www.OrOliveSur.com. Web usage mining is realized as a case study on an Indian e-learning site. The objective of Mahajan et al (2014) is the analysis of the web log data of an eLearning system and the deduction of useful conclusions. Nowadays, using data mining techniques to extract knowledge from web log files has become a necessity. Elena (2011) describes the effective and complete preprocessing of access stream before actual mining process can be performed using an example. Web usage analysis requires data abstraction for pattern discovery. Lakshmi et al (2013) presents different formats of web server log files and how web server log data is preprocesses for web usage analysis. Data preprocessing is an important activity for discovering behavioral patterns.

Reddy et al (2012) focuses on the preprocessing techniques implemented on a specially designed Web Sift (WebIS) tool on an IIS web server and also proposes some efficient heuristics and techniques. The focus of Pathak et al (2015) is to establish an algorithm for pattern discovery based on the association between the users's accessed web pages. The biggest constrain for mining web usage patterns are computation overhead and memory overhead. Web mining has-been explored to a vast degree and different techniques have been proposed for a variety of applications that includes Web Search, Classification and Personalization etc. Raju and Suresh (2015) highlight the significance of studying the evolving nature of the Web personalization. E-Commerce websites are considered as face or representatives of their respective companies. In this regard Verma et al (2015) are going to discuss improved mining strategies which are required to mai ntain optimized website structure which in turn is helpful for businesses to increase their revenues, to keep check on competitor's websites, comparison of various brands, attracting new customers and to retain the old customers. Web Usage Mining (WUM) refers to extraction of knowledge from the web log data by application of data mining techniques. Web Log Preprocessing Methods to efficiently identify users and user sessions have been implemented and results have been analyzed by Shivaprasad et al (2015).

III. RESEARCH METHODOLOGY

A new architecture is proposed, which is to predict user future requests. The model is being partitioned into two interleaved phases; off-line and online. In spite of being separated in the model, the off-line phase strongly affects the online phase.

DOI: 10.15680/IJIRCCE.2017. 0504032

Longest common subsequence (LCS)

The Longest Common Subsequence algorithm is to classify user navigation patterns and predicts users' future requests. The main aim of LCS is to find the longest subsequence common to all sequences in a set of sequences. The algorithm works with two features. The first property states that if two sequences X and Y both end with the same element, then their LCS will be found by removing the last element and then finding LCS of the shortened sequence. The second property is used when the two sequences X and Y does not end with the same symbol. Then, the LCS of X and Y is the longest sequence of LCS (X_n, Y_{m-1}) and LCS (X_{n-1}, Y_m).

Recommender systems are alternative, user-centric, promising approaches to tackle the problem of information overload by adapting the content and structure of websites to the needs of the users by taking advantage of the knowledge acquired from the analysis of the users' access behaviors. Fast LCS algorithm gives the opportunity to handle the degenerate strings in all the variants of the LCS problem.

LCS Problem with Fixed Gap

In this section present a naive algorithm for FIG. then show how to improve this algorithm with some non-trivial modifications. Note that, in FIG, due to the gap constraint, a continuing common sequence may have to stop at an arbitrary T [i, j] because the next match is not within the gap constraint. Here the idea is to determine the longest common subsequences for all possible prefix combinations of the input strings. The recurrence relation for extending the length of LCS for each prefix pair (X[1..i], Y [1..j]), i.e. r(X[1..i], Y [1..j]), is as follows:

$$T[i,j] = \begin{cases} 0 & if\ i = 0\ or\ j = 0, \\ T[i-1, j-1] + 1 & if\ X[i] = Y[j], \\ \max\left(T[i-1,j], T[i,j-1]\right) & if\ X[i] \neq Y[j]. \end{cases} \quad (1)$$

Here have used the tabular notion T [i, j] to denote r(X[1..i], Y [1..j]). After the table has been filled, r(X, Y) can be found in T [n, n] and lcs(X, Y) can be found by backtracking from T [n, n]. Unfortunately, the attempt to generalize this algorithm in a straightforward way doesn't give us an algorithm to solve our problems. Note that, in FIG, due to the gap constraint, a continuing common sequence may have to stop at an arbitrary T [i, j] because the next match is not within the gap constraint. In order to cope with this situation what we do is as follows. For each tabular entry T [i, j], (i, j) 2 M we calculate and store two values namely Tlocal[i,j] and Tglobal[i,j]. For all other (i,j), Tlocal[i,j] is irrelevant and, hence, is undefined. The recurrence relations are defined below:

$$T_{local}[i,j] = \begin{cases} Undefined & if (i,j) \notin M, \\ max_{\substack{i-1-K \le l_i < i \\ j-1-K \le l_j < j \\ (l_i, l_j) \in M}} \left(T_{local}[l_i, l_j]\right) + 1 & if (i,j) \in M \end{cases} \quad (2)$$

```
for $a \in \Sigma$ do
    Insert the positions of a in X in $L_X$ [a] in sorted order
    Insert the positions of a in Y in $L_Y$ [a] in sorted order
end for
$E_P$ = $\in$ {initializing a vEB data structure}
for $a \in \Sigma$ do
    for $i \in L_X[a]$ do
        for $j \in L_Y[a]$ do
            insert (i, j) in $E_P$ according to $(i * (n-1) + j)$
        end for
    end for
end for
return E
```

Algorithm 1 Pre-processing step to get M in the prescribed order

IV. EXPERIMENTAL RESULT

The experimental evaluation of the proposed method LCS Problem with Fixed Gap is discussed in this session. This experiment is evaluated by using two different datasets such as CTI and MSNBC is used for the performance; the average length of a user session is about three pages. Since we still have almost half of all the sessions, we then choose this value as the minimum length for an active session to be classified. The experimental results show that our approach improved the quality of clustering for user navigation pattern and the quality of recommendations for both CTI and MSNBC datasets.

Table 1: Accuracy and Execution Time for proposed method

Methods	**Accuracy (%)**	**Execution Time (Seconds)**
Longest Common Subsequence	81	22
LCS Problem with Fixed Gap	90	15

Table 1 shows the accuracy and execution time for longest common subsequence and LCS Problem with Fixed Gap techniques. The proposed method of LCS Problem with Fixed Gap method has high accuracy and less execution time when compare with standard longest common subsequence technique.

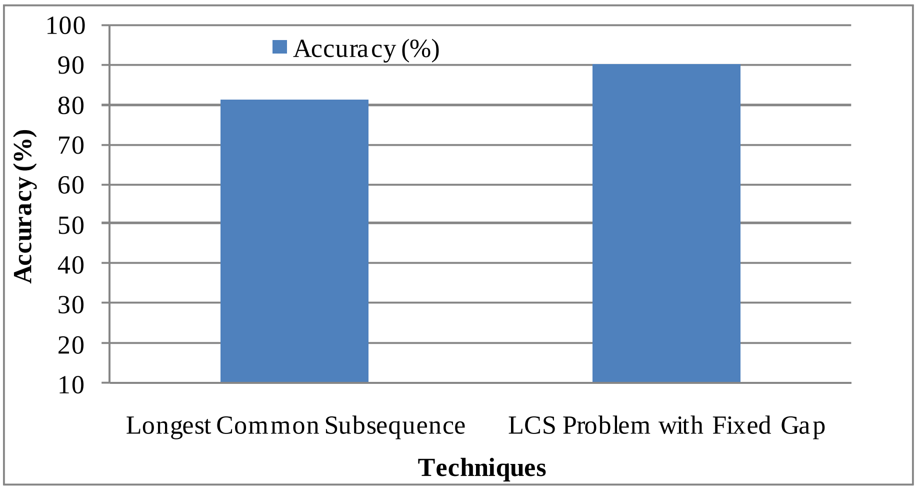

Figure 2: Accuracy for recommendation

Figure 2 shows the accuracy for LCS and LCS Problem with Fixed Gap techniques. Thus the proposed method of LCS Problem with Fixed Gap has high accuracy when compare with another technique.

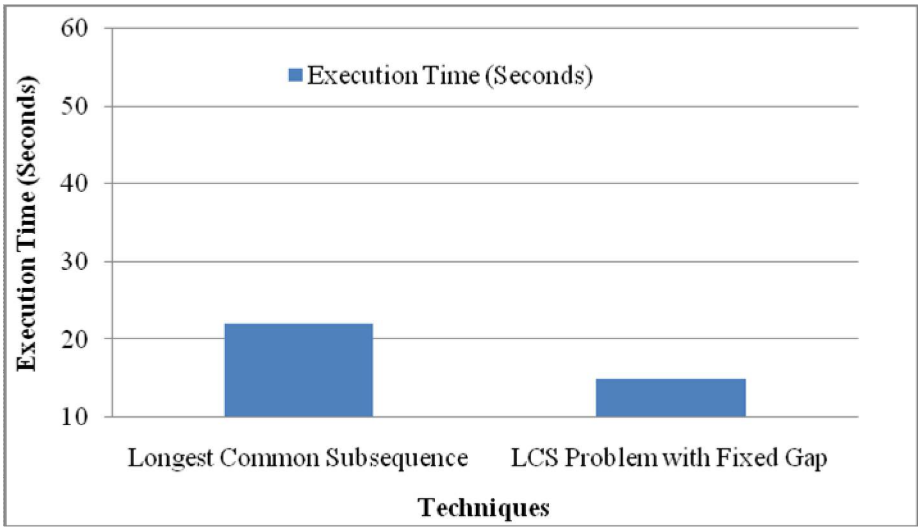

Figure 3: Execution time for recommendation

Figure 3 shows the execution time for LCS and LCS Problem with Fixed Gap, thus the proposed method of LCS Problem with Fixed Gap has less execution time when compare with the longest common subsequence technique. So, thus the LCS Problem with Fixed Gap performs better function than other technique.

V. CONCLUSION

Web usage mining is indeed one of the emerging area of research and important sub-domain of data mining and its techniques. In order to take full advantage of web usage mining and its all techniques, it is important to carry out preprocessing stage efficiently and effectively. Log files are the best source to know user behavior. But the raw log files contains unnecessary details like image access, failed entries etc., which will affect the accuracy of pattern discovery and analysis. So preprocessing stage is an important work in mining to make efficient pattern analysis. To get accurate mining results user's session details are to be known. Once preprocessing stage is well-performed, we can apply data mining techniques like clustering, association, classification etc for applications of web usage mining such as business intelligence, e-commerce, e-learning, personalization, etc.

REFERENCES

1. Li, Xiang-ying. "Data Preprocessing in Web Usage Mining." In *The 19th International Conference on Industrial Engineering and Engineering Management*, pp. 257-266. Springer Berlin Heidelberg, 2013.
2. Jagan, S., and S. P. Rajagopalan. "A Survey on Web Personalization of Web Usage Mining." (2015).
3. Aye, Theint Theint. "Web log cleaning for mining of web usage patterns." In*Computer Research and Development (ICCRD), 2011 3rd International Conference on*, vol. 2, pp. 490-494. IEEE, 2011.
4. Srivastava, Mitali, Rakhi Garg, and P. K. Mishra. "Analysis of Data Extraction and Data Cleaning in Web Usage Mining." In *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015)*, p. 13. ACM, 2015.
5. Parmar, Devendra. "SURVEY ON WEB USAGE MINING AND PRE-FETCHING." *Development* 2, no. 3 (2015).
6. Shamsi, Arshi, and Rahul Nayak. "Web Usage Mining by Data Preprocessing 1." (2012).
7. Carmona, Cristóbal J., S. Ramírez-Gallego, F. Torres, E. Bernal, M. Jose del Jesus, and Salvador García. "Web usage mining to improve the design of an e-commerce website: OrOliveSur. com." *Expert Systems with Applications* 39, no. 12 (2012): 11243-11249.
8. Mahajan, Renuka, J. S. Sodhi, and Vishal Mahajan. "Web Usage Mining for Building an Adaptive e-Learning Site: A Case Study." *International Journal of e-Education, e-Business, e-Management and e-Learning* 4, no. 4 (2014): 283.
9. Elena, Dinuca Claudia. "The process of data preprocessing for Web Usage Data Mining through a complete example." *Ovidius University Annals, Economic Sciences Series* 11, no. 1 (2011): 610-612.
10. Lakshmi, Naga, Raja Sekhara Rao, and Sai Satyanarayana Reddy. "An Overview of Preprocessing on Web Log Data for Web Usage Analysis."*International Journal of Computer Applications. India* 2, no. 4 (2013): 274-279.
11. Reddy, K. Sudheer, G. Varma, and I. Ramesh Babu. "Preprocessing the web server logs: an illustrative approach for effective usage mining." *ACM SIGSOFT Software Engineering Notes* 37, no. 3 (2012): 1-5.

12. Pathak, Nisarg, Viral Shah, and Chandramohan Ajmeera. "A Memory Efficient Algorithm with Enhance Preprocessing Technique for Web Usage Mining." In*Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2*, pp. 601-608. Springer International Publishing, 2015.
13. Raju, Y., and D. Suresh Babu. "A NOVEL APPROACHES IN WEB MINING TECHNIQUES IN CASE OF WEB PERSONALIZATION." (2015).
14. Verma, Neha, and Jatinder Singh. "Improved Web Mining for E-commerce Website Restructuring." In *Computational Intelligence & Communication Technology (CICT), 2015 IEEE International Conference on*, pp. 155-160. IEEE, 2015.
15. Shivaprasad, G., NV Subba Reddy, and U. Dinesh Acharya. "Knowledge Discovery from Web Usage Data: An Efficient Implementation of Web Log Preprocessing Techniques." *International Journal of Computer Applications*111, no. 13 (2015).