International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org, editorijettcs@gmail.com
Volume 4, Issue 3, May-June 2015 ISSN 2278-6856

Improved Log Miner for Frequent Items Generation

Rajdeep Marathe[1], Mrs. Dhanashree Phalke[2]

[1]Department of Computer Engineering
D Y Patil COE, Akurdi Pune, India

[2]Mrs. Dhanashree Phalke
Department of Computer Engineering
D Y Patil COE, Akurdi, Pune, India

Abstract

Web log mining is the newest technology of data mining. There are various web related activities that are taken into consideration. Such data are mostly structured in nature as they are collected from various web pages and other web logs that are maintained in the server. Web Mining is divided into three types web content mining, web usage mining and web structure mining. In case of Web usage minings, the main aim and area is to focus on Web users and to learn the way they interacts with various Web sites available. As web log data are mostly noisy and extremely ambiguous, still there is a way where we can discover useful information and structure in the way the users interacts with a web site. The main objective of using mining is to quickly and automatically identify users from the vast log data. We can identify information such as frequent access paths, frequent access page groups and then cluster the users. With the help of web usage mining algorithms, the web application server logs, registration information, the user interest and other data such as user access patterns can be mined which will be helpful in laying foundation for decision making of organizations.

Keywords: Web Mining, candidate sets, framing, Improved Apriori algorithm, AprioriAll algorithm, E-Web Miner algorithm, Web Log Analysis

1. INTRODUCTION

Due to the rapid development and expansion of the World Wide Web, the increase in use of new automated Web-mining techniques to discover useful, relevant information has become an increasingly important research area. Depending upon the hits the website gets or on its popularity, a Web Application server can generate web logs which can hold records in thousands or tens of thousands depending upon the requests it gets every day. Web logs record are found mainly on the web server. Whenever user surfs the internet a request is send and recorded in a log and a entry is made, which contains various types of information, including the IP address of the computer from which the request is been made, date and time the user accessed the document or any content like image video, audio or any link, and so on.

A session is defined as a series of requests that are made by a any user for a single or particular navigation purpose. Depending upon the number of times the user accesses the website during a period of time, the session are classified as single session or multiple sessions. There is a need to find useful patterns (such as association rules or sequential patterns) from this vast amount of information. The various requests (or log entries) that come to the website need to be grouped into usage sessions.

Once these sessions are identified then the common usage patterns among sessions can be discovered by Web usage mining algorithms. The process of finding the patterns may involve pre-processing, in which the original data is integrating from multiple sources, and then transforming the integrated data into a form suitable for input which is feasible for the algorithm to perform the mining operation [1]. The data that is brought from multiple sources has to be initially collected and then the preprocessing part is been done. Preprocessing include data cleaning, pageview, user, session identification, and clickstream data with other data sources [2]. Data Cleaning or preprocessing is very important stage of data

mining technique [3]. Log which is been brought from various sources is noisy and to increase the efficiency of algorithm preprocessing is done using various algorithms such as Field extraction algorithm and data cleaning algorithm [4]. Log files provide access patterns of the users, the typical behavior of users (profiles), the operating system being used, error handling invoked and the time period of web usage for a particular successful/ unsuccessful transaction [4]. All these information that are tabulated in a predefined format; for example, a log file of Microsoft Internet Information Server(IIS)5.0 having a format of W3C extension norm. Analysis of these web log files give out the result depending upon the perspective chosen; from the client point of view or the server point of view. Server side web log analysis reveals information about availability of this servers, vulnerability of servers, security loop holes of servers, user unfriendliness of the web site etc., and a web site designer can gain a lot from web log analysis for the desired improvement of the services and web site design. The client are better served if a web log analysis for clients reveals information about frequency of the usage

of particular web page etc. for performing pre fetching and caching of pages.

This paper intends to show that the Proposed Hybrid Algorithm generated the result in short span as compared to others and confirms the correct result that are obtained.

2. RELATED WORK

A number of web mining algorithms have been evolved over the last decade to cater various clients and server side needs.

A. The AprioriAll Algorithm

The algorithm happens to be a modification of Apriori Algorithm. The modification allows to put the data in correct order by using UserID and time-stamp sort. The main difference that differentiates between AprioriAll and Apriori is that AprioriAll uses full join in order to generate candidate sets. If Apriori is considered, it is only forth joined. Hence, as compared AprioriAll is more appropriate than Apriori for web usage mining. Tong and Pi-lian, et.al suggested the new Improved AprioriAll Algorithm by saying that there is sizeable reduction in the size of candidate sets [5]. The number of scanning database is reduced when generating the large set.

B. Social Recommendation

Recently, as the explosive growth of Web 2.0 applications, social-based applications gain lots of traffics on the Web. Social recommendation, which produces recommendations by incorporating users' social network information, is becoming to be an indispensable feature for the next generation of Web applications. Overall, the concept of a social network is quite simple and can be described as a definite set of individuals, by sociologists called actors, who are the nodes of the network, and ties that are the linkages between them. In other words, social network indicates the ways in which actors are related.

C. Cookie Picker

Cookies especially HTTP cookies, are typically used for recording session state, personalization, authenticating and tracking user behaviors. Cookies keep all the important information about the user. Cookies can be also be very harmful as they are exploited by web site to track and build user profiles exploiting users privacy and violating security of communication. Hence there was a need for a cookie management scheme. Yue, Xie and Ways introduced Cookie Picker, a system that can automatically validate the usefulness of cookies from a web site. It could also set the cookies usage permission on behalf of users [5].

D. E-Web Miner

E-Web Miner is the web mining algorithm that removes the flaws of Improved AprioriAll algorithm and improve upon the time complexity of the earlier AprioriAll algorithm. It provides an improved candidate set pruning as well. It has been shown successfully that it mines correct result of candidate set where as the Improved AprioriAll algorithm fails to deliver the correct result [5]

E. Collaborative Filtering

Two types of collaborative filtering approaches are widely studied: neighborhood based and model-based. The neighborhood-based approaches are the most popular prediction methods and are widely adopted in commercial collaborative filtering systems [10]. The most analyzed examples of neighborhood-based collaborative filtering include user-based approaches and item-based approaches [10]. User-based approaches predict the ratings of active users based on the ratings of their similar users, and item-based approaches predict the ratings of active users based on the computed information of items similar to those chosen by the active user. In the model-based approaches, training data sets are used to train a predefined model.

3. IMPLEMENTATION DETAILS

Now a days the use of internet is increasing rapidly. The user use the internet in order to do most of their work. There is a very vast amount of information that is available on internet. The current web is very large and unorganized. Whenever user logs on the internet and search for any data on websites. he is been tracked.

Whatever he does is stored in the log of server. The each and every actions that are performed by the user are stored with all the details such as IP address, Log on time, Items accessed, date, etc.

Product discovery is an important step in users interaction with an E-commerce website. Log mining play a very useful part in the discovery of products. As it is difficult for users to express their intent through well formed queries with increasing size of the catalog.

3.1System Overview

Collect the log file from the server. Preprocessing is done on the log file. Give the log file input to the three algorithms. Each algorithm mines and generates frequent items. Showing the results based on the input given ie. log file. Modify the system or E-Commerce website as per the results are produces for enhancement. Clearly the proposed hybrid Algorithm shows correct and fast results as compared to the other two algorithm.

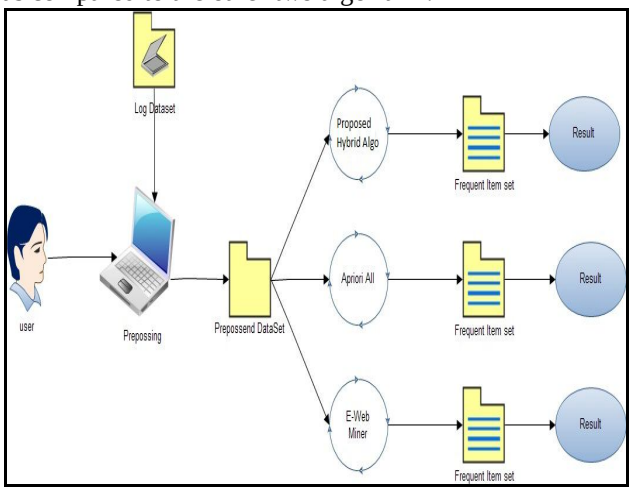


Figure 1: Architecture of the proposed system

In the above figure, the User who wants to extract the frequent patterns uses the algorithm. He first takes the logs in this case the logs could be any standard logs been provided from Web application Servers. In this scenario we have used as standard data set from NASA. The logs undergo preprocessing . In the preprocessing part the logs are been cleaned, where the unique URL's are added as unique numbering. So if once the URL is repeated it gets the same number as its original one. And then the log file is been passed to the respective algorithm to generate frequent pattern.

The ultimate aim of all the algorithm is to generate the frequent pattern. The webminer algorithm works on Minimum support and confidence. where we need to manually add the parameters. E-Webminer algorithm works on MAX count. Its scans the database and finds out the MAX count and then compares the result with other. The proposed Hybrid Algorithm works in two phases, one half part of FP growth Algorithm and the results are passed to Apriori Algorithm. The first half is mainly generation of a tree and then the generated results are passed to Apriori algorithm for further processing.

3.2 ALOGITHM USED

Algorithm: Proposed Hybrid Algorithm Pass 1:

1: Scan the data(logs) and support for each item.
2: Then discard infrequent items.
3: Sort the frequent items in decreasing order based on their support.
4: Use this order when building the FP-Tree, so common prexes can be shared
5: This first part is mainly used for tree generation and the
output is given to next algorithm.

Algorithm 2 Proposed Hybrid Algorithm

1: State itemsets in Lk-1
2: Join Lk-1p with Lk-1q, as follows:
3: insert into Ck
4: select p.item1, p.item2, . . . , p.itemk-1, q.itemk-1 from Lk-1 p, Lk-1q
5: where p.item1 = q.item1, . . . p.itemk-2 = q.itemk-2, p.itemk-1 ¡ q.itemk-1
6: Generate all (k-1)-subsets from the candidate itemsets in
Ck
7: Prune all candidate itemsets from Ck where some (k-1)-subset of the candidate itemset is not in the frequent itemset Lk-1
8: Scan the transaction database to determine the support for each candidate itemset in Ck.

3.3 DATA SET

The data set that is been used as an input to the algorithm,
is a standard dataset. The dataset contain two month's worth of all HTTP requests to the NASA Kennedy Space Center WWW server in Florida.

Format:

The logs are an ASCII file with one line per request, with the following columns:

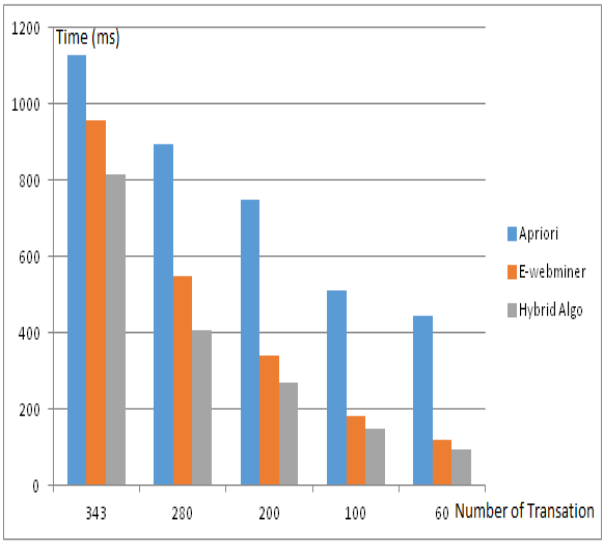
1) Any host making any particular request.
2) The host is identified mostly by IP address.
3) Timestamp in the format "DAY MON DD HH:MM:SS YYYY"
4) Request given in quotes.
5) HTTP reply code.
6) Bytes in the reply.

3.4 RESULT IN GRAPH

The below data contains the results obtained by applying the respective algorithm on a standard data set. The logs are first preprocessed and then forwarded to retrieve the valuable data from it. The number of transaction represent the number of logs in the log file. (i.e. Unique URL's). The results clearly show that the proposed Algorithm generates the results in much shorter time.

Table 1. Result Table (in ms)

Number of tran	Apriori Algorithm	E- Web Miner Algorithm	Proposed Hybrid Algo
343	1128	957	815
280	893	550	408
200	750	341	270
100	510	180	150
60	443	120	96



4. CONCLUSION AND FUTURE SCOPE

In this paper, a framework implemented for finding out user interest and personalizing user preference and generating patterns is implemented over the hybrid algorithm. It is different from the previous work on ApriroiAll and E-WebMiner algorithm. Here it is shown that the algorithm used is useful to enhance results for user preference and obtaining the more relevant results. In the future, this work can be extended to work on more huge log files. New technique cloud be found to find out user preference, Also one can do log mining on mobile logs.

References

[1] B. Kotiyal, A. Kumar, B. Pant, R. Goudar, S. Chauhan, and S. Junee, "User behavior analysis in web log through comparative study of eclat and apriori," in Intelligent Systems and Control (ISCO), 2013 7th International Conference on, 2013, pp. 421–426.

[2] D. Sisodia and S. Verma, "Web usage pattern analysis through web logs: A review," in Computer Science and Software Engineering (JCSSE), 2012 International Joint Conference on, May 2012, pp. 49–53.

[3] C. Zhang and J. Ruan, "A modified apriori algorithm with its application in instituting cross-selling strategies of the retail industry," in Electronic Commerce and Business Intelligence, 2009. ECBI 2009. International Conference on, June 2009, pp. 515–518.

[4] X. Xia, Q. Pei, Y. Liu, J. Wu, and C. Liu, "Multi-level logs based web performance evaluation and analysis," in Computer pplication and System Modeling (ICCASM), 2010 International Conference on, vol. 4, 2010, pp. V4–37–V4–41.

[5] M. Yadav, P. Keserwani, and S. Samaddar, "An efficient web mining algorithm for web log analysis: E-web miner," in Recent Advances in Information Technology (RAIT), 2012 1st International Conference on, 2012, pp. 607–613.

[6] Y. Liu and Y. Guan, "Fp-growth algorithm for application in research of market basket analysis," in Computational Cybernetics, 2008. ICCC 2008. IEEE International Conference on, 2008, pp. 269–272.

[7] L. juan Huang, "Fp-growth apriori algorithm's application in the design for individualized virtual shop on the internet," in Machine Learning and Cybernetics, 2007 International Conference on, vol. 7, 2007, pp. 3800–3804.

[8] K. Park, T. Lee, S. Jung, H. Lim, and S. Nam, "Extracting search intentions from web search logs," in Information Technology Convergence and Services (ITCS), 2010 2nd International Conference on, 2010, pp. 1–6.

[9] P. Hernandez, I. Garrigos, and J. Mazon, "Modeling web logs to enhance the analysis of web usage data," in Database and Expert Systems Applications (DEXA), 2010 Workshop on, 2010, pp. 297–301.

[10] M. Eltahir and A. Dafa-Alla, "Extracting knowledge from web server logs using web usage mining," in Computing, Electrical and Electronics Engineering (ICCEEE), 2013 International Conference on, 2013, pp. 413–417.

[11] C. Varnagar, N. Madhak, T. Kodinariya, and J. Rathod, "Web usage mining: A review on process, methods and techniques," in Information Communication and Embedded Systems (ICICES), 2013 International Conference on, 2013, pp. 40–46.

[12] K. Sudheer Reddy, M. Kantha Reddy, and V. Sitaramulu, "An effective data preprocessing method for web usage mining," in Information Communication and Embedded Systems (ICICES), 2013 International Conference on, 2013, pp. 7–10.

[13] J. Jojo and N. Sugana, "User profile creation based on navigation pattern for modeling user behaviour with personalised search," in Current Trends in Engineering and Technology (ICCTET), 2013 International Conference on, 2013, pp. 371–374.

[14] P. Hernandez, I. Garrigos, and J. Mazon, "Modeling web logs to enhance the analysis of web usage data," in Database and Expert Systems Applications (DEXA), 2010 Workshop on, 2010, pp. 297–301.

[15] Q. Zhou and L. Liu, "Understanding requirements for online services based on users' behavioural data analysis," in Computer Software and Applications Conference Workshops COMPSACW), 2013 IEEE 37th Annual, 2013, pp. 27–34.

[16] A. Yadav and S. Jain, "Analyses of web usage mining techniques to enhance the capabilities of e-learning environment," in Emerging Trends in Networks and Computer Communications (ETNCC), 2011 International Conference on, 2011, pp. 223–225.

AUTHOR

Rajdeep Marathe : Pursuing ME in Computer Engineering from D Y Patil College of Engineering, Akurdi through Savitribai Phule Pune University. Completed BE in computer Engineering from Gharda Insstitute of Technolgy, Lavel (MUmbai University. Area of intrest in data mining. Current working with It firm on SAP profile, Pune Maharashtra.

Mrs. D. A. Phalke: Pursuing PhD in Computer Science and Information Technology. Completed ME in Computer Engineering from D Y Patil College of Engineering, Akurdi through University of Pune and is currently working as Assistant Professor in the Department of Computer Engineering at D. Y. Patil College of Engineering and has 13.5 years of experience in the field.