



User Identification and Page Error Detection in Web Server Logs

¹Dr. Sanjeev Dhawan, ²Mamta Lathwal, ³Karambir

^{1,3}Asst. Professor, CSE Deptt. , University Institute of Engineering & Technology, KUK, India.

²M.Tech Scholar, CSE Deptt. , University Institute of Engineering & Technology, KUK, India.

Abstract: Web mining is the application of data mining techniques to data repositories. Web mining is divided into three parts: content mining, structure mining and log mining. Web log mining is used for extraction of interesting patterns in web access logs. Web log mining is generally divided into three phases: preprocessing, pattern discovery and pattern analysis. Among these phases data preprocessing is most time- consuming. Results of data preprocessing are used in next steps such as path analysis, transaction identification, association rule mining, sequential pattern mining etc. In this paper a novel data preprocessing technique is proposed for data cleaning, unique user identification and system error identification.

Keywords: Web log mining, data cleaning, User identification, field extraction, error detection.

I. Introduction:

The rapid expansion of the Web has provided a great opportunity to study user and system behavior by exploring Web access logs. The WWW is serving as a huge widely distributed global information service center for technical information, advertisement, e-commerce and other information service. WWW is an open, dynamic and heterogeneous global distribution network. There is too huge information but Web pages lack a unified structure. So information retrieval in web pages is difficult. Web mining, which is the application of data mining technologies in Web, can extract useful and interesting pattern and potential knowledge from relevant record. Web mining consists of three main research areas: Web content mining, Web structure mining and Web usage mining. Web Content Mining is the process of extracting useful information from Web documents. The content data is the collection of facts, a Web page is designed to convey to its users. It may consist of images, text, audio, video, or lists and tables. Web structure mining is a method used to identify the relationship between Web pages linked by information or direct link connection. This data can be discovered by using the web structure schema through database techniques. The link connection allows a search engine to pull data relating to a search query directly to the linking Web page. Web usage mining, also called web log mining, is the process of extraction of interesting patterns in web access logs [1].

II. Web Log Mining

Web log mining is also called web usage mining. Whenever a user requests for resources, the web server of a website stores the data about user interaction in the log file that serves as a valuable pool of information. Analyzing the web access logs of different websites can help understand the user behavior and web structure thus improving the website design. Web log mining consists of three phases as preprocessing, pattern discovery and pattern analysis. These three phases are connected to each other to form a complete web log mining methodology.

A. Data Preprocessing

Due to large amount of irrelevant information in the Web log, the original log can't be directly used in the Web log mining procedure. By data cleaning, user identification, session identification and path completion the information in the Web log can be used as transaction database for mining procedure.

B. Pattern Discovery

In this phase of log mining statistical methods as well as data mining methods (path analysis, Association rule, Sequential patterns, and cluster and classification rules) are applied in order to detect interesting patterns. The objective of mining process is to discover sequential association rules. This knowledge will form the knowledge base which can be used in recommendation and personalization systems [2].

C. Pattern Analysis

The patterns discovered are analyzed using knowledge query management mechanism, OLAP tools, and intelligent agent to filter out the uninteresting rules/patterns. The result of such analysis might include:

1. the frequency of visits per document,
2. most recent visit per document,
3. who is visiting which documents,
4. the frequency of use of hyperlinks,
5. most recent use of hyperlinks.

III. Web Data Preprocessing:

The most important task of the web log mining process is data preparation. The success of the website is highly correlated to how well the data preparation task is executed. The data preprocessing of Web usage mining is

usually complex. Purpose of data preprocessing is to offer structural, reliable and integrated data source to pattern discovery. It consists of four steps: data cleaning, user identification, session identification, path completion [3].

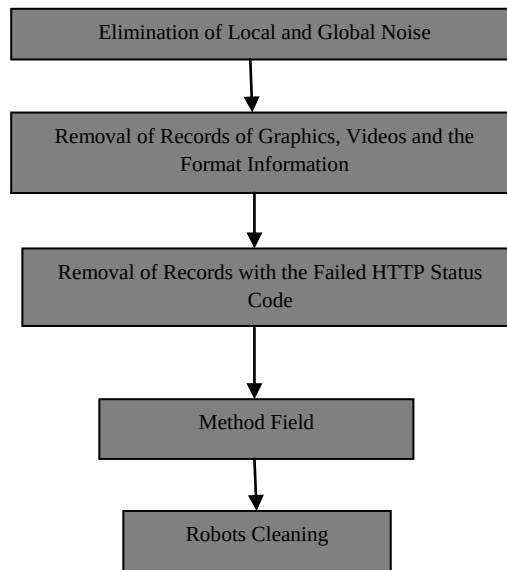


Fig. 1 Steps in Data Cleaning

A. Data Cleaning

Data cleaning refers to the process of deleting the data irrelevant to mining log algorithms in web server logs. This is necessary for improving the mining efficiency. Data cleaning includes elimination of local and global Noise, removal of records of graphics, videos and the format information; removal of records with the failed HTTP status code, robots cleaning [4].

B. User Identification

User identification is the process of identifying each user accessing Web site, whose goal is to mine every user's access characteristic, and then make user clustering and provide personal service for the users.

C. Session Identification

For logs that span long periods of time, it is very common that users will visit the same Web site more than once. The purpose of session identification is to divide the page accesses of each user into individual sessions. A session is a series of web pages that a user browse in a single access. The simplest method of deciding session is through a timeout. If the time between page requests exceeds a certain time limit, it is assumed that the user is starting a new session. Many commercial products use 30 minutes as a default timeout. Because this method is easy and has been tested by some experiments [5].

D. Path Completion

It is necessary to determine the existence of important accesses that are not recorded in the access log. Path completion refers to the inclusion of important page accesses that are missing in the access log due to browser and proxy server caching. Similar to user identification, the heuristic assumes that if a page that is requested by the user is not directly linked to the previous page accessed by the same user, the referrer log can be referred to see from which page the request came. If the page is in the user's recent click history, it is assumed that the user browsed back with the "back" button, using cached sessions of the pages. So each session reflects a full path, including the pages that have been backtracked [6].

IV. Proposed Work

The process of User Identification and analysis of web server logs consists of following steps:

- (i) Extract the web logs that collect the data in the web server.
- (ii) Clean the web logs and remove the redundant information.
- (iii) Parse the data and put it in a relational database or data warehouse and data is reduced to be used in frequency analysis to create summary reports.
- (iv) Identify the page errors and error types.

First of all the data present in web server file is to be extracted and to be stored in the relational database. After that the process of data cleaning and user identification is performed. The server errors and type of those errors are also identified in order to improve the effectiveness of the website.

A. Field Extraction

The web log file contains various fields which we have to separate out for task the preprocessing. The process of separating field from the single line of the log file is known as field extraction. The server uses different characters which work as separators. The most used separator character is or 'space' character. The following algorithm is used for field extraction.

Input: Log File

Output: DB

Begin

Open a DB connection

Create a table to store log data

Open Log File

Read all fields contain in Log File

Separate the Attributes in the string Log

Extract all fields and Add them into the log_table

Close a DB connection and Log File

End

B. Data Cleaning

The purpose of data cleaning is to eliminate irrelevant or unnecessary items in the analyzed data. The records with failed HTTP status codes are also included in log data. Data cleaning is usually site specific and involves extraneous references to embedded objects that may not be important for purpose of analysis, including graphics or sound files, references to style files. Therefore the useless entries are removed from the log files. By Data cleaning, errors and inconsistencies will be detected and removed to improve the quality of data. An algorithm for cleaning the entries of server logs is presented below:

Input: log_table

Output: refine_log_table

'' = access pages consist of objects like .jpg, .gif, etc*

*'**' =successful status codes and requested methods*

Begin

Read records in log_table

For each record in log_table

Read fields (Status code)

*If Status code= '**'*

Then

Get all fields.

If URL.suffix= {.gif, *.jpg, *.jpeg, *.css, *.ico} Then*

Remove URL

Save fields in new table.

End if

Else

Next record

End if

End

C. User Identification

IP address, User agents and referring URL fields of log file are used to identify user. There are some problems which can arise in user identification. We may assume that each user has unique IP address and each IP address represents one user. But in fact there are three conditions: (1) Some user has unique IP address. (2) Some user has two or more IP addresses. (3) Due to proxy server, some user may share one IP address[7].

D. Identification of page errors and error types:

We have analyzed the log files of a web server to get information about visitors; top errors which can be utilized by system administrator and web designer to increase the effectiveness of the web site.

V. Experimental Results:

In this study, we have analyzed the log files of a Web server. First of all, all the fields are extracted from the log file and stored into the database. After that data cleaning and user identification tasks are performed. Then we determined different types of errors that occurred in web surfing. Statistics about detected errors and error types are shown in table 4 and 5. It is clear from the table that 404 is most frequently occurred error. Some other types of client and server errors are shown in Table 5.

Log File	Size (kb)	No. of Records
A.log	1154	4236

Table 1 Input Log File

Usage Data Before Data cleaning		After data cleaning
No. of Unique Users (Without agent)	175	133
No. of Unique Users (With agent)	184	139

Table 2 Number of unique Users identified before and after data cleaning

Log File A.log							
Status	200	206	301	304	403	404	406
Counter	3105	242	13	316	2	557	1

File Type	Txt	Pdf	jpg	gif	Ico	Css
Counter	21	268	968	679	85	527

Table 3 Summary of Statistical Report

Error Code	Error text	Hits
403	Forbidden	2
404	Not Found	557
406	Not Acceptable	1

Table 4 Client Errors

Error Code	URL	Hits
403	/download/	2
404	/admission.html	1
404	/alumni/images/logo.jpg	1
404	/Computer	1
404	/download/admnotice30082011.pdf	1
404	/download/ECE+syllabus+full.pdf	1
404	/download/Mechanical+Engineering.pdf	1
404	/download/BTIL2.pdf	1
404	/download/InsituteLevel.pdf	2
404	/images/footer_bg.gif	287
404	/priyanka.jangra@gmail.com	1
404	/RcMgQ/reetakuk@gmail.com	1
404	/robots.txt	21
404	/slideshow2.css	237
404	/tap	1

406	/images/dsc_2786.jpg	1
-----	----------------------	---

Table 5 Error codes With URL**VI. Conclusion**

Data preprocessing is an important task of web usage mining application. Therefore, data must be processed before applying data mining techniques to discover user access patterns from web log. The data preparation process is often the most time consuming. This project presents two algorithms for field extraction and data cleaning. This system removes accesses to irrelevant items and failed requests in data cleaning. After that important items remain for purpose of analysis. Speed up extraction time when users' interested information is retrieved and users' accessed pages is discovered from log data. The information in these records is sufficient to obtain session information. Another task performed in this project is determination of different types of errors that occurred in web surfing. Statistics about hits, page views, visitors are determined. In order to make a website popular among its visitors, System administrator and web designer should try to increase its effectiveness because web pages are one of the most important advertisement tools in international market for business. The results obtained from the study can be used by system administrator or web designer to arrange their system by determining occurred system errors.

References

- [1] Fang Yuan, Li-Juan Wang, Ge Yu, "Study On Data Preprocessing Algorithm In Web Log Mining", Proceedings of the Second International Conference on Machine Learning and Cybernetics, November 2003 .
- [2] Paweł Weichbroth , Mieczysław Owoc and Michał Pleszkun "Web User Navigation Patterns Discovery from WWW Server Log Files", Proceedings Of The Fedcsis. Wrocław, 2012.
- [3] Liu Kewen," Analysis of Preprocessing Methods for Web Usage Data" IEEE 2012.
- [4] P.Nithya and Dr.P.Sumathi "Novel Pre-Processing Technique for Web Log Mining by Removing Global Noise and Web Robots", National Conference on Computing and Communication Systems, IEEE 2012.
- [5] Zhang Huiying and Liang Wei , " An Intelligent Algorithm of Data Pre-processing in Web Usage Mining", Proceedings of the 5th World Congress an Intelligent Control and Automation, IEEE 2004.
- [6] V.Chitraa and Dr. Antony Selvadoss Davamani, "An Efficient Path Completion Technique for web log mining", IEEE International Conference on Computational Intelligence and Computing Research, 2010.
- [7] Ma Shu-yue, Liu Wen-cai and Wang Shuo "Study On Data Preprocessing Algorithm In Web Log Mining".