An approach for data cleaning for Web Usage Mining

JayantiMehra[1] ,Dr.R.S.Thakur[2]
MANIT Bhopal, India[1][2]
mehra.jayanti109@Gmail.com[1] ,ramthakur2000@yahoo.com[2]

Abstract: - In the day- today-life, the information in World Wide Web is increasing in an explosive way and simultaneously the usage of the Web was also growing in an enormous way. For each and every usage or accessing of the web pages, it created a separate log entry in the web log file so that the log file also increased correspondingly. From the web log file we got some interesting information about the users' previous access sequence. Using that interested information, it was possible to predict the users' future access sequence and also to personalize the interested. The working of WUM involves three steps – preprocessing, pattern discovery and analysis. The first step in WUM - Preprocessing of data is a necessary action which will help to improve the quality of the data and consecutively the mining results. This research paper explains and presents different data preparation techniques for web log data. This techniques converted raw data into usable data.

Keywords:. Web Usage Mining, Data Preprocessing, Web log, User Session, Path completion

I. Introduction

Data mining is the techniques to discover patterns in huge volumes of raw data. Web mining can be referred as the revolution of the data mining techniques to web data. Web mining has three parts – content, structure and usage mining of web data. Web Usage Mining (WUM.) is all concerning identifying user browsing patterns over web, with the help of knowledge acquired from web logs. The outcomes of the WUM we can use in web personalization, improving the performance of the system, change of the site, business intelligence, usage classification etc [1][2].

The data preprocessing of WUM is focused research field today. This research paper studies the preprocessing of data in Web usage mining. This research paper has prepared in various sections, includes – related research in this area, descriptions about the preprocessing of the usage mining and the proposed algorithms, investigation study to verify the productivity and effectiveness of the algorithms suggested and finally, conclusions and future scope of study has been proposed.[10][12]

DATA PREPROCESSING

The information in the World Wide Web is growing exponentially and also there is no limit to access the information from the WWW Nowadays Web mining is one of the intensive research areas and in particular Web usage mining is an interested research area in the field of the Data mining [4]. Web Usage mining is the process of finding the activities of the users when they are browsing through the web from the web log data[10]. The web log data can be collected from Client side, Server side (or) Proxy servers and we collected web logs from server side [5][6][16]. For each and every access of the web page it records a log entry in web log file.[15] Whenever a page is clicked the corresponding data is recorded in the log file. The log file contain Date & Time stamp, IP Address, URL address of the accessed item, Result status and Byte transferred .The server log files acts as a main data sources in Web usage mining, which include - access logs of the web server and application server logs.[16] The important task in the preprocessing phase is field extraction. The log files containing log entries which represents the single click stream. The programming logic was employed to separate various fields from the log files.[21] All the log files collected from the data source are sorted and joined together in a single log

file. Due to different server setting parameters, there exists several types of web logs, but typically these log files (a log file is a simple text file), share the basic information, such as: user IP address, request time, requested Uniform Resource Locator, HTTP status code, referrer, and so on. Data sets, which has web log records for 1727 users were collected from NASA website. Generally, data cleaning, identification of user, session and path completion are the various steps involved in preprocessing.
Various steps of data preprocessing

A. Data Cleaning

Data cleaning means remove the irrelevant information from the original Web log file and restore the Web log as database which is suitable for user identification, session identification and path complement.[14][17]

B. User Identification

The aim of the user identification process is to discover the diverse users from the web access log file. Different users are organism illustrious by using their Internet Protocol (IP) addresses. The method used for this process is a referrer-based method. User recognition is multifaceted due to the attendance of resident caches, firewalls and proxy servers. To deal this problem, the WUM methods were employed that rely on user cooperation. However, it's not easy because of high security and privacy. [8]

PROPOSED WORK

1. Data Storage

Web logs are initially present in the form of text files. a variety of fields in this text file are divided using a different character such as space character or comma character, this step in preprocessing is known as Field Extraction. Once field extraction to store these web logs in the knowledge base a table is needed to be created. The fields in the log file are to be stored in individual columns of the table. The SQL query for creating a table with name Test Log and having the columns as fields in the Apache common log format is shown below in Fig 1:-

D. Session Identification

The goal of the user session identification is to discover the dissimilar user sessions from the web access log file. A set of user clicks regularly referred to as a click stream, across Web servers is distinct as a user session. The user session recognition involves - separating the page accesses of every user into separate sessions. At present, the methods that are at present obtainable will recognize user session mostly contain break mechanism.[9]

```
"199.72.81.55 - - [01/Jul/1995:00:00:01 -0400] ""GET /history/apollo/
HTTP/1.0"" 200 6245"
"unicomp6.unicomp.net - - [01/Jul/1995:00:00:06 -0400] ""GET
/shuttle/countdown/ HTTP/1.0"" 200 3985"
"199.120.110.21 - - [01/Jul/1995:00:00:09 -0400] ""GET
/shuttle/missions/sts-73/mission-sts-73.html HTTP/1.0"" 200 4085"
"burger.letters.com - - [01/Jul/1995:00:00:11 -0400] ""GET
/shuttle/countdown/liftoff.html HTTP/1.0"" 304 0"
"199.120.110.21 - - [01/Jul/1995:00:00:11 -0400] ""GET
/shuttle/missions/sts-73/sts-73-patch-small.gif HTTP/1.0"" 200 4179"
"burger.letters.com - - [01/Jul/1995:00:00:12 -0400] ""GET /images/NASA-
logosmall.gif HTTP/1.0"" 304 0"
"burger.letters.com - - [01/Jul/1995:00:00:12 -0400] ""GET
/shuttle/countdown/video/livevideo.gif HTTP/1.0"" 200 0"
"205.212.115.106 - - [01/Jul/1995:00:00:12 -0400] ""GET
/shuttle/countdown/countdown.html HTTP/1.0"" 200 3985"
"d104.aa.net - - [01/Jul/1995:00:00:13 -0400] ""GET /shuttle/countdown/
HTTP/1.0"" 200 3985"
"129.94.144.152 - - [01/Jul/1995:00:00:13 -0400] ""GET / HTTP/1.0"" 200
7074"
"unicomp6.unicomp.net - - [01/Jul/1995:00:00:14 -0400] ""GET
/shuttle/countdown/count.gif HTTP/1.0"" 200 40310"
"unicomp6.unicomp.net - - [01/Jul/1995:00:00:14 -0400] ""GET
/images/NASA-logosmall.gif HTTP/1.0"" 200 786"
"unicomp6.unicomp.net - - [01/Jul/1995:00:00:14 -0400] ""GET /images/KSC-
logosmall.gif HTTP/1.0"" 200 1204"
"d104.aa.net - - [01/Jul/1995:00:00:15 -0400] ""GET
/shuttle/countdown/count.gif HTTP/1.0"" 200 40310"
```

Filed	Meaning
Date	Date at which the web page were accessed
Time	The time at which the web page were accessed
IP Address	From the address only the user accessed the server to get the desired web page
URL	The accessed web pages of the web site by the user
Browser	The type of the browser used by the client
OS	The type of operating system used by the client
Code	Status code about the Accessed web page

PROPOSED WORK

1. Data Storage

Web logs are initially present in the form of text files. a variety of fields in this text file are divided using a different character such as space character or comma character, this step in preprocessing is known as Field Extraction. Once field extraction to store these web logs in the knowledge base a table is needed to be created. The fields in the log file are to be stored in individual columns of the table. The

SQL query for creating a table with name Test Log and having the columns as fields in the Apache common log format is shown below in Fig 1:-

IP Address	Date/Time	URL	Status Code	File Size(Byte)
in24.inetnebr.com	[01/Aug/1995:00:00:01	GET /shuttle/missions/sts-68/news/sts-68-mcc-05.txt HTTP/1.0	200	1839
uplherc.upl.com	[01/Aug/1995:00:00:07	GET / HTTP/1.0	304	0
uplherc.upl.com	[01/Aug/1995:00:00:08	GET /images/ksclogo-medium.gif HTTP/1.0	304	0
uplherc.upl.com	[01/Aug/1995:00:00:08	GET /images/MOSAIC-logosmall.gif HTTP/1.0	304	0
uplherc.upl.com	[01/Aug/1995:00:00:08	GET /images/USA-logosmall.gif HTTP/1.0	304	0
ix-esc-ca2-07.ix.netcom.com	[01/Aug/1995:00:00:09	GET /images/launch-logo.gif HTTP/1.0	200	1713
uplherc.upl.com	[01/Aug/1995:00:00:10	GET /images/WORLD-logosmall.gif HTTP/1.0	304	0
slppp6.intermind.net	[01/Aug/1995:00:00:10	GET /history/skylab/skylab.html HTTP/1.0	200	1687
piweba4y.prodigy.com	[01/Aug/1995:00:00:10	GET /images/launchmedium.gif HTTP/1.0	200	11853
slppp6.intermind.net	[01/Aug/1995:00:00:11	GET /history/skylab/skylab-small.gif HTTP/1.0	200	9202
slppp6.intermind.net	[01/Aug/1995:00:00:12	GET /images/ksclogosmall.gif HTTP/1.0	200	3635
ix-esc-ca2-07.ix.netcom.com	[01/Aug/1995:00:00:12	GET /history/apollo/images/apollo-logo1.gif HTTP/1.0	200	1173
slppp6.intermind.net	[01/Aug/1995:00:00:13	GET /history/apollo/images/apollo-logo.gif HTTP/1.0	200	3047
uplherc.upl.com	[01/Aug/1995:00:00:14	GET /images/NASA-logosmall.gif HTTP/1.0	304	0
133.43.96.45	[01/Aug/1995:00:00:16	GET /shuttle/missions/sts-69/mission-sts-69.html HTTP/1.0	200	10566
kgtyk4.kj.yamagata-u.ac.jp	[01/Aug/1995:00:00:17	GET / HTTP/1.0	200	7280
kgtyk4.kj.yamagata-u.ac.jp	[01/Aug/1995:00:00:18	GET /images/ksclogo-medium.gif HTTP/1.0	200	5866
d0ucr6.fnal.gov	[01/Aug/1995:00:00:19	GET /history/apollo/apollo-16/apollo-16.html HTTP/1.0	200	2743
ix-esc-ca2-07.ix.netcom.com	[01/Aug/1995:00:00:19	GET /shuttle/resources/orbiters/discovery.html HTTP/1.0	200	6849
d0ucr6.fnal.gov	[01/Aug/1995:00:00:20	GET /history/apollo/apollo-16/apollo-16-patch-small.gif HTTP/1.0	200	14897
kgtyk4.kj.yamagata-u.ac.jp	[01/Aug/1995:00:00:21	GET /images/NASA-logosmall.gif HTTP/1.0	304	0
kgtyk4.kj.yamagata-u.ac.jp	[01/Aug/1995:00:00:21	GET /images/MOSAIC-logosmall.gif HTTP/1.0	304	0

Algorithm for data storage: -

Input: Log file in text format **Output:** Log Table

1. Open SQL Server Management Studio and connect to an instance.
2. Create a table with SQL query as shown above with appropriate columns to store the log data.

3. Insert data in bulk from the text file.
4. Close the database connection.

2. Data Cleaning

The implementation of the algorithm is done in Java programming language. It makes use of some of Java's inbuilt classes and methods. It is assumed that space character is acting as the separator. The log file is read character by character up to the end and then by using the methods of String Tokenizer class the data fields are broken into tokens and saved in an array. This algorithm will remove the accessorial entries like jpg, css, gif, png files, entries with status code other than 200 and entries.

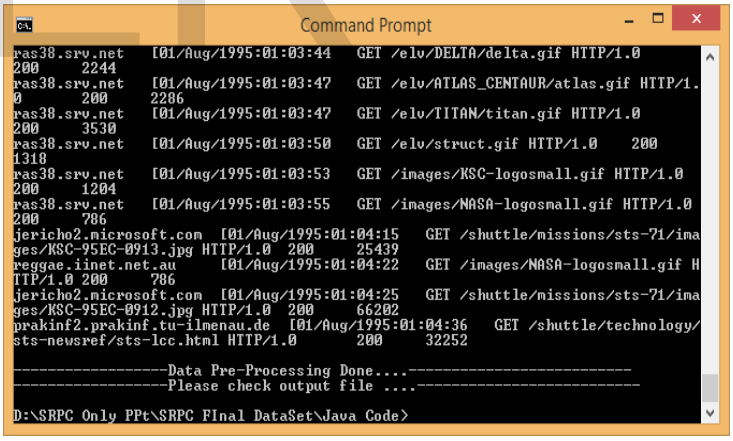
Data Cleaning Algorithm Input: Log Table

Output: Summarized Log Table

1) Declare filename, method, ip address, file extension, hostname, username, timestamp, offset, protocol, bytes, and status code.
2) Open a database connection.
3) Create an object of Prepared Statement to read each record in the log table.
4) For each record read from the log table
   a) Read status_code
   b) Read method
   c) Read filename
   d) Read ip_address

   f) If (status_code=200 and method=GET)
      i) Read hostname, username, timestamp, offset, protocol, bytes.
      ii) Extract file_extension from filename.
      iii) If file_extension !={.png, .jpg, .gif, .css}
      Insert data into summarized log table.
      iv) Else
      v) Remove the entry.

6) Close the connection
7) End



4	d0ucr6.fnal.gov	[01/Aug/1995:00:00:19	GET /history/apollo/apollo-16/apollo-16.html HTTP/1.0	200 2743
5	ix-esc-ca2-07.ix.netcom.com	[01/Aug/1995:00:00:19	GET /shuttle/resources/orbiters/discovery.html HTTP/1.0	200 6849
6	slppp6.intermind.net	[01/Aug/1995:00:00:32	GET /history/skylab/skylab-1.html HTTP/1.0	200 1659
7	uplherc.upl.com	[01/Aug/1995:00:00:43	GET /shuttle/missions/sts-71/mission-sts-71.html HTTP/1.0	200 13450
8	133.43.96.45	[01/Aug/1995:00:00:46	GET /shuttle/resources/orbiters/endeavour.html HTTP/1.0	200 6168
9	uplherc.upl.com	[01/Aug/1995:00:00:55	GET /shuttle/resources/orbiters/atlantis.html HTTP/1.0	200 7025
10	uplherc.upl.com	[01/Aug/1995:00:01:13	GET /shuttle/resources/orbiters/challenger.html HTTP/1.0	200 8089
11	uplherc.upl.com	[01/Aug/1995:00:01:17	GET /history/apollo/apollo-17/apollo-17.html HTTP/1.0	200 2732
12	ip-pdx6-54.teleport.com	[01/Aug/1995:00:01:17	GET /history/history.html HTTP/1.0	200 1602
13	piweba4y.prodigy.com	[01/Aug/1995:00:01:32	GET /history/history.html HTTP/1.0	200 1602
14	uplherc.upl.com	[01/Aug/1995:00:01:38	GET /shuttle/missions/sts-71/images/images.html HTTP/1.0	200 8529
15	133.43.96.45	[01/Aug/1995:00:01:39	GET /shuttle/missions/sts-72/mission-sts-72.html HTTP/1.0	200 3804
16	haraway.ucet.ufl.edu	[01/Aug/1995:00:01:43	GET /facilities/lc39a.html HTTP/1.0	200 7008
17	133.68.18.180	[01/Aug/1995:00:01:48	GET /persons/nasa-cm/jmd.html HTTP/1.0	200 4067
18	ip-pdx6-54.teleport.com	[01/Aug/1995:00:01:48	GET /history/apollo/apollo.html HTTP/1.0	200 3260
19	www-c3.proxy.aol.com	[01/Aug/1995:00:01:48	GET /shuttle/countdown/count.html HTTP/1.0	200 73231
20	endeavor.fujitsu.co.jp	[01/Aug/1995:00:01:51	GET /shuttle/missions/sts-68/ksc-srl-image.html HTTP/1.0	200 1404
21	www-d3.proxy.aol.com	[01/Aug/1995:00:01:52	GET /shuttle/missions/sts-71/mission-sts-71.html HTTP/1.0	200 13450
22	205.163.36.61	[01/Aug/1995:00:01:55	GET /shuttle/countdown/countdown.html HTTP/1.0	200 4324
23	rpgopher.aist.go.jp	[01/Aug/1995:00:01:58	GET /ksc.html HTTP/1.0	200 7280
24	139.230.35.135	[01/Aug/1995:00:02:02	GET /shuttle/missions/sts-49/mission-sts-49.html HTTP/1.0	200 9271
25	ip-pdx6-54.teleport.com	[01/Aug/1995:00:02:03	GET /history/apollo/apollo-13/apollo-13.html HTTP/1.0	200 10556
26	piweba4y.prodigy.com	[01/Aug/1995:00:02:04	GET /history/apollo/apollo.html HTTP/1.0	200 3260

NEW Web log Data Sheet1 Sheet2 Sheet3

Before Data Cleaning 1727
After Data Cleaning 380

EXPERIMENTAL RESULTS

The sample log file used for the work was in raw log format. Size of the file before cleaning was 164 KB with 1727 entries. When cleaning was performed size of file after cleaning was 38 KB with 380 entries.

CONCLUSION

A data preprocessing action system for web usage mining has been investigated and executed for log data. It has different steps such as data cleaning, user identification, session identification and clustering. Dissimilar from usual implementations records are cleaned effectively by removing from some algorithms. This paper has described various information about data preprocessing behavior that is essential to execute Web Usage Mining. In every stage of the data preprocessing, a few rules given to plan and apply them simply and efficiently. Our experiments have approximation data preprocessing importance and our methodology's effectiveness. It is not only to reduce the size of the log file but also increases the quality of the data available.

REFERENCES

[1] Communication of the ACM. Oren Etzioni. The World Wide Web quagmire or gold mine. 1996,39(11)

[2] Xia Huosong. Data warehouse and data mining technology [M],Beijing: Scienc Press, 2004

[3] Long Yinxiang. Gong Weihua, Yang Lianghuai, Jin Rong. Ding Weilong. user interest domain algorithm based on theme. Journal of communication, 2011,11:32- 01

[4] V. Chitraa and A. S. Davamani , "An Efficient Path Completion Technique For Web Mining," in *International Conference on Computational Intelligence and Computing Research*, 2010.

[5] S. Anand and R. R. Aggarwal, "An Efficient Algorithm for Data Cleaning of Log File using File Extensions," *International Journal of Computer Applications*, pp. 13-18, 2012.

[6] S. Langhnoja, M. Barot and D. Mehta, "Pre-Processing: Procedure on Web Log File for Web Usage Mining," *International Journal of Emerging Technology and Advanced Engineering*, pp. 419-423, 2012.

[7] T. T. Aye , "Web log cleaning for mining of web usage patterns," in *Computer Research and Development (ICCRD), 2011 3rd International Conference*, Shanghai, 2011.

[8] M. Srivastava, R. Garg and P. K. Mishra, "Preprocessing Techniques in Web Usage Mining: A survey," *International Journal of Computer Applications*, pp. 1-9, 2014.

[9] "Robot Ip Address," [Online]. Available: http://chceme.info/ips/.

[10] "IP Addresses of

Search Engine Spiders," [Online]. Available: http://www.iplists.com/.

[11] K. Nigam et al., "Text Classification from Labeled and Unlabeled Documents using EM," Machine Learning, vol. 39, no. 2–3, 2000, pp. 103–134.

[12] J. Srivastava et al.,"Web Usage Mining:Discovery and Applications of Usage Patterns from Web Data,"SIGKDD Explorations,vol. 1, no. 2, 2000, pp. 12–23.

[13] R. Kosala and H. Blockeel, "Web Mining Research:A Survey,"SIGKDD Explorations, vol. 2, no. 1, 2000, pp. 1–15.

[14] R. Agrawal et al.,"Fast Discovery of Associ- ation Rules," Advances in Knowledge Dis- covery and Data Mining,U.M. Fayyad et al., eds.,AAAI/MIT Press, 1996, pp. 307–328. MARCH/APRIL 2006 www.computer.org/intelligent

[15] W. Wang, J. Yang, and R. Muntz, "Sting: A statistical information grid approach to spatia data mining," 1997.

[16] R.Agrawal, J.Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of highdimensional data for data mining applications," pp. 94– 105, 1998.

[17] G. Sheikholeslami, S. Chatterjee, and A. Zhang, "WaveCluster: A multi-resolution clustering approach for very large spatial databases," in Proc. 24th Int. Conf. Very Large Data Bases, VLDB, pp. 428– 439, 24–27 1998.

[18] C. Fraley and A. Raftery, "Mclust: Software for model-based cluster and discriminant analysis," 1999.

[19] J. C. Bezdeck, R. Ehrlich, and W. Full, "Fcm: Fuzzy c-means algorithm," Computers and Geoscience, vol. 10, no. 2-3, pp. 191–203, 1984.

[20] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise." in K pp. 226–231, 1996.

[21] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sa "Optics: ordering points to identify the clustering structure," in SIGMOD '99: Proceedin the 1999 ACM SIGMOD international conference on Management of data, (New York, USA), pp. 49–60, ACM Press, 1999.

[22] Dempster, A. P. A Generalization of Bayesian Infer J. Roy. Stat. Soc. B, 30(1968), 205-247.