ijecs
open access www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 7 July 2015, Page No. 13196-13200

Optimized Distributed Association Mining (ODAM) Algorithm for detecting Web Robots

Akshata Dilip Jagtap, Prof. Vijayalaxmi Kadroli

M.E. Information Technology
Terna Engineering College, navi mumbai
Nerul, India
akshatajagtap333@gmail.com
Information Technology
Terna Engineering College, navi mumbai
Nerul, India
udachanv@gmail.com

Abstract— As a result of the tremendous growth of the World Wide Web, the raw web data has become a vast source of information. Nowadays, the usages of search engines are considerably high by all types of users. The information required by different group of users are collected and represented by different types of search engines from many other sources. Web robot is one of the strategies which is used by many search engines as well as many other websites for collecting their data from the respective websites. These web robots are useful for many application, these are also dangerous for websites due to extracting the information of a particular site in an unauthorized way.

In this paper, We are suggesting the association rule mining algorithm named as "Optimized Distributed Association Mining" Algorithm for identifying the web robots which trying to traverse through website's confidential page or content and want to extract that content and after identifying ODAM will remove that web robot program. In such a way ODAM will help to prevent the websites from unauthorized access and other malicious programs.

Keywords — Web crawler, Association rule mining, web logs, Frequent logs.

INTRODUCTION

As the World Wide Web growing rapidly, the raw web data has become a vast source of information. To use and extract this data, various data mining techniques are developed. Web mining referred to as the application of data mining technologies to the web data.

Consequently, this has turned researcher’s attention towards the use of data mining techniques to this data. Web Mining is referred to as the application of data mining technologies to the web data. Nowadays, the usages of search engines are considerably high by all types of users, users of different age, profession, and purposes. The information required by different group of users are collected and represented by different types of search engines from many other sources. All search engines collect that information from their particular website using various approaches and show those results to user just in some seconds.

Web robots also called as spider, Web Wanderers, or Crawlers. Search engines such as Google use them to index the web content, spammers use them to scan for email addresses, and they have many other uses. If these web robots are useful for many applications, these are also dangerous for websites due to extracting the information of a particular site in an unauthorized way. Every website has some rules or

access permission for every external access that which contents of websites are confidential and not accessible. These access permissions are written in one plain text document called "robots.txt" file and which can be easily accessible for external users. Search engines or some other websites which want to extract the information are suppose to follow those rules and not try to access or mine the confidential information but most of the users ignore this rule and try to get the prohibited data by using some tools or programs. Web robot is one of those programs.

Whenever any kind of web robot tries to gain access a particular website, it makes an entry in that particular web site’s server. It is called as web log. By examining this web log the website administrator gets all the entries of programs or web robots which tried to access the information of that website. But, the problem is there are lots of entries could be entered in that web server so that the administrator can not examine each and every entry in particular span of time manually. In this case, the website requires a special automated tool for examining all these web logs in less time also it should find out the irrelevant as well as harmful program for the server and notifies the administrator about that. So that administrator can block all those entries in future.

RELATED WORK

The associations rule mining technique was first introduced by R. Aggrawal, where it was original proposed in terms of transactional databases. These rules were able to predict the items that can be purchased within the same transaction. Such rules have a great impact on making decisions about which item should be put on sale or which items should be placed near to each other.

Mining association rules is costly process and the complexity grows exponentially with the number of items presented in the database.

Many algorithms have been developed to generate association rules, each of them adopting a different optimization technique that applies either to the structure of the data or to the algorithm that traverses the search space. Among these algorithms, as apriori, FP-growth. The association involves the sensor's values at that interval formulate the transaction to be stored in the database. Each different value of a sensor is regarded as single element and it is assumed that sensors take on a finite number of discrete states. Each transaction is associated with a weight value indicating the validity of the transaction. The support of the pattern occurs.

Most of the data mining techniques applied to sensor data is based on centralized data extraction, where the data is collected at a single site and then the mining technique is applied. ODAM is the distributed technique which use distributed nature of sensors; we are proposing a distributed framework for building a classifier. In this framework, most of the work is pushed to the sensors themselves to build local models.

WEB ROBOTS

Web search engines, digital libraries, and many other web applications depend on robots to acquire documents. Web robots, also called "spiders", "crawlers", "bots" or "harvesters", are self-acting agents that continuously navigate through the hyperlinks of the Web, harvesting topical resources without significant human management cost. Web robots are highly automated and seldom regulated manually. With the increasing importance of information access on the Web, online marketing, and social networking, the functions and activities of Web robots have become extremely diverse. These functions and activities include not only regular crawls of web pages for general-purpose indexing, but also different types of specialized activity such as extraction of email and personal identity information and service attacks.

Even general-purpose web page crawls can lead to unexpected results for Web servers. For example, robots may overload the bandwidth of a small website such that normal user access is impeded. Robot-generated visits can also affect log statistics significantly so that real user traffic is overestimated. Robot activities can be regulated from the server side by deploying the Robots Exclusion Protocol in a file called robots.txt in the root directory of a web site. The Robots Exclusion Protocol1 (REP1) is a convention that allows website administrators to indicate to visiting robots which parts of their site should not be visited.

If there is no robots.txt file on a website, robots are free to crawl all content. A file named "robots.txt" with internet media type "text/plain" is placed under the root directory of a Web server. Each line in the robots.txt file has the format:

< Field> : < optional space >< value >< optional space >.

How to detect Web robots

1. Check the User Agent and IP Address of the session
2. Check for sessions that access the robots.txt file.
3. Check for sessions with unusually large number of HEAD requests.
4. Check for sessions with unusually large number of requests with empty referrers.

Web Logs

The use of internet and World Wide Web is increasing in a dense manner. Everyday tremendous volumes of user browser detail are stored in the form of web log files in the web server. So, careful investigations on the web server log are important to analyze the user behavior and their actions for accessing the data.

But it is complex task to handle the numerous volumes of web logs without preprocess them. The server logs are increased in the dense manner because every day number of users using the internet. The server logs are stored in the web server in the form of unformatted text files. It is too complex to manipulate the web logs with properly arrange them in some order. Preprocessing is applied in the web logs to reduce the volume of web log files and eliminate the unwanted data in the log files. It is always better to group the web logs for applying any kind of operation. In data mining terminology, this grouping is called clustering.

Association Rule Mining (ARM) algorithm

World Wide Web is a very fertile area for data mining research, with huge amount of information available on it. From its very beginning, the potential of extracting valuable knowledge from the Web has been quite evident. The term Web mining has been used in two different ways. The first, called Web content mining and the second, called Web usage mining. The web content mining is the process of information discovery from sources across the World Wide Web. Web usage mining is the process of mining for user browsing and access patterns. Interest in Web mining has grown rapidly in its short existence, both in the research and practitioner communities.

It is very important to analyze and extract useful patterns. For performing these tasks various Association Rule Mining algorithms can be used. The examples are apriori algorithm, fp - growth. Apriori algorithm is most commonly used algorithm in association mining but it has some limitations in terms of processing speed and time as well as cost compare to ODAM.

Challenges in Parallel Pattern Discovery using Apriori Algorithm

Most parallel and distributed ARM algorithms are based on sequential apriori, because of its success in sequential setting. Hence, directly adapting an apriori algorithm won't significantly improve performance over frequently itemsets generation or overall DARM performance. To perform better than apriori algorithms, it is important to focus on their problems.

The performance of apriori ARM algorithms degrades for various reasons.

1. It requires n number of database scans to generate a frequent n-itemset. Furthermore, it doesn't recognize transactions in the data set with identical itemsets if that data set is not loaded into main memory.
2. Therefore, it unnecessarily occupies resources for repeatedly generating itemsets from such identical transactions.

In this paper, I suggest the Optimized Distribution Algorithm (ODAM) for performing association mining with improved response time and cost.

PERFORMANCE EVALUATION

The Optimized Distributed Association Rule Mining algorithm is to be used for identifying and removing the web robots through the web browser information. The system will prevent the "robots.txt" file which created by website administrator for access permissions from external web robots.

- The first step is to define the specific platform for which the log of data will be collected and worked upon.
- Next is the data cleaning process is a vital task and consumes the most amount of time. Unless you have the correct pre-processed data, it is difficult to achieve good results. The next important step is feature extraction, where one needs to think out of the box keeping in mind the project goal.
- After the data is processed, the other important step is to visualize the results obtained from the mining of the dataset.
- In order to achieve all of this, the most important decision is to choose the right set of tools that will support all the necessary requirements like handling of huge amount of data, data pre-processing, mining algorithms, and data visualization.

1. Information Discovery Process:

The knowledge discovery process refers to the series of steps involved in the usage mining process.

1. The initial step is to define the problem, for which this project can be referred to as analyzing the access logs and obtaining reasonable results that can be used to improvise the success of the identifying log entries.

2. The next step in the process is to prepare the data for mining. This is the most time consuming process and requires significant information about the domain and the focus should always remain on the project goals.

3. Once we have the preprocessed data, we can use this data for mining analysis. The analysis basically refers to discovering the navigation patterns of the users by analyzing the user sessions. The mining algorithms like clustering, classification and trees can be run on this dataset and different results can be achieved. The interpretation of these results is then based on the patterns discovered and with respect to the background knowledge of the web site.

4. The results obtained by performing the data mining techniques on the processed data, should lead to concrete and meaningful suggestions for the betterment of the web site (full or part). In order to verify the impact of the changes suggested for improvising the success of the web site, the access logs of the re-designed web site should be analyzed.

2. **Data Preprocessing:**

1. Data fusion
2. Data Cleaning
3. User Identification
4. Session Identification
5. Formatting
6. Data summarization.

3. Feature Extraction

The last step in the pre-processing task is to extract features from the available transactions in the log. For mining the associated patterns and frequent patterns the Optimized Distributed Association Mining is proposed.

4. **Pattern Discoveries and analysis of Web Logs**

The discovery of user access patterns from the use access logs, referrer logs, user registration logs etc is the main purpose of the Web Usage Mining activity. Pattern discovery is performed only after cleaning the data and after the identification of user transactions and sessions from the access logs. The analysis of the pre-processed data is very beneficial to all the organizations performing different businesses over the web.

The tools used for this process use techniques based on AI, data mining algorithms, psychology, and information theory. The different systems are needed for the Web Usage Mining process have introduced different algorithms for finding the maximal forward reference, large reference sequence, which can be used to analyze the traversal path of a user. The different kinds of mining algorithms that can be performed on the preprocessed data include path analysis, association rules, sequential patterns, clustering and classification. It totally depends on the requirement of the analyst to determine which mining techniques to make use of.

1. Association Rules:

This technique is generally applied to a database of transactions consisting information. This rule implies some kind of association between the user activities in the database. It is important to discover the associations and correlations between these set of activities. In the web data set, the activities consist of the number of URL visits by the client, to the web site. It is very important to define the parameter support, while performing the association rule technique on the transactions. This helps in reducing the unnecessary transactions from the database. Support defines the number of occurrences of user transactions within the transaction log. The discovery of such rules from the access log can be of tremendous help in reorganizing the structure of the web site. The frequently accessed web pages should be organized in their order of importance and be easily accessible to the users.

2. The clustering and classification:

The clustering and classification discovery rules allow grouping the items with similar attributes together. Therefore, when new data is added to the database, it can be classified on the basis of its attributes. In the web transaction data set, the clustering can result in forming clients with similar interests or clients that visit the specific web page based on their demographic information and access patterns. This can help organizations to become client centric by serving to the interests of their clients and developing a one-to-one relationship with their clients.

Working of ODAM

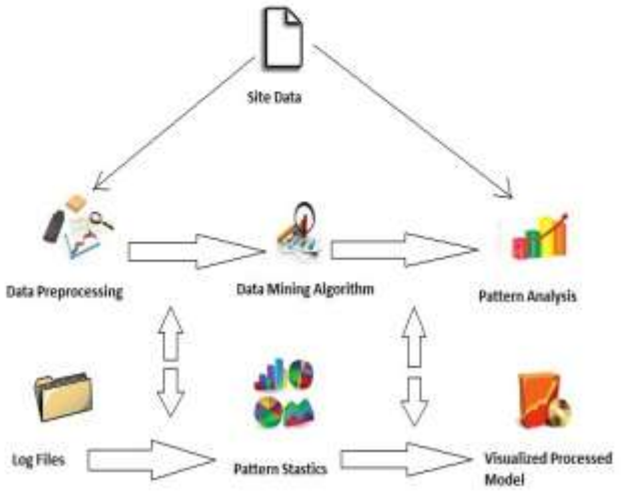
ODAM is a type of Distributed mining algorithm based on following technique. Distributed association discovers rules from various geographically distributed data sets. However, the network connection between those data sets isn’t as fast as in a parallel environment, so distributed mining usually aims to minimize communication costs. Researchers proposed the Fast Distributed Mining algorithm to mine rules from distributed data sets partitioned among different sites. In each site, FDM finds the local support counts and prunes all infrequent local support counts. After completing local pruning, each site broadcasts messages containing all the remaining candidate sets to all other sites to request their support counts. It then decides whether large itemsets are globally frequent and generates the candidate itemsets from those globally frequent itemsets.

ODAM ALGORITHM

We assume each ODAM site has the same tasks as sequential association mining, except it broadcasts support counts of candidate itemsets after every pass.

```
Nf = {Non_frequent global 1-itemset}
for all transaction {
    for all 2-subsets s of t
        if(s Є c2) s.sup++
        t' = delete_nonfrequent_items(t)
        Table.add(t');
}

Send_to_receiver(c2);
/*Global frequent support counts from receiver
F2 = receive_from receiver (fq);
C3 = {candidate itemset};
T = Table.getTransacations;

while ( ck ≠ { }) {
for all transaction
    for all k_subsets s of t
     if ( s Є Ck) s.sup++
       k++;

send_to_receiver (Ck);
/* Generating candidate Itemset of k+1 pass
k+1 pass;

Ck+1 = { candidate itemset }
}
```

PROPOSED ARCHITECTURE



We implemented ODAM using JAVA. We conduct an experiment in a single site with different support, by comparing total execution time between ODAM and CD. We compared the execution process of ODAM with Apriori algorithm with same data set.

RESULTS AND DISCUSSION

Figure1 shows ODAM and Apriori’s total processing time for calculating frequent itemsets of different lengths. We have taken the same data sets for both of algorithms. ODAM requires less time for execution compare to Apriori. Both algorithms take more time for calculating frequent itemsets as the longer candidate itemsets. But the difference between two consecutive iteration of Apriori is greater than ODAM. ODAM removes a significant number of infrequent 1-itemsets from every transaction after first pass, so it finds a significant number of identical transactions.

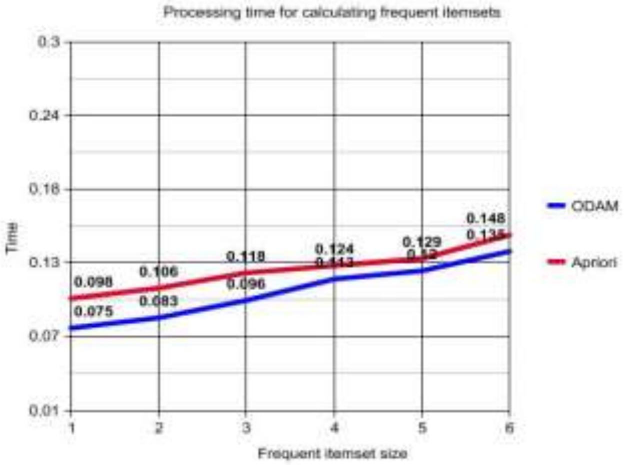


Fig.1. Processing time required for calculating frequent itemsets using ODAM and Apriori with 0.3% of support

Figure 2 shows the total size of messages (in number of bytes) that ODAM and CD (Apriori) transmit to generate frequent itemsets with different support values. For comparing the number of messages that ODAM and CD exchange among various sensors to calculate globally frequent itemsets in a distributed environment, we partition the original data set into four partitions. The number of identical transactions among different partitions can be low because each one contains only 20 percent of the original data sets. As the result shows, ODAM reduces communication overhead and cost by 30 to 40 percent compared to CD. In each site, CD exchanges message with all other sites after every pass, consequently the message exchange size increases when we increase the number of sites (sensors).

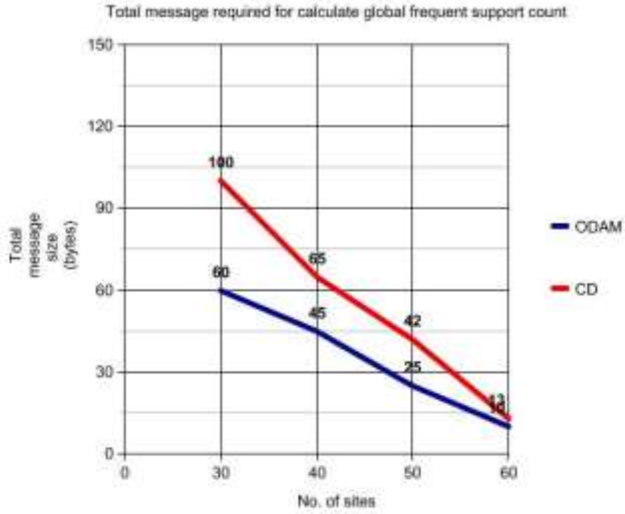


Figure 2 Total message size required to calculate global support count using ODAM and CD.

CONCLUSION

In this paper, I proposed ODAM algorithm for finding out frequent logs and suspicious logs. ODAM is association rule mining algorithm works effectively on distributed environment. Unlike traditional sequential algorithms, ODAM finds frequent items in minimum passes. ODAM doesn’t generate candidate set after first pass. In first pass, ODAM prunes all infrequent items and load remaining items in another table. It then delete original data set from memory and perform same procedure on updated data set. Because of this technique, it requires very less memory and time than sequential apriori algorithm.

For finding web robots, we scanned frequent logs from previous frequent itemsets and check whether they contain the HEAD methods or accessing robots.txt files. Also, it checks they are having empty referrer fields. The result contains the suspicious logs (Web robots).

REFERENCES

[1] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Database," Conf. Very Large Databases (VLDB 94), Morgan Kaufman, 1994, pp. 407-419.

[2] Mafruz Zaman Ashrafi, Monash University, David Taniar, Monash University, Kate Smith, Monash University,"ODAM: An Optimized Distributed Association Rule Mining Algorithm" IEEE DISTRIBUTED SYSTEMS ONLINE 1541-4922 © 2004 Published by the IEEE Computer Society Vol. 5, No. 3; March 2004.

[3] ZHANG Yuzhou, WANG Jianyong, ZHOU Lizhu, " Parallel Frequent Pattern Discovery: Challenges and Methodology" TSINGHUA SCIENCE AND TECHNOLOGY ISSN□ 1007-0214□ 15/20□ pp719-728 Volume 12, Number 6, December 2007.

[4] Anand S. Lalani, "Data Mining of Web Access Logs", School of Computer Science and Information Technology Royal Melbourne Institute of Technology Melbourne, Victoria, Australia, July, 2003

[5] Anjan Das, “A Novel Association Rule Mining Mechanism in Wireless Sensor Networks”, Department of Computer Science,St. Anthony’s college, Shillong, India. 978-1-4