Web Page Pattern Prediction Model Based on Dynamic Apriori Algorithm

Nisha Soni[1] Pushpednra Kumar Verma[2]
[1,2]Department of Computer Science & Engineering
[1,2]Chhatrapati Shivaji Institute of Technology Durg, Chhattisgarh

Abstract— Web usage mining deals with understanding user behavior in interacting with a Web site or with the Web. The goal of Web usage mining is to discover Web page navigation pattern, which predicts the path of Website visitors. This navigation patterns are discovered using various techniques available for Web log mining. The accuracy of this navigation patterns are dependent on the quality of the Web log data hence, pre-processing of Web log files are necessary before application of specific technique for Web log mining. In this paper we propose a Web page navigation pattern mining approach using Apriori algorithm with dynamic programming approach. This dynamic programming approach of Apriori algorithms helps in utilizing memory space and minimizing execution time.

Key words: Web Usage Mining, Navigation Pattern, Pre-Processing, Dynamic Programming Approach

I. INTRODUCTION

Nowadays, vast amount of data would be easily generated and collected from the Web environment because of the growth of Internet [1]. Due to the significant and rapid growth in the data and the number of users, Web users are facing the problems of information overload and drowning. Hence, how to extract the useful knowledge and information efficiently from this huge amount of Web data become a more important issue. The solution to this problem is Web log mining. Web usage mining is also called Web log mining since Web usage patterns are discovered from Web logs [2].

Web usage mining can be described as discovery and analysis of user accessibility patterns. Web navigation patterns are useful to understand and predict visitors' browsing behaviour and intentions [3]. The process of discovering patterns from access logs is known as Web usage mining or Web log mining [3]. Web usage mining process basically involves three steps: First step is pre-processing or data cleaning which filters duplicate and unwanted data, second is navigation pattern discovery which uses this filtered data for discovery of frequent pattern which predicts the path of Website visitor and final step is analyzing of discovered patterns, this analyzed patterns are used for various applications of Web log mining such as Web service improvement, business intelligence, personalization etc.

In this paper we proposed a new framework of Web usage mining for Web page navigation pattern mining using Apriori algorithm with dynamic programming methodology. The main focus of this paper is on generating more accurate navigation patterns from Web usage logs. The framework we have proposed is based on three major steps. First is pre-processing step, in which filtration operation is performed for cleaning of unwanted and duplicate data to reduce the size of Web log files. Then this cleaned log files are used for user and session identification. In the second step, clusters are created on the basis of their user and session identification. In the final step Dynamic Apriori algorithm is used for navigation pattern discovery and generate association rule for frequent pattern.

II. LITERATURE REVIEW

Guerbas A., Addam O., Zaarour O., Nagi M., Elhajj A. and Ridley M. [4] proposed a method for effective Web log mining and efficient online navigation pattern prediction. In their paper, they have included a refined time-out based heuristic for session identification. Secondly, they suggest the usage of a specific density based algorithm for navigational pattern discovery. A new approach for efficient online prediction is also suggested.

Jilhedar N. P. and Shirgave S. K. [5] proposed an approach for generation of frequent pattern using semantic related frequent patterns. The quality of Web usage pattern generated is measured with standards methods for evaluation. The methodology includes preprocessing, rule and pattern generation and results evaluation measures. Preprocessing involves pruning, extraction of navigation history and mapping to ontology instances. Rule and pattern generation incorporates sequential association rule mining, the frequent patterns generated tend to maintain the sequence relation between the set of items discovered. Results evaluation measures used to evaluate generated patterns, preference is given to build recommendation engines.

III. PROPOSED SYSTEM

A. Data Pre-Processing

The Web log file generated as a result of users' communication with the Web server is provided as input to the pre-processing step [2]. Data pre-processing is essential stage that improves the quality of data which can be done by data collection, data cleaning, user identification, session identification, path completion, transaction identification and formatting [6].

1) Data Collection

Generally, in web log files huge number of records is inserted at server which may be created different log files day wise or month wise. So records of all log files are gathered into one log file in beginning of the data pre-processing [6].

2) Data Cleaning

Cleaning is necessary at initial stage. The data cleaning is process that filter irrelevant information/fields/records that are not required for mining.

3) User Identification

User identification means identifying unique users by observing their IP address. Following rules to identify unique users: 1) If there is new IP address then there is a new client; 2) If the IP Address is same but the operating

system or browsing software are different, a reasonable assumption is that each different agent type for an IP address represents a different client; 3)If the IP address, operating system and browsers are all the same , the client can be a new one by identifying whether the requesting page can be reached by the pages accessed before, according to the topology of the site [6].

4) Session Identification

A set of pages visited by the same user within the duration of one particular visit to a web-site can be defined as a user session. A user may have a single or multiple sessions during a period, once a user was identified, the click stream of each user is portioned into logical clusters [7].

B. Clustering

In this step user clusters are created on the basis of the user identification and session identification for further application of dynamic Apriori algorithm to discover navigation patterns.

C. Pattern Discovery

Pattern discovery is the major step in Web usage mining, which provides navigation patterns for Website visitors. There are various methods are available for pattern discovery such as association rules, clustering, classification, sequential pattern analysis etc. In this paper we use association rule mining technique. Many algorithms for generating association rules were presented over time. Some well known algorithms are Apriori, Eclat and FP-Growth. Apriori algorithm requires threshold values minimum support and minimum confidence to find out frequent patterns from database [8]. There are two main functions available in Apriori to find out association rules. First, based on minimum support count, it finds frequent item set. After that association rules between frequent items are find out on the basis of minimum confidence [8].

Dynamic programming is one of the techniques to design an efficient algorithm [8]. This technique store the previous solutions in a table, so when the same problem reappear no need to calculate again, it can be directly access from table which store the solutions without creating more overhead[8][9]. Many problems solve in an optimal way by dynamic programming approach. i.e. Matrix Chain Multiplications, Longest Common Sequence[8][9].

IV. EXPERIMENTAL RESULTS

In this project, the system discovers navigation patterns for all genuine users which predict the path of respected user. The system takes the log file as input and the overall result of the system shows navigation patterns for genuine users and association among frequent pages.

Here we compare our proposed system with longest common sequence (LCS) algorithm. We take different sizes of Web log datasets and implement algorithms on it. And then take the experimental results which are shown in following figure. We compare the experimental results on the basis of:

- Accuracy,
- Precision and
- Execution time

Here Table 1 shows the computed values of accuracy for both proposed system and LCS approach for various sizes of datasets.

After the computation process result shown that proposed system provides higher accuracy than LCS approach

Various sizes of datasets	Accuracy	
	Proposed (AprioriDP)	LCS
1000	65.36	60.7
2000	67.57	61.02
3000	71.61	61.23
4000	77.21	61.32

Table 1: Accuracy of Proposed & LCS for Various Sizes of Datasets

Fig. 1 shows the line graph of the accuracy values provided by proposed and LCS approach for various sizes of datasets. Here the blue line shows the accuracy values for AprioriDP and the green line shows the accuracy values for LCS. We can see that LCS approach produces lower accuracy value than AprioriDP approach for each segment of datasets.

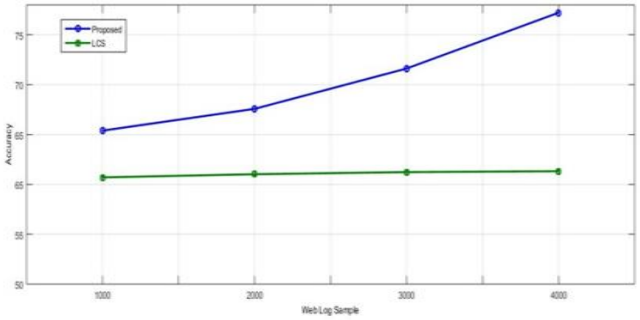

Fig. 1: Accuracy for Various Sizes of Datasets

Here Table 2 shows the computed values of precision rate for both proposed system and LCS approach for various sizes of datasets.

After the computation process result shown that proposed system provides higher precision rate than LCS approach.

Various sizes of datasets	Precision in %	
	Proposed (AprioriDP)	LCS
1000	80	70
2000	85	75
3000	83.33	70
4000	82.5	72.5
5000	84	78

Table 2: Precision in percentage of Proposed & LCS for Various Sizes of Datasets

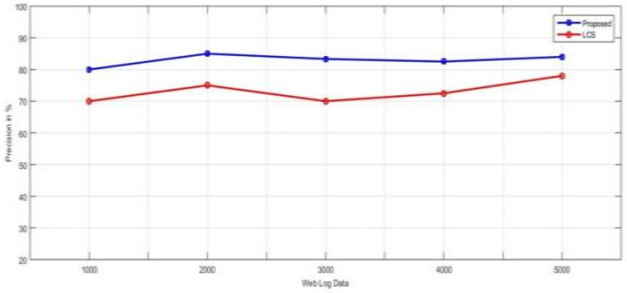

Fig. 2: Precision in % for Various Sizes of Datasets

Fig. 2 shows the line graph of the precision values generated by proposed and LCS approach for various sizes of datasets. Here the blue line shows the precision rate for

AprioriDP and the red line shows the precision rate for LCS. We can see that LCS approach produces lower precision rate than AprioriDP approach for each segment of datasets.

Here Table 3 shows the computed values of execution time for both proposed system and LCS approach for various sizes of datasets.

After the computation process result shown that initially both proposed and LCS approach takes approximately same execution time for small sizes of datasets but when size of datasets increases proposed approach takes less execution time than LCS approach.

Various sizes of datasets	Execution time in seconds	
	Proposed (AprioriDP)	LCS
1000	0.19	0.25
2000	0.33	0.33
3000	0.67	0.69
4000	0.81	0.85
5000	1.01	1.12

Table 3: Execution Time of Proposed & LCS for Various Sizes of Datasets

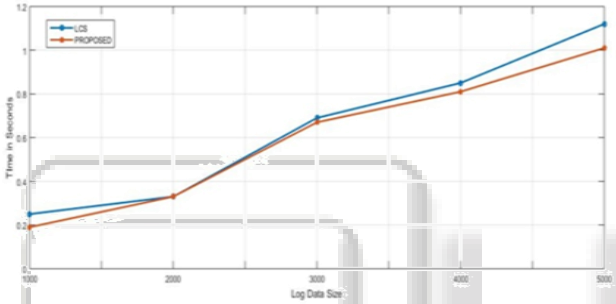

Fig. 3: Execution time for Various Sizes of Datasets

V. CONCLUSION

Web usage mining is basically a process to discover hidden and interesting user navigation patterns that helps the Website administrators to better serve the needs of their Website users. We proposed a system which is based on three steps: pre-processing, clustering and pattern discovery. Our main focus is on generating navigation patterns of users and association among frequent Web pages on the basis of minimum support and minimum confidence.

REFERENCES

[1] Jia-Ching, Ying, Chu-Yu, Chin & Tseng, Vincent S. 2012. Mining Web Navigation Patterns with Dynamic Thresholds for Navigation Prediction. International Conference on Granular Computing (pp.614-619)IEEE. DOI:10.1109/GrC.2012.6468696

[2] Bhargav, A., & Bhargav, M. 2014. Pattern Discovery and Users Classification Through Web Usage Mining. International Conference on Control, Instrumentation, Communication and Computational Technologies (pp. 632-636). IEEE. DOI: 10.1109/ICCICCT. 2014. 6993038

[3] Yao-Te & Anthony J.T. 2011. Mining Web navigation patterns with a path traversal graph. Journal of Expert Systems with Applications, 38 (6), 7112-7122. ELSEVIER

[4] Guerbas, A., Addam, O., Zaarour, O., Nagi, M., Elhajj, A., & Ridley, M. 2013. Effective web log mining and online navigational pattern prediction. Journal of Knowledge-Based Systems, 49, 50-62. ELSEVIER.

[5] Jilhedar, N.P., & Shirgave, S. K. 2014. User Web Usage Mining for Navigation Improvisation Using Semantic Related Frequent Patterns. International Conference on Computer and Communications Technologies (pp. 1-5). IEEE. DOI: 10.1109/ICCCT2.2014.7066697

[6] Upadhyay J. B. & Patel S. V. 2015. A Review Analysis of Preprocessing Techniques in Web usage Mining. International Journal of Engineering Research & Technology, 4(4), 1160-1166.

[7] Chitraa V. & Davamani A. S. 2010. A Survey on Preprocessing Methods for Web Usage Data. International Journal of Computer Science and Information Security, 7(3), 78-83.

[8] Murjani S. & Rajput I. 2014. Finding Frequent Items Dynamically. International Journal of Computer Science and Information Technologies, 5(3), 3690-3694.

[9] Bhalodiya, D., Patel, K. M., & Patel, C. 2013. An Efficient way to Find Frequent Pattern with Dynamic Programming Approach. Nirma University International Conference on Engineering (pp. 1-5). IEEE. DOI: 10.1109/NUiCONE.2013.6780102

[10] Gupta, G. K. (2009). Introduction to Data Mining with Case Studies. Clayton, Australia: Prentice Hall Of India Pvt. Limited.

