Proceedings of the 3rd National Conference; INDIACom-2009
Computing For Nation Development, February 26 – 27, 2009
Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi

Preprocessing For Click Stream Data Mining

S. Shrivastava[1], O. P. Vyas[2], Clemens H. Cap[3] and **Anne Gutschmidt[4]**
[1,2]School of Studies in Computer Science, Pandit Ravishankar Shukla University, Raipur, C.G.
[3, 4]Department of Computer Science, University of Rostock (Germany)
{shishir.contact, dropvyas }@gmail.com, {clemens.cap, anne.gutschmidt }@uni-rostock.de

ABSTRACT
Web mining involves a wide range of applications that aims at discovering and extracting hidden information in data stored on the Web. The focus of this paper is to provide an overview of data preparation techniques and algorithms that can be used in order to convert raw log data for click stream data mining. The data in the log files about the actions of the users can not generally be used for mining purposes in the same form as it is stored. For this reason a preprocessing step must be performed before the pattern discovering phase. Here we performed some basic operations for data preparation, which contains three separate phases. Firstly, the collected data must be cleaned, Secondly, the different sessions belonging to different users should be identified, and The third step is to convert the data into the format needed by the mining algorithms. In our experiments we used one web server log file, and perform all the three steps and get results accordingly.

KEYWORDS
Preprocessing web log, weblog mining, web usage mining, Click Stream

INTRODUCTION

The expansion of the World Wide Web (Web for short) has resulted in a large amount of data that is now in general freely available for user access. The different types of data have to be managed and organized in such a way that they can be accessed by different users efficiently. Therefore, the application of data mining techniques on the web is now the focusing area of an increasing number of researchers. Several data mining methods are used to discover the hidden information in the Web[3]. However, Web mining does not only mean applying data mining techniques to the data stored in the Web. The mining algorithms have to be modified such that they better suit the requirements of the web user. New approaches should be used which better fit the characteristics of Web data. Thus, Web mining has been developed into an autonomous research area.

Web mining involves a wide range of applications that aims at discovering and extracting hidden information in data stored on the Web. Another important purpose of Web mining is to provide a mechanism to make the data access more efficiently and adequately. The third interesting approach is to discover the information which can be derived from the activities of users, which are stored in log files. Thus, Web mining can be categorized into three different classes based on which part of the Web is to be mined[4]. These three categories are (i) Web content mining, (ii) Web structure mining and (iii) Web usage mining. Data preprocessing assume special significance in the click stream data mining, as the data logged by the user actions are mouse move, mouse clicks along with various log file attributes and is not directly usable by data mining methods.

In this paper, we propose a complete preprocessing methodology that allows the analyst to transform any collection of web server log to a simple text file for data mining, The log files from different Web sites of the same organization are merged to apprehend the behaviors of the users that navigate in a transparent way. Afterwards, this file is cleaned by removing all unnecessary requests, basically this work is for click stream web log so we clean un-necessary sequences Then, the remaining requests are grouped by user, user sessions. And then conversion for mining data is performed on the data . We have provided filters to filter the unwanted, irrelevant, and unused data. The objective is to considerably reduce the large quantity of Web usage data available and,at the same time, to increase its quality by structuring it and providing additional aggregated variables for the data mining analysis that follow.

PROBLEM DEFINITION

Consider the set R = {r1, r2…, rn} of all Web resources from a Web site. If U = {u1, u2…, um} is the set of all the users who have accessed that site, then the log entry is defined as li =< ui,p, typ,t,d >, where ui ∈ U; p, represents the project, typ represents the event type, t represents the time and, d represents the referring page. Li = {li1, li2…, lin} arranged in the ascending order, constitutes a Web server log. In the case of N Web servers, the set of log files is Log = {L1, L2…, LN}. Using these notations, the preprocessing problem is formulated as follows – "Given a set of log files Log of Web site – extract the Users, User sessions, visits, and click streams of the Web site users with a given time interval Δt".

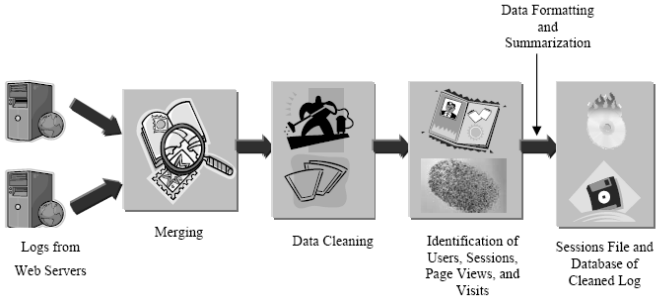
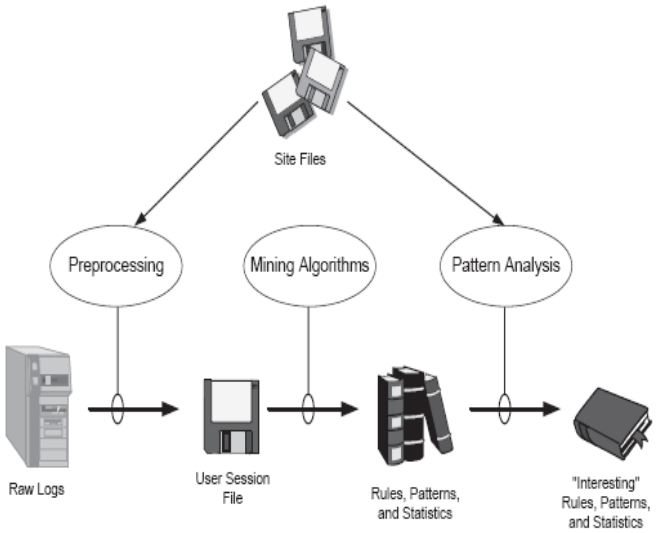


Fig 1 : Preprocessing Processes

DATA PREPROCESSING

The data in the log files of the server about the actions of the users can not be used for mining purposes in the form as it is



stored. For this reason a preprocessing step must be performed before the pattern discovering phase[16].

Fig 2 : Data Mining Processes

It comprises fallowing steps Merging of Log files from Different Web Servers, Data cleaning, Identification of Users, Data formatting and Summarization

1) **Merging**

In this step we merge all log files in one log, after merging all log we sort them by time and then we got our main file in which we perform rest of the processes. Algorithm for merging log is as follows

Step 1 :- create a new file named merge_log
Step 2 :- initialize it's curser
Step 3:- initialize i=1
Step3 :- read the log entries from log file Li
Step 4:- append to merge_log
Step 5:- repeat 3 and 4 until i <=N
Step 6:- Sort the merge_log entries in ascending order based on access time
Step 7:- return merge_log

2) **Data Cleaning**

Second step consist of cleaning useless entries from the log file. Since all the log entries are not valid we need to eliminate the irrelevant entries. Usually, this process removes requests concerning non-analyzed resources such as images, multimedia files, and page style files. But in our case we also have to remove all those entries which are generated during moving on page such as mouse move events occurs several time repeatedly. By filtering out useless data, we can reduce the log file size to use less storage space and to facilitate upcoming tasks.

In our case after data cleaning our file reduce size more than 50%. After data cleaning we have log file not only smaller in size but also less complex. We must perform this task otherwise one click stream like mouse move or mouse over may dominate other stream and give uncertain and meaningless results. This is necessary to provide mining full results in pattern analysis phase

3) **Session Identification**

the different sessions belonging to different users should be identified. A session is understood as a group of activities performed by a user when he is navigating through a given site. To identify the sessions from the raw data is a complex step, because the server logs do not always contain all the information needed. If there is no any user identification available we can use IP address of System although we know that one IP address is used by several users through multiuser system such as a system in cyber café, or in a library. But in our log file we have a user identification number so we don't have to bother regarding this and we get the sequence of requests made by a single user over a certain navigation period.

Session identification task is dependent open the pattern analysis process if we perform sequence analysis than the row must be converted such that they represent sequence. But one row represents single item in item set so we convert and group user id to find sequence of click stream.

And in the item set pattern we need not to do lots of work after data cleaning this step of preprocessing is completed already A row is converted into an itemset by omitting the duplicates of the pages, and sorting them regarding their time. After session identification log file contain only one unique record for a each user

4) **Data Formatting**

This step is required only when we need to identify and group events by their category we can also give them a unique identification code to summarize if required. Here we provide the table for identification of event according to over log file

Code	Group	Event
1	Mouse Event	Mouse Move
		Mouse Over
		Mouse Click
		Mouse In
		Mouse Out
2	Tab event	Open
		Select
		Close
3	Page Event	Show
		Hide
4	Browser Event	Reload
		Open
		Stop
5	Scrolling	Scroll
6	Keystroke	Keypress

Fig 3

EXPERIMENTAL RESULTS

We have conducted experiments in log files. Through this experiment we show that our preprocessing methodology reduces and structured given log files for pattern discovery. Fig 4 shows GUI for preprocessing and the Fig 5 is the log file, log files attributes are owner(user id), project (site name), type (Event), time, detail (visited link, mouse position etc.), and the Fig 6 shows the resultant output log file for pattern discovery.

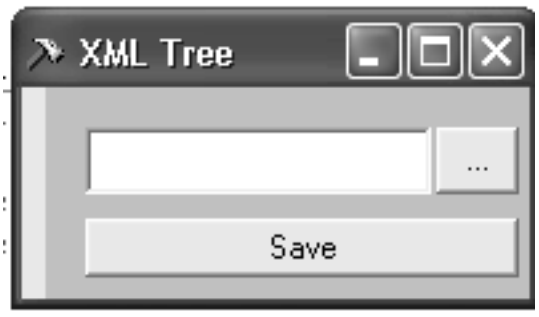


Fig 4 : GUI

```
1206370692984 spiegel mouseclick 1206370723703 menuitemAu
1206370692984 spiegel mouseout 1206370723703 menuitemAufr
1206370692984 spiegel mousemove 1206370723718 704155
1206370692984 spiegel mousemove 1206370723734 845155
1206370692984 spiegel mousemove 1206370723749 913156
1206370692984 spiegel mousemove 1206370723765 954156
1206370692984 spiegel mousemove 1206370723765 995156
1206370692984 spiegel mousemove 1206370723781 1038156
1206370692984 spiegel mousemove 1206370723781 1082156
1206370692984 spiegel mousemove 1206370723796 1123157
1206370692984 spiegel mousemove 1206370723796 1166157
1206370692984 spiegel mousemove 1206370723812 1204158
1206370692984 spiegel mousemove 1206370723812 1238160
1206370692984 spiegel mousemove 1206370723828 1265161
1206370692984 spiegel mousemove 1206370723828 1279163
1206370692984 spiegel mousemove 1206370723843 1279165
1206370692984 spiegel mousemove 1206370723843 1279166
1206370692984 spiegel mousemove 1206370723874 1279167
1206370692984 spiegel mousemove 1206370723921 1279168
1206370692984 spiegel mousemove 1206370723937 1279169
1206370692984 spiegel mousemove 1206370723937 1279171
1206370692984 spiegel mousemove 1206370723953 1279173
1206370692984 spiegel mousemove 1206370723953 1279174
1206370692984 spiegel mousemove 1206370723968 1278175
1206370692984 spiegel mousemove 1206370723984 1277176
1206370692984 spiegel mousemove 1206370723999 1276178
1206370692984 spiegel mousemove 1206370723999 1275179
1206370692984 spiegel mousemove 1206370724015 1273181
1206370692984 spiegel mousemove 1206370724015 1272183
1206370692984 spiegel mousemove 1206370724031 1268188
1206370692984 spiegel mousemove 1206370724031 1267191
1206370692984 spiegel mousemove 1206370724046 1267193
1206370692984 spiegel mousemove 1206370724046 1265197
1206370692984 spiegel mousemove 1206370724062 1264200
1206370692984 spiegel mousemove 1206370724062 1263202
1206370692984 spiegel mousemove 1206370724078 1262205
1206370692984 spiegel mousemove 1206370724093 1260211
1206370692984 spiegel mousemove 1206370724093 1260215
1206370692984 spiegel mousemove 1206370724109 1258221
1206370692984 spiegel mousemove 1206370724109 1257226
1206370692984 spiegel mousemove 1206370724124 1256229
1206370692984 spiegel mousemove 1206370724124 1255233
1206370692984 spiegel mousemove 1206370724140 1255234
1206370692984 spiegel mousemove 1206370724203 1254235
1206370692984 spiegel mousemove 1206370724218 1253237
1206370692984 spiegel mousemove 1206370724406 1252237
1206370692984 spiegel mousemove 1206370724421 1251237
1206370692984 spiegel mousemove 1206370724421 1250237
1206370692984 spiegel mousemove 1206370724437 1249237
1206370692984 spiegel mousemove 1206370724437 1248237
```

Fig 5 : Log File

```
1206370692984 starttask finishtask starttask finishtask mouseclick mouseout mousemove mouseclick mousemove mouseover mousemove
1206371114659 mouseout mousemove mouseover mousemove mouseout mousemove mouseover mousemove mouseout mouseover mousemove mouse
1206371124656 mouseout mousemove mouseover mousemove mouseout mousemove mouseover mousemove mouseout mousemove mouseover mouse
1206371135609 mouseout mousemove mouseover mousemove mouseout mouseover mousemove mouseout mouseover mousemove mouseout mousec
1206371183937 mouseout mouseover mousemove mouseout mouseover mousemove mouseout mouseover mousemove mouseout mousemove mousec
1206371416624 mouseout mouseover mousemove mouseout mousemove mouseover mousemove mouseout mousemove mouseover mousemove mouse
```

Fig 6 : Output Log

RELATED WORK

After illustrating the use of our preprocessing methodology through different processes, we present in this section the main related works in this domain. In the recent years, there has been much research on Web usage mining [3,4,5,6,7,8,9,10,11,12,13,14,15,16]. However, as described below, data preprocessing in KDWUD has received far less attention than it deserves. Methods for user identification, sessionizing, all other preprocessing works basically focused on web log file and the pages but in our case we try to preprocesses clickstream data.

CONCLUSION

This paper has presented the details of preprocessing tasks that are necessary for performing Click stream Data Mining, the application of data mining and knowledge discovery techniques for Web log data collected by web servers. The contribution of the paper is to introduce the process of preparing web log data for different processes of click stream data mining, and then we can use web log for mining. The experimental results presented

in section 4, illustrates the importance of the data preprocessing step and the effectiveness of our methodology, by reducing not only the size of the log file but also increasing the quality of the data available through the new data structures that we obtained. The process itself does not fully guarantee that we identify correctly all the interactions (i.e. user sessions & visits). This can be due to the data format used in the initial log file therefore, we need a systematic procedure that guarantees the quality and the accuracy of the data obtained at the end of data preprocessing. In conclusion, our methodology yields useful results because:

- It offers the approach to analyze several click stream data in the form of xml file
- It employs the effective cleaning of unnecessary data
- It gives the formatted log file in a text file format

FUTURE SCOPE

Here we have performed an simple laboratory based experiment which may be deferent for Real life mass implementations of the browser mechanism, it may be modified with some more large scale experimentation. We are experimenting for more click stream log and realistic and sophisticated data. There may be some changes occurs on web log data. Some new phases may be introduced for time series data.

REFERENCES

[1] A. Banerjee and J. Ghosh. Clickstream clustering using weighted longest common subsequences. In Proc. of the Wokshop on Web Mining, SIAM Conference on Data Mining, pages 33{40, 2001. Chicago, IL.

[2] G T Raju and P S Satyanarayana Knowledge Discovery from Web Usage Data: Complete Preprocessing Methodology IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.1, January 2008 ,page 179-186

[3] J. Srivastava, R. Cooley, M. Deshpande, P.-N. Tan, Web usage mining: discovery and applications of usage patterns from web data, SIGKDD Explorations, 1(2), 2000, 12–23

[4] R. Kosala, H. Blockeel, Web mining research: a survey, SIGKDD: SIGKDD explorations: newsletter of the special interest group (SIG) on knowledge discovery & data mining, ACM 2 (1), 2000, 1–15

[5] R. Kohavi, R. Parekh, Ten supplementary analyses to improve e-commerce web sites, in: Proceedings of the Fifth WEBKDD workshop, 2003.

[6] B. Mobasher R. Cooley, and J. Srivastava, Creating Adaptive Web Sites through usage based clustering of URLs, in IEEE knowledge & Data Engg work shop (KDEX'99), 1999

[7] Bettina Berendt, Web usage mining, site semantics, and the support of navigation, in Proceedings of the Workshop "WEBKDD'2000 - Web Mining for E-Commerce - Challenges and Opportunities", 6th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2000, Boston, MA

[8] B. Berendt and M. Spiliopoulou. Analysis of Navigation Behaviour in Web Sites Integrating Multiple Information Systems. VLDB, 9(1), 2000, 56-75

[9] A. Joshi and R. Krishnapuram. On Mining Web Access Logs. In ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pages 2000, 63-69

[10] C. Shahabi and F. B. Kashani. A Framework for Efficient and Anonymous Web Usage Mining Based on Client-Side Tracking. In WEBKDD 2001 - Mining Web Log Data Across All Customers Touch Points, Third International Workshop, San Francisco, CA, USA, August 26, 2001, Revised Papers, volume 2356 of LNCS, Springer, 2002, 113-144

[11] Y. Fu, K. Sandhu, and M. Shih. A Generalization-Based Approach to Clustering of Web Usage Sessions. In Proceedings of the 1999 KDD Workshop on Web Mining, San Diego, CA, vol. 1836 of LNAI,. Springer, 2000, 21-38

[12] M. S. Chen, J. S. Park, and P. S. Yu. Efficient Data Mining for Path Traversal Patterns. Knowledge and Data Engineering, 10(2), 1998, 209-221

[13] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization. Data Mining and Knowledge Discovery, 6(1), 2002, 61-82,

[14] M. El-Sayed, C. Ruiz, and E. A. Rundensteiner. FS-Miner: Efficient and Incremental Mining of Frequent Sequence Patterns in Web Logs. In Proceedings of the Sixth Annual ACM International Workshop on Web Information and Data Management (WIDM '04), ACM Press, 2004, 128-135

[15] B. Berendt, B. Mobasher, M. Nakagawa, and M. Spiliopoulou. The Impact of Site Structure and User Environment on Session reconstruction in Web Usage Analysis. In Proceedings of the Forth Web KDD 2002 Workshop, at the ACM-SIGKDD Conference on Knowledge Discovery in Databases (KDD'2002), Edmonton, Alberta, Canada, 2002

[16] C. Marquardt, K. Becker, and D. Ruiz. A Pre-Processing Tool for Web Usage Mining in the Distance Education Domain. In Proceedings of the International Database Engineering and Applications Symposium (IDEAS'04), 2004, 78-87