An Efficient Technique for Web Log Preprocessing using Microsoft Excel

Chintan H. Makwana
P. G. Student,
Department of Computer Engineering, C. U. Shah College of Engineering and Technology, Gujarat Technological University, Gujarat, India.

Kirit R. Rathod
Assistant Professor
Department of Computer Engineering, C. U. Shah College of Engineering and Technology, Gujarat Technological University, Gujarat, India.

ABSTRACT

Web log file shows the behavior of user when they access the website. It is in text format and automatically generated by web server whenever user accesses the website. Entry of particular web page with date, time, cs-method, cs-uri-stem and other information is generated automatically in text file for every access of web page. Web Usage Mining is different mining technique apply on weblog file to discover the different pattern which is useful for efficient design of website, performance enhancement of server etc. Mining of web log file consist of three steps Data Preprocessing, Pattern Discovery and Pattern Analysis. Data preprocessing task convert the web log file in database by applying data extraction, data storage and data cleaning technique. In existing technique data extraction from the log file is perform line by line and store in multidimensional array and then extracted data field will be store in database and then data cleaning is performed. In this research paper, a new technique is being proposed for Data Preprocessing of web log using Microsoft Excel. The experimental results show that the proposed new technique using Microsoft excel file is an effective compare to existing technique.

Keywords

Web Log, Web Usage Mining, Preprocessing, Pattern Discovery, Pattern Analysis.

1. INTRODUCTION

Web is become the imperative medium of information dissemination. Determining size of the World Wide Web is extremely difficult [6]. According to a Survey by Netcraft, the growth of web sites is multiplying day by day [4]. Survey of February 2012 results shows that there are approximately 614,000,000 Web sites available. Where in December 2011 it was 556,000,000. In August 2011 survey 463,00,317 web sites available which has been doubled when compared with August 2010 survey which have 213,458,815 [4]. January 2008 survey, it was 156,000,000. The web can be viewed as the largest database available and presents a challenging task for effective design and access [6].

Web mining is applying different data mining techniques like association rule, classification, clustering or associative classification on data related to the World Wide Web. Here data may be present in web page or data related to Web activity like web log file.

There are three Web Mining Techniques.

1. Web Content Mining

2. Web Structure Mining

3. Web Usage Mining

Web Content Mining involves efficiently extracting useful and relevant information form the millions of web sites and databases.

Web Structure Mining involves the techniques used to study the web pages schema of a collection of hyper-links.

Web Usage Mining involves the analysis and discovery of user access patterns from the web servers logs in order to better serve the user's needs.

The remaining of this paper is organized as follows. In section 2, we discussed about the Web Log Mining. In section 3, we describe the related work and section 4 describe the proposed work and algorithm for data preprocessing. Finally, section 5 and 6 shows the experiments perform on log file and result analysis of it. In last section 7, conclusion gives the summarization of paper.

2. WEB LOG MINING

The Web log mining is a technique in which different data mining techniques apply on the log files.

Web Log File:

Input of Web Log Mining or Web Usage Mining is web log file. Web Log file is collection of user click stream data. It shows the behavior of user when they are accessing particular website. It is text file generated over web server having "*.log" extension. Web log file contains hundreds of entry per day. Size of it is between 1kb to 100MB. There are two format of web log file to store data are NCSA and W3C. Here we are using W3C format of web log and data field in are store by space as separator.

It contains tuples having data fields date, time, s-sitename, s-computername, s-ip, cs-method, cs-uri-stem, cs-uri-query, s-port, cs-username, c-ip, cs-version, cs(User-Agent), cs(Cookie), cs(Referer), cs-host, sc-status, sc-substatus, sc-win32-status, sc-bytes, cs-bytes, time-taken.

Here s-server, c-client, cs-client to server, sc-server to client

A tuple of Log file:

"2013-04-03 01:06:23 W3SVC1 SERVER 220.225.146.19 POST /student/Default.aspx - 80 - 49.201.72.90 HTTP/1.1 Mozilla/5.0+(Windows+NT+5.1)+AppleWebKit/535.19+(KHTML,+like+Gecko)+Chrome/18.0.1025.151+Safari/535.19

ASP.NET_SessionId=ytnhw04510tgyqzxrg3lvf55 http://220.225.146.19/student/ 220.225.146.19 200 0 0 99010 3497 45640"

Where,

date	2013-04-03
time	01:06:23
s-sitename	W3SVC1
s-computername	SERVER
s-ip	220.225.146.19
cs-method	POST
cs-uri-stem	/student/Default.aspx
cs-uri-query	-
s-port	80
cs-username	-
c-ip	49.201.72.90
cs-version	HTTP/1.1
cs(User-Agent)	Mozilla/5.0+(Windows+NT+5.1)+AppleWebKit/535.19+(KHTML,+like+Gecko)+Chrome/18.0.1025.151+Safari/535.19
cs(Cookie)	ASP.NET_SessionId=ytnhw04510tgyqzxrg3lvf55
cs(Referer)	http://220.225.146.19/student/
cs-host	220.225.146.19
sc-status	200
sc-substatus	0
sc-win32-status	0
sc-bytes	99010
cs-bytes	3497
time-taken	45640

Web Usage Mining involves the three phases (see figure 1): Data Preprocessing, Pattern Discovery, and Pattern Analysis.

Data Preprocessing is an important step because of the complex nature of the Web architecture which takes 80% in mining process [1]. In this phase different task are data filed extraction, data storage, data cleaning, user identification, session identification, path completion and transaction identification are performed.

Pattern Discovery phase different data mining techniques association rule, classification, clustering and associative classification apply on the data collected during the Data Preprocessing task in order to find the hidden pattern with in the log data, classify the web page according to user, clustering the user etc.

Pattern Analysis phase removes the uninteresting patterns identified during the pattern discovery phase. There are two most common approaches for the pattern analysis: SQL query mechanism and constructing multi-dimensional data cube to perform OLAP Operation [4].

3. RELETED WORK

In this section, we introduce some related work in data preprocessing. In recent year there are lots of research done on the Web Usage Mining. However Data Preprocessing which take the almost 80% percent of Web Usage Mining but then also it has received far less attention than what it deserve. Surbhi Anand and Rinkie Rani Aggarwat [4] uses java programming language they have given three algorithm data field extraction, data storage and data cleaning. They have first performed first data extraction operation for that they are using the array to store the token means data field extracted, then data storage operation for store the data and at last data cleaning operation for removing unnecessary data. Ms. Dipa Dixit and Ms. M Kiruthika [7] have given two approaches for data preprocessing using XML DOM tree and using text file.

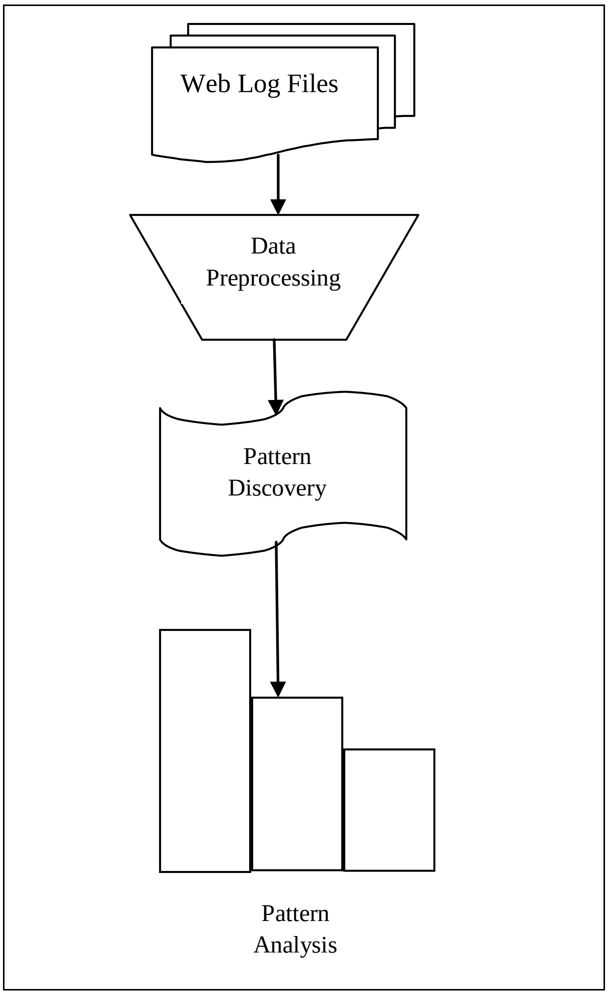


Figure 1. Phases of Web Log Mining

4. PROPOSED WORK

4.1 DATA PREPROCESSING

Data Preprocessing is among the most time consuming task. It almost take 80% time of the overall web log mining. Every time a web browser downloads a HTML document on the internet the images are also downloaded and stored in the log file. This is because though a user does not explicitly request graphics that are on a web page, they are automatically downloaded due to HTML tags. The process of data cleaning is to remove irrelevant data. In this phase data from the "*.log" format is converted to database "*.sql" format that can be useful for the data mining.

Following operations are performed during the data preprocessing are Data Extraction, Data Cleaning and Data Storage.

Algorithm

Input: Web Log File, Separator=space

Output: Clean database table

Procedure Data_Preprocessing

1. Open web log file in excel sheet through asp.net.
2. In excel sheet find and Replace "#field, date, time, s-sitename, s-computername, s-ip, cs-method, cs-uri-stem, cs-uri-query, s-port, cs-username, c-ip, cs-version, cs(User-Agent), cs(Cookie), cs(Referer), cs-host, sc-status, sc-substatus, sc-win32-status, sc-bytes, cs-bytes, time-taken" word with "Space" character.
3. Delete top 3 row
4. Pass Space delimited for column separation
5. Save as .xsl file.
6. Fetch Last access date if it is null then start form first starting position of excel sheet file for conversion of excel to database table.
7. Data cleaning is apply on the database table. All unrelated entries of multimedia file extension, css uri etc are removed, data other than status_code>=200 and status_code<=290 are removed.
8. Every tuple which are clean will be store in the database table line by line.
9. Close excel file and Close the operation

End.

The implementation of the algorithm is done in ASP.net with C# programming language and Microsoft SQL Server 2008 use for data storage.

A. Data Extraction

A server web log file consist of various data fields separated by space character. In the existing technique data field extracted character by character and space character work as data field separator, extracted data field are store in array data structure.

In an efficient novel technique data field are extracted directly using the Microsoft excel. Log file in text file are directly open in excel and pass the space as data field separator and save the excel file which very efficient compare to existing technique of array structure.

B. Data Selection and Storage

The status code return by the server is three digit number. There are four class of status code: Success (2XX series), Redirect (3XX Series), Failure (4XX Series), and Server Error (5XX Series). Other than 2XX series Success HTTP Status code means other than 200 to 299 are useless for analysis process and therefore they're cleaned form the log file. Data from the Excel file are selected which have the status code between 200 to 299 and store in Microsoft SQL SERVER 2008 database.

C. Data Cleaning

During the data cleaning phase from database table record having filename suffixes style files (*.css), multimedia files (*.jpg, *.gif, *.png, *.mp3, *.flv, *.mpeg) which can be found in cs-uri-stem field are removed.

After data extraction, data selection and storage, Data cleaning other preprocessing operations like user identification, session identification, and path completion will be performed.

5. EXPERIMENT RESULT

All experiment performed on a 64-bit PC having 2.4GHz speed with 4GB as main memory, running Microsoft window 7.

The log file used for experiment was of size 3160 KB having 7858 tuples. In existing technique it take near about 10 second for data preprocessing but the novel efficient approach technique takes 5.69 second.

6. RESULT ANALYSIS

Experiment performed on five different web log file's result shows that the novel approach technique is efficient and reduces the processing time compare to exciting approach. Table and graphically representation of required time in exciting approach and novel approach is shown below.

No	Total Entries	Processing Time
1	7858	5.69 Sec
2	11050	7.15 Sec
3	17160	9.9 Sec
4	28210	15.58 Sec
5	36068	20.83 Sec

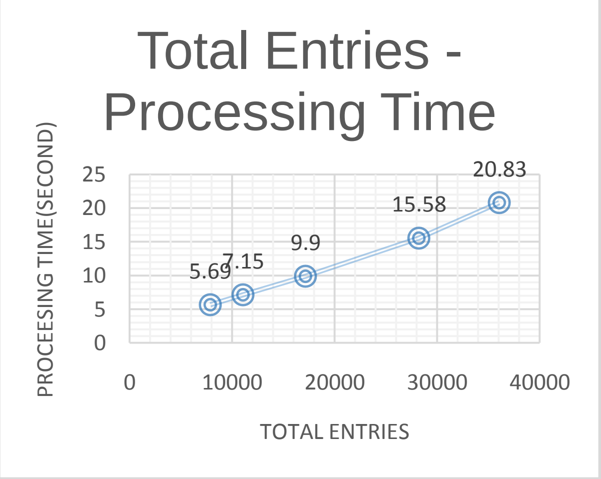


Table and graphically representation of Size of log file before cleaning and after cleaning is shown below.

No	Total Entries	Size Before Cleaning	Size After Cleaning
1	7858	3160 KB	1288 KB
2	11050	4162 KB	1264 KB
3	17160	7017 KB	2560 KB
4	28210	11179 KB	3823 KB
5	36068	14338 KB	5103 KB

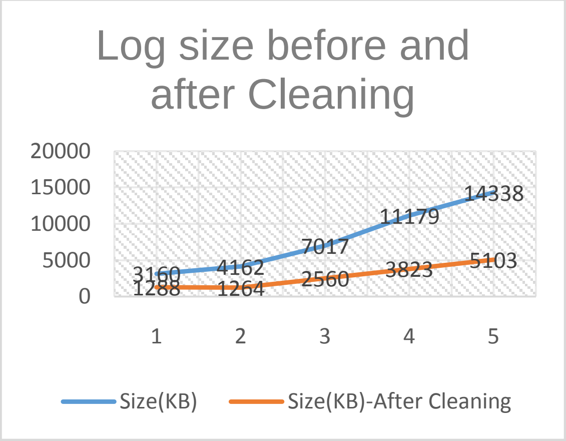
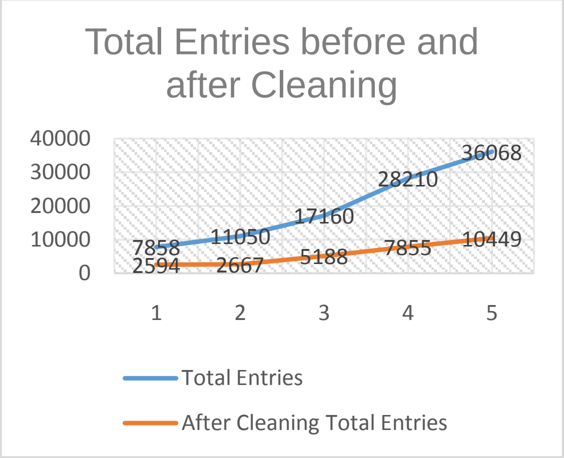


Table and graphically representation of total entries before cleaning and after cleaning is shown below.

No	Total Entries Before Cleaning	Total Entries After Cleaning
1	7858	2594
2	11050	2667
3	17160	5188
4	28210	7855
5	36068	10449



7. CONCLUSION

Data Preprocessing is an important step in the Web Log Mining because of two reasons, one is that it takes almost 80% time of Web Log Mining and other is quality decisions are based on quality of data. In this paper, we have discussed Novel approach for Data Preprocessing using Microsoft Excel file and compare its efficiency with exciting technique of array for store the extracted data field. Experiment performed on both the approaches of Data Preprocessing and Result analysis shows the comparison of time required in both technique. Result shows that Novel technique is better compare to existing technique. Result analysis also shows comparison of the log size before and after cleaning and comparison of the total entries before cleaning and after cleaning.

8. REFERENCES

[1] P.Nithya and Dr.P.Sumathi, Novel Pre-Processing Technique for Web Log Mining by Removing Global Noise and Web Robots, *IEEE National Conference on Computing and Communication Systems (NCCCS), 2012*

[2] S. Chitra and Dr. B. Kalpana, A Novel Preprocessing Mixed Ancestral Graph Technique for Session Construction, *IEEE International Conference on Computer Communication and Informatics (ICCCI-2013)*, Jan. 04 – 06, 2013, Coimbatore, INDIA

[3] Manisha Valera and Kirit Rathod, A Novel Approach of Mining Frequent Sequential Pattern from Customized Web Log Preprocessing, *International Journal of Engineering Research and Applications*, Vol. 3, Issue 1, February 2013, 269-380

[4] Surbhi Anand and Rinkie Rani Aggarwal, An Efficient Algorithm for Data Cleaning of Log File using File Extenstions, *International Journal of Computer Applications*, 48(8), June 2012, 13-18

[5] V. Chitraa and Dr. Antony Selvadoss Thanamani, A Novel Techniques for Sessions Identification in Web Usage Mining Preprocessing, *International Journal of Computer Applications*, 34(9), November 2011, 23-27

[6] Margaret H. Dunham, *Data Mining Introductory and Advanced Topics*, 8th Impression, Person, 2012, 193-218.

[7] Ms. Dipa Dixit and Ms. M Kiruthika, PREPROCESSING OF LOG FILES, *International Journal on Computer Science and Engineering(IJCSE)*, Vol. 02, No. 07, 2010, 2447-2452.

[8] Vijayashri Losarwar, Dr. Madhuri Joshi, “Data Preprocessing in Web Usage Mining”, *International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012)* July 15-16, 2012 Singapore.