Review of an Efficient and Optimized Web Mining Technique for Web Log Analysis

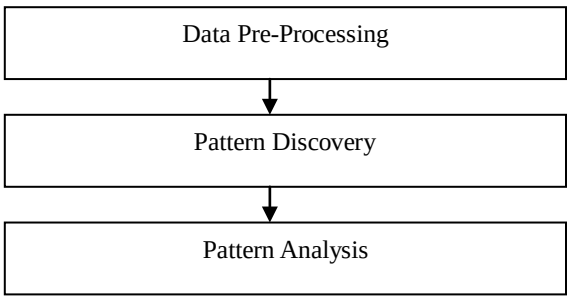
Mr.Nishant Barsainyan, Mr. Anurag Jain

Abstract— In this paper we reviewed an efficient and optimized algorithm for the web mining for weblog analysis. We are approaching to design a system whose results may applied to the relevant class of issues from search engine, this is to identify the concerned contexts which are directly associated with the securities and marketing of the e-commerce websites it also helpful for the designing of the e-commerce web portals. The concept of the paper is incepted from the Apriori Algorithm and then study various implementation of Apriori algorithm in weblog analysis and page ranking and we compare all the earlier versions of Apriori algorithm. we also study and review the performance of the algorithms implemented for the web mining for weblog analysis and study the time and space complexity with this study we are going to design an technique which can reduce the reduce the data base for the scanning and in result the smaller be the candidate sets in stage wise comparison. We are also trying to make a technique which is much generalized and can be implemented in any type of web log analysis.

Index Terms— Web Mining, candidate sets, framing, Improved Aprori algorithm, Web Miner Algorithm, Web Log Analysis, Pre Processing, Data Mining.

1 INTRODUCTION

The World Wide Web (WWW) is one of the most important medium that provides an interface to store, share and distribute information. At present, the figure for Google is index of 8 billion Web pages [1]. The extreme use of the Web has provided an opportunity to study user and system behavior by exploring Web access logs. Web mining is a suitable technique to discover and extract interesting knowledge/patterns from Web. The information gathered can be classified into three broad categories of web mining namely: Web structure mining, Web content mining and Web usage mining. Web Structure mining focuses on improvement in structural design of a website. Web content mining focus on the contents of the webpage and the web usage mining is concerned with the knowledge discovery of usage of websites by an individual or a group of individuals. The web usage mining process can be broken down in three steps as shown in the figure below:



2.1 About Web Log

Logs are basically the collection of the data so whenever we talk about the log of something then it means the details in a very crisp format in a very frequent manner which will trace the each and every step of the any process an these logs will be used in the future for the analysis of the processes concerned similarly when we talk about the Web Logs then it remind us about the log file which contains the web server access in the each and every link and in each and every click web logs provides us the information's that will be use full to generate report for the analysis of the user behavior and the access track of the web of the user. Normally the log format is known as the CLF Common Log format which is not very rich in nature shown in the

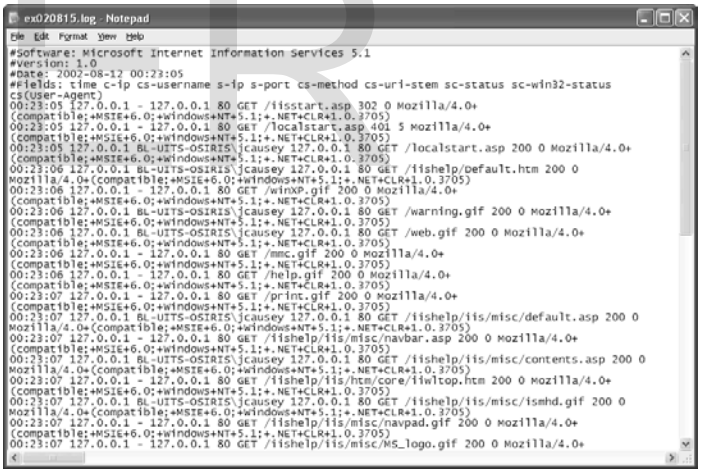


figure 1.1 and the Fig 1.2 shows the sample Detailed Log file.

Fig 1.1

date time	10/6/2014 8:06
c-ip	172.16.241.98
cs-username	Nishant
s-ip	192.168.136.35
s-port	80
cs-method	GET
cs-uri-stem	/html/Index.asp
cs-uri-query	ad=banners
sc-status	200
sc-bytes	20313

cs-bytes	544
time-taken	0
cs-host	www.xyz.com
cs(User-Agent)	(compati-ble;+MSIE+6.0;+Windows+NT+5.1 + Mozilla/4.0)
cs(Referer)	http://www.bing.com

Fig 1.2

The retrieval of the information for the analysis of the data is known as the Web data mining and the analysis is known as the web analytics the web log mining can provide us the significance in the following aspects

- Web Personalization
- System Improvement
- Website structure design improvement
- Business Decision Making Support
- Optimization of Search Engine

Now the main thing is to understand how these server logs generated generally whenever any use access any webpage then the request is generated and sent from the client computer to the Server
Computer where the targeted website is stored these targeted files webpage files that is any HTML file, Image File, or any script file and these request saved in the servers side as the desired format and this format is known as the web log. Most Commonly used log formats are

- Common Log Format (CLF)
- Extended Common Log Format (ECLF)
- W3C Extended Log File ExLF. (Extended Log File Format)

3 Web Analysis

Interpreting these weblogs is known as web analysis or web analytics. web analytics may be defined as the collection of the data which is further measured or analyzed behavior of it and generates the reports as we need to understand the web usage pattern this will shares some common methodology and theoretical characteristics in all types of logs like fig 3.1 shows the strategic flow of web analytics.

1) intranet Logs

2) system logs

3) OPAC Logs

4) Search Logs

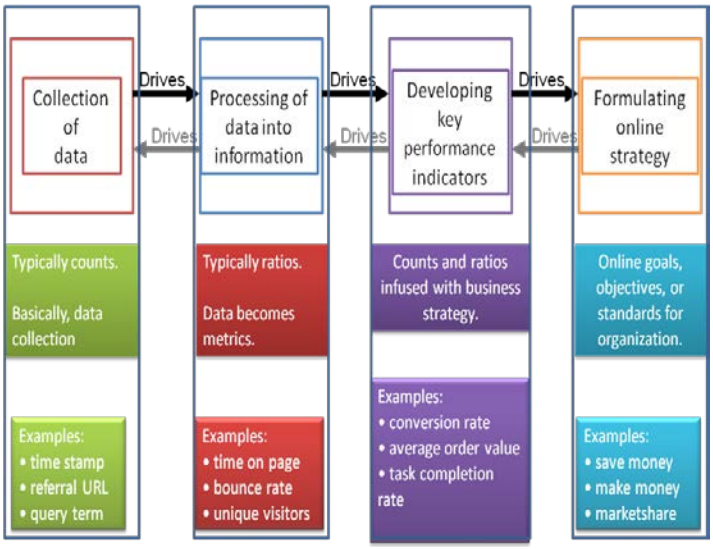


Fig 3.1

3.1 Type of Web Analysis

Basically when we categories the web analytics then it will be break in main two categories that is

1) On Site Type
2) Off Site Type

Offsite type of Web analytics: - It basically includes the Calculation of web pages potential of audience i.e. (Opportunity), share of Visibility (Voice), and Comments (buzz) these all are running happenings on the world of Internet.

Onsite type of Web Analytics: - On site type web analytics mainly calculates the presentation rate of the website in a techno-commercial context. This generated data is then compared with the "key performance indicators" for rating the performance of the website, in the web analysis we analyze or measuring the website date to generates the data the factors on which we generally measures the data is known as the metrics these metrics are divided in the following types

1) Count
2) Ratio
3) KPI (Key Performance Indicator)
4) Dimension

- *Nishant Barsainyan is currently pursuing masters degree program in Computer Science Engineering in RGPV University, MP,. E-mail: nishi033@rediffmail.com*
- *Anurag Jain is currently HOD of Computer Science Deptt., RITS College,RGPV University. E-mailanurag.akjain@gmail.com*

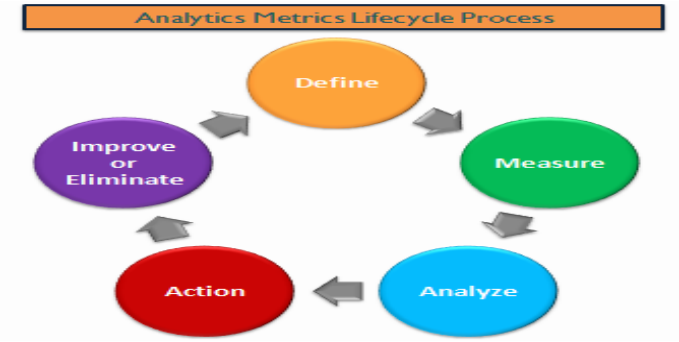


Fig 3.2

These metrics in the web analytics for the analysis on the three levels of the granularity these three levels are

1) Aggregate :- it defined as traffic on site for the particular duration of time
2) Segmented :- it is defined as monitoring of a small portion of traffic for the particular duration of tome
3) Individual:- it is defined as the monitoring of the user for the particular duration of time

Classification of Metrics

1) Building Block – Metrics of Foundation
2) Visit Characterization – it is defined as the calculation of visits, it either of the aggregated or single
3) Content Characterization – it is understanding of the content and its use
4) Conversion – calculations of links of visitors or contents

4 Web Minning Algorithm

When we fetch and analyze the discovered data, or important information from the World Wide Web i.e. www. It is said to be a web mining and it is basically categorize in three important parts that are,

1) web content mining
2) web structure mining
3) web usage mining

Each of the above types can work or analyze the following points

1) Records of weblog
2) structure of hyperlink\
3) documents of text and/or multimedia
4) usage mining of web
5) structure mining of web
6) content mining of web
7) web mining

Different types of Algorithms

1) Association Rule Mining Algorithms
2) Clustering Algorithms
3) Classification
4) Sequential Patterns

5 Literature Survey

The paper [1] is trying to generate a competent web mining algorithm for analysis of web log. In the attempt results generated from the analysis adapted for problems of the search engine in the classes to find the context to develop the demanding secured web apps of e-commerce on the basis of the association. As per the Paper [1] application they are successfully applied them in any web log analysis , including information centric network design[1] .Paper [1] take into consideration the support and the confidence of any sequential pattern of web pages of users. This may provide further refinement in the result of candidate set pruning.

Paper [3] is also the attempt to dig in the field of the web mining by adopting the another technique i.e. clustering for the weblog analysis basically this paper explains the clustering and its usage in the data mining and clustering algorithms they provide the solution for the data clusters for nominal and the numerical data in the Paper[3] explains the differences in the approach of the Clustering Algorithm and the Apriori algorithm how they are different and how the Apriori algorithm is not use full for the nominal and the numerical data. Paper [3] is also attempts to explain how the average value calculation is easier fast and faster than the traditional approach of the log values calculation. Paper [3] represents the improved technique in the field of web Usage Mining by discovers the log files for the users at single place.

Approach of the paper [4] is very different from the others; Paper [4] tries to work on the personalization of the web content by using the data mining techniques and web log analysis. Paper [4] works on the usage of the internet users and on the basis of the usage it generates the pattern and classifies them in the groups or classes and maintains the cache for them. “Term set analysis and direction cosine distance approach” is used in the searching of the data to generate the desired set on the basis of the query. So the approach of the Paper [4] is different from the traditional log analysis since it approach is very users perspective by generating personalized content results.

Paper [5] approaches to the study of the Web usage mining technique in the much elaborated manner for the weblog analysis. The different approaches and the patterns in the web log mining are the Page Sets, Page Sequence and Page Graphs. Paper [5] gives us the use full information and tells how to use the web usage mining techniques to

discover the patterns and works on the weblog data to understand the navigation behavior of the user and obtains the useful information.

Paper [6] focuses more on the web structure mining as compared to the web usage mining they concerns to the page rank and the weighted page rank to propose a new algorithm moreover on the basis of this they are also focus on the study and the analysis on the basis of the topics. Paper [6] is also focuses on the issues that the search or queries are done by topics so if the pages being ranked on the basis of the topic then it will better respond to the users query and the search.

Paper [7] is trying to represent the different approach by adapting the "Markov Predictors " to design the algorithm Paper [7] focuses more on the how to expand the candidates so that we can get the much improved and efficient result as in the others for achieving this they design the new algorithm known as the WM0 by adapting this algorithm they just generalized the existing algorithm to achieve the targeted results and by adapting this WM0 the can smartly improves the performance of the web and designs the system.

Paper [8] is the very focused paper for the working in the field of the WB Log Mining in this paper [8] we found that how the "prefix Span" and there algorithm "LAPIN WEB : Last Position Induction for Web

Log" works using the concept of the Web Log Mining System i.e. data Processing, Sequential Pattern Mining and visualization. In the proposed system of the paper [8] they works on the sequential mining algorithm to find the access patterns of the users from the web logs this analysis were finaly visually shown to the user as per the query .

6 PROPOSED METHODOLOGY

Mining of web usage is Process for integrating no of stages involved in the data mining it includes the following stages

1) pre processing of web log
2) Pattern Discovery
3) Analysis of Patterns

For any technique of web mining, first of all the making of appropriate source of data is an very important step since the characteristic web log from sources are distributed and complex in structure. Before we applying any techniques of mining to web log data, cleansing of data collected from resource, then integrated after that transformed. in the modern days we use the tools which are very intelligent in itself they help web miner to extract resource data then filter to evaluate smartly to generate required information for modification. For this task it is required to differentiate the accesses created by physical users and web based search engines. Later on to process the data of weblog we can try to apply any or all techniques of web. Now the final step of usage mining of web is analysis of pattern which discovers the Patterns then find the unique patterns of interest. The proposed concept will use FIESP (Fast Intelligent & Efficient System of Pre-processing) for generating efficient results. Base paper has provide the concept of E-web miner algorithm which calculate following parameters

- Number of Transactions
- Number of Items
- Time

Following parameters will be calculated in proposed system by which we increase the efficiency of the system

- Total no of pages which retrieved in a session of web.
- Total no of image pages which retrieved in a session of web.
- Total of time that spent by visitor.
- The same webpage requested twice or more than then once in a session of web.
- Requests in Percentage.

After consideration of the all the above mentioned points we will design the proposed methodology in such an efficient manner that make compression in a existing methodology and generates results in much less time.

Fast and Efficient System of Pre-processing: FESP

The main objective of FESP system is to divide the physical user and Web based search engine accesses. This system gets the raw data of web log as an input and removes the access of search engine automatically within less time. To analyze behavior of user browsing we must remove the accesses done by web based search engines. After removing the access of search engine from data of web logs, the left out data in web log is adapted as physical accesses. This physical user access data acquires cleansing as a pre processing, identification of user, session identification of path completion. This system can be categorized broadly two types

1) Server side intelligent systems
2) Client side intelligent systems.

In preparation of web data mostly use client side systems. These systems are also called session initialization based heuristic functions.
The FIESP is the collective architecture, which concentrate on very critically on pre- processing of weblog data. The main advantages with FESP are.

- Improvement in the efficiency of web log pre processing.
- It separates Physical user and accesses web based search engine automatically, in very less time.
- Rate of error of learning algorithm is also reduces.

7 CONCLUSION

This presents review that we are going to make an automatic web log information mining by Web Page Collection algorithm and is been proved to be more effective as compare to the other algorithms. It stands above other web mining algorithms. With the mined results, the web applications are developed and provides adaptive user interface. We are also planning graphical presentation of the output which is convincing enough to prove that the planned algorithm stands much above the other web mining algorithm - a sample of which is represented By New Improved Apriori All algorithm

REFERENCES

[1] Mahendra Pratap Yadav,Pankaj Kumar Keserwani and Shefalika Ghosh Samaddar "An Efficient Web Mining Algorithm for Web Log Analysis: E-Web Miner" 1st Int'l Conf. on Recent Advances in Information Technology | RAIT-2012 |

[2] R.Shanthi1, Dr.S.P.Rajagopalan2 "An Efficient Web Mining Algorithm To Mine Web Log Information" International Journal of Innovative Research in Computer and Communication Engineering Vol. 1, Issue 7, September 2013

[3] Pooja Sharma and Asst. Prof. Rupali Bhartiya "An Efficient Algorithm for Improved Web Usage Mining" Pooja Sharma et al ,Int.J.Computer Technology & Applications,Vol 3 (2), 766-769, ISSN:2229-6093

[4] Indrajit Mukherjee, V. Bhattacharya and Samudra Banerjee, Pradeep Kumar Gupta "Efficient Web Information Retrieval based on U sage Mining"].1 Infl Conf. on Recent Advances in Information Technology I RAIT-20121

[5] Renáta Iváncsy, István Vajk "Frequent Pattern Mining in Web Log Data" Acta Polytechnica Hungarica Vol. 3, No. 1, 2006

[6] Shesh Narayan Mishra*,Alka Jaiswal,Asha Ambhaikar "An Effective Algorithm for Web Mining Based on Topic Sensitive Link Analysis" International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 4, April 2012 ISSN: 2277 128X

[7] Alexandros Nanopoulos, Dimitrios Katsaros, and Yannis Manolopoulos "A Data Mining Algorithm for Generalized Web Prefetching" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 15, NO. 5, SEPTEMBER/OCTOBER 2003

[8]] Zhenglu Yang, Yitong Wang and Masaru Kitsuregawa An Effective System for Mining Web Log

[9] Priyanka Patil1 and Ujwala Patil2 Preprocessing of web server log file for web mining World Journal of Science and Technology 2012, 2(3):14-18

[10] Edith Cohen_, Mayur Datar, Shinji Fujiwara, Aristides Gionis, Piotr Indyk, Rajeev Motwani, Jeffrey D. Ullman, Cheng Yang Finding Interesting Associations without Support Pruning Knowledge and Data Engineering, IEEE Transactions on (Volume:13 , Issue: 1) Jan/Feb 2001

[11] L. Hubert and P. Arabie, "Comparing Partitions," *J. Classification,* vol. 2, no. 4, pp. 193-218, Apr. 1985. (Journal or magazine citation)

[12] Rahul Neve, K.P Adhiya, Comparative Study of Web Mining Algorithms for Web Page Prediction in Recommendation System International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 1, January 2013

[13] WEN-HAI GAO RESEARCH ON CLIENT BEHAVIOR PATTERN RECOGNITION SYSTEM BASED ON WEB LOG MINING Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, Qingdao, 11-14 July 2010

[14]] Sathya Babu Korra, Saroj Kumar Panigrahy, Sanjay Kumar Jena, Web Usage Mining: An Implementation View Springer Berlin Heidelberg Advances in Computing, Communication and Control Communications in Computer and Information Science Volume 125, 2011

[15] Arumugam, P. and Christy Advanced Web Usage Mining Algorithm using Neural Network and Principal Component Analysis Christy V et al , International Journal of Computer Science & Communication Networks,Vol 3(3),168-172

[16] S.Veeramalai, N.Jaisankar, A.Kannan Efficient Web Log Mining Using Enhanced Apriori Algorithm with Hash Tree and Fuzzy International journal of computer science & information Technology (IJCSIT) Vol.2, No.4, August 2010