IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified

Vol. 5, Issue 11, November 2016

An Improved Preprocessing and Clustering Using Web Log Data

Bharat Chauhan[1], Hemant Kumar[2], Mihul Singh[3], Piyush Kumar[4], Ms. Sakshi Hooda[5]

Bachelor of Technology in Computer Science & Engineering, Maharaja Surajmal Institute of Technology, C-4, Janak Puri, New Delhi, India[1, 2, 3, 4]

Assistant Professor, Department of Computer Science &Engineering, Maharaja Surajmal Institute of Technology, C-4, Janak Puri, New Delhi, India[5]

Abstract: The process of web usage mining is implemented in an attempt to discover the patterns of user browsing from the web log data available on the server side of almost each and every website or in the user's cache. By analyzing the browsing behavior of various users, next web page predictions can be made which is an important aspect of most of the websites around these days. But the prediction of future requests comes with its own set of issues which make it unreliable at times; there are some concerns with accuracy and efficiency. In this paper, we have designed a custom algorithm for the Clustering process. The main aim of this algorithm is to provide more efficient and accurate results as compared to the other Clustering algorithms.

Index Terms: Web Usage Mining, Recommendation, Prediction, Browsing Patterns, Patterns, Clustering, Preprocessing.

1. INTRODUCTION

The internet has become a global necessity. We use the internet for various tasks such as watching a video, sharing information, banking, shopping, socializing and much more. As a result, the amount of Web Data produced throughout the internet has escalated tremendously. This Web Data comes in large quantities and is seriously unorganized and unstructured in nature.

In this plethora of data, some room must be made for personalization as well, in order to make the internet a little more feasible for the user. Web personalization can be viewed as a process which aims to make the web experience of a user more adaptive to the user's behavior.

Web Data Mining is an application which can be used for the discovery of valuable information or knowledge from the user's data that has been collected or cached over a certain period of time. Web mining allows us to look for patterns in data through content mining, structure mining, and usage mining. Content mining is used to examine data collected by search engines and Web spiders.

Structure mining is used to examine data related to the structure of a particular Web site and usage mining is used to examine data related to a particular user's browser as well as data gathered by forms the user may have submitted during Web transactions. Web structure mining is a tool that is used for recognizing the connection between web pages linked by information. Web Usage Mining mainly deals with the extraction of information from the server log files. Server Log files contains textual data which is available in standard formats.

Web Usage Mining has two main applications which are termed as Personalization and Recommendation. Personalization is something that the users need; it is a tool which aims to make the browsing experience for the users a little more personal. The main elements in web personalization are the users and the web objects. User profiles which are based on the user’s navigation activities are used to determine the Personalization actions. Since various aspects of Personalization are used for the purpose of customization, this presents a model for fulfilling an individual’s needs and to provide sensible and accurate results to the users.

Web Usage Mining processes the web log files, which contain the user’s navigational information. Using these log files which store the URL, IP Address, Timestamp as well as the user’s current navigation pattern, recommendations can be generated for the next web files to the user in form of a smart recommendation list. In the process of Web Usage Mining, prediction is implemented in two stages - Training and Testing. Under training, the knowledge base is prepared by making use of the web log files captured from the web servers, proxy servers or the client’s system cache. While under testing, prediction is done by using the previously prepared knowledge base and including the current navigation pattern so as to recommend the next web page to the users.

2. RELATED WORKS

Web Usage Mining is a developing field in the research area. It was Etzioni [1] who first coined the term Web

DOI 10.17148/IJARCCE.2016.51120

Mining. Etzioni started by making a hypothesis that the information on the Internet is sufficiently structured and outlines the subtasks of Web mining. His paper describes the Web mining processes. The first paper that we know that noticed the confusion in the Web Mining research is [2]. It gives a Web mining taxonomy but restricted to Web content, it also gives a survey on Web usage mining. In [3], the classification of Web mining into three categories is briefly explained along with identifying additional user-centered Web mining processes and providing new perspectives for the Web mining categories.

All this research on Web Usage Mining and recommendation is done to improve the accuracy and efficiency of the system. However, some performance issues are there. Regression Analysis is an accurate method for prediction that is applicable to numeric values. However, our proposed algorithm shall improve the accuracy and efficiency of prediction.

3. WEB MINING

Web mining is the implementation of data mining techniques to consequently discover and extract valuable information from the Web documents and services. The reason why this research is so extensive today is mainly due to the interests of various research communities, the tremendous advancement in the sources of information available on the Web and the recent interest in online shopping websites.

Web Mining can be distributed into these subtasks, namely:

- Resource finding: The effort to retrieve intended Web based documents.
- Information selection and pre-processing: The natural selection and pre-processing of specific data from various retrieved Web resources.
- Generalization: Discovering general patterns on unique individual Web sites as well as across multiple sites automatically.
- Analysis: Authorization or clarification of the freshly discovered patterns.

Resource finding is the mechanism of fetching the data that is either online or offline from the text sources available on the Web such as newsletters, the text contents of HTML documents obtained by removing HTML tags and also the manual selection of Web resources. We also incorporate text sources that were originally not accessible from the Web but are now made accessible; these include online texts made for research purposes, text databases etc.

Such transfigurations could be either a kind of pre-processing that are mentioned above such as removing stop words or a pre-processing aimed at achieving the desired representation such as finding expressions in the training corpus, transforming the representation to relational or first order logic form. In the third step above, data mining techniques or machine learning are typically used for the generalization. We should also take note that humans play an important role in the information or knowledge discovery process on the Web, since the Web is a correlative medium. Hence, interactive query-triggered knowledge discovery is just as important as the more automatic data-triggered knowledge discovery. However, we exclude the knowledge discovery done manually by humans. In this way, Web Mining refers to the mechanism of determining potentially useful and formerly unknown information or knowledge from the Web data. It essentially covers the accepted process of knowledge discovery in databases. There is a close relationship between machine learning, data mining and advanced data analysis.

3.1 PREPROCESSING

Proposed work is done on web log file generated by web server. First we have to collect weblog file then preprocessing is implemented on weblog file. In preprocessing module unformatted web log data is converted into formatted web log data. Formatted data obtained through preprocessing is used for further processing. The Preprocessing includes three steps which are cleaning, user identification and session identification. All entries which will have no use during mining are removed in cleaning. Users are identified on the basis of Ip address in user identification. In Session identification process sessions are identified by taking threshold value of time.

Find and remove all entries which has accessed server file•

• Find and remove all entries with visiting time of access as midnight (commonly used as the network activity at that time is light) Remove entry when access mode is HEAD instead of POST or GET•

•Calculate browsing speed and remove all entries whose speed exceeds a threshold T1 and number of visited pages exceeds a threshold T2.•

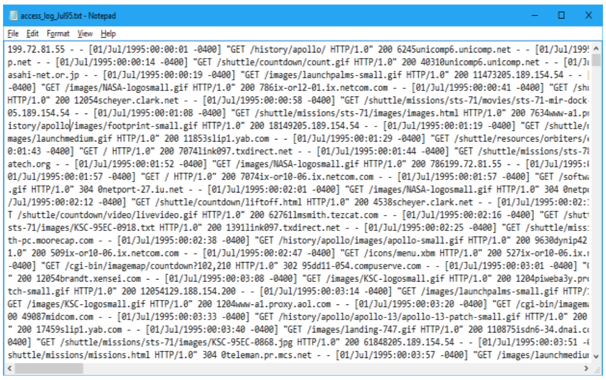
The process of Preprocessing is implemented on the web log files generated by the web server. In the beginning, we have to collect the web log files after which Preprocessing is implemented upon them. In the Preprocessing module [4], the unformatted and unstructured web log data is converted into properly formatted web log data. Formatted data obtained through Preprocessing is used for further processing functions. Preprocessing includes three steps which are Data Cleaning, User Identification and Session Identification. All attributes in the web Log files which have no use during information mining are removed in the cleaning process. Users are identified on the basis of their IP Address or the MAC Address in some cases. In the Session identification process, the following steps are followed:

- Find and remove all attributes which have accessed the server files.

DOI 10.17148/IJARCCE.2016.51120

- Filter out all attributes that have time stamps containing odd visiting hours.
- Calculate browsing speed and remove all entries whose speed and number of visited pages exceeds a particular threshold.

The following images are the preprocessing process that was used in this research.

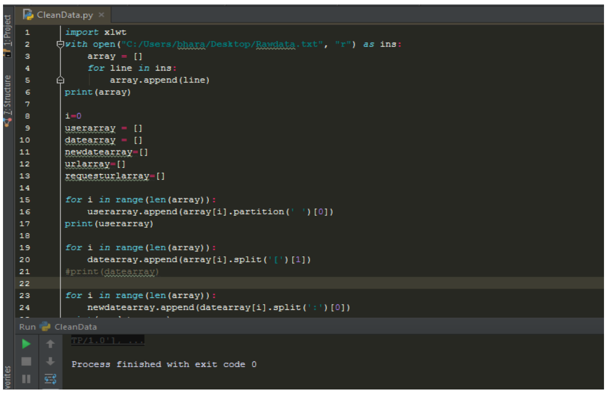


access_log_Jul95.txt - Notepad

File Edit Format View Help

```
199.72.81.55 - - [01/Jul/1995:00:00:01 -0400] "GET /history/apollo/ HTTP/1.0" 200 6245unicomp6.unicomp.net - - [01/Jul/199
p.net - - [01/Jul/1995:00:00:14 -0400] "GET /shuttle/countdown/count.gif HTTP/1.0" 200 40310unicomp6.unicomp.net - - [01/Ju
asahi-net.or.jp - - [01/Jul/1995:00:00:19 -0400] "GET /images/launchpalms-small.gif HTTP/1.0" 200 11473205.189.154.54 - -
-0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 200 786ix-orl2-01.ix.netcom.com - - [01/Jul/1995:00:00:41 -0400] "GET /shu
HTTP/1.0" 200 12054scheyer.clark.net - - [01/Jul/1995:00:00:58 -0400] "GET /shuttle/missions/sts-71/movies/sts-71-mir-dock-
05.189.154.54 - - [01/Jul/1995:00:01:08 -0400] "GET /shuttle/missions/sts-71/images/images.html HTTP/1.0" 200 7634www-a1.pr
istory/apollo/images/footprint-small.gif HTTP/1.0" 200 18149205.189.154.54 - - [01/Jul/1995:00:01:19 -0400] "GET /shuttle/r
mages/launchmedium.gif HTTP/1.0" 200 11853slip1.yab.com - - [01/Jul/1995:00:01:29 -0400] "GET /shuttle/resources/orbiters/
0:01:43 -0400] "GET / HTTP/1.0" 200 7074link097.txdirect.net - - [01/Jul/1995:00:01:44 -0400] "GET /shuttle/missions/sts-7
atech.org - - [01/Jul/1995:00:01:52 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 200 786199.72.81.55 - - [01/Jul/1995:
01/Jul/1995:00:01:57 -0400] "GET / HTTP/1.0" 200 7074ix-or10-06.ix.netcom.com - - [01/Jul/1995:00:01:57 -0400] "GET /softw
.gif HTTP/1.0" 304 0netport-27.iu.net - - [01/Jul/1995:00:02:01 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 304 0netp
/Jul/1995:00:02:12 -0400] "GET /shuttle/countdown/liftoff.html HTTP/1.0" 200 4538scheyer.clark.net - - [01/Jul/1995:00:02:
T /shuttle/countdown/video/livevideo.gif HTTP/1.0" 200 62761lmsmith.tezcat.com - - [01/Jul/1995:00:02:16 -0400] "GET /shut
sts-71/images/KSC-95EC-0918.txt HTTP/1.0" 200 1391link097.txdirect.net - - [01/Jul/1995:00:02:25 -0400] "GET /shuttle/miss
th-pc.moorecap.com - - [01/Jul/1995:00:02:38 -0400] "GET /history/apollo/images/apollo-small.gif HTTP/1.0" 200 9630dynip42
1.0" 200 509ix-or10-06.ix.netcom.com - - [01/Jul/1995:00:02:47 -0400] "GET /icons/menu.xbm HTTP/1.0" 200 527ix-or10-06.ix.
-0400] "GET /cgi-bin/imagemap/countdown?102,210 HTTP/1.0" 302 95dd11-054.compuserve.com - - [01/Jul/1995:00:03:01 -0400] "
" 200 12054brandt.xensei.com - - [01/Jul/1995:00:03:08 -0400] "GET /images/KSC-logosmall.gif HTTP/1.0" 200 1204piweba3y.pr
tch-small.gif HTTP/1.0" 200 12054129.188.154.200 - - [01/Jul/1995:00:03:14 -0400] "GET /images/launchpalms-small.gif HTTP/
GET /images/KSC-logosmall.gif HTTP/1.0" 200 1204www-a1.proxy.aol.com - - [01/Jul/1995:00:03:20 -0400] "GET /cgi-bin/imagem
00 49087midcom.com - - [01/Jul/1995:00:03:33 -0400] "GET /history/apollo/apollo-13/apollo-13-patch-small.gif HTTP/1.0" 200
" 200 17459slip1.yab.com - - [01/Jul/1995:00:03:40 -0400] "GET /images/landing-747.gif HTTP/1.0" 200 110875isdn6-34.dnai.c
0400] "GET /shuttle/missions/sts-71/images/KSC-95EC-0868.jpg HTTP/1.0" 200 61848205.189.154.54 - - [01/Jul/1995:00:03:51 -
shuttle/missions/missions.html HTTP/1.0" 304 0teleman.pr.mcs.net - - [01/Jul/1995:00:03:57 -0400] "GET /images/launchmediu
```

Fig. 1: Raw web log file used in this research.



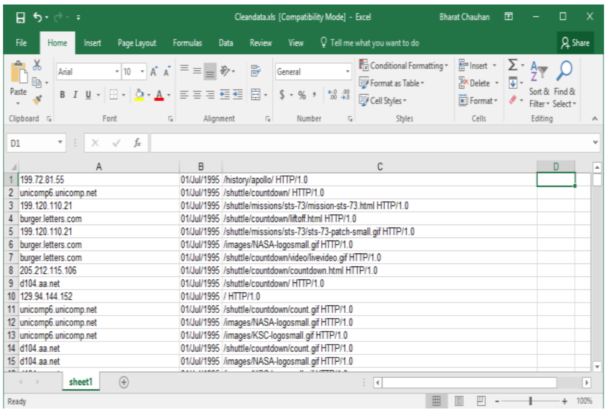
CleanData.py

```python
import xlwt
with open("C:/Users/bhara/Desktop/Rawdata.txt", "r") as ins:
    array = []
    for line in ins:
        array.append(line)
print(array)

i=0
userarray = []
datearray = []
newdatearray=[]
urlarray=[]
requesturlarray=[]

for i in range(len(array)):
    userarray.append(array[i].partition(' ')[0])
print(userarray)

for i in range(len(array)):
    datearray.append(array[i].split('[')[1])
#print(datearray)

for i in range(len(array)):
    newdatearray.append(datearray[i].split(':')[0])
```

Run CleanData

```
TP/1.0'], ...

Process finished with exit code 0
```

Fig. 2: Data Preprocessing using Python.



Cleandata.xls [Compatibility Mode] - Excel

	A	B	C	D
1	199.72.81.55	01/Jul/1995	/history/apollo/ HTTP/1.0	
2	unicomp6.unicomp.net	01/Jul/1995	/shuttle/countdown/ HTTP/1.0	
3	199.120.110.21	01/Jul/1995	/shuttle/missions/sts-73/mission-sts-73.html HTTP/1.0	
4	burger.letters.com	01/Jul/1995	/shuttle/countdown/liftoff.html HTTP/1.0	
5	199.120.110.21	01/Jul/1995	/shuttle/missions/sts-73/sts-73-patch-small.gif HTTP/1.0	
6	burger.letters.com	01/Jul/1995	/images/NASA-logosmall.gif HTTP/1.0	
7	burger.letters.com	01/Jul/1995	/shuttle/countdown/video/livevideo.gif HTTP/1.0	
8	205.212.115.106	01/Jul/1995	/shuttle/countdown/countdown.html HTTP/1.0	
9	d104.aa.net	01/Jul/1995	/shuttle/countdown/ HTTP/1.0	
10	129.94.144.152	01/Jul/1995	/ HTTP/1.0	
11	unicomp6.unicomp.net	01/Jul/1995	/shuttle/countdown/count.gif HTTP/1.0	
12	unicomp6.unicomp.net	01/Jul/1995	/images/NASA-logosmall.gif HTTP/1.0	
13	unicomp6.unicomp.net	01/Jul/1995	/images/KSC-logosmall.gif HTTP/1.0	
14	d104.aa.net	01/Jul/1995	/shuttle/countdown/count.gif HTTP/1.0	
15	d104.aa.net	01/Jul/1995	/images/NASA-logosmall.gif HTTP/1.0	

Fig. 3: Structured Data obtained in the Preprocessing process.

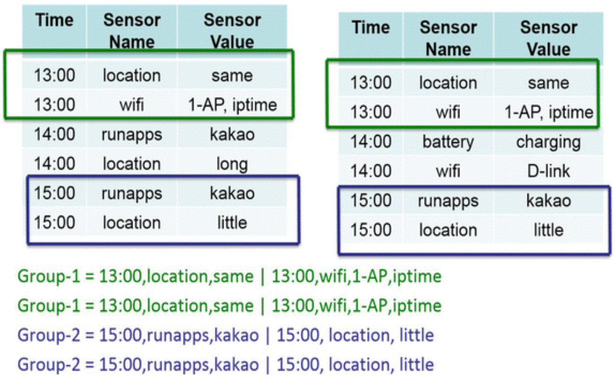
3.2 CLUSTERING PROCESS

A unique algorithm designed by us is used for the Clustering process. It aims to produce efficient and accurate results. The main motive of the algorithm is to group users that have similar navigation patterns.

```
Algorithm: while (D in w) for all of D do
        datavalue←D.current.day
        datavaluenext←D.next.day
        grpTemp←findingSimilarPatterns (datavalue, datavaluenext)
        if (grpTemp in grpPrevious) then
                grpNew← merge(grpPrevious , grpTemp)
                grpAll← add (grpNew)
        else
                grpAll← add (grpTemp)
```

An explanation of this algorithm is as follows:
datavalue←D.current.day
datavaluenext←D.next.day

Where data value and datavalue next are the variables that are used to store the values of the data, D.current.day and D.next.day are the parameters used for present day and next day respectively while D is used for data.



Time	Sensor Name	Sensor Value
13:00	location	same
13:00	wifi	1-AP, iptime
14:00	runapps	kakao
14:00	location	long
15:00	runapps	kakao
15:00	location	little

Time	Sensor Name	Sensor Value
13:00	location	same
13:00	wifi	1-AP, iptime
14:00	battery	charging
14:00	wifi	D-link
15:00	runapps	kakao
15:00	location	little

Group-1 = 13:00,location,same | 13:00,wifi,1-AP,iptime
Group-1 = 13:00,location,same | 13:00,wifi,1-AP,iptime
Group-2 = 15:00,runapps,kakao | 15:00, location, little
Group-2 = 15:00,runapps,kakao | 15:00, location, little

Fig. 4: Finding relations in data for clustering by the algorithm used in this research.

Step by step explanation of Algorithm.

Step 1: Data Assignment. Every data is assigned according to the day it is being generated. These outcomes in grouping the information.
grpTemp←findingSimilarPatterns(datavalue, datavaluenext)

Step 2: Grouping. Values of both the parameters are matched and are grouped in a temporary variablegrpTempfor further processing of the data.
If (gsrpTemp in grpPrevious) then
    grpNew←merge (grpPrevious, grpTemp)
    grpAll←add(grpNew)

Step 3: Now check the group made in step 2 with the previously made groups. If the group is matched with one of the previous groups, then merge the group with its match.
else
    grpAll← add (grpTemp)

Step 4: If the group made in step 2 is not matched with any of the groups previously made then add the group in the group all table and continue.

DOI 10.17148/IJARCCE.2016.51120

4. EXPERIMENTAL RESULTS

The following results were obtained in this research.

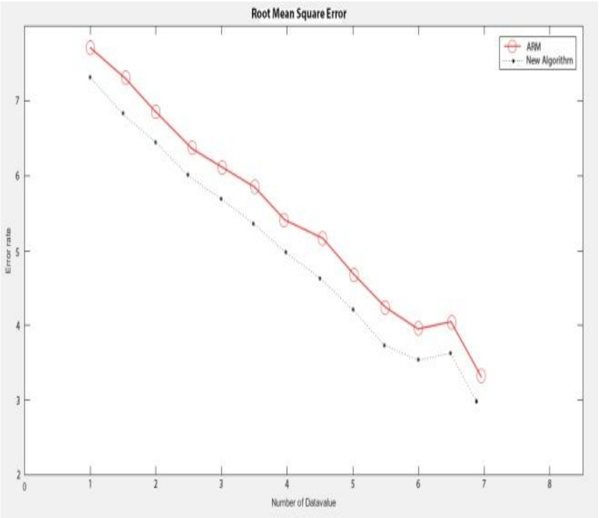


Fig. 5: Root Mean Square Error Graph between Ant Colony Optimization & the New Algorithm which shows that by applying association rules on clustered data, we get more accurate results having less rate as compared to apply it on preprocessed datasets.

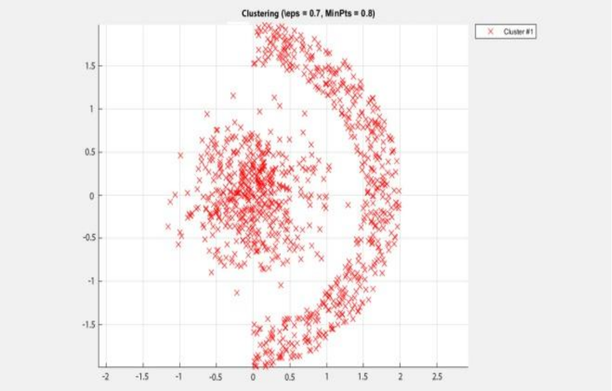


Fig. 6: Clustered data not having any noise. Result obtained by using the new algorithm.

5. CONCLUSION

In this paper, the proposed program does the function of generating smart recommendations and website predictions through the process of collecting the Web Log data files from the server side of a particular website and then Pre-Processing that Log file in order to clean and give structure to the unorganized data. A program designed using Python is implemented for Pre-Processing. Once this is accomplished, the formatted data is used in the Clustering process where the data is analyzed and clusters are made by collecting the similar data.

For the Clustering process, we implement your own unique algorithm. The main aim of this algorithm is to speed up the clustering process, improve reliability and reduce the noise generated in the process. Our algorithm analyzes the data based on days as compared to the other algorithms which implement time instead. We also compare the results that we obtained with the outcome obtained in those cases when Ant Colony Optimization was used. The given model has potential for improvement when it comes to data evaluation time required.

REFERENCES

[1] O. Etzioni. "The world wide web: Quagmire or Gold Mine. Communications of the ACM", 39(11):65–68, 1996.

[2] R. Cooley, B. Mobasher, and J. Srivastava. "Web mining: Information and Pattern Discovery on the World Wide Web". In Proceedings of the 9th IEEE International Conference on "Tools with Artificial Intelligence" (ICTAI'97), 1997.

[3] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data". SIGKDD Explorations, 1(2), 2000.

[4] Vedpriya Dongre, Jagdish Raikwal, "An Improved User Browsing Behavior Prediction Using Web Log Analysis", International Journal of Advanced Research in Computer Engineering and technology (IJARCET), Vol. 4, Issue 5, 2015.

DOI 10.17148/IJARCCE.2016.51120