International Journal of Technology Research and Management
ISSN (Online): 2348-9006
Vol 3 Issue 12 December 2016

An Enhanced Approach for Web Log Mining

Paridhi Nigam[1] and Rajesh Kumar Chakrawarti[3]
Computer Science and Engineering, Shri Vaishnav Institute of Technology & Science, Indore, M.P, India[1]
Computer Science and Engineering, Shri Vaishnav Institute of Technology & Science, Indore, M.P, India[2]
pari.nigam31@gmail.com[1], rajesh_kr_chakra@yahoo.com[2]

Abstract: *In this paper, we present a review of new web mining algorithms. Mining of logs present in the web is the application of data mining technique. At present web is enormous collection of data & it will keep growing up with growing of internet technologies, so web mining is a computationally costly task because it becomes difficult to website holder to present proper information to the users. So still there is a need to update and enhance the existing web mining techniques so that we can get the more efficient methods for the same task. Visit web item set mining is a famous issue of research for a few specialists throughout the years. In this paper, we have built up an energetic system to discover visit web item sets from the web exchange database. The proposed procedure is quick in correspondence to more established calculations. Likewise it consumes less memory space for computation purpose.*

Keywords: *Web mining, web log mining, frequent itemset mining.*

1. INTRODUCTION

The internet is becomes the most bursting forth area of information gathering. While working on the internet, web users leave many of the records. So proper mining process are required for the huge amount of data for gathering knowledge and information**.** By utilizing information mining strategy, web mining is utilized to concentrate data from web [1, 2] likewise used to discover the patterns of information which are available on the web with the assistance of client's conduct. Web mining are for the most part separated into three sections. These are web content mining, In web content mining procedure of social event helpful data from web substance is characterized as web substance mining, Another class of web mining is web structure mining it is utilized to dissect the structure of site or progressive system of site, and the web use mining is the third classification of web mining, it comprise of disorderly information like picture, video, sound, organized hyperlinks.

Web Usage Mining is the technique for information mining and is utilized to findout a fascinating utilization designs from Web, with a specific end goal to comprehend and better serve the requirements of Web-based applications [3]. Web use mining is further separated into three classes.

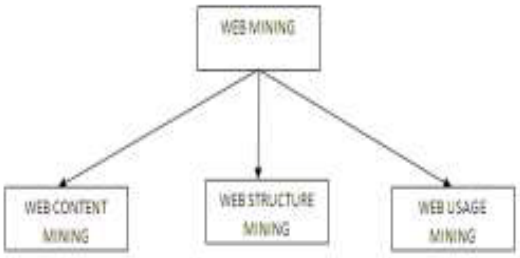

Figure 1: Dissemination of web mining

2. BACKGROUND

The authors Alexandra's Nanopoulos [4] proposed a web mining method based on the concept of web perfecting. It helped in the reducing the user latency perception ratio.

Mathis Gerry [5] proposed three different web mining approaches. These three methods are based on association rules, frequent sequences, and frequent generalized sequences.

The authors have developed and implemented the algorithms for all three methods.

Association rule learning [6] is a very common method from discovery of useful patterns from data & also for representing the useful patterns in form of a rule.

In 2010 author Varun Kumar [7] proposed "A Real Time Stream Mining Algorithm" which has an ability to grip the numerous sizes of the batch rather than the fixed in other. The time has been fixed for isolate the Batches. In the earlier algorithms the rare items were eliminated. Later if those items become frequent then the data cannot be bagged again. Also they concentrated only on the frequent item sets, but not on the extracting knowledge from it. Such kind of problems are clearly solved by this paper. Proposed work uses an extension of tree structure with the log-time window as its data structure. Method constitute three columns namely tilted-time, frequency & size of the batch. Recent data holds big space whereas the old one holds the less only. The work follows two different types of tail pruning in examining whether the superset needs to be dropped or not based on the different batch sizes & time.

In 2010 the author ZHOU [8] proposed algorithm "A More Accurate Space Saving Algorithm for Finding the Frequent Items**"** by considering the space as an important factor. Authors used an improved LRU (Least Recently Used) based algorithm. Proposed algorithm omits the rare items before taken for the processing. Method increases the stability & the performance. Method is used to find out the frequent items as well as the frequency of those items.

In 2011 author Mahmood Deypir [9] proposed "A New Adaptive Algorithm for Frequent Pattern Mining over Data Streams" algorithm based on the different kind of sliding window based model. Method don't need entire data that are in streaming. Method takes an advantage of the already existing item sets. To intensify the feature of sliding window concept. Also it reduces the amount of space occupying and time taken to calculate based on the fixed size of the window.

3. PROPOSED TECHNIQUES

Techniques used in web usage mining consist of some steps. These are.

3.1 Preprocessing

Pre-processing is the most important phase in data mining. This is a data mining technique that involves converting raw data into an understandable format. Data which are present in the real world are noisy (containing errors), inconsistent, no quality data. To remove the noisy data pre processing is used. Data collection should be done before the preprocessing phase Preprocessed files consists of data such as the page is gather by whom, what page are accessed and how long the user accessed that page, which user accessed which page, time of access, date of access date**,** duration of access etc. The preprocessing consist of further steps.

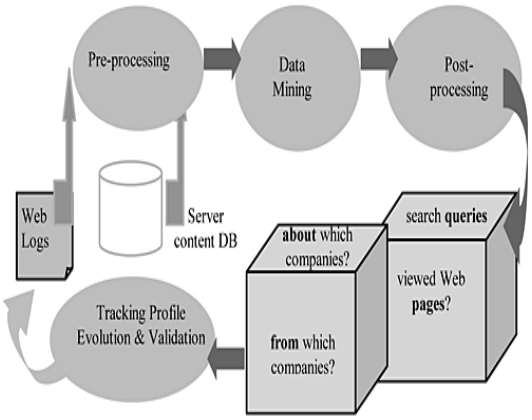

Figure 2: Strategy for preprocessing

3.1.1 Data Cleaning

In proposed methodology data cleaning method use to removes dirty data, for example data with incomplete fields, missing or wrong values, in the preprocessing stage. The clean data is then reduced and/or transformed so that the data is represented by the useful features and actionable dimensions. The user performs the required mining functions to finding the patterns which include summarization/ generalization of data characteristics, classification or

clustering of data for future prediction, association finding, etc.

3.1.2 User & Session Identification

In user identification phase users are discovered, that who access the web pages, who contact the web server, requesting for some resources on the web. In session identification, the session of individual user is discovered. If the time exceeds a certain limit that implies user is started a new session.

3.1.3 Path Completion

In this phase the new paths are added and also identified paths by the user.

3.2 Pattern Discovery

After the raw data have been preprocessed data mining techniques can be applied for finding the patterns [10] Pattern discovery phase used to find out the frequent patterns which are bring forth by server. Pattern discovery describes the type of mining technique that has been applied to the Web domain. In Web Usage Mining, session of server is an ordered sequence of pages requested by a user. Furthermore, due to the difficulty in identifying unique sessions, additional prior knowledge is required. Some important techniques of pattern discovery are frequent itemset mining, clustering, statistical analysis, classification and sequential analysis used to recognize the new useful patterns.

3.2.1 Frequent Itemset Mining

Frequent item set mining is used to find out the association of pages which are frequently accessed by user. This method is also used to discover the information like the paths which are frequently accessed by the web users, clubbed pages accessed by users repeatedly. Frequent item set mining is one of the most important issues faced by the knowledge discovery & data mining community. It plays an important role in data mining fields as association rule in data mining fields as association rule [11, 12]. For finding the frequent itemsets, the support of each item set must be computed by scanning each transaction in data with the explosive growth of data mining information & knowledge from large database has become one of the major challenges for data management & mining community. Frequent item set mining plays an important role in many data mining tasks that try to find interesting patterns from databases such as association rules, correlations, sequences, classifiers, clusters and many more of which the mining of association rules is one of the most popular problems.

In proposed methodology the problem of mining frequent itemsets are essentially used , to discover all rules, from the given transactional database D that have support greater than or equal to the user specified minimum support.

- To analyze the various existing techniques and find their strengths and weakness.
- To develop an algorithm for frequent item set mining. The proposed algorithm will be faster in comparison to present algorithms
- Validate the implementation by desired input.

Association mining rule can be defined formally as Association rule is an implication of the form X -> Y where X,Y subset of I are the sets of items called Item sets and $X \cap Y = \Phi$. Association rules show attributes value conditions that occur frequently together in a given dataset[13]. This rules provide information in the form of "if-then" statements. These rules are computed from the data and, unlike the if-then rules of logic, association rules are probabilistic in nature. Support for an association rule X->Y is the percentage of transaction in database that contains X U Y .The other number is known as the Confidence of the rule. Confidence or Strength for an association rule X U Y is the ratio of number of transactions that contains X U Y to number of transaction that contains X. An itemset (or a pattern) is frequent if its support is equal to or more than a user specified minimum support (a statement of generality of the discovered association rules). Association rule mining is to identify all rules meeting user-specified constraints such as minimum

support and minimum confidence (a statement of predictive ability of the discovered rules) [14]. One key step of association mining is frequent itemset (pattern) mining, which is to mine all itemsets satisfying user specified minimum support. However a large number of these rules will be pruned after applying the support and confidence thresholds. Therefore the previous computations will be wasted. To avoid this problem and to improve the performance of the rule discovery algorithm, mining association rules may be decomposed into two phases: Discover the large itemsets, i.e., the sets of items that have transaction support above a predetermined minimum threshold known as frequent itemsets.

- Use the large itemsets to generate the association rules for the database that have confidence above a predetermined minimum threshold.

3.2 Pattern Analysis

In pattern analysis phase, the patterns which are extracted from pattern discovery phase are preprocessed to get frequent patterns from web. Pattern analysis makes prediction of new data which are coming from the same source. Data are present in many forms like audio, video, images etc.

4. PROPOSED ALGORITHM

Algorithm:

STEP 1:

Start

STEP 2:

Transaction Data Set & Minimum Support Threshold

STEP 3:

First the algorithms scans the transaction Data Base and calculate the support of each single size item.

STEP 4:

In this step, the transaction Data Base is transformed into a new compressed data structure based table by pruning of all those items from the transaction database, whose support is lesser then the minimum support threshold because they will not appear in any frequent patterns.

Step 5:

Call algorithm recursively to generate bigger frequent patterns by using the Union OR expansion of lower size items.

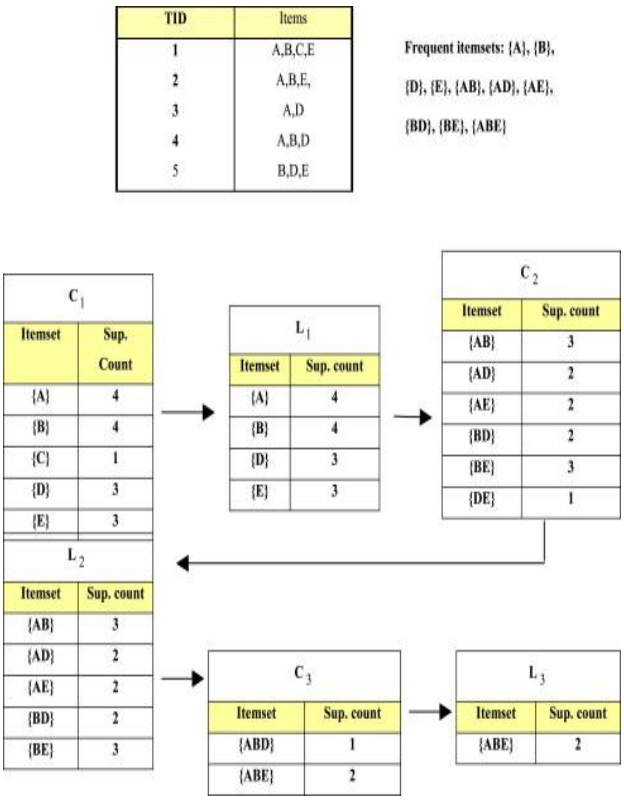

Figure 3: Generation of frequent patterns

The overall performance of mining association rules is determined primarily by the first step. The second step is easy. After the large itemsets are identified, the corresponding association rules can be derived in straightforward manner.

5. ADVANTAGES

The following are the advantages of proposed algorithm.

- It take less time as compared to the existing methods for web log mining
- Space consumption also be lesser.
- Companies can understand the needs of the customer better and they can act in response to customer needs faster.
- The Organizations can establish a good and healthy relationship with customer by giving them the exact information that they want.

6. DISADVANTAGES

- The information collected by the companies can be used some other purpose which leads to the violation of privacy of users.
- Privacy is considered lost, when this technology is used on data of that might be a cause of concerns.

7. RESULT ANALYSIS

The experimental results are shown in Figure 4, Figure 5, & Figure 6. We compare the number of itemsets found, time consumed, and memory used by the datasets in the database. The result shows that the number of itemsets found after scanning the data sets is equal, but the time taken by the old algorithm is more as compared to the novel algorithm, as well as novel algorithm take less memory space.

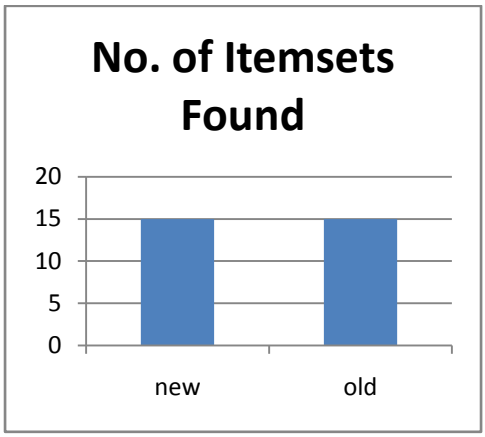

Figure 4: Result Comparison

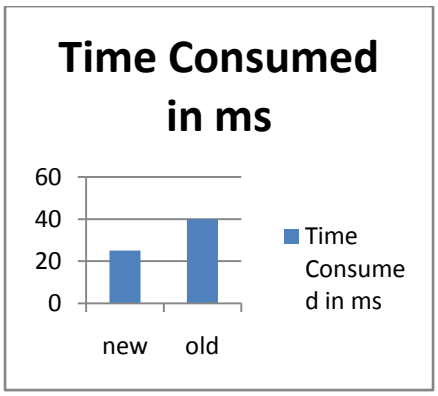

Figure 5: Time Consumption Comparison

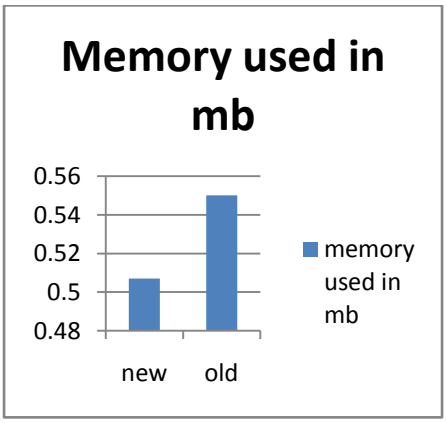

Figure 6: Memory Consumption Comparison

8. CONCLUSION

In this paper, we surveyed the list of existing web mining techniques. We restricted ourselves to the classic web mining problem. It is the generation of all frequent item sets that exists in market basket like data with respect to minimal thresholds for support & confidence.

In this paper, we present a novel algorithm for mining web log data sets. Frequent mining of data mining is used for that purpose. Frequent item set mining is crucial for association rule mining. We have evaluated the performance of our proposed algorithm. It is fast. Also it is taking less main memory for computation in comparison to previous algorithm.

9. FUTURE SCOPE

Our proposed algorithm works for the normal data set. The same algorithm can be extended to work for the uncertain data set. More compact data structure can be proposed to reduce space consumption. One limitation though data mining can help reveal patterns and relationships, it does not tell the user the value or significance of these patterns. It does not tell the users which patterns are sensitive and which are not. It can be said that software privacy failures can be direct result of one or more of the following points that are taken from risk management.

10. ACKNOWLEDGEMENTS

I am highly thankful to my guide Rajesh Kumar Chakrawarti (Reader, Deparment of Computer Science and Department). His suggestion and interest have helped me in integrating the work. His accomodating nature tolerates my persistent queries and provided the best solution to my problem. I am thankful to Mr. Anand Rajavat (Head of Department of Computer Science) for providing all facilities and resources needed for this research paper.

REFERENCES

[1] Ashok Kumar, D. Loraine Charlet Annie M. C., "web log mining using K-Apriori Algorithm", volume 41, March -2012.

[2] Jian Pei, Jiawei Han, Behzad Mortazavi-asl, Hua Zhu, "Mining Access Pattern Efficient from Web Logs".

[3] B. Santhosh Kumar, K. V. Rukmani, "Implementation of Web Usage Mining Using Apriori and FP-Growth Algorithms", volume: 01, Issue: 06, Pages: 400-404(2010).

[4] Aggrawal. R, Imielinski. t, Swami. A., "Mining Association Rules between Sets of Items in Large Databases", In Proc. Int'l Conf. of the 1993 ACM SIGMOD Conference Washington DC, USA.

[5] Han. J, Pei. J, and Yin. Y., "Mining frequent patterns without candidate generation", In Proc. ACM-SIGMOD Int'l Conf. Management of Data (SIGMOD), 2000.

[6] Toivonen. H., "Sampling large databases for association rules", In Proc. Int'l Conf. Very Large Data Bases (VLDB), Sept. 1996, Bombay, India,

[7] Varun Kumar, Rajanish Dass. Proceedings of the 43rd Hawaii International Conference on System Sciences, 2010 IEEE,

[8] Sonali Shukla, Sushil Kumar, Bhupendra Verma, "A Linear Regression-Based Frequent Itemset Forecast Algorithm for Stream Data", International Conference on Methods and Models in Computer Science, 2009.

[9] Mahmood Deypir, Mohammad Hadi Sadreddini, "A New Adaptive Algorithm for Frequent Pattern Mining over Data Streams", ICCKE,2011, 230-235 FLEX Chip Signal Processor (MC68175/D), Motorola,1996.

[10] Harish Kumar and Anil Kumar, "Clustering Algorithm Employ in Web Usage Mining: An Overview", INDIA Com publication, Edition 2011.

[11] Varun Kumar, Rajanish Dass. Proceedings of the 43rd Hawaii International Conference on System Sciences, 2010 IEEE, 978-0-7695-3869-3.

[12] Sonali Shukla, Sushil Kumar, Bhupendra Verma, "A Linear Regression-Based Frequent Itemset Forecast Algorithm for Stream Data", International Conference on Methods and Models in Computer Science, 2009.

[13] ZHOU Jun, CHEN Ming, XIONG Huan, "A More Accurate Space Saving Algorithm for Finding the Frequent Items", IEEE-2010.

[14] Yong-gong Ren, Zhi-dong Hu, Jian Wang, "An Algorithm for Predicting Frequent Patterns over Data Streams Based on Associated Matrix", Ninth Web Information Systems and Applications Conference, 2012. 95-98.

[15] Mahmood Deypir, Mohammad Hadi Sadreddini, "A New Adaptive Algorithm for Frequent Pattern Mining over Data Streams", ICCKE,2011, 230-235 FLEX Chip Signal Processor (MC68175/D), Motorola, 1996.