A LITERATURE REVIEW OF VARIOUS TECHNIQUES FOR MINING FREQUENT PATTERNS FROM A STANDARD DATA SET

Chetan Bhagat[1], Durjoy Datta[2], Ravinder Singh[3]
[1] Author & Motivational Speaker; Alumnus of IIT Delhi and IIM Ahmedabad, India
[2] Author and Screenwriter; Co-founder of Grapevine India, India
[3] Author and Publisher; MBA from ISB Hyderabad; Founder of Black Ink Publications, India

ABSTRACT

Visit thing set mining has been a heart most loved topic for information digging specialists for over 10 years. A lot of writing has been committed to this exploration and enormous advancement has been made, going from productive and versatile calculations for incessant item set mining in exchange databases to various research wildernesses, for example, consecutive example mining, organized example mining, relationship mining, acquainted order, and continuous example based bunching, just as their expansive applications. In this paper, a writing survey of different most recent procedures for mining regular things from an exchange information base are introduced in basic way.

Keywords: *Data Mining, Frequent Pattern Mining, Support, Confidence, Apriori.*

I. INTRODUCTION

Information mining [1] is the way toward separating concealed examples from information. As more information is accumulated, with the measure of information multiplying at regular intervals, information mining is turning into an undeniably imperative apparatus to change this information into learning. It is regularly utilized in a wide scope of uses, for example, promoting, misrepresentation location and logical disclosure. Information mining can be connected to informational collections of any size, and keeping in mind that it tends to be utilized to reveal shrouded designs, it can't uncover designs which are not effectively exhibit in the informational collection. The found knowledge [2][3] can be utilized from various perspectives in comparing applications. For instance, recognizing the as often as possible showed up sets of things in a retail database can be utilized to improve the basic leadership of stock situation or deals advancement. Finding examples of client perusing and buying (from either client records or Web traversals) may help the displaying of client practices for client maintenance or customized administrations. Given the ideal databases, regardless of whether social, value-based, spatial, fleeting, or sight and sound ones, we may get helpful data after the learning revelation process if suitable mining strategies are utilized. A typical process of knowledge discovery in databases is illustrated in Fig. 1.

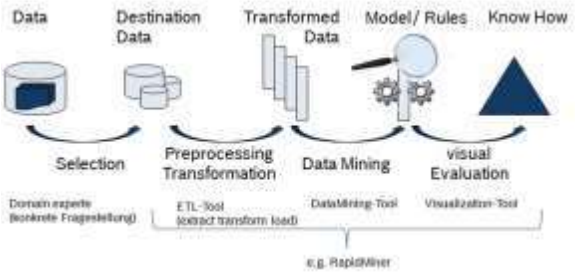

Fig. 1. The process of knowledge discovery in databases [1]

Knowledge discovery in databases is a complex process, which covers many interrelated steps. Key steps in the knowledge discovery process are:

- Data Cleaning: remove noise and inconsistent data.
- Data Integration: combine multiple data sources.
- Data Selection: select the parts of the data that are relevant for the problem.
- Data Transformation: transform the data into a suitable format.

- Data Mining: apply data mining algorithms and techniques.
- Pattern Evaluation: evaluate whether they found patterns meet the requirements
- Knowledge Presentation: present the mined knowledge to the user (e.g., Visualization).

The key step of association mining is frequent item set (pattern) mining which is to mine all item sets satisfying user specified minimum support [5] 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:

- Generate the large item sets: the sets of items that have transaction support above a predetermined minimum threshold known as frequent Item sets.
- Using the large item sets to generate the association rules for the databases that have confidence above a predetermined minimum threshold.

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.

II. LITERATURE SURVEY

The most popular frequent item set mining called the FP-Growth algorithm was introduced by [5]. The main aim of this algorithm was to remove the bottlenecks of the Apriori-Algorithm in generating and testing candidate set. The problem of Apriori algorithm was dealt with, by introducing a novel, compact data structure, called frequent pattern tree, or FP-tree then based on this structure an FP-treebased pattern fragment growth method was developed. FP-growth uses a combination of the vertical and horizontal database layout to store the database in main memory. Instead of storing the cover for every item in the database, it stores the actual transactions from the database in a tree structure and every item has a linked list going through all transactions that contain that item. This new data structure is denoted by FP-tree (Frequent- Pattern tree) [4]. Essentially, all transactions are stored in a tree data structure.

In 2009 the authors Ling Chen et al. [6] proposed that the fading factor model can be used to compute the frequent itemsets. This fading factor lm contributes more to the recent items than the older. The fading factor ranges between $0<lm<1$, where lm is frequency. Value near to 1 is considered to be most frequent item. This technique has two major advantages. Firstly, It takes the all old data items based on the frequency and the other is changes in frequency varies by a small values. In 2009 the work done by Cai-xia Meng [7] proposed the efficient algorithm for mining frequent item sets over a high speed data streams. The frequent pattern mining algorithms pose two steps. This involves calculations behind the arrival of every frequency of new item sets and formatting them into the output. In this algorithm these two steps are blended together to reduce too much of time that arrives in Lossy Counting (LC) and FDPM. That loss of data occurs in other algorithms is sorted out here. This method uses Defer Counting (DC) method which delays the frequency calculation and provides gap between step 1 and step 2 to avoid the transaction missing problem. In step 1the DC includes the information that achieved the threshold value. In step 2 it frequent pattern from the stored information, the data structures used are List and Trie. The List is for the frequent items and Trie for the frequent item sets. Data structure List is used with three-fields stating, the id of each item, the frequency of that item and error between the actual and the estimated frequency. The Trie composed of two fields in which one points to counter in the list. The other field is to compute the frequency of the item sets.

In 2010 author Varun Kumar et al. [8] proposed this algorithm which has an ability to hold the various sizes of the batch rather than the fixed in other. The time has been fixed for segregating the Batches. In the previous algorithms the infrequent items were removed. 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 is clearly solved by this paper. Proposed work uses an extension of trie structure with the logtime window as its data structure. Method constitute three columns namely tilted-time, frequency and 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 and time. In 2009work of Sonali Shukla et al. [9] proposed this algorithm with the regression based methodology to find out the frequent item sets continuously that are streaming regularly. In this method the 2-Dimensional stream data is preprocessed and converted into sampling value. Regression analysis is carried out with these values. Method bags the data using sliding window

model and then it applies FIM- 2DS algorithm to compute with the item set. It is processed to the sampling value for the further process with least square method. Every data is paired (mi, ni) to find the arrival time difference between them. Data is calculated like t, t-1, t-2, t-3…tn. if the pair is (m,n) then it is mean that m is an independent variable of n and n is dependent variable on m. By taking the help of the pairs of Data Sets dependent and independent variable values are calculated. After that the regression line is drawn from fragment slope values. Also the regression analysis is also used to find out the functional relationships between the paired data items. In 2010 the author ZHOU Jun et al. [10] proposed this algorithm by considering the space as an important factor. Authors used an improved LRU (Least Recently Used) based algorithm. Proposed algorithm omits the infrequent items before taken for the processing. Method increases the stability and the performance. Method is used to find out the frequent items as well as the frequency of those items.

In 2012 work of Yong-gong Ren et al. [11] proposed this algorithm in order to predict the future data based on the new method called AMFP-Stream known as Associated Matrix Frequent Pattern-Stream. It predicts the frequently occurred item sets over data streams efficiently. Proposed work also has a capability to predict that which item set will be frequent with high potential. Method takes the data in the form of 0-1 matrix and then it updates the values by doing logical bit operations. Then on this it will find out the item sets that will frequently occur in the future. Method uses the associated matrix for the further manipulation. Experimental results says that this algorithm is how much feasible. In 2011 author Mahmood Deypir et al. [12] proposed this algorithm based on the different kind of sliding window based model. I Method don‟t need entire data that are in streaming. Method takes an advantage of the already existing item sets. To enhance 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.

Frequent pattern mining is a favorite topic of many researchers across the globe. Frequent item set mining has a wide range of real world applicat ions. It affects decision making of many industries. This paper presented a comprehensive survey of latest techniques for mining frequent patterns from a standard data set. This review will be useful for future researchers of frequent pattern mining.

REFERENCES

1. *Tan P.-N., Steinbach M., and Kumar V. "Introduction to data mining, Addison Wesley Publishers". 2006*
2. *Nizar R.Mabrouken, C.I.Ezeife. Taxonomy of Sequential Pattern Mining Algorithm". In Proc. in ACM Computing Surveys, Vol 43, No 1, Article 3, November 2010.*
3. *A.M.Said, P.P.Dominic, A.B. Abdullah. "A Comparative Study of FP-Growth Variations". In Proc. International Journal of Computer Science and Network Security, VOL.9 No.5 may 2009.*
4. *Brin.S, Motwani. R, Ullman. J.D, and S. Tsur. "Dynamic itemset counting and implication rules for market basket analysis". In Proc. ACMSIGMOD Int'l Conf. Management of Data (SIGMOD), May 1997, pages 255–264.*
5. *C. Borgelt. "An Implementation of the FP growth Algorithm". Proc. Workshop Open Software for Data Mining, 1–5.ACMPress, New York, NY, USA 2005.*
6. *Ling Chen, Shan Zhang,Li Tu, "An Algorithm for Mining Frequent Items on Data Stream Using Fading Factor".33rd Annual IEEE International Computer Software and Applications Conference.172-179,2009.*
7. *Cai-xia Meng, An Efficient Algorithm for Mining Frequent Patterns over High Speed Dat Streams. World Congress on Software Engineering,nIEEE 2009, 319-323.*
8. *Varun Kumar,Rajanish Dass.Proceedings of the 43rd Hawaii International Conference on System Sciences, 2010 IEEE, 978-0-7695-3869-3.*
9. *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.*
10. *ZHOU Jun, CHEN Ming, XIONG Huan A More Accurate Space Saving Algorithm for Finding the Frequent Items.IEEE-2010.*
11. *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.*
12. *Mahmood Deypir, Mohammad Hadi Sadreddini, A New Adaptive Algorithm for Frequent Pattern Mining over Data Streams, ICCKE,2011, 230-235 FLEXChip Signal Processor (MC68175/D), Motorola, 1996.*