Classification of Web Log Data to Identify Interested Users

Herit Trivedi[1], Mr. Narendra Limbad[2]

[1] M.E. Student, Computer Engineering, L.J. Institute of Engineering & Technology, Gujarat, India
[2] Assistant Prof., Computer Engineering, L.J. Institute of Engineering & Technology, Gujarat, India

ABSTRACT

—With the increasing demand of internet more number of website are used for getting required information and thus more usage of web-based data. Whereas the data that is stored in different type of format in form of web log file. This log file should be maintained as these data are in unsorted manner and it is done through preprocessing. Web usage mining focuses on discovering useful knowledge or information. Web log file is automatically generated by web server whenever user accesses the resource like webpage of website. Web Usage Mining consists of three steps, Data Preprocessing, Pattern Discovery and Pattern Analysis. Data Preprocessing extracts text format data form log file and store clean data into database. Pattern Discovery finds pattern, Classify data by applying mining techniques. Pattern analysis finds knowledge from the discovered pattern.The main objective of this thesis is instead of spending high amount of time in tracking the behaviour of overall users to redesign the web site, spend less amount of time in focusing interested group of users only.The existing model used Naive Bayesian Classification to identify interested group of users from web log data. In this we propose Classification based on Predictive Association Rules Mining (CPAR) algorithm to identify interested group of users and also we present a comparative study of Naive Bayesian with CPAR.

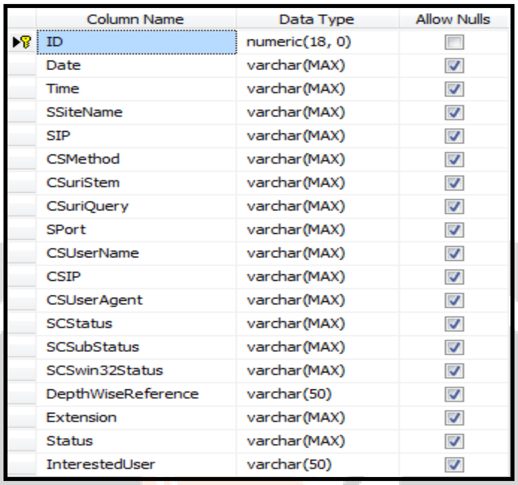
Keyword: - *Web Mining, Web Data Mining, Web Usage Mining, Data Preprocessing, Log File Analysis, Web Log Mining, Associative Classification.*

1. INTRODUCTION

Web usage mining (WUM) term is derived from the term web mining. Whereas web is a collection of inter-related files on one or more web servers and web mining is the application of data mining technique to extract knowledge from web data. Web usage mining is the application that uses data mining to analyze and discover interesting patterns of user's usage data on web. Ex. http logs, app server logs, etc. To deal with this web log files they must be first cleaned and preprocessed and then their pattern is discovered and then analyzed and finally with the rule generation is derived with interested and not interested ones. For discovering the pattern and its analysis classification based association algorithm is been used comparing with the naïve Bayesian algorithm. For finding interested and not-interested ones it undergoes three process and are as derived below[1].

1.1 Preprocessing of web log files

It includes usage preprocessing, content preprocessing and structure preprocessing. Other steps that are involved in preprocessing are data cleaning, efficient user identification, session identification, path completion and transaction identification. Here unwanted files such as robot files, media files such as image files, etc. are been removed/cleaned and finally provides cleaned data with only the required web log files such as .aspx,.java,.php, etc files are kept and further their pattern usage are discovered.



Column Name	Data Type	Allow Nulls
ID	numeric(18, 0)	☐
Date	varchar(MAX)	☑
Time	varchar(MAX)	☑
SSiteName	varchar(MAX)	☑
SIP	varchar(MAX)	☑
CSMethod	varchar(MAX)	☑
CSuriStem	varchar(MAX)	☑
CSuriQuery	varchar(MAX)	☑
SPort	varchar(MAX)	☑
CSUserName	varchar(MAX)	☑
CSIP	varchar(MAX)	☑
CSUserAgent	varchar(MAX)	☑
SCStatus	varchar(MAX)	☑
SCSubStatus	varchar(MAX)	☑
SCSwin32Status	varchar(MAX)	☑
DepthWiseReference	varchar(50)	☑
Extension	varchar(MAX)	☑
Status	varchar(MAX)	☑
InterestedUser	varchar(50)	☑

Fig-1 Preprocessing selecting the fields

1.2 Pattern Discovery with Classification based on Association rules

Pattern discovery techniques used to extract knowledge from preprocessed data. Some of the techniques used in pattern discovery are Statistical Analysis, Association rules, Classification, Clustering, Sequential Patterns, and Associative Classification. Using the Asp.net application and Classification based on Predictive Association Rule Mining (CPAR) rule generated from the preprocessed data for identifying interested users and not interested users

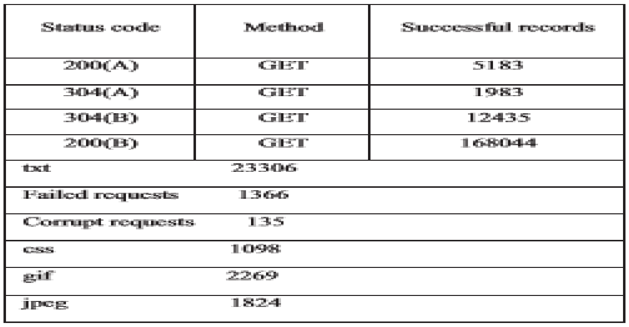
1.1 Pattern analysis from pattern discovery

Different pattern generated using CPAR method has been analyzed and patterns not interested users are removed. We discover knowledge which is set of rules for Interested Users.

2. RELATED WORK

2.1 Preprocessing: Cleaning the log files

With preprocessing the unused files such as image files or robot files is been removed and it shows the no. of entries that are present in an particular log file and it shows the total cleaned file as well as their size before cleaning and after cleaning.



Status code	Method	Successful records
200(A)	GET	5183
304(A)	GET	1983
304(B)	GET	12435
200(B)	GET	168044
txt	23306	
Failed requests	1366	
Corrupt requests	135	
css	1098	
gif	2269	
jpeg	1824	

Fig-2 Preprocessing: Data Cleaning [2]

2.2 Naïve Bayesian Algorithm

It works while performing the probabilistic prediction, i.e., predicts class membership probabilities.

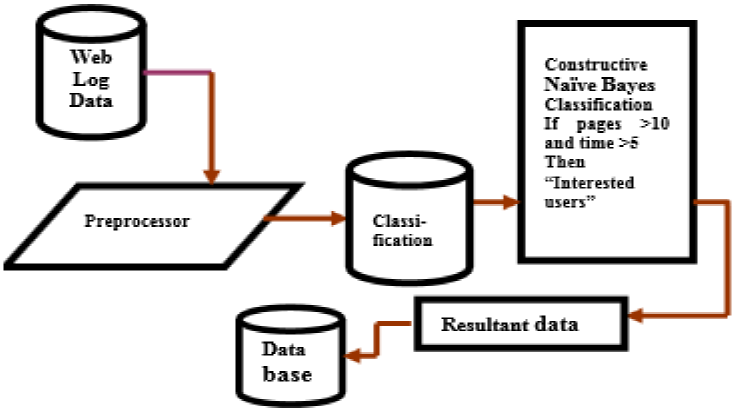

Web Log Data
Preprocessor
Classi-fication
Constructive Naïve Bayes Classification If pages >10 and time >5 Then "Interested users"
Resultant data
Data base

Fig -3 Naïve Bayesian algorithm [4]

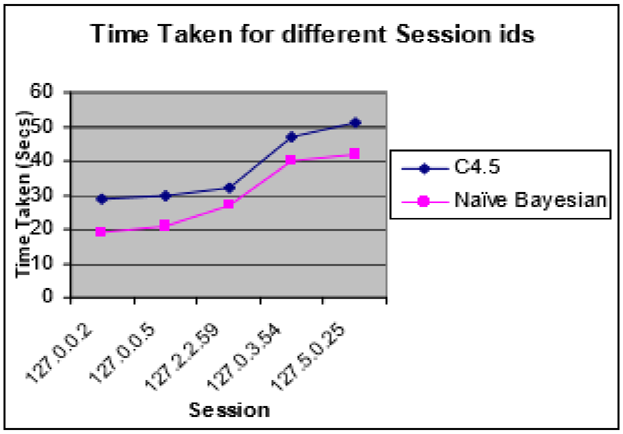
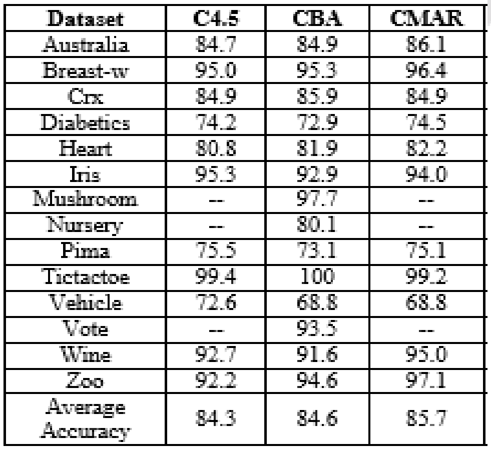


Fig -3 Naïve Bayesian Comparison [4]

2.3 Associative Classification

There are various classification based on association rules mining and done mostly done with c4.5, CBA, CMAR, CPAR and here with the real datasets they are compared and their accuracy are been measured.



Dataset	C4.5	CBA	CMAR
Australia	84.7	84.9	86.1
Breast-w	95.0	95.3	96.4
Crx	84.9	85.9	84.9
Diabetics	74.2	72.9	74.5
Heart	80.8	81.9	82.2
Iris	95.3	92.9	94.0
Mushroom	--	97.7	--
Nursery	--	80.1	--
Pima	75.5	73.1	75.1
Tictactoe	99.4	100	99.2
Vehicle	72.6	68.8	68.8
Vote	--	93.5	--
Wine	92.7	91.6	95.0
Zoo	92.2	94.6	97.1
Average Accuracy	84.3	84.6	85.7

Fig -4 Accuracy compared on real datasets [6]

3. PROPOSED WORK

With the preprocessing the data is been cleaned and the processing time is compared with the conventional approach as compared to novel approach it consumes more time and then the result are split into different set of rules are derived with the help of CPAR associative classification and then its accuracy are compared with the past approaches to distinguish the different set of rules.

4. RESULTS

With the help of different algorithm they are implemented by the means of programming language here the c# is used to derive and then with some modification with this final result of cleaning is done as before and after. And the same dataset are compared with naïve Bayesian and the CPAR as their results are as below.

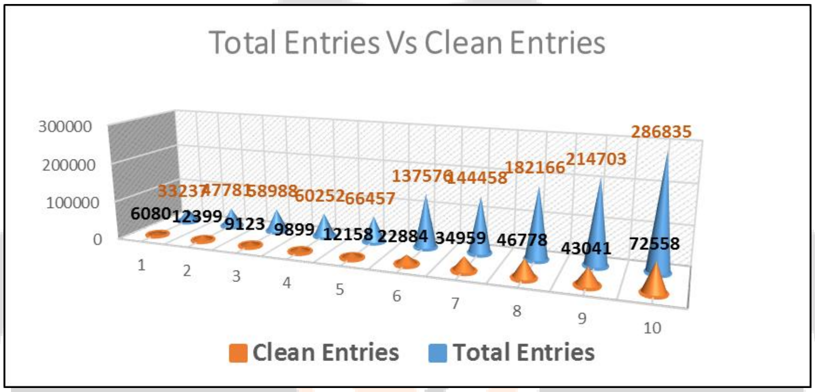


Chart -1: Total entries vs cleaned entries

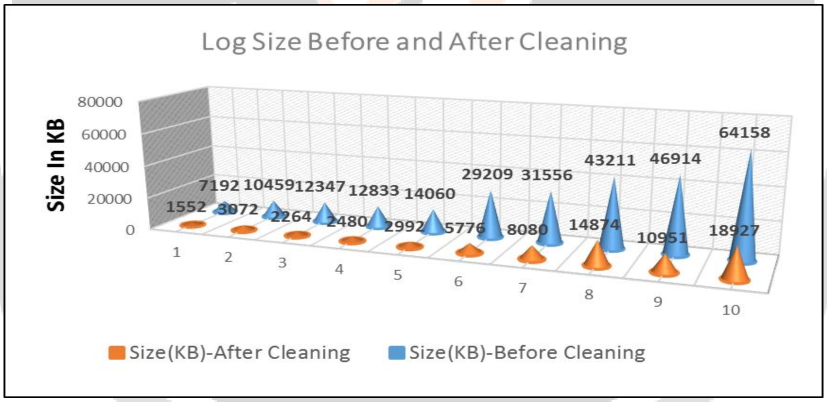


Chart -2: Log size before and after cleaning

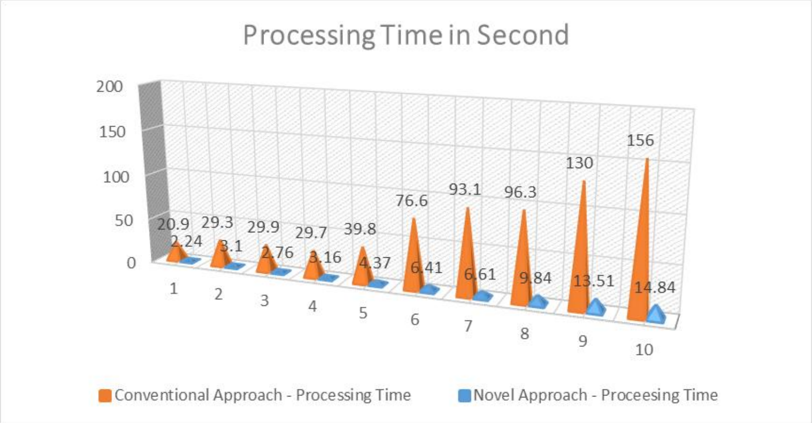


Chart -3: Comparison between conventional and novel approach

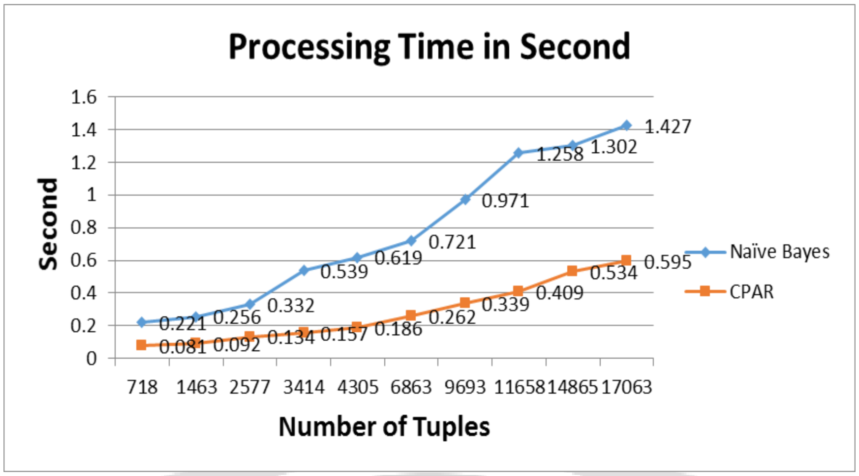


Chart -4: Comparison of processing time between naïve Bayesian and CPAR

4. CONCLUSIONS

Web usage mining is useful to extract out knowledge from the unstructured data. With the help of web log data the useful data can be sorted out and one can judge its popularity by deriving the interested and not interested ones. Classification based on Predictive Association Rule Mining (CPAR) is efficient compared to the Naive Bayesian Classification. From the experiments that are been conducted, many attributes are not used for classifying them as they are irrelevant. With the help of classification the number of irrelevant attributes can be reduced so that the performance can also be proved efficiently.

5. REFERENCES

[1] K. Sudheer Reddy, M. Kantha Reddy andV. Sitaramulu , "An effective Data Preprocessing method for Web Usage Mining", in IEEE , pp. 01-04, June. 2013.

[2] Theint Theint Aye , "Web Log Cleaning for Mining of Web Usage Patterns ", in IEEE, ISSN- 978-1-61284-840-2 ,pp. 490-494, 2011.

[3] Anshul Bhargav and Munish Bhargav "Pattern Discovery and Users Classification Through Web Usage Mining" ,IEEE International Conference on Control, Instrumentation, Communication and Computational Technologies,ISSN- 978-1-4799-4190-2, pp 632-636,2014.

[4] A. K. Santra and S. Jayasudha, "Classification of Web Log Data to Identify Interested Users Using Naïve Bayesian Classification " IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, ISSN-1694-0814, pp.381-387,Jan 2012.

[5] Bing Liu, Wynne Hsu and Yiming Ma, "Integrating Classification and Association Rule Mining" in AIII,ISSN-1694-0814 ,pp. 550-554, 2008.

[6] K prasanna Lakshmi , Dr.C.R.K.Reddy "Fast Rule-Based Prediction of Data Streams using Associative Classification Mining" IEEE 2015 International Conference on innovations Information Technology, ISBN 978-1-4673-6537-6 pp.29-34,Aug 2015.