ISSN: 2321-7782 (Online)
Volume 2, Issue 12, December 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

An Effective Approach for Pattern Discovery in Web Usage Mining

Kirti. V. Deshpande[1]
PG student
Department of computer Engineering
JSPM'S R.S.C.O.E, Tathwade
Pune, Maharshtra, India

Dr. A. B. Bagwan[2]
Professor & HOD
Department of computer Engineering
JSPM'S R.S.C.O.E, Tathwade
Pune, Maharashtra, India

Dr. P. K. Deshmukh[3]
Professor
Department of computer Engineering
JSPM'S R.S.C.O.E, Tathwade
Pune, Maharshtra, India

Abstract: ***The intended work is an attempt to apply an efficient web mining algorithm for web log analysis. This type of web mining allows for the collection of Web access information for Web pages. This usage data provides the paths leading to accessed Web pages. This information is often gathered automatically into access logs via the Web server. Due to tremendous use of web, web log files tend to grow faster resulting in noisy and confusing data files. It is essential to pre-process such a web log file before mining. It can discover the browsing patterns of user and some kind of correlations between the web pages. Web usage mining helps in the support for the framework of website developing, identifying useful content of site etc. Web mining applies the data mining to the web data and traces user's uniqueness, and then extracts the patterns according to user pattern. Our results based no: of data base scanning are reduced and also candidate set are found to be smaller as compared to basic apriori. So it is applied for web log analysis purpose any type user who is connected to internet.***

Keywords: candidate set, web log analysis, usage mining, apriori algorithm, navigation Pattern analysis

I. INTRODUCTION

With the explosive growth of information sources available on the World Wide Web and the rapidly increasing pace of adoption to Internet commerce, the Internet has evolved into a gold mine that contains or dynamically generates information that is beneficial to E-businesses. Web mining can be broadly defined as discovery and analysis of useful information from the World Wide Web. Based on the different emphasis and different ways to obtain information, web mining can be divided into two major parts: Web Contents Mining and Web Usage Mining. Web Contents Mining can be described as the automatic search and retrieval of information and resources available from millions of sites and on-line databases though search engines / web spiders. Web Usage Mining can be described as the discovery and analysis of user access patterns, through the mining of log files and associated data from a particular Web site.

Web usage Mining system should able to

» Collect useful usage data thoroughly.

» Scan and filter out useful usage data.

» Prepare actual usage data.

» Discover interesting patterns.

» Analyze and interpret navigation patterns correctly.

» Apply mining results effectively.

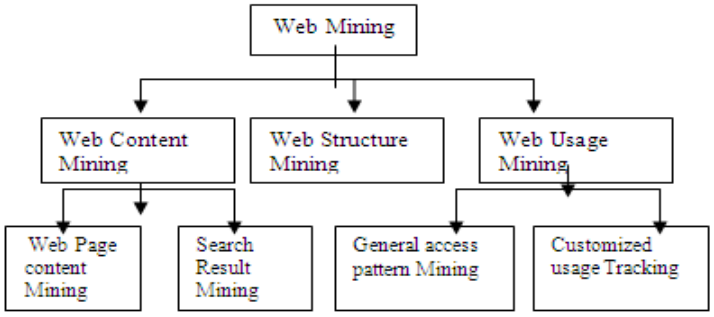


Fig 1.1 Hierarchy of Web Mining

Web usage Mining is used for predicting user behavior when he interact with WWW using different data mining techniques. When user hit the websites it generates different footprint in the different format at many places kike in the browser history option or etc. These traces are collected in appropriate way, but it may have some conflicts or noise then ,Some data mining algorithms applied directly on them so Five major steps followed in web usage mining are:

1. Data collection – Web log files, which keeps track of visits of all the visitors.
2. Data Integration – Integrate multiple log files into a single file.
3. Data preprocessing – Cleaning and structuring data to prepare for pattern extraction
4. Pattern extraction – Extracting interesting patterns.
5. Pattern analysis – Analyze the extracted pattern.

This work intends to show that the mining algorithm has lower complexity of time and space than Improved AprioriAll Algorithm and confirms the correctness of result obtained by providing a trace back route for candidate set pruning for the algorithms

II. LITERATURE SURVEY

A no: of algorithms have been invented over last decade for web log analysis according to client and server point of view. Apriori algorithm invented by Dunham in 2003 then some changes are allows to sorting the data according to user-id and timestamp which available in log data base related to each user worked with WWW.Basic difference between apriori and AprioriAll is that AprioriAll algorithm uses full join of candidate set and in case of Apriori forth join is used so AprioriAll is more used in usage mining.

Wang tong HE Pi-lian in their paper[1], showed that the possibility and importance about applying Data Mining in Web log mining and showed some problems in the conventional searching engines. Then it offers an improved algorithm based on the original AprioriAll algorithm, which has been used in Weblogs mining widely. Test results show the improved algorithm has a lower complexity of time and space.

Gui-Rong Xue et al [2] proposes a novel re-ranking method based on user logs within websites. With the help of website taxonomy, they mine for generalized association rules and abstract access patterns of different levels. Mike Thelwall in [3] stated that web log files are a important input for predicting user behavior and to some extent demographics.

Zhenglu Yang et al. [4] consist an effective web log-mining system consists of data preprocessing, sequential pattern mining and visualization Magdalini Eirinaki and Michalis Vazirgiannis [5] give a survey of the use of web mining for web

personalization Yongjian Fu and Ming-Yi Shih [6]they stated a framework to mine Web usage data on client side as a complement to the server side Web usage mining. sandip sing Rawat and Lakshmi Rajmani [7] proposed that custom built Apriori algorithm which analyzed on educational log file.Its rules are helps in decision making for developing the website.

Jianli Duan,, Shuxia Liu[8] invented a Web mining tool which is make easy to analyzing web log files and user patterns on the other side some limitation in it i.e user actions cannot simulate exactly.

G. Sudhamathy, C. Jothi Venkateswaran [11]The HFPA approach is compared with the traditional FPA approach and found that the new HFPA approach outperformed the existing FPA approach in many aspects like run time, memory usage, rules pruned, rules produced and accuracy percentage. There are scopes for future work in this proposal, like applying the same for different web sites to confirm the results and other interestingness measures can be explored to see if they give better results.

III. Implementation Details

a) System Design:

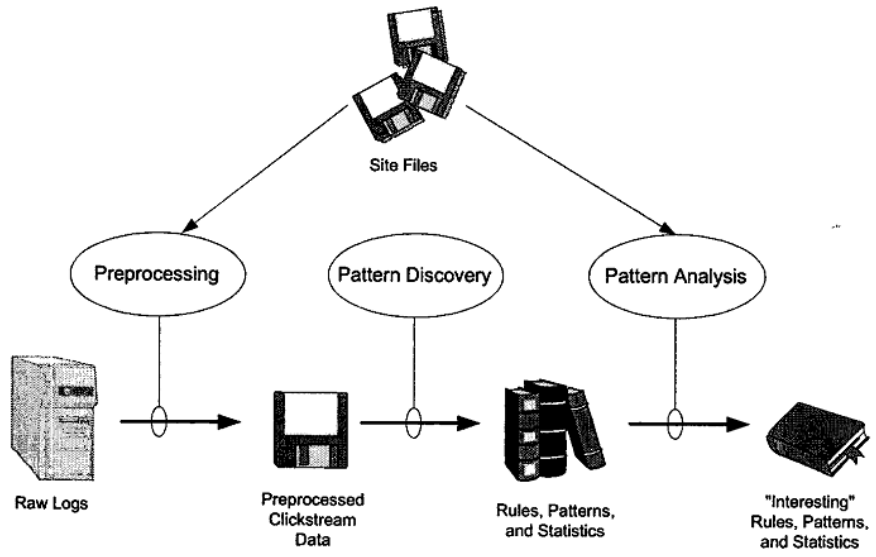


Fig.1.2 High level Web usage Mining

b) Approach for Web Mining

In Web Log Mining basic and most useful source is Log files which is available in different formats like web server log files, client side log files,acess logs, agent logs. This web log data is created automatically by web server when it service user reuests,when it contains all information about web visitors activity.

Data Preprocessing

Data preparation is very crucial and time consuming task because remaining two steps is totally depending upon cleaned data so if data is more qualitative then better results are achieved .data processing again having subtasks like data cleaning and feature selction,user identification, session identification and etc .

Pattern discovery

The pattern discovery has three main operations of interest: association (i.e. which pages to be accessed together), clustering (i.e. finding groups of users, transactions, pages, etc.), and sequential analysis (the order in which web pages tend to be accessed).

Pattern Analysis

Pattern analysis the motivation is to filter out uninteresting rules or patterns found in the previous phase. Visualization techniques are useful to help application domains expert analyze the discovered patterns.

c) Mathematical Model

1. $0<T_s<=30$ min

whereT_s=time-stamp of session

2. $D=\{ds_1,ds_2,ds_3,\ldots ds_k\}$

Where D is set of webpage's hited on different IP address.

ds1={U}

where U is Data set of WebPages on each IP Address

$U=\{U_1,U_2,U_3\ldots U_k\}$,$U1=\{url_1,url_2,url_3\ldots.url_m\}$

Where U1=Data set of WebPages hited by user1 to..user m on particular IP address.

3. $X[i]=U_i$ store all web pages set in array I >=0

4. C=0;MC=0;

5. calculate $\int_{i=0}^{n} G(x)$

G(x)= {

a. Calculate $\int_{j=0}^{n-1} f(x)$

b. f(x)= $|c++ \quad substring(X[i],X[j]==1|$

Y[i]=C

c. if Max<=C

Max=C

}

6. compute L=frequently occurred substring.

d) Algorithm

Step1:Decide session time which is less than or equal to 30 min.

Step2:Arrange the web page set which is available on different IP addresses in increasing order.

Step3:store all web page sets from IP adresswise in string array X.

Step4:initialize count=0,Max=0;

Step5:for (i=0 to n)

{

For(j=0 to n-1)

{

If substring(x[i],x[j])

Increment the count by1

End if

Y[i]=count;

}

If Max <=count

Max=count;

End if

}

Step6: determine all places in array Y where value is equal to Max and select corresponding substring from X.

Step7: produce output of all substring with their position which is intended output.

IV. INPUT DATASET

A. Input Datasets

Input={dS1dS2,…..dSk}//database of session collected from each IP address.

DS={i_1,i_2,i_3…i_n} //database of item set

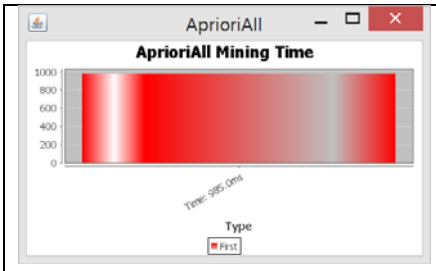
Output: sequential data pattern D'

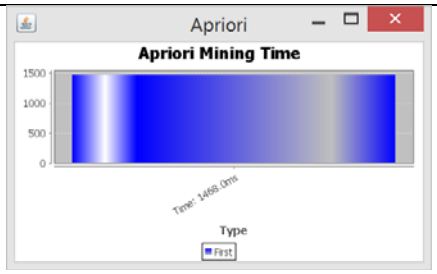
D'=sort D according to IP address and timestamp of first reference page on each IP address for all clients. Find L1 in D'

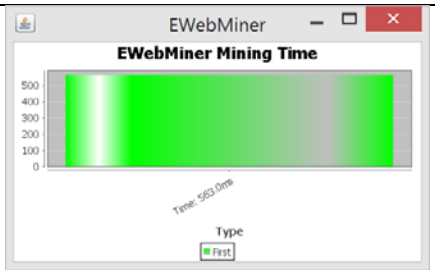
L=WM(D',DS,L1);

Find Maximal reference sequence from L.

B. Results.

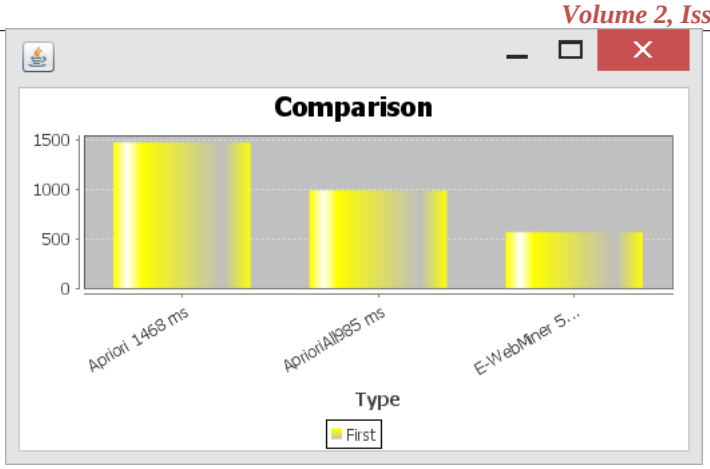






The graphical representation of the system is convincing enough to prove that the New Web Mining algorithm stands much better than apriori algorithm. The graph is drawn as no: of transaction against time .In the graph on X- axis time factor is placed and on Y-axis No: of transactions are placed.

Comparative analyis of new web mining algorithm with previous techniques are shown with the help of following graph.



V. Conclusion

The main aim is to reduce the number of elements in every candidate set and repetitive scanning of database. This work intends to show that the mining algorithm has lower complexity of time and space than Apriori Algorithm. The study is performed on only limited no: of data. If more number of web log record increases CPU performance get affected and database server required more CPU time. This may provide further refinement in the result of candidate set pruning.

Acknowledgement

We take the opportunity to thank Dr. A.B.Bagwan, our Head of Computer Engineering Department, Dr.P.K.Deshmukh our PG Coordinator, all the teaching and non teaching staff of Computer engineering department for their encouragement, support and untiring cooperation.

References

1. Tong, Wang and Pi-lian , He, Web Log Mining by an Improved AprioriAll Algorithm World Academy of Science, Engineering and Technology, Vol 4 2005 pp 97-100
2. Gui-Rong Xue1, Hua-Jun Zeng, Zheng Chen, Wei-Ying Ma and Chao-Jun Lu, “Log Mining to Improve the Performance of Site Search” Computer Science and Engineering Shanghai Jiao-Tong University, Shanghai 200030, P.R.China.
3. Mike Thelwall, ““Web Log File Analysis: Backlinks and Queries” School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton, WV11SB, UK.
4. Zhenglu Yang, Yitong Wang and Masaru Kitsuregawa, “An Effective System for Mining Web Log”,Institute of Industrial Science, The University of Tokyo 4-6-1, 153-8305, Japan.
5.]Magdalini Eirinaki, Michalis Vazirgiannis, “Web Mining for Web Personalization”, Department of Informatics Athens University of Economics and Business Patision 76, Athens, 10434, GREECE, (c) ACM, [2003], Vol. 3, No. 1, February 2003.
6. Yongjian Fu, Ming-Yi Shih published, “A Framework for Personal Web Usage Mining”, Department of Computer Science Department of Computer Science, University of Missouri-Rolla University of Missouri-Rolla Rolla, MO 65409-0350.
7. Sandeep Singh Rawat1 and Lakshmi Rajamani”Discovering Potential User Behaviors Using Custom Built Apriori Algorithm” Department of Computer Science & Engineering,Guru Nanak Institute of Technology,
8. Dunham., Margaret H., Data Mining Introductory and Advanced Topics. Beijing:Tsinghua University Press, 2003, p195-220
9. Han Jiawei and Kamber Micheline Data Mining Concepts and Techniques[M].Beijing: China Machine Press, 2001, p290-297.
10. Jianli Duan,, Shuxia Liu, “Research on web log mining analysis”, School of Science Qingdao Technological University Qingdao, China, International symposium on Instrumentation and Measurement, sensor Network & Automation(IMSNA)2012.
11. G. Sudhamathy , C. Jothi Venkateswaran “An Efficient Hierarchical Frequent Pattern Analysis Approach for Web Usage Mining” International Journal of Computer Applications (0975– 8887)Volume 43,No.1, April 2012.

