International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 10, Issue 3 (March 2014), PP.01-05

Profit and Preferable Itemsets- A Study

Goguri. Rashmitha[1], L.Vandana[2], Saturi. Rajesh[3], Mahipal Reddy Pulyala[4]
[1,2,3]*Assistant Professor, Dept Of Cse, Nalla Narasimha Reddy Educational Society's*
Group Of Institutions, Hyderabad, Ap, India.
[4]*Assistant Professor,Dept Of Cse, Vaagdevi College Of Engineering,Bollikunta,Warangal,Ap,India.*

Abstract:- High utility itemset mining is a research area of utility based descriptive data mining, aimed at finding itemset that contributes most to the total utility. A specialized form of high utility itemset mining is utility-frequent itemset mining, which – in addition to subjectively defined utility – also takes into account itemset frequencies. This paper presents a profitable and preferable within the given utility and support constraints threshold. An extensive performance study using both synthetic and real data sets is reported to verify the effectiveness and efficiency of proposed algorithms.

Keywords:- item-set, frequent, preferable, data mining;

I. INTRODUCTION

Data flow analysis is an emerging topic extensively studied in recent decade. A data stream is a continuously ordered sequence of transactions that arrives sequentially in real-time manner. There are many applications of data stream mining, such as knowledge discovery from online e-business or transaction flows, analysis of network flows, monitoring of sensor data, and so on. Different from traditional databases, data streams have some special properties, namely continuous, unbounded, high speed and time-varying data distribution. Therefore, some limitations are posed in data stream mining as follows [1-9]. First, since the infinite transactions could not be stored, multi-scan algorithms are no more allowed. Second, in order to capture the information of the high speed data streams, the algorithms must be as fast as possible [1-9].

Otherwise, the accuracy of the mining results will be reduced. Third, the data distribution of data streams should be kept to avoid concept drifting problem. Fourth, incremental processes are needed for mining data streams in order to make processing with the old data as little as possible. Therefore, efficient one-pass methods and compact data structures for characterizing the data flow are needed [1-9]. In recent years, utility mining emerges as a new research issue, which is to find the itemset with high utilities, i.e., the itemset whose utility values are greater than or equal to the user-specified minimum utility threshold. There exist rich applications of utility mining, such as business promotion, webpage organization and catalog design. Unlike traditional association rule mining, utility mining can find profitable itemset which may not appear frequently in databases. By the advantages mentioned above, we can see that it is important to push utility mining into data stream mining, such as for the streaming data of the chain hypermarkets. In view of this, we aim at finding maximal utility itemset from data streams with the purpose of reducing the number of patterns in this paper. This is motivated by the observation that there exists no study that explores advanced utility itemset patterns like maximal utility itemset

II. LITERATURE

Frequent itemset mining as a research area came into being in the nineties. The seminal paper appeared in 1994 [10-20]. In the subsequent decade numerous papers were published. The basic problem in *frequent itemset mining* is, given a series of sets, to find all subsets that are contained in at least *minsup* of them; here *minsup* is a user-specified threshold. The problem of frequent itemset mining plays an important role in several data mining fields [10-20], such as association rules [10-20]warehousing [10-20], correlations [10-20] and classification [10-20]. The subject is also related to rough sets [10-20] and logical analysis of data [10-20]. Moreover, frequent item-sets have many application areas, amongst others customer relationship management, fraud detection, product assortment decisions [10-20] episode mining [10-20], functional dependency discovery [10-20], etc. In the literature on frequent item-sets the algorithms are usually studied from a practical viewpoint. Almost any paper focuses on speed. To achieve an optimal running time, specific implementation tricks are applied. In the current paper, the theory of frequent item-sets is discussed.

III. SYSTEM STRUCTURE

Skyline queries have been studied since 1960s in the theory field where skyline points are known as Pareto sets and admissible points or maximal vectors. However, earlier algorithms such as are inefficient when there are many data points in a high-dimensional space. Skyline queries in database were first studied by Borzsonyi in 2001[21]. After that, various techniques were proposed to accelerate the computation of skyline and its variations. Here, we briefly summarize some of them. Some representative methods include a bitmap method a nearest neighbor (NN) algorithm and a branch-and-bound skylines (BBS) method. Disadvantages of this are the existing systems take much time complexities. These systems directly cannot be applied for finding the profitable products and popular products.

In this paper, we identify and tackle the problem of finding top-k preferable products, which has not been studied before. We study two instances of preferable products, namely profitable products and popular products. We propose methods to find top-k profitable products and top-k popular products efficiently. An extensive performance study using both synthetic and real data sets is reported to verify its effectiveness and efficiency. As future work, we will study other instances of the problem of finding top-k preferable products by setting the utility function to other meaningful objective functions. One promising utility function is the function which returns the sum of the unit profits of the selected products multiplied by the number of customers interested in these products and the structure is shown in Fig 1. Advantages of this are, the proposed approach takes lesser computational cost and directly can be applied for finding profitable and popular products.

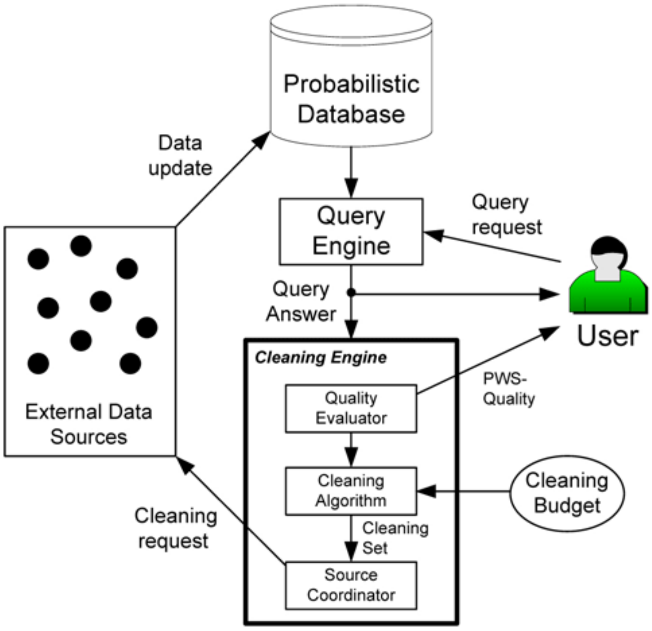

Fig 1 System structure

IV. MODULES IMPLEMENTED

The proposed system consists of set of stages, these are discussed here

Popular Products Dataset: The dataset should be prepared based on the existing products in the market by various vendors. For example: For experimenting on car products, the existing cars, features, cost etc. has to be gathered from different vendors

Frequent Feature Set: Frequent feature set identification on the dataset of popular products. The high utility mining refers to the feature set extraction, by satisfying certain conditions. The condition in the work is high profitability. Considering the high profitability, identify the feature set that occurred more frequently by the other vendors.

Finding the Smaller Frequency Sets: The Frequency set provides various levels of sets. Example:level1: with single item, level2: with two combinations of items, level3: with three combinations of items. Have to build the least sets to build a product.

Price Correlation: For the price Correlation among the optimal feature set ‘t’ and the existing feature and their costs. Here NP-Hard with Greedy Approach is applied.

Finding Top-k Popular Products: For this problem the build set ‘t’ in the previous module is the input. The set is cross checked against the user interest ‘dataset2’ which is collected from the home website log.

Web Log Preprocessing: The access log can be obtained from the web server.The log consists of all the request sends by the users. The attribute values can be obtained from the URI of the web logs.

Converting Features to Binary Access Table: in these modules the set‘t’ to be converted to ‘0’ and ‘1’. The frequent item set cross checking will be done in the Brute force approach with linear fashion. If ‘t’ is popular then sustains in the market. The process is iterative until we get the top-k sets, which provide profitability.

V. EXAMPLE OF ITEMSETS

This section describes a sample execution of the some of the algorithm. The transaction data of the transaction database D are given in Table 1; the minimum support is 0.4; n=5 is the number of items, and m=5 is the number of transactions. Therefore, the minimum support number minsupsh=2. The transaction database D is transformed into the Boolean matrix A_{5*5} as shown in the Figure 2.

Transaction data of the transaction database d

TID	Item sets
T1	**A,D**
T2	**B,C,E**
T3	**A,B,C,E**
T4	**B,E**
T5	**A,B,C**

$$\begin{array}{cc} & \begin{array}{ccccc} \mathbf{A} & \mathbf{B} & \mathbf{C} & \mathbf{D} & \mathbf{E} \end{array} \\ \begin{array}{c} \mathbf{T1} \\ \mathbf{T2} \\ \mathbf{T3} \\ \mathbf{T4} \\ \mathbf{T5} \end{array} & \begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix} \end{array}$$

Figure 2. The Boolean matrix A_{5*5}

We compute the sum of the element values of each column in the Boolean matrix A_{5*5} and the set of frequent 1-itemset is: L1= {{A},{B},{C},{D}}

The fourth column of the Boolean matrix A_{5*5} is deleted because the support number of item D is smaller than the minimum support number 2. We then compute the sum of the element values of each row in the Boolean matrix and delete all rows where the sum of the element values is smaller than 2. Finally, the Boolean matrix A4*4 is generated as shown in figure 3.

$$\begin{array}{cc} & \begin{array}{cccc} \mathbf{A} & \mathbf{B} & \mathbf{C} & \mathbf{E} \end{array} \\ \begin{array}{c} \mathbf{T2} \\ \mathbf{T3} \\ \mathbf{T4} \\ \mathbf{T5} \end{array} & \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \end{array}$$

Figure 3 The Boolean matrix A_{4*4}

The operation of 2-supports is executed for the all columns of the Boolean matrix A_{4*4}, and the set of frequent 2-itemset is: L2={{A,B},{A,C},{B,C},{B,E},{C,E}}

In pruning the Boolean matrix A_{4*4} by the set of frequent 2-itemsets L2 , the third row of the Boolean matrix A_{4*4} is deleted because sum of its element values is smaller than 3. Finally, the Boolean matrix A_{3*4} is generated as shown in figure 3.

$$\begin{array}{cc} & \begin{array}{cccc} \mathbf{A} & \mathbf{B} & \mathbf{C} & \mathbf{E} \end{array} \\ \begin{array}{c} \mathbf{T2} \\ \mathbf{T3} \\ \mathbf{T5} \end{array} & \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \end{array}$$

Figure 4The Boolean matrix A_{3*4}

The operation of 3-supports is executed for all columns of the Boolean matrix A_{3*4}, and the set of frequent 3-itemset is: L3= {{A,B,C},{B,C,E}}

According to Proposition 3, the proposed algorithm is terminated because there are two frequent 3-itemsets in the set of frequent 3-itemset L3.

VI. COMPARATIVE STUDY

First, we performed our method as well as the baseline method on the literature repositories. We extracted 10 common topics from the six sequences. For each topic, 10 topical words with highest probability p(w|z). We can see that all topics extracted by our method (sync) were meaningful and easy to understand. For example, #7 includes research topics such as data mining, high-dimensional/multidimensional data, data warehouse, association rule, workflow, etc., while #10 includes sensor network, privacy preserving, classification, ontology, top-k query, etc. All of these topical words accurately suggest most important research topics in the database area. Comparing the topics extracted by our method to those by the baseline method (no sync), we can see that our method provided highly discriminative topics. As a contrast, the baseline method suffered from the synchronism in the sequences and extracted many duplicated topical words (see Fig. 4). In asynchronous sequences, documents related to different topics may be indexed by the same time stamp, and documents related to the same topic may appear at different time stamps. As a result, common topics discovered by the conventional method contain redundant information, whereas our method is able to fix the synchronism and discover highly discriminative topics. To further prove that our time synchronization technique helped to generate more discriminative topics, we computed the pair wise KL-divergence between topics as follows:

$$KL(z_1, z_2) = \sum_{\mathbf{w}} p(w|z_1) \log \frac{p(w|z_1)}{p(w|z_2)}.$$

Note that larger KL-divergence indicates that the two topics are more discriminative to each other and 0 divergence means that two topics are identical. We present the results, where darker blocks mean smaller KL-divergence values. We can see that our method extracted much more discriminative topics than those extracted by the baseline method. As discussed above, this was due to the fact that our method successfully fixed the synchronism in the data set. Note that we used p(t|z), which can be computed from p(z|t)

VII. CONCLUSIONS

Data mining can be used extensively in the enterprise based applications with business intelligence characteristics to provide a deeper kind of analysis while meeting strict requirements for administration management and security. Business intelligence is information about a company's past performance that is used to help predict the company's future performance. In this paper, study two instances of preferable products, namely profitable products and popular products. This study will help the researchers who are begun to study in the related areas.

REFERENCES

[1]. Chan, R., Yang, Q. and Shen, Y. Mining high utility item-sets. In Proc. of Third IEEE Int'l Conf. on Data Mining, Nov., 2003, 19-26.

[2]. Chi, Y., Wang, H., Yu, P. S. and Muntz, R. R. Moment: maintaining closed frequent item-sets over a stream sliding window. In Proc. of the IEEE Int'l Conf. on Data Mining, 2004.

[3]. Gaber, M. M., Zaslavsky, A. B. and Krishnaswamy, S. Mining data streams: a review. ACM SIGMOD Record, Vol.34, No.2, 2005, 18-26.

[4]. Golab, L. and Ozsu, M. T. Issues in data stream management. In ACM SIGMOD Record, Vol. 32, No. 2, 2003.

[5]. Gouda, K. and Zaki, M. J. efficiently mining maximal frequent item-sets. In Proc. of the IEEE Int'l Conf. on Data Mining, 2001.

[6]. Han, J., Pei, J., and Yin, Y. Mining frequent patterns without candidate generation. In Proc. of the ACM-SIGMOD Int'l Conf. on Management of Data, 2000, 1-12.

[7]. Jiang, N. and Gruenwald, L. CFI-Stream: mining closed frequent item-sets in data streams. In Proc. of the Utility-Based Data Mining Workshop, ACM KDD, 2006.

[8]. Jiang, N. and Gruenwald, L. Research issues in data stream association rule mining. In ACM SIGMOD Record, Vol. 35, No. 1, 2006.

[9]. S.Rajesh, “Research methodologies for data mining”, GJCAT, Vol 2 (3), 2012, 1131-1137

[10]. Lee, D. and Lee, W. Finding maximal frequent item-sets over online data streams adaptively. In Proc. of 5th IEEE Int'l Conf. on Data Mining, 2005.

[11]. Agrawal, R. and R. Srikant (1994), Fast algorithms for mining association rules, in: Proceedings 20th International Conference on Very Large Data Bases (VLDB'94), Morgan Kaufmann.

[12]. S.Rajesh, “Pattern discovery through web mining”, Global Journal of Computer Application and Technology, GJCAT, Vol 2 (3), 2012, 1093-1098, ISSN: 2249-1945.

[13]. Tan, P.-N., M. Steinbach, and V. Kumar (2006), Introduction to data mining, Addison Wesley.

[14]. Wu, C. (2006), Applying frequent itemset mining to identify a small itemset that satisfies a large percentage of orders in a warehouse, Computers and Operations Research 33, 3161-3170.

[15]. Brin, S., R. Motwani, and C. Silverstein (1998), beyond market baskets: Generalizing association rules to correlations, Data Mining and Knowledge Discovery 2, 39-68.

[16]. Hu, K., Y. Lu, L. Zhou, and C. Shi (1999), Integrating classification and association rule mining: A concept lattice framework, in: Proceedings of the Seventh International Workshop on New Directions in Rough Sets, Data Mining and Granular Soft Computing, Springer.

[17]. Kosters, W. A., E. Marchiori, and A. J. Oerlemans (1999), Mining clusters with association rules, in: Proceedings of the Third International Symposium on Advances in Intelligent Data Analysis (IDA'99), Springer Lecture Notes in Computer Science 1642

[18]. Pawlak, Z. (2000), Rough Sets, Theoretical Aspects of Reasoning about Data, Morgan Kaufman.

[19]. Boros, E., P. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz, and I. Muchnik (2000), An Implementation of Logical Analysis of Data, IEEE Transactions on Knowledge and Data Engineering 12, 292-306.

[20]. Brijs, T., G. Swinnen, K. Vanhoof, and G. Wets (1999), Using association rules for product assortment decisions: A case study, in: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 1999).

[21]. Huhtala, Y., J. KÄarkkÄainen, P. Porkka, and H. Toivonen (1999), TANE: An efficient algorithm for discovering functional and approximate dependencies, The Computer Journal 42, 100-111.

[22]. Mannila, H., H. Toivonen, and A. Verkamo (1995), discover frequent episodes in sequences, in: Proceedings of the First International Conference on Knowledge Discovery and Data Mining.

[23]. S. Borzsonyi, D. Kossmann, and K. Stocker, The Skyline Operator. In ICDE 2001

ABOUT AUTHORS

G. Rasmitha Reddy working in Nalla Narasimha Reddy Education Society’s Group of Institutions as Assistant Professor, Computer Science Department. Having 3+ Years of experience in teaching and published papers in various Journals.

L.Vandana working in Nalla Narasimha Reddy Education Society’s Group of Institutions as Assistant Professor, Computer Science Department. Having 14 Years of experience in teaching and published papers in various Journals.

S.RAJESH, Asst.prof, Dept of CSE, NNRESGI, Hyd, AP, INDIA. Had 6 years of experience in teaching and Attended and presented papers in international and national conferences like IEEE, Springer etc. And published many papers in peer-reviewed national and international journals. Has done M.tech from university college of engineering, osmania university. His research interests include cloud computing, data mining and data bases

Mr Mahipal Reddy Pulyala, Post Graduated in Computer science and Engineering (M.Tech) , JNTUH , Hyderabad in 2011 and Post Graduated in Master of Computer Applications(MCA),JNTUH , Hyderabad in 2009, He is working as an Assistant Professor in Department of Computer Science & Engineering in Vaagdevi College Of Engineering, Warangal Dist, AP, and India. He has 4 years of Teaching Experience. His Research Interests Include Data Mining