PATH COMPLETION FOR EFFECTIVE WEB PAGE ANALYSIS

*Jothish Chembath**

Abstract

Data preprocessing is required to do a meaningful analysis of web log files. It helps in adding support and efficiency during analysis of data. Preprocessing make the data more reliable and meaningful. Incomplete paths, when it occurs, have to be gently filled. This paper presents an algorithm for path completion. The proposed path completion algorithm efficiently completes the loss of information to improve the reliability of data sets.

Keywords: LCS, Preprocessing, Data cleaning, Path Completion.

I. INTRODUCTION

Users visit many web sites consisting of numerous pages for obtaining information. Transactions involved need to be identified in a session. A transaction usually contains a collection of pages, but not all pages of a session. Users who visit or browse the site with no actual interest in the content are considered as irrelevant. These irrelevant users should be removed. Removal of irrelevant users reduces dataset size and improves the time and effectiveness of analysis. Removing the irrelevant users also means filling the significant loop holes that exist when user path is not completed. Then path completion step is completed for obtaining all the transactions in the session.

II. LITERATURE REVIEW

In [1], Jalali et al. has proposed the longest common subsequence (LCS) algorithm for classifying user activities. Their results showed an enhancement of quality recommendations. In [2],Jothish Chembath, S.K. Mahendran has proposed an identification algorithm to separate a session into a trail of significant pages. They have introduced an algorithm that effectively condensed web pages to identify transactions. In [3], R Padmapriya, D. Maheswari has quoted that data in web access log are assimilated by local caching which needs to be organized. They have also proposed that path completion algorithm needs be focused for any proper analysis to be completed.

III. Methodology

This is performed when the count of URLs falls below the actual number of pages visited[4]. Figure 1.0 shows the sequence of unfinished transactions

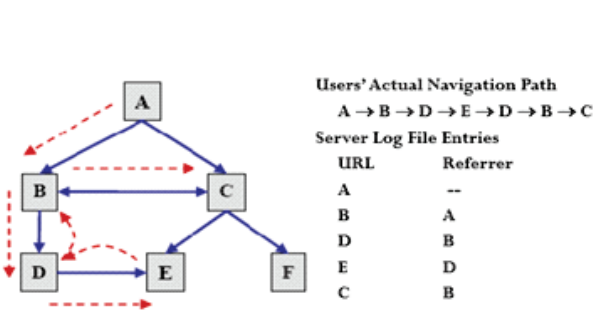

Figure 1.0 Unfinished Transaction

METHOD OF PATH FILLING

Subsequent to path identification:

1. If the URL address, indicated in the referred URL is not matching the address in the preceding record[5], then this web address replaces the recent record in the session for completing the trailing web page.

2. Then reference length of new attached pages is determined.

Department of Computer Science,
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India
*Corresponding Author

3. Reference length of neighboring ones are modified.

4. The reference lengths of neighboring pages are then regulated.

IV. EXPERIMENTAL RESULTS

Performance is evaluated making use of four types of web log after path filling (Table T1). The trials are planned to investigate the efficiency of the scheduled algorithms on the model proposed by Jalali et al. 2010. The metrics used to check the effectiveness of the algorithm are accuracy and coverage.

Table T1: Data Sets used Performance Evaluation

Dataset	Jargon	Period	File Size in MB	Record size
NASA http://ita.ee.lbl.gov/html/contrib)	NASA	01-07-1995 to 31-08-1995	205	34,61,612
University of Saskatchewan http://ita.ee.lbl.gov/html/contrib/Sask-HTTP.html	SASK	01-06-1995 to 31-12-1995	233	24,08,625
ClarkNet http://ita.ee.lbl.gov/html/contrib/ClarkNet-HTTP.html	CN	24-08-1995 to 10-09-1995	171	33,28,587
University of Calgary's, department of Computer Science http://ita.ee.lbl.gov/html/contrib/Calgary-HTTP.html	CL	24-10-1994 to 11-10-1994	52.3	7,26,739

EFFECT OF PREPROCESSING ALGORITHMS ON PREDICTION

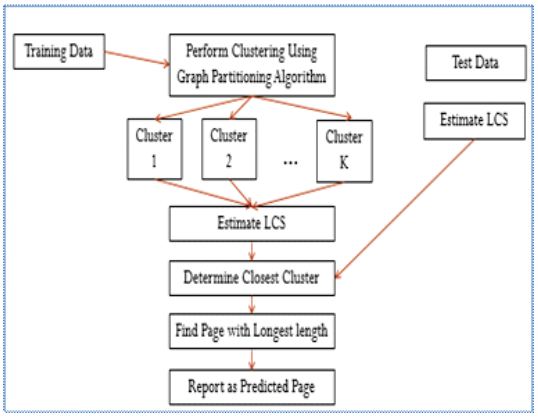

Fig.2 Effect of preprocessing algorithms on prediction

In order to gauge the effect of the projected preprocessing algorithm, the model suggested by Jalali *et al*. 2010 was used. This algorithm used the Longest Common Subsequence (LCS) during prediction. This preprocessing algorithm labeled as LCSPA consists of the following steps as given in the Figure 1.1.

Table T2 : Coding Scheme

Jargon used	Narrative
LCSPA	**Longest Common Subsequence Prediction Algorithm**
LPAC	**LcsPA Contemporary**
LPAPF	**LcsPA Path Filled**

Table T3 : Validity of Preprocessing Algorithm in Predicting Accuracy

Prediction Model	NASA	SASK	CN	CL
LPA	85.62	84.79	86.25	86.94
LPAC	88.63	87.32	88.76	89.08
LPAPF	**90.35**	**88.26**	**89.45**	**90.49**

From the above Table T3, it can be concluded that the path filling operation LPAPF reigns supreme over LCS and Contemporary LPA in giving a better performance in terms of prediction accuracy.

Preprocess Classifier Information analysis by WEKA from the above data

Instances: 3
Attributes: 5
Prediction Model
NASA
SASK
CN
CL
Test mode: evaluation of data

=== Synopsis as analyzed by WEKA===

Exactly Classified Instances	1	33.3333%
Inaccurately Classified Instances	2	66.6667%
Kappa statistic	0	
Mean absolute error	0.444	
Root mean squared error	0.4714	
Relative absolute error	100 %	
Root relative squared error	100 %	
Total Number of Instances	3	

V. COVERAGE

Table T4: Effect of Preprocessing Algorithm on Prediction in terms of Coverage

Prediction Model	NASA	SASK	CN	CL
LPA	1.6325	1.6518	1.6163	1.6068
LPAC	1.3049	1.2996	1.2617	1.2172
LPAPF	**1.2184**	**1.2169**	**1.1897**	**1.1302**

These results are given in table T4 further confirms the effect of preprocessing algorithm LPAPF in terms of coverage performance metric also.

VI. CONCLUSION

Data preprocessing is considered as a vital phase in web mining. Web log analysis is crucial to understand web traffic [6]. Web log analysis makes making easy and accurate precise prediction possible. Clustering step is completed before prediction, which allows partitioning sessions into similar groups. This step gave impetus to the construction of competitive prediction models. Hence it has reduced the complexities in making decisions and helped to minimize the scalability issue to improve the reliability of prediction. Extracting information from the large sets of data is the beckoning task in data mining. At this juncture, clustering needs to be carried out to group the data based on its similarities. Hence separating

the stream of data is required for proper and efficient mining of data.

REFERENCES

[1] Jalali, M., Mustapha, N., Sulaiman, N. and Mamat, A. (2010) WebPUM: A web-based recommendation system to predict user future movements, Expert systems with applications, Vol. 37, Pp. 6201-6212.

[2] JothishChembath and S.K.Mahendran (2015), Transaction Identification Algorithm Enhanced With User Pruning and Combined Maximal Forward Reference and Reference Length Approach for Improving Prediction of Next Web Page from Web Log entries, www.allsubjectjournal.com e-ISSN: 2349-4182,p-ISSN: 2349-5979,Impact Factor: 3.762

[3] R.Padmapriya, D.Maheswari(2017),A Novel Technique for Path Completion in Web Usage Mining, International Journal of Advance Research, Ideas and Innovations in Technology,2017. Volume 3

[4] Yan Li, Boqin Feng, School of Electronics and Information Engineering, Xi'an Jiaotong University, Shaanxi, China, "Research on Path Completion Technique in Web Usage Mining", IEEE Xplore: 30 December 2008

[5] Nirali Honest ,Dr. Atul Patel, Dr. Bankim Patel, Smt. Chandaben Mohanbhai Patel Insitute of Computer, "A study of Path Completion Techniques in Web Usage Mining", 2015 IEEE International Conference on Computational Intelligence & Communication Technology.

[6] Mr. Pratik V. Pande , Mr. N.M. Tarbani , Mr. Pavan V. Ingalkar, Department of Computer Science & Engineering, Prof. Ram Meghe College of Engineering, Amravati, "A Study of Web Traffic Analysis", International Journal of Computer Science and Mobile Computing, Vol. 3, Issue. 3, March 2014, pg.900 – 907.