ISSN: 2348 - 2117

International Journal of Engineering Technology and Computer Research (IJETCR)
Available Online at www.ijetcr.org
Volume 5; Issue 3; May-June: 2017; Page No. 161-167
Journal Approved by UGC

A study on process of Multi Data Agent of Web Log Files

Keshav Dev Gupta, Dr. Kavita

[1] Research Scholar, [2]Associate Professor

Jayoti Vidyapeeth Women's University Jaipur

Abstract

Web, the largest resource for data from several fields, systematically keeps and maintains all the information in the form of web sites. The lots of information which is available in the web exhibits the necessity of web data log to observe and locate all the transactions or communication made possible through the web. Generally, the information about the working users' web accesses gets stored generally in the web server or in the intermediate proxy server in the form of log entries. Web mining helps to Retrieve and analyze relevant information on the web with the use of log files. This paper presents general steps and algorithms of some of the important steps which can be used for data mining from log files.

Keywords: *Data mining, log files, retrieving of data*

Introduction

Web mining, is a methodology of data mining with a practice of identifying the behavioral patterns of the web users. In other words, it helps to extract and analyze helpful information from net. Web usage mining is a kind of web mining that is applied to assure the patterns of web usage actually derived from the data stored in the web log file. It is essential to know that, the web log files actually contain deafening, indecisive and uncertain data. So, it becomes compulsory to clean the log file to develop actual user access sequences. As fact, a user's session exhibits a collection of requests fired by a particular user during a particular navigation or a time. Perhaps, a user would have either single or more than single session during a particular period of time. Web mining uses data mining terminology to find out patterns from web register file, as inferred through the data mining aspect. Actually, the web log files are processed to assure the web users, their access to the URLs, the time used up by a user in all session, besides, the complete path of their web access. Lots of algorithms are proposed in for web log mining which can be used for merging of web log files, representing the merged log file in CLF(Common Log File), cleaning the merged log file and identifying the web users and their corresponding sessions. The web log files in general processed in two steps namely, pre-processing the log file and differentiate various web users and their respective sessions exclusively.

Generating the Web Log File

Initially, the log entries generated from different web servers are collected with a specific purpose to merge and

Corresponding author: Keshav Dev Gupta

clean the web log files. If log files {Log_1, Log_2, …, Log_n} are considered, where n is number of log files. Consecutively, same log records are merged into a single log file. Further, the entries in the log file can be considered where Log_i have {$Log_{i,1}$, $Log_{i,2}$, …, $Log_{i,m}$}, 1≤ i ≤ n and m = number of log entries in the i^{th} log file. The subsequent algorithm is used to combine the log records into a on its own log folder.

A log entry in the web log file perhaps, acts as an anthology of fields and the consolidated

fields are separated by space. The unused field is marked as "-"

An Algorithm to generate a compound File

```
Input : dissimilar Log file
Output : The compound Log file CLF
begin
    let num = amount of log files for a from
        1 to num do
            for every log entry b in the a^th log file do
                combine the log entry b in the file CLF
                end
        end
    organize the log entries based on the entrée time.
    return (CLF)
End
```

Because of difference in format of various log files, It to be compulsory to create the ordinary log file, where the log entries are controlled under the ordinary set of variables.

Pre-processing of CLF

The pre-processing step starts with the removal of all the data tracked in web logs that are apparently futile for the mining point of view, For example, graphics page data requests like jpeg (Joint photographic expert group) and gif (graphical interchange format), imagery, desires for some other file that may not be integrated in a web page, even browsing sessions by robots and web spiders. In this research, the pre-processing steps are applied in the log files based on the format, content and purpose of the log file. The algorithm applied in the cleaning of the log file is given below:**An Algorithm to Clean the CLF File**

```
Input : CLF file which have inventory of log entries Output:
Cleaned CLF file
begin
    let initlog  = deign log entries for input purpose
```

outlog = design the ouput clean of all log entries scode

= condition code of the log entry

exten = extend of the familiar URL

met = access function

Number 1 Step: for every log entry x in initlog do

Tokenize the log entry x using StringTokenizer stock up the tokens in outlog

Number 2 Step: for each token i in the outlog do

if ((scode>=300 && scode<=399) && (met is either POST or GET) && (exten is not equals to image formats and automated programs))

print the token x in outlog

End

User Identification

Initially, using the fields, IP addresses and browsing agents, the users are identified with the inference obtained from the cleaned CLF web log file. The algorithm stated subsequently is used to identify the users in log file:

Algorithm to find out the user from Pre-processed CLF File Input : Cleaned Pre-processed CLF log file.

Output: List of users start

Suppose d= Date object have the user's navigation initial date and time d1= Date object have the user's navigation last date and time user = user valuable information

Number 1 Step: for x from 1 to number of log entries do

Assign flag[x] is false

Number 2 Step: for every log entry x in the CLF log file do if (flag[x] is false) k=0

Allocate the URL from the log entry to user[x][k]

Allocate user x initial time into d1; Change flag[i] is true; for each log entry j from i+1 to n do

if(IP address and browsing agent of x^{th} and y^{th} log entries are same) Assign the y^{th} log entry to the same user x; Change flag[y] is true Assign user x ending time into d2

Locate the time taken by the user as dateDifference (d1, d2) Increment x for the next user

End

Session Identification of the Users

A group of requests complete by a single user for a single routing is titled as a session. The algorithm elucidated in this section identifies the sessions for each user. Here, the maximum time taken per session is noticed to be a specific time like 30 minutes.

Algorithm to Identify the Sessions for each User from the Log File Input : Log entries user-wise

Output: Sessions from each user and the visited URL begin

Suppose p = amount of users

referrer = String have the referrer URL for x from 1 to p

do

for x from 1 to number of entries for x^{th} user

if (referrer. equals(„-") || time differ(d1,d2)>30) Store it in a fresh Session

else

backward: for k from 0 to y-1 // backward reference if (referrer. equals (referrer of k^{th} log entry))

The log entry y belongs to same user session and break backward

end

Generation of Frequent Patterns

The generation of frequent patterns is keenly observed through all the tetra-steps dealt in detail, namely,

- Analysis of the Query Log File
- Identification of Users and their Sessions
- Architecture to Identify the Relationship between the Users
- Generation of Query Access Pattern

Analysis of the Query log file

When a user submits the query to the searching process, the following events may occur.

- User clicks any one of the URL in the search result.
- User clicks any one of the image formats in the search result.
- User is idle for some period of time.

User is satisfied with the information available in the web snippets, which is returned by the search engine.

Algorithm QLF-cleaning

Input : Query log file

Output: Cleaned Query log file begin

Let inlog = Input query log entries outlog = Cleaned
query log entries
for each query log entry i in inlog do
if an URL extension is an image format or URL is „-„ remove the
query log entry from inlog
else
assign the query log entry to outlog
end end

Identification of Users and their Sessions

Generation of Query Access Pattern

Generation of query access pattern is achieved by the execution of the subsequent algorithm titled, QUERY_PATTERN.

Algorithm QUERY_PATTERN Input : Cleaned
Query log file Output : Query and access patterns
begin

Step 1: for each user U_i do [where i=1 to number of users] Group the query
log entries based on query term

Step 2: for each Query term for user U_i do

Generate the access pattern for each session by concatenating the URLs

end

URL Click Count From Search Engine Query Logs

The search engines must have an approach to find the intent of the user with respect to their queries and click-through data. The proposed system in this section mines the query log file and generates the relationship between the queries by using click count of the URL.

Finding URL Click Count

Algorithm SEQ_URLCOUNT

Input : User and Query-wise URL access pattern Output: User
and Query-wise URL click count begin

Step 1: for every user U_i do

Classify the query log entries based on the query term for every query term Q_j do

Combine the URLs for each query term Q_j Step 2: for each query cluster pattern Q_j do

Find the URL count for the unique URL u_k Store the userid, query term, URL and its count

end

Processing of Query Log file

To start with, an analysis of the log entries from the search engine‟s query log are pursued in connection to query log search or report. The striking factor here is the organization of the search available in the log file has its firm foot on the defined general and different attributes. Generally, the response of a search engine to the user‟s query Q is edified through the web snippets. The prominent recurrence of a keyword in the top retrieved documents of the web snippets add adequate significance to the concept corresponding to that particular query.

REFERENCES

1. S. K. Pani Et El(2011), Web Usage Mining: A Survey On Pattern Extraction From Web Logs, Web Usage Mining: A Survey On Pattern Extraction From Web Logs, International Journal Of Instrumentation, Control & Automation (Ijica), Volume 1, Issue 1,
2. S. K. Pani Et El(2011), Web Usage Mining: A Survey On Pattern Extraction From Web Logs, Web Usage Mining: A Survey On Pattern Extraction From Web Logs, International Journal Of Instrumentation, Control & Automation (Ijica), Volume 1, Issue 1,
3. Mehak(2013), Web Usage Mining: An Analysis, Journal Of Emerging Technologies In Web Intelligence, Vol. 5, No. 3,
4. Govind Murari(2013), Web Content Mining: Its Techniques and Uses, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 11
5. Priyanka Verma(2014), Web Usage Mining Framework For Data Cleaning And Ip Address Identification, International Journal Of Advanced Studies In Computer Science And Engineering Ijascse, Volume 3, Issue 8
6. Shiva Asadianfam And Masoud Mohammadi(2014), Identify Navigational Patterns Of Web Users, International Journal Of Computer-Aided Technologies (Ijcax) Vol.1,No.1,
7. Dr. Girish S. Katkar(2014), Use Of Log Data For Predictive Analytics Through Data Mining, Current Trends In Technology And Science Volume: 3, Issue: 3
8. Pooja Kherwa (2015). Data Preprocessing: A Milestone Of Web Usage Mining , International Journal Of Engineering Science And Innovative Technology (Ijesit) Volume 4, Issue 2,
9. Shivaprasad G.(2015), Knowledge Discovery From Web Usage Data: An Efficient Implementation Of Web Log Preprocessing Techniques, International Journal Of Computer Applications (0975 – 8887) Volume 111 – No 13
10. V. Bharanipriya & V. Kamakshi Prasad WEB CONTENT MINING TOOLS: A COMPARATIVE STUDY.
11. M.Santhanakumar(2015), Web Usage Based Analysis Of Web Pages Using Rapidminer, Wseas Transactions On Computers, Volume 14

12. Chandana S. Khatavkar(2015), A Hybrid Approach For Clustering Weblog(2015), International Journal Of Advanced Research In Computer Science And Software Engineering, Volume 5, Issue 3

13. Durga Choudhary(2015), Review Paper on Web Content Mining International Journal of Research in Engineering and Applied Sciences