International Journal of Technology Research and Management
ISSN (Online): 2348-9006
Vol 6 Issue 5 May 2019

A Tree Based Time Efficient Technique for Mining Web Log Data from Internet Usage Data Set

Varsha Sharma[1], Prof. Jyotsana Goyal[2], Prof. Monika Date[3]
LNCT, Indore, MP[1, 2, 3]
varshasharma517@gmail.com[1]

Abstract: *In this paper, we are presenting an overview of existing web log mining algorithms. Nowadays web log mining is a very popular and computationally expensive task. The web usage mining is the task of applying data mining techniques to discover usage patterns from Web data in order to understand and better serve the needs of users navigating on the Web. It is used to discover the navigation patterns from web data, predicts the user behavior while the user interacts with the web and also it helps to improve large collection of resources.*

The frequent item set mining techniques are used for mining web log data. We have also explained the fundamentals of frequent item set mining. We have described today's approaches for frequent item set mining. From the broad variety of frequent item set mining algorithms that have been developed we will compare the most important ones. An updated POCU tree based technique is also proposed. This technique is proposed for speedy execution.

Keywords: *Data Mining, KDD, Web log mining, minimum support.*

1. INTRODUCTION

The web mining is used to extract the useful information from the World Wide Web by using data mining techniques. Web mining can be classified into three categories: (i) web content mining, (ii) web structure mining and (iii) web usage mining. The web content mining is used to search the web pages via content. The web structure mining is used discover the web page structure and hierarchy of hyper links in the web site.

The web utilization mining is the undertaking of applying information mining procedures to find use designs from Web information so as to comprehend and better serve the requirements of clients exploring on the Web. It is utilized to find the route designs from web information, predicts the client conduct while the client connects with the web and furthermore it improves expansive accumulation of assets.

Web usage mining consists of three main steps: (i) preprocessing, (ii) pattern discovery and (iii) pattern analysis. The data extracted from the web log files are preprocessed to remove the noisy data. The data collection should be done before the preprocessing phase.

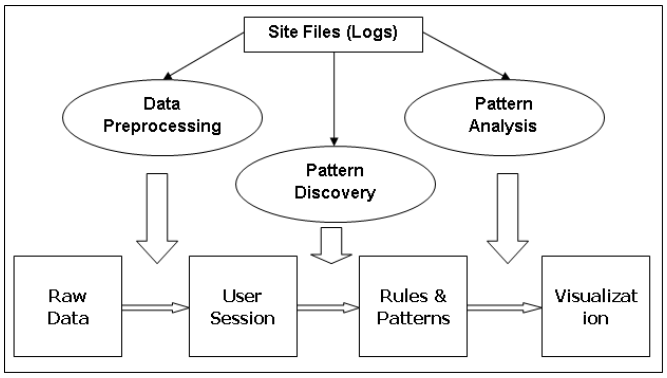

Figure 1: Web Usage Mining [5]

In data preprocessing phase the data are cleaned and the frequent patterns are mined. The preprocessed file consists of information such as who accessed the web site, what pages where accessed and how long the user accessed that page. In pattern discovery phase the activities of the users on the web are discovered. The frequent patterns discovery phase needs only the Web pages visited by a given user. In this stage the sequences of the pages are irrelevant. Also the duplicates of

the same pages are omitted, and the pages are ordered in a predefined order. In pattern analysis phase the patterns extracted from the pattern discovery phase are processed to get most frequent pattern.

With the expansion in Information Technology, the measure of the databases made by the associations because of the accessibility of ease stockpiling and the advancement in the datacapturing innovations is additionally expanding,. These association segments incorporate retail, oil, media communications, utilities, fabricating, transportation, charge cards, protection, banking and numerous others, separating the profitable information, it important to investigate the databases totally and productively. Learning disclosure in databases (KDD) serves to distinguishing valuable data in such enormous databases. This significant data can help the leader to settle on precise future choices. KDD applications convey quantifiable advantages, including decreased expense of working together, upgraded gainfulness, and improved nature of administration. In this way Knowledge Discovery in Databases has turned out to be a standout amongst the most dynamic and energizing examination territories in the database network.

As of late the span of database has expanded quickly. This has prompted a developing enthusiasm for the improvement of devices competent in the programmed extraction of learning from information. The term information mining or learning disclosure in database has been received for a field of research managing the programmed revelation of verifiable data or learning inside the databases. The verifiable data inside databases, predominantly the fascinating affiliation connections among sets of items that lead to affiliation guidelines may uncover helpful examples for choice help, monetary estimate, promoting approaches, even restorative determination and numerous different applications.

2. LITERATURE SURVEY

Mining frequent item sets is an important problem in data mining and is also the first step of deriving association rules [2]. Hence many efficient item set mining algorithms (e.g., Apriori [2] and FP-growth [10]) have been proposed. While all these algorithms work well for databases with precise values but it is not clear how they can be used to mine probabilistic data.

For uncertain databases the Aggarwal [1] and Chui [9] developed efficient frequent pattern mining algorithms based on the expected support counts of the patterns. However Bernecker et al. [3] Sun[14] and Yiu [16] found that the use of expected support may render important patterns missing. Hence they proposed to compute the probability that a pattern is frequent and introduced the notion of PFI. In work done in [3] the dynamic programming based solutions were developed to retrieve PFIs from attribute uncertain databases. However their algorithms compute exact probabilities and verify that an item set is a PFI in O(n2) time. The proposed model-based algorithms avoid the use of dynamic programming and are able to verify a PFI much faster.

In [16] the approximate algorithms for deriving threshold-based PFIs from tuple-uncertain data streams were developed. The Zhang et al. [16] only considered the extraction of singletons (i.e., sets of single items) our solution discovers patterns with more than one item. Recently Sun [14] developed an exact threshold based PFI mining algorithm. However it does not support attribute-uncertain data considered in this paper. In a preliminary version of this paper [15] we examined a model-based approach for mining PFIs. we study how this algorithm can be extended to support the mining of evolving data.

All the other works on the retrieval of frequent patterns from imprecise data includes [4], it studied approximate frequent patterns on noisy data then the [11], it examined association rules on fuzzy sets and [13], proposed the notion of a vague association rule. However none of these solutions are developed on the uncertainty models studied here.

For evolving databases there are a few incremental mining algorithms that work for exact data have been developed. Just For example in [6] the Fast Update algorithm (FUP) was proposed to efficiently maintain frequent item set & for a database to which new tuples are inserted. The proposed incremental mining framework is inspired by FUP. In [7] the FUP2 algorithm was developed to handle both addition and deletion of tuples. The work done by ZIGZAG [11] also examines the efficient maintenance of maximal frequent item sets for databases that are constantly changing. In [8] a data structure called (CATS Tree) was introduced to maintain frequent item sets in evolving databases. Another data structure called CanTree [12] arranges tree nodes in an order that is not affected by changes in item frequency. This data structure is used to support mining on a changing database.

3. PROPOSED METHODOLOGY

Definition 1. POCU-tree is a tree structure:

(1) It consists of one root labeled as ‘‘null’’, and a set of item prefix subtrees as the children of the root.

(2) Each node in the item prefix subtree consists of five fields: item-name, children-list, pre-order. item-name registers which item this node represents. children-list registers all children of the node. preorder is the pre-order rank of the node.

Step 1: input: 1. a web log database.
2. User defined Threshold.
Step 2: The web log database is scanned once and the count of each item is found.
Step 3: If count of any item of step 2 is less than user defined threshold then eliminate the infrequent item.
Step 4: Now arrange the frequent we log items found in step3 in decreasing order of their count. It will be used in construction of the POCU tree
Step 5: Construct POCU tree by reading one transaction at a time.
Step 6: Extract a sub tree ending in an item (Suppose X).
Step 7:

- Check that the item of step 6 is frequent or not.
- If it is frequent then extract it as frequent item.
- New item X is frequent so now find the other frequent items ending with X.
- Continue this recursive procedure until no item found.

Step 8: Arrange the frequent item sets in the decreasing order of their size.
Step 9: For each frequent item sets having support more than the MST.

- Find all the super sets(S) of the frequent item stets.
- If any super set of the frequent item set is not having the same support as the frequent item set then add both in FIS list.
- Otherwise add only superset in the FIS list.

Step 10: Delete duplicate items from the FIS list if any.
Step 11: Return FIS.

4. RESULT ANALYSIS

We ran the comparison algorithms on several parts of real datasets, which are common datasets from previous frequent item set mining studies. This dataset was downloaded from the UCI machinery data bank (https://archive.ics.uci.edu/ml/datasets/Internet+Usage+Data) . In our experiment, internet usage data set is used. Total 4000 instances of data set are used in the experiment. Minimum threshold is 40 percent.

The results for retail data set are shown below in graphs:

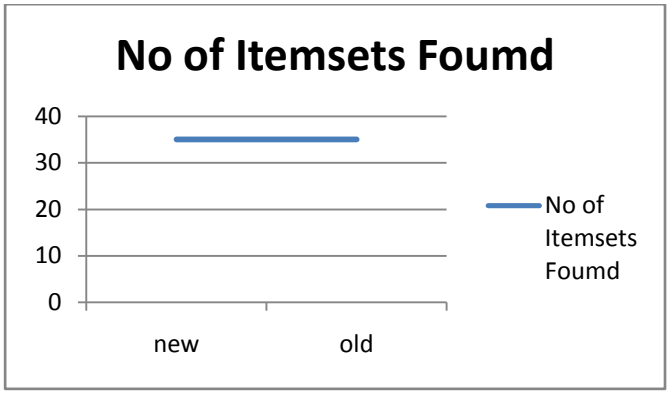

Figure 2: No of Item sets Found

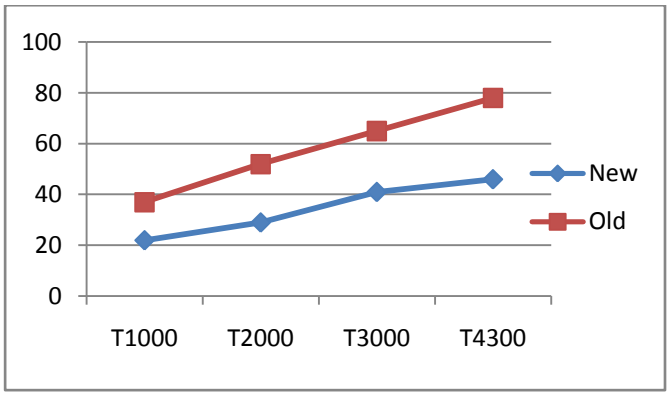

Figure 3: Time Comparison

5. CONCLUSION

This paper elaborates the concept of web usage mining in lucrative manner. The overall concept of web usage mining is described in brief. A set of techniques, which is used in modern days, is examined. Their working along with advantages and drawbacks has been analyzed. This paper presented a comprehensive critical review of web log mining techniques. Also a POCU tree based technique for mining frequent items is proposed. The proposed technique is taking less time.

REFERENCES

[1] C. Aggarwal, Y. Li, J. Wang, and J. Wang, "Frequent Pattern Mining with Uncertain Data," Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), 2009.
[2] R. Agrawal, T. Imieli_nski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," Proc. ACM SIGMOD Int'l Conf. Management of Data, 1993.
[3] T. Bernecker, H. Kriegel, M. Renz, F. Verhein, and A. Zuefle, "Probabilistic Frequent Itemset Mining in Uncertain Databases," Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), 2009.
[4] H. Cheng, P. Yu, and J. Han, "Approximate Frequent Itemset Mining in the Presence of Random Noise," Proc. Soft Computing for Knowledge Discovery and Data Mining, pp. 363-389, 2008.
[5] R. Cheng, D. Kalashnikov, and S. Prabhakar, "Evaluating Probabilistic Queries over Imprecise Data," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2003.
[6] D. Cheung, J. Han, V. Ng, and C. Wong, "Maintenance of Discovered Association Rules in Large Databases: An Incremental Updating Technique," Proc. 12th Int'l Conf. Data Eng. (ICDE), 1996.
[7] D. Cheung, S.D. Lee, and B. Kao, "A General Incremental Technique for Maintaining Discovered Association Rules," Proc. Fifth Int'l Conf. Database Systems for Advanced Applications (DASFAA), 1997.
[8] W. Cheung and O.R. Zaı¨ane, "Incremental Mining of Frequent Patterns without Candidate Generation or Support Constraint," Proc. Seventh Int'l Database Eng. and Applications Symp. (IDEAS), 2003.
[9] C.K. Chui, B. Kao, and E. Hung, "Mining Frequent Itemsets from Uncertain Data," Proc. 11th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD), 2007.
[10] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2000.
[11] C. Kuok, A. Fu, and M. Wong, "Mining Fuzzy Association Rules in Databases," SIGMOD Record, vol. 27, no. 1, pp. 41-46, 1998.
[12] C.K.-S. Leung, Q.I. Khan, and T. Hoque, "Cantree: A Tree Structure for Efficient Incremental Mining of Frequent Patterns," Proc. IEEE Fifth Int'l Conf. Data Mining (ICDM), 2005.
[13] A. Lu, Y. Ke, J. Cheng, and W. Ng, "Mining Vague Association Rules," Proc. 12th Int'l Conf. Database Systems for Advanced Applications (DASFAA), 2007.
[14] L. Sun, R. Cheng, D.W. Cheung, and J. Cheng, "Mining Uncertain Data with Probabilistic Guarantees," Proc. 16th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2010.
[15] L. Wang, R. Cheng, S.D. Lee, and D. Cheung, "Accelerating Probabilistic Frequent Itemset Mining: A Model-Based Approach," Proc. 19th ACM Int'l Conf. Information and Knowledge Management (CIKM), 2010.
[16] Q. Zhang, F. Li, and K. Yi, "Finding Frequent Items in Probabilistic Data," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2008.