International Journal of Technology Research and Management
ISSN (Online): 2348-9006
Vol 6 Issue 3 March 2019
NCETRCSE-2019 Conference Proceedings, CSE Dept. LNCT, Indore

A Literature Survey of Diffrent Techniques for Web Log Mining

Varsha Sharma[1], Jyotshna Goyal[2]
Computer Science & Engineering, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Indore, India
Lakshmi Narain College of Technology, Indore[1, 2]
varshasharma517@gmail.com[1], JyotsanaGoyal13@gmail.com[2]

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 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.*

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 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.

2. TABLES, FIGURES AND EQUATIONS

2.1 Tables and Figures

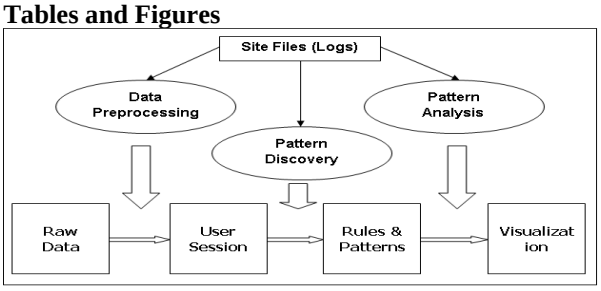

Figure 1: Web Usage Mining

2.2 Equations

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.

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 increase in Information and communication Technology the size of the databases created by the organizations due to the availability of low-cost storage and the evolution in the data capturing technologies is also increasing. It included retail, credit cards, insurance, banking and many others, for extracting the valuable data, it is necessary to explore the databases completely and efficiently. The Knowledge discovery in databases (KDD) helps to identifying precious information in such huge databases. Such valuable information can help the decision maker to make accurate future decisions. The KDD applications deliver measurable benefits including reduced cost of doing

business enhanced profitability and improved quality of service. That's why Knowledge Discovery in Databases has become one of the most active and exciting research areas in the database community.

[1] Defined the problem of finding the association rules from the database. In this section, the basic concepts of frequent pattern mining for discovery of interesting associations and correlations between item sets in transactional and relational database. Association rule mining is defined formally as follows:

Generally, a large number of these rules will be pruned after applying the support and confidence thresholds. Therefore most of the previous computations will be wasted. To overcome this problem and to improve the performance of the rule discovery algorithm, the association rule may be decomposed into two phases:

1. Generate the large item sets: the sets of items that have transaction support above a predetermined minimum threshold known as frequent Item sets.

2. Using the large item sets to generate the association rules for the database that has confidence above a predetermined minimum threshold.

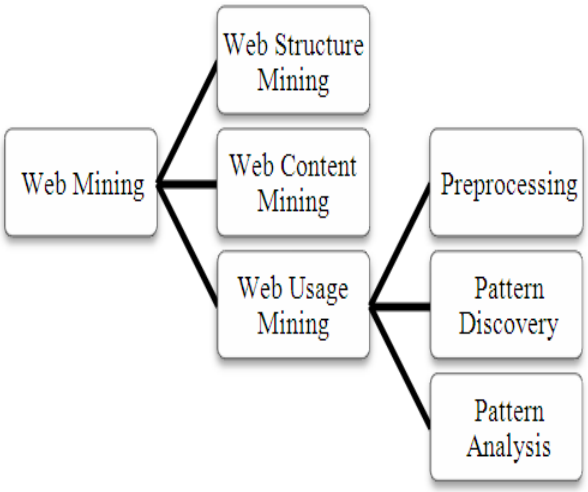

Figure 2: Web Mining [3]

The overall performance of mining association rules is depends primarily by the first step. The second step is easy. Once the large item sets are identified the corresponding association rules can be derived in straightforward manner. The main consideration of the thesis is First step i.e. to find the extraction of frequent item sets.

3. 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 Can Tree [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.

The developments of computed technology in last few decades are used to handle large scale data that includes large transaction financial data, bulletins, emails etc. Hence information has become a power that made possible for user to voice their opinions and interact. As a result revolves around the practice, data mining [17] come into sites. Association rule mining is one of the Data Mining techniques used in distributed database. In distributed database the data may be partitioned into fragments and each fragment is assigned to one site. The issue of privacy arises when the data is distributed among multiple sites and no other party wishes to provide their private data to their sites but their main goal is to know the global result obtained by the mining process. However privacy preserving data mining came into the picture. As the database is distributed, different users can access it without interfering with one another. In distributed environment, database is partitioned into disjoint fragments and each site consists of only one fragment.

Data can be partitioned in three different ways that is, like horizontally partitioned data, vertically partitioned data or mixed partitioned data.

Horizontal partitioning: - The data can be partitioned horizontally where each fragment consists of a subset of the records of relation R. Horizontal partitioning [20] [22] [23] [24] divides a table into several tables. The tables have been partitioned in such a way that query references are done by using least number of tables else excessive UNION queries are used to merge the tables sensibly at query time that can affect the performance.

Vertical partitioning: - The data can be divided into a set of small physical files each having the subset of the original relation, the relation is the database transaction that normally requires the subsets of the attributes.

Mixed partitioning: - The data is first partitioned horizontally and each partitioned fragment is further partitioned into vertical fragments and vice versa.

The market basket analysis used association rule mining [20][21] in distributed environment. Association rule mining [18][19][17] is used to find rules that will predict the occurrence of an item and based on the occurrences of other items in the transaction, search patterns gave association rules where the support will be counted as the fraction of transaction that contains an item X and an item Y and confidence can be measured in a transaction the item i appear in transaction that Also contains an item X

Privacy preserving distributed mining of association rule [21][17] for a horizontally partitioned dataset across multiple sites are computed. The basis of this algorithm [21][17] is the apriori algorithm that uses K-1 frequent sets. The problem of generation size of one item set may be carried out with secure computation on multiple sites by generating the candidate set, the pruning method, finding the union of large item set.

In [25], the authors conducted a comparative study to analyze the performance of FP-Growth & other frequent item set mining algorithms. The time complexity is the only performance metric used in this study. We have used the adult data set as the input data set.

4. 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.

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.
[17] Han, J. Kamber, M, "Data Mining Concepts and Techniques". Morgan Kaufmann, San Francisco, 2006.
[18] Agrawal, R., et al "Mining association rules between sets of items in large database". In: Proc. of ACM SIGMOD'93, D.C, ACM Press, Washington, pp.207-216, 1993.
[19] Agarwal, R., Imielinski, T., Swamy, A. "Mining Association Rules between Sets of Items in Large Databases", Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207-210, 1993.
[20] Srikant, R., Agrawal, R "Mining generalized association rules", In: VLDB'95, pp.479-488, 1994.
[21] Kantarcioglu, M., Clifton, C, "Privacy-Preserving distributed mining of association rules on horizontally partitioned data", In IEEE Transactions on Knowledge and Data Engineering Journal, IEEE Press, Vol 16(9), pp.1026-1037, 2004.
[22] Sugumar, Jayakumar, R., Rengarajan, C "Design a Secure Multi Site Computation System for Privacy Preserving Data Mining". International Journal of Computer Science and Telecommunications, Vol 3, pp.101-105. 2012.
[23] N V Muthu Lakshmi, Dr. K Sandhya Rani ,"Privacy Preserving Association Rule Mining without Trusted Site for Horizontal Partitioned database", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, pp.17-29, 2012.
[24] N V Muthu lakshmi, Dr. K Sandhya Rani, "Privacy Preserving Association Rule Mining in Horizontally Partitioned Databases Using Cryptography Techniques", International Journal of Computer Science and Information Technologies(IJCSIT), Vol. 3 (1) , PP. 3176 – 3182, 2012.
[25] Pramod S, O P Vyas, "Survey on Frequent Item Sets Mining Algorithms", International Journal of Computer Applications (IJCA), Vol 1(15), PP. 86-91.