Web Usage Mining: Dynamic Methodology to Preprocessing Web Logs

^{*1}Mahesh Manchanda, ²Dr. Neena Gupta

¹Graphic Era Hill University, Dehradun

²Gurukul Kangri Vishwavidyalaya, Haridwar

Email: ^{*}manchandamahesh@gmail.com, neena71@hotmail.com

Received: 09th July 2018, Accepted: 14th August 2018, Published: 31st August 2018

Abstract

Internet is a huge source of massive information for retrieving information and searching knowledge from WWW, leading to increase network traffic, access delay & server overload, which results in poor web services. With the use of Web-caching & web prefetching techniques to enhance the performance of web services where web mining techniques play an important role to decide which web object should be pre-fetched from server and stored in proxy cache memory so that the web object with high probability of request, in the next couple of days, serves as the base of the proxy cache. But for efficient web mining and to extract meaningful usage access pattern, the raw log file must be transformed into a meaningful & formatted file. This paper proposed a new dynamic preprocess technique to create a dynamic training dataset for prediction model using web mining, and Graph based substructure Pattern Mining (GSPAN) for improved preprocessing using proxy log. The proposed model would help in minimizing the cache size by 40% thus improving the overall performance.

Keywords: Web Usage Mining, Data Cleaning, URL Rank, GSPAN, Web Pre-Fetching

Introduction

The dependency on the web services for day to day usage has swiftly increased in the last two decades. Web services are frequently used for many purposes like shopping, banking, filling application forms, exams etc. User access to these services recorded in log files stored in three places client browser, proxy server and finally in origin web server. These different sources of logs have its own advantages and limitation but their importance to record the user access event for web usage mining is pertinent [1]. Genuine access pattern of individual client can be depicted from browser log document [2]. It is as one-to-numerous connections of customer and sites went to by that specific client. Proxy log also contained user access entries in log file. It is in many-to-numerous relationships. On the other side, Server log files having numerous-to-one relationship. Server log recorded the access pattern of several users visiting a particular web site. Server log caters only to a group

of users accessing the same web site. Real motivation behind choosing the proxy server log file for creating a training dataset for our new dynamic model, since proxy server is an intermediary server, sits between the client (browser) and the origin Web server. As the proxy server resides on the same network as the user, resulting in a much faster load page, reduction in bandwidth usage, network congestion and enhance the performance for group of users. In Literature review, we notice that maximum of researcher considered server log for mining purpose but it is also evident that server log does not record the cached pages' requests hence not considered for web mining[4]. The growth of web services and excessive use of these web services increase the size of web log file rapidly. The data which is stored in web log files will be consisting of huge amount of information with some kind of incomplete, noisy and unwanted data too. It is difficult to deal with entire data which is huge in size. So, unwanted or irrelevant data must be removed by using efficient preprocessing technique. Thus, preprocessing plays a vital role in web mining to reduce the size of a log file by removing irrelevant entries from the log files. The quality preprocessed log file decides the quality of the accurate pattern analysis required for efficient Web prefetching. The process of discovering hidden knowledge from log files and finds the interesting patterns is known as web usage mining.

Web mining classifies into three categories depending upon which portion of the web page to be mined are specifically *structure mining*, *web usage mining* and *web content mining*. *Web structure mining* is the method for discovering information the hyperlink structure of the web pages. *Web Usage Mining* (WUM) is a method of discovering user navigation pattern and predict user's access behavior from the proxy server Log. *Web content mining* is a process of searching useful information and knowledge of Web page content. Here in this paper web usage mining (WUM) is used with web log files for studying user access pattern which user interacts with a different web page. The process of WUM is classified into 3 phases: *Preprocessing*, *pattern discovery* and *pattern analysis*. The classification of WUM is depicted in the following figure 1.

Figure 1: Classification of Web Usage Mining

In preprocessing step, the raw log file is transform into document having only visited URLs and hit counts, necessary step for further applying the mining algorithm. First, the log file is sanitized via casting off irrelevant and inaccurate entries. In the subsequent step, filter the visited URLs and their hit counts. After preprocessing phase, the pattern discovery converge the algorithms and data mining techniques from the preprocessed log file. Preprocessing is the time consuming process, deals with the translation of raw log file into meaning format. In this research paper we focus on the following task:

- Clean the original log file & retain only visited URLs and their hit counts.
- Fetch all the keywords online by visiting each URL, arrange all the similar keywords together with their URLs and hit counts.

This paper is organized as follows: section2 demonstrated the related works, section3 describe Motivation behind this paper, section 4 describes Proposed Methodology and algorithm section5 provides the analysis and results and finally conclusion of our research.

Related Work

In literature review, we study and analyze the available preprocessing techniques. As we know that, the raw web log file is used directly in web mining gives an ambiguous result. Because original weblog contains the large volume of inappropriate records, hence preprocessing of web log file becomes significant. The subsequent segment discusses the studies of numerous authors: Suneetha et. al [1] discussed different sources and types of log files with their structure in detail. Authors performed two preprocessed technique but data integration part was not considered in their research.

In data cleaning process, noisy, error and other unwanted entries from weblog are removed. Weblog cleaning is mandatory and most commonly used preprocessing technique. Researcher generally choose web server log for preprocessing task because it has least issues as compared to browser and proxy server weblog document. Only few researchers executed their experiment on client log. Like Murata et. al [2] used page rank algorithm using a graph to reveal the user interest.

While studying the literature of preprocess of weblog. We observed that *data cleansing* and *filtering* are the common techniques used prior to Pattern discovery and pattern analysis. We also found different types of attributes from log file used in different preprocessing techniques. Frequent used attributes are *IP Address*, *Date*, *Time*, *URL*, and *User-agent*.

Pabarskaite, [9] and Wahab, et al., [10] used the structure of web page and source code for data filtering as a sub- step of data cleaning. Similarly, Castellano, et al, [6] used least visited pages for cleaning purposes. D. R. Krishnamoorthi, et al. [1], Chan [5], and Wang, et al. et al., [4] used client IP, browser application and operating system attribute from weblog to identify the individual user's access pattern. But Stermsek. et al., [3] used only IP Address to identify the users. Satyanarayana et. al. [7]. Feng Li et al [11] has implements Optimized User Identification and Optimized Session Identification to identify the complex user sessions.

The optimized data preprocessing technology was used for the improvement of the technology and betters the quality of data preprocessing results. Sumathi, C.P et al, discussed about the implementation of how data fusion, data cleaning, user identification and session identification can applies on different types of log files. Authors used CLF log file for their experiment and finally they illustrate how the log files is transformed into user session.

Motivation

There are various techniques used at preprocessing levels of web usage mining. Data cleaning, data filtering, and data integration are different techniques applying at preprocessing level. We proposed a new keyword based preprocessing model to remove all the noisy or irrelevant log entries and filter all the URLs's having common keywords. A weblog file, used as an input in the preprocessing phase. Original log file is vast in their size, records the number of erroneous, noisy, referrer and unwanted records. As a result, log file cannot be used for Web Mining. Preprocessing of log file is complex and back

breaking task and it takes most of the total time of web usage mining process as whole. After studying the merits and demerits of preprocessing, we come to the conclusion that, we cannot ignore the importance of preprocessing task in WUM process. Considering preprocessing step, to improving the quality of log [13], it also improves the efficiency and effectiveness of future steps of WUM such as pattern discovery and pattern analysis.

Proposed Methodology

Preprocessing

Main objective of using preprocessing to reduce the size of weblog file and improve the quality of filter data for better pattern discovery and analysis to serve hint for proxy cache. We proposed a new keyword based preprocessing model to remove all the noisy or irrelevant log entries and filter all the URLs's having common keywords. The original log file used as the input for preprocessing phase. Three standard layouts to represents log file such as: Common Log format (CLF), Extended Common Log Format (ECLF) and IIS. All web log formats having common attribute about the user like client IP address, date and time of visited URL, Visited URL address, Hit count, client name, date, time, type of service, server name and IP address, latency time, user bytes requested, server bytes sent, status code etc.

Data Cleaning

Data cleaning is the first & essential step of pre-processing phase. It extracts the unwanted or erroneous entries from the weblog file. The original log files are very large in their size and it takes 80% of the entire process of preprocessing, then only the further processes of pattern discovery are performed effectively. During data cleaning phase, we have to remove the irrelevant entries', such as:

Remove entries related to image file extension like .gif, jpg and png etc. The log entries with failure status code which is to be removed from the weblogs. e.g. Error code within the range of 100 to 199 represent that the request is continue. 300-399 range indicates about the bad request, page not found and authorization required etc. and 500 series indicate about service not available, gateway timeout and other server related errors. The weblogs also having entries for automated search engine such as robots, spider and crawler files removed from web logs. In this paper proposed algorithm first removes all the above mention errors entries from the raw log files except visited URL & their hit count. Secondly, it visits every URL & find all the keyword used in that particular page automatically and makes a cluster URL having common keywords. After that, we remove all the preposition from keywords. Finally we

get file having only visited URL with their common keywords and sorts all the keywords with their corresponding URLs.

Proposed Algorithm

Data Preprocessing Algorithm

Input: Proxy server log File

Output: Log file with 60% less records

Step1. Read log file from server

Step2. if (PDF) then

 loop till end of file

 extract lines using PDF API's

 extract or write line to a text file

 end loop

end if

Step3. read text file

Step4. loop till end of file

 clean the text using Pattern Matching

 retain URL and Hit count only

 urlText.txt ← write to text file

end loop

Step5. Open urlText.txt

Step6. loop till end of urlText.txt

 read URL

 connect with URL

 kwd ← read meta keywords of the URL

 kwd ← write URL and keywords to a text

file

 end loop

Step7. close all text files

Graph Creation

Step8. Open kwd

Step9. graphcache ← create object for Graph Database Service

Step10. node ← declare nodes

Step11. relation ← declare relationships

Step12. tx ← begin graph transaction

Step13. create head node

Step14. set properties

Step15. loop till end of kwd

 create child

 set properties of the child

 URL

 Keywords

 Hit count

 end loop

Step16. Close kwd

Experimental Result

In this experiment the one week dataset for log files are collected from proxy server of ICF AI University and

Graphic Era Hill University, Dehradun, India. There are total 6391 web log entries recorded in the day1 log file, which contains a page files, image files, error files, and other files. The Sample of proxy server log file using Common Log file Format is shown in the Fig 2.

```

1502596834.936 558 172.16.1.249 TCP_MISS/200 11190 GET http://15.taboola.com/t? - DIRECT/151.101.101.249 text/html
1502596835.136 34 172.16.1.249 TCP_MISS/200 3185 GET http://14.zedo.com/vtels/vbidding/v42bvw.js? - DIRECT/23
1502596835.281 225 172.16.1.249 TCP_MISS/200 4959 GET http://10.crowdnews.com/v1/content/AaTat? - DIRECT/23.1
1502596835.266 265 172.16.1.249 TCP_MISS/200 516 GET http://101statics.crowdnews.com/tb? - DIRECT/119.81.17.85
1502596835.309 661 172.16.1.249 TCP_MISS/200 387 POST http://webtest11.dmail.com/SpamResolver/G/SpamResolve
1502596835.422 5 172.16.1.249 TCP_MISS/200 5851 GET http://www.credentia.com/vpaidunits/15_2_3/assets/cse
1502596835.436 24 172.16.1.249 TCP_HIT/200 53765 GET http://www.credentia.com/vpaidunits/15_2_3/assets/cse
1502596835.534 138 172.16.1.249 TCP_REFRESH_HIT/200 515 GET http://miles.quantcount.com/rules-p-101/101PKX.js
1502596835.536 244 172.16.1.249 TCP_MISS/200 2124 POST http://ss.symcd.com - DIRECT/23.57.219.27 application/js
1502596835.653 41 172.16.1.249 TCP_HIT/200 42945 GET http://media2.intoday.in/aajtak/Images/Photo_gallery/102016/1
1502596835.653 90 172.16.1.249 TCP_HIT/200 52156 GET http://media2.intoday.in/aajtak/Images/Photo_gallery/102016/1
1502596835.653 48 172.16.1.249 TCP_HIT/200 50616 GET http://media2.intoday.in/aajtak/Images/Photo_gallery/102016/1
1502596835.686 77 172.16.1.249 TCP_REFRESH_MISS/200 23325 GET http://widget.crowdnews.com/verticalcardst/101/
1502596835.690 6 172.16.1.249 TCP_HIT/200 21149 GET http://media2.intoday.in/aajtak/Images/Photo_gallery/102016/1
1502596835.753 3 172.16.1.249 TCP_HIT/200 28525 GET http://media2.intoday.in/aajtak/Images/Photo_gallery/102016/1
1502596835.760 4 172.16.1.249 TCP_HIT/200 10009 GET http://media2.intoday.in/aajtak/Images/Photo_gallery/102016/1
1502596835.761 3 172.16.1.249 TCP_HIT/200 29829 GET http://media2.intoday.in/aajtak/Images/Photo_gallery/102016/1
1502596835.763 6 172.16.1.249 TCP_HIT/200 21282 GET http://media2.intoday.in/aajtak/Images/Photo_gallery/102016/1
1502596835.765 6 172.16.1.249 TCP_HIT/200 21308 GET http://media2.intoday.in/aajtak/Images/Photo_gallery/102016/1
1502596835.767 6 172.16.1.249 TCP_HIT/200 10000 GET http://media2.intoday.in/aajtak/Images/Photo_gallery/102016/1
1502596835.769 11 172.16.1.249 TCP_HIT/200 31363 GET http://media2.intoday.in/aajtak/Images/Photo_gallery/102016/1
1502596835.936 90 172.16.1.249 TCP_MISS/200 460 GET http://icon.hovr.me/images/right.svg - DIRECT/119.81.86.22

```

Fig 2: Sample Web Log File Format

1. According to the proposed methodology, read the day1 pdf of web log from the designated folder and convert them into text file using *PDF Table Extractor* using java PDF API.

2. From text file we extract all the visited URLs and hit count and create a new log file named *urlTXT.txt* having

Only visited URLS and their Hit Count. Here we remove all type of irrelative file extension, and error code such as

- Image files - .jpg, .jpeg, .png
- Error code of series -100, 200, 400,
- Other files - .js, .css, .log,

These unwanted log entries will be removed from the weblogs (using java pattern matching API) and retain only the visited URLs, number of time a particular URL visited (Hit Count) and User list only. Here cleaning of log file is done. Below Figure 3 showing the sample of Web log file after cleaning process.

```

URL ;HITS ;USER
http://www.amarujala.com ;4087 ;Ashok Kuri
Sub Total ;4163
http://www.bhaskar.com ;264 ;Shobha Jain
Sub Total ;264 ;
http://www.timeslog.indiatimes.com ;195 ;Q
Sub Total ;249 ;
http://www.facebook.com ;1289 ;Handa sc,
adserver.adtech.de ;1059 ;Kumar.ashok
209.85.153.83 ;917 ;Singh, rajkumar,dhir
t1.gstatic.com ;476 ;Prajapati, Deepak,kush
Sub Total ;1 ;Kumar, vivek
t2.gstatic.com ;442 ;Kaur, navjot

```

Fig 3: Log with URL, Hit Count and their User

In the next three steps preprocessing work will be implementing at the web log.

3. It reading *urlTXT.txt*, fetch all the keywords online by visiting each URL from file using *java.net. URL Connection API*. (Here a new file *URLkwd.txt* created) Result shown in figure 4

```

http://www.amarujala.com
Hindi news,news in hindi,latest hindi news,
123
http://www.bhaskar.com
News, Hindi News, Latest News in Hindi, Ne
39
http://www.scorecardresearch.com
surveys, sweepstakes, earn cash, opinions,
65
http://www.indeed.co.in
job search, jobs, work, employment, career,
145

```

Fig 4: URL, Keyword & Hit Count

4. Create a Graph from the file *URLkwd.txt* having URLs, Keywords and Hit count by using neo4J a machine Learning based graph dataset tool. Graph created shown in Figure 5A and Figure 5B.

Figure 5A: Root Node of Graph Database in Neo4j Eclipse

Figure 5B: Detail Graph Database view in Neo4j Eclipse

5. Traverse the Graph using GSPAN traversing and extract all the keywords, URL and Hit count together and stored them in an array. GSPAN is a new approach for frequent graph-based pattern mining in graph datasets.

6. Arrange all the similar keywords together with their URLs and hit counts.

Finally, the web log is cleaned & Pre-process effectively now the log files have only 2556 relevant web log entries for effective web mining. The below table1 summarizes the number of weblogs entries in the cleaning process

Raw Log Entries	Number of Error Entries Removed	Number of Image Links Removed	Number of Spider Entries Removed	Cleaned Logs Entries
6391	1217	783	1835	2556

Table 1 Entries of Log File

Finally, we get web log file having only relevant 2556 entries as compare to original web log with 6391 entries for further analysis decision making.

Figure 6 (60% less Log)

Figure 6 shows the result of final pre-processed weblog that contains only 2556 entries

Conclusion

Preprocessing of web log file is important and a required task before filtering web log. Removing irrelevant log entries before preprocessing is a firm base for identifying used access behavior. We have used java based pattern matching technique to remove and filter the irrelevant log entries. By applying efficient preprocessing method we can ensure the quality of pattern mining and pattern analysis. For future work we should explore

Dynamic, Robust and customizable preprocessing techniques, consider all the attributes of web log file, and apply to all types of web logs to discover the quality usage pattern. The authors are working on designing a Machine learning algorithm for the entire process so that during certain intervals the process automatically cleans and updates the log files.

References

1. Suneetha, K. R. and D. R. Krishnamoorthi (2009). "Identifying User Behavior by Analyzing Web Server Access Log File." IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009
2. Murata, T. and K. Saito (2006). Extracting Users' Interests from Web Log Data. Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06) 0-7695-2747-7/06.
3. Stermsek, G., M. Strembeck, et al. (2007). A User Profile Derivation Approach based on Log-File Analysis. IKE 2007: 258-264.
4. Yuan, F., L.-J. Wang, et al. (2003). Study on Data Preprocessing Algorithm in Web Log Mining. Proceedings of the Second International Conference on Machine Learning and Cybernetics, Wan, 2-5 November 2003
5. Khasawneh, N. and C.-C. Chan (2006). Active User-Based and Ontology-Based Web Log Data Preprocessing for Web Usage Mining. Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings) (WI'06) 0-7695-2747-7/06 © 2006.
6. Castellano, G., A. M. Fanelli, et al. (2007). LODAP: A Log DATA Preprocessor for mining Web browsing patterns. Proceedings of the 6th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Corfu Island, Greece, February 16-19, 2007.
7. Raju, G. T. and Satyanarayana. P. S., 2008, "Knowledge Discovery from Web Usage Data: Complete Preprocessing Methodology", IJCSNS International Journal of Computer Science and Network Security, V|OL. 8 No. 1, January 2008.
8. S. Khan, Y. Singh and A. Kumar. Sachan, "Web Mining Approach in Analysing User Behavior and Interest for Website Modification", International Journal of Advanced Research in Computer Science and Software Engineering, Vol 5, 2015.
9. Pabarskaite, Z. (2002). Implementing Advanced Cleaning and End-User Interpretability Technologies in Web Log Mining. 24th Int. Conf.

information Technology Interfaces /TI 2002, June 24-27, 2002, Cavtat, Croatia.

10. Wahab, M. H. A., M. N. H. Mohd, et al. (2008). Data Preprocessing on Web Server Logs for Generalized Association Rules Mining Algorithm. World Academy of Science, Engineering and Technology 48 2008.

11. Ling Zheng Hui Gui. Feng Li, “Optimized Data Preprocessing Technology for Web Log Mining”, International Conference On Computer Design And Applications (ICCD A 2010)

12. Sumathi, C. P., et al., 2011. An Overview of Preprocessing Of Web Log Files For Web Usage Mining. Journal of Theoretical and Applied Information Technology, ISSN: 1992-8645.

13. Nisarg Pathak,Viral Shah, Chandramohan Ajmeera,"A Efficient Algorithm with Enhance Preprocessing Technique for Web Usage Mining",ICTCS '14 Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies,ACM,2014.

14. Srivastava, Mitali, Rakhi Garg, and P. K. Mishra. "Analysis of Data Extraction and Data Cleaning in Web Usage Mining. " Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015). ACM, 2015.