International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal | www.ijtsrd.com
ISSN No: 2456 - 6470 | Volume - 2 | Issue – 6 | Sep – Oct 2018

A Compression Based Methodology to Mine All Frequent Items

Rajendra Chouhan[1], Khushboo Sawant[2], Dr. Harish Patidar[3]
[1]M.Tech Scholar, [2]Assistant Professor, [3]HOD, Associate Professor
Department of Computer Science and Engineering, LNCT, Indore, Madhya Pradesh, India

ABSTRACT
Data mining is not new. People who first discovered how to start fire and that the earth is round also discovered knowledge which is the main idea of Data mining. Data Mining, also called knowledge Discovery in Database, is one of the latest research area, which has emerged in response to the Tsunami data or the flood of data, world is facing nowadays. It has taken up the challenge to develop techniques that can help humans to discover useful patterns in assive data. One such important technique is frequent pattern mining. This paper will present an compression based technique for mining frequent items from a transaction data set.

Keyword: *Data Mining, KDD Process, Frequent Pattern Mining, Minimum support, Data Compression.*

1. INTRODUCTION
The term data mining refers to the extraction or 'mining' of valuable knowledge from large amounts of data, in analogy to industrial mining where small sets of valuable nuggets (e.g. gold) are extracted from a great deal of raw material (e.g. rocks).

It is the combination of Multiple Disciplines. Figure 1 shows the different disciplines that take part in data mining.

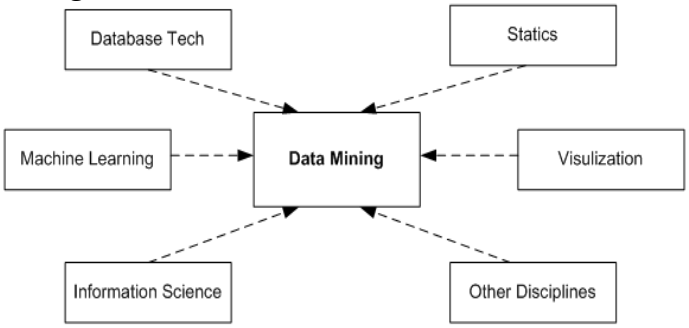

Figure 1 Show the Multiple Disciplines for data mining

Even before technologies were used for Data mining, statisticians were using probability and regressing techniques to model historical data [1]. Today technology allows to capture and store vast quantities of data. Finding and summarizing the patterns, trends, and anomalies in these data sets is one of the big challenges in today's information age. "With the unprecedented growth-rate at which data is being collected [2] and stored electronically today in almost all fields of human endeavor, the efficient extraction of useful information from the data available is becoming an increasing scientific challenge and a massive economic need" [3].

In Data Mining process, selection and transformation are forms of preprocessing, where one selects part of the complete database and possibly transforms it into a certain form required for the next step. Often data cleaning and data integration are also part of this initial phase of data preparation. The resulting data is the input for the data mining phase, which in its turn results in discovered patterns. The interesting patterns are then presented to the user. As these patterns can be stored as new knowledge in a knowledge base, they can, in turn, again be considered as input for another knowledge discovery process [4]. All patterns which have support no less than the user-specified min sup value are mined as frequent patterns.

TID	List of Item_Ids
T100	I1, I2, I5
T200	I2, I4
T300	I2,I3
T400	I1,I2, I4
T500	I1, I3
T600	I2,I3
T700	I1, I3
T800	I1, I2, I3,I5
T900	I1, I2, I3

Item set	Support count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

Figure 2: Frequent Pattern Mining

Frequent pattern mining is used to prune the search space and limit the number of association rules being generated. Many algorithms discussed in the literature use "single min sup framework" to discover the complete set of frequent patterns. The reason for the popular usage of "single min sup framework" is that frequent patterns discovered with this framework satisfy downward closure property, i.e., all non-empty subsets of a frequent pattern must also be frequent. The downward closure property makes association rule mining practical in real-world applications [7] [8]. The two popular algorithms to discover frequent patterns are: Apriori [6] and Frequent Pattern-growth (FP-growth) [9] algorithms. The Apriori algorithm employs breadth-first search (or candidate-generate-and-test) technique to discover the complete set of frequent patterns. The FP-growth algorithm employs depth-first search (or pattern-growth) technique to discover the complete set of frequent patterns. It has been shown in the literature that FP-growth algorithm is relatively efficient than the Apriori algorithm [9].

2. RELATED WORK

As a result additional interestingness measures, such as lift, correlation and all-confidence, have been proposed in the literature to address the interestingness of an association rule [10]. Each measure has its own selection bias that justifies the rationale for preferring a set of association rules over another [11]. As a result, selecting a right interestingness measure for mining association rules is a tricky problem. To confront this problem, a framework has been suggested in for selecting a right measure. In this framework, authors have discussed various properties of a measure and suggested to choose a measure depending on the properties interesting to the user.

Generally, frequent-pattern mining results in a huge number of patterns of which most can be found to be insignificant according to application and/or user requirements. As a result, there have been efforts in the literature to mine constraint-based and/or user-interest based frequent patterns [12], [13], [14], [15]. In recent times, temporal periodicity of frequent patterns has been used as an interestingness criterion to discover a class of user-interest based frequent patterns, called periodic-frequent patterns [16].

3. PROPOSED METHOD

STEP1: START

STEP2: INPUT TRANSACTION DATA SET & MINIMUM SUPPORT THRESHOLD

STEP3: FIRST THE PROPOSED ALGORITHM SCANS THE TRANSACTION DATA BASE AND CALULATES THE SUPPORT OF EACH SINGLE SIZE ITEM.

STEP4: IN THIS STEP A LIST OF FREQUENT ITEM AND INFREQUENT ITEM IS PREPARED ON THE BASIS OF MINIMUM SUPPORT THRESHOLD.

IF AN ITEM IS HAVING SUPPORT GREATER THAN THE MINIMUM SUPPORT THRESHOLD THEN ITEM IS PLACED IN FREQUENT ITEM LIST AND ALSO IN EXPANSION LIST. OTHERWISE IT IS PLACED IN INFREQUENT ITEM LIST

STEP5: REMOVE THE TRANSACTION WHICH DOES NOT CONTAIN ANY FREQUENT ITEM

STEP 6: IN THIS STEP, ALL THE MEMBERS OF THE INFREQUENT ITEM LIST ARE REMOVED FROM THE TRANSACTION DATA BASE BECAUSE THEY WILL NOT APPEAR IN ANY FREQUENT ITEM SET. IN THIS WAY, THE ORIGINAL TRANSACTION DATA BASE IS CONVERTED INTO REDUCED SIZE DATA BASE. NOW THIS REDUCED DATA BASE WILL BE USED IN THE CALCULATION OF LARGER SIZE FREQUENT ITEM SETS.

STEP7: WHILE EXPANSION LIST IS NOT EMPTY PERFORM LEFT EXPANSION OF SMALLER SIZE ITEMS TO GENERATE HIGHER SIZE ITEMS AND THEN REPEAT STEP 4 FOR THEM

OR

PERFORM RIGHT EXPANSION OF ELEMENTS AND THEN REPEAT STEP4 FORTHEM.

STEP8: WRITE THE LIST OF FREQUENT ITEM SETS

STEP9: STOP

4. RESULT

The proposed and existing algorithms are implemented in Java. The retail data set is used as the input data set. The time consumed by both algorithms (For minimum support 40 percent) is shown below

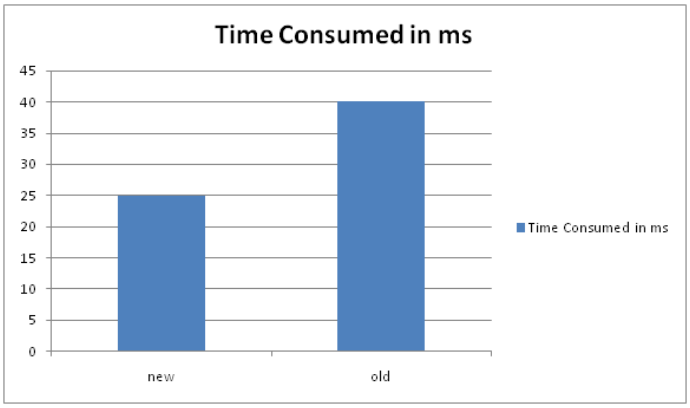

Figure 3: Time Consumption Comparison

5. CONCLUSION:

Frequent pattern mining has a wide range of real world applications. That's why it is one of the most favorite topic of research. Frequent mining helps in mining of items which are worthy. This paper proposed an updated method to find frequent item sets from a transaction data set. The proposed method makes use of data compression for data reduction. Useless items are eliminated in the initial stage of the mining process.

REFERENCES

1. Groth Robert. "Data Mining: A Hands-on Approach for Business Professionals". Prentice Hall PTR, 1998.
2. Witten Ian and Eibe. Frank, "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations". San Francisco: Morgan Kaufmann Publishers, 2000.
3. Zaki Mohammed and Ho Ching-Tien, "Large-Scale Parallel Data Mining". Berlin: Springer, 2000.
4. From Data Mining to Knowledge Discovery in Databases Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth.
5. Omiecinski E., "Alternative interest measures for mining associations in databases", IEEE Trans. Knowl. Data Eng., 15(1):57–69, 2003.
6. Agrawal R., Imielinski T., and Swami A. N., "Mining association rules between sets of items in large databases", In SIGMOD Conference, pages 207–216, 1993.
7. Chen Chun-Hao , Hong Tzung-Pei and Tseng Vincent S. , "Genetic-fuzzy mining with multiple minimum supports based on fuzzy clustering", 2319-2333, Springer 2011.
8. Weimin Ouyang , Qinhua Huang , "Mining direct and indirect fuzzy association rules with multiple minimum supports in large transaction databases", 947 – 951, IEEE 2011.
9. Han J., Pei J., Yin Y., and Mao R., "Mining frequent patterns without candidate generation: A frequent-pattern tree approach", Data Min. Knowledge. Discovery. 8(1):53–87, 2004.
10. Haiying Ma,Dong Gang, "Generalized association rules and decision tree", Page(s) 4600 – 4603, IEEE, 2011.
11. Kim Hyea Kyeong, Kim Jae Kyeong , "A product network analysis for extending the market basket analysis", Pages 7403–7410, Elsevier 2012.
12. Tan P.-N., Kumar V., and Srivastava J. "Selecting the right interestingness measure for association patterns". In KDD, pages 32–41, 2002.
13. Vaillant B., Lenca P. , and Lallich S. "A clustering of interestingness measures". Pages 290–297. Springer, 2004.
14. Hu T., Sung S. Y., Xiong H., and Fu Q., "Discovery of maximum length frequent itemsets". Inf. Sci., 178:69–87, January 2008.
15. Schmidt Jana and Kramer Stefan, "The Augmented Itemset Tree: A Data Structure for Online Maximum Frequent Pattern Mining", pp 277-291 Springer 2011.
16. Tanbeer S. K., Ahmed C. F., Jeong B.-S., and Lee Y.-K., "Discovering periodic-frequent patterns in transactional databases". In PAKDD '09: Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, pages 242–253, Berlin, Heidelberg, Springer-Verlag 2009.