Efficient Pattern Mining Algorithms: A Study

Amit Verma[1], Raman Kumar[2]
[1]Department of Computer Science and Engineering,
IKGPTU, Kapurthala, Punjab, India
E-mail: [1]verma0152@gmail.com, [2]er.ramankumar@aol.in

***Abstract*—Data mining is a process of discovering usable data/information from the set of huge raw databases. By using data mining usable patterns are extracted from large batches of data which helps the business to learn more about which helps in developing more efficient business strategies. In this paper we will study some of the techniques of extracting patterns from transactional database. Frequent itemsets helps to find the items which are bought frequently and together by the customers. It helps to understand the customer's choices more. Utility Mining associates a utility value which refers to some quantitative representation of the customer preference. The process of mining the high utility item sets from some transactional database is basically a process of extracting new patterns based on some utility like profit. Diverse pattern mining discovers itemsets on the bases of their categories from the transactional database. The itemsets are low diverse if they belong to same category for e.g items milk and butter belongs to category diary items. We will study all these methods with suitable examples and present a review of our study.**

Keywords: Data Mining, Frequent Patterns, Utility Mining, High Utility Mining, Diverse Pattern Mining.

I. Introduction

The term pollution as Data Mining is an activity that derives some new non superficial information from huge databases [2]. Earlier data mining methods concentrate mainly on finding the relations (fluctuations in sales) among the items having good frequency in the transaction databases. It is normally known as frequent data mining [4], such techniques are based on the logic that items or item set which have high frequency are of high interest for the user as it can be taken as a good parameter for business. However, frequency is one notion of interestingness, there has been defined other notions such as high-utility [3,6], diversity etc. In this paper we throw light upon integration of existing interesting measures in different ways to derive new meaningful measures. High-utility item set considers both factors together, the frequency of the item sets as well as the utility related with every item [10]. On the other hand, notion of diversity tries to capture the relevance of pattern in terms of a category set coverage, i.e., how many categories from a category set do the items of an item set cover? An item set with large coverage would be considered as more relevant. Algorithms have been proposed for each individual interestingness measure and a very little work exist that focus on combination of different measures. We propose to study and define new interesting measures that can be derived using combinations of existing measures. We also plan to design efficient algorithms for finding patterns according to the new defined measures. The motivation of defining new measures and algorithms come from the issue of not clear understanding of the notion of relevance of a pattern, i.e., a frequent or diverse pattern [18] may not be relevant for all application domains. However, for an application domain, a frequent as well diverse pattern may be relevant.

The ultimate goal of research is to design and implementation of algorithms for new interestingness measures for pattern mining over transaction databases. No algorithm exists till date that mines diverse and high utility patterns. This section presents the literature survey about various techniques and aspects of above area such as; data mining, frequent item set mining (explain with help of example), utility mining and diverse frequent mining.

II. Data Mining

Data mining is related with filtration of data from large volume of databases and producing analytical reports which can be helpful in finding out the regularities or relationships which further can lead to provide good understanding of the present participle processes [8, 9]. The main aim of the mining process is to detect all the hidden and unexpected t patterns and trends from the database. It uses combination of various methods and techniques like data base technologies, artificial intelligence, machine learning etc.

Data mining has a great use in creating the analysis report of daily user transactions from some retail research and can be called market basket analysis. And this market basket analysis has also good application in clinical research, genetics, medicines, bioinformatics etc. We below discuss the basics of market basket algorithms such as frequent item set mining [11].

A. Frequent Item Set Mining

A non-empty collection of items can be termed as an item set. A k-item set can be defined as item set

 ISBN: 978-93-91355-11-1

with k different items. For e.g., combination of (milk, butter, bread) may denote a 3-item set from a database of supermarket transactions. Agrawal et al [1] gives the concept of frequent item sets. Item sets which occur frequently in the transactions are termed as frequent item sets. The main aim of this approach is to find all items with good frequency from transaction database and identify them as frequent item sets [10, 14]. Frequent item set mining have great importance in theory and practically have very good application in various data mining processes like long patterns search, dependency rules [1,9] emerging patterns etc. It has found application in the areas like text analysis, census analysis and telecommunications.

The principle of being frequent can be conveyed in terms of support value of the item sets. It is basically the percentage of transactions that include the item set.

Our given below example explains the sales report of some transaction database. Here we show below the sales data values for 3 items along with 10 transactions that contains the sales data along with associated profit. Each cell of the table shows the unit sold items in respective transaction.

Table 1: Transaction Database.

Transaction	Quantity of Sold Items		
	X	Y	Z
T-1	2	0	1
T-2	4	0	2
T-3	4	1	0
T-4	0	1	1
T-5	5	1	2
T-6	10	1	5
T-7	4	0	2
T-8	1	0	0
T-9	3	0	0
T-10	5	0	0

Table 2: Unit Profit Calculated (Associated to Sale or Items)

Item	Profit/ Unit in INR
X	5
Y	100
Z	40

Let us take the item set XY. As it is having only three 3 transactions (T-3, T-5 & T-6) that includes the item set derived from the given 10 transactions, here the support calculated of for item set will be

Support (XY) = 3 / 10 * 100 = 30 %

As T-3 includes four units of item X and one unit of the item Y, here profit calculated from sales of the item set XY in transaction T-3 will be:

Profit (XY, T-3) = 4 * profit(X) + 1 * profit(Y)= 4*5 + 1*100 = 120

As XY also present in transactions T-3, T-5 and T-6, so the complete transactions set of 10 gives the total profit associated with item set XY is:

Profit (XY) = profit (XY, T-3) + profit (XY, T-5) +profit (XY, T-6) (4*5+1*100) + (5*5+1*100) + (10*5+1*100) = 395

In the same way support values for different item sets can be calculated along with the profit from the sale of those item sets from complete transactions set as shown in below table.

Table 3: Describing Support (%) and Profit.

Item Set	Support in %	Profit Calculated (INR)
X	90	190
Y	40	400
Z	60	520
XY	30	395
XZ	50	605
YZ	30	620
XYZ	20	555

In our example if we take the minimum sup value equals to 40% the we find that item sets X, Y, Z and XZ are the only which qualifies the criteria of the frequent item set mining as all these are having the very good value of the support which is greater than the min sup threshold value defined.

But in case if we also consider the associated profit then we found that item sets Z, XZ, YZ and XYZ are the most profitable item sets but out of these only two are frequent item sets. And the item sets YZ and XYZ are not frequent but these are gaining good profits in comparison to some frequent item sets like X and Y. This is only because unit profit of item sets is deviating.

The calculated support will not always reflect the statistical correlations without affecting their semantic significance which is here represented as associated profit. It is clear from the example that frequent item set mining will not always satisfy all goals of an application. Similarly, if an application demands to prefer item set that contains items from diverse set of categories, frequency and utility notions of interestingness would not capture it. It requires definition of new measures and hence new algorithms to mine them. An algorithm for frequent item set mining would not be efficient if used directly to mine high-utility pattern or diverse patterns.

B. Utility Mining

We see that there are few limitations introduced by the frequent pattern approach and then the researchers made some efforts to find genuine solutions of those

limitations and defined a new utility-based technique that considers the usefulness of all the item sets and define a term called utility value which refers to some quantitative representation of the customer preference. In this approach basically some utility threshold value [14] is defined and then to detect all the item sets that have the utility value [15] greater than the threshold value defined. It is basically calculated by considering the importance of the item set from the user's perspective. For example, in a retail store an analyst can find that which items of the store can give more revenue and then he or she can define the utility of that item set in terms of monetary profit that the store can earn by sale of that particular item set [10,11].

In this approach the frequency of item set in various transactions is not of much interest, but the focus is on revenue generated collectively by the all set of transactions that include the item set. Any item set can have the utility value defined for itself and it can be termed as in the form of profit, page rank, popularity or other aspects such as design or beauty or any other measure of user's preference can also be included. If any item set let us say S satisfies some utility constraint, then it is useful i.e., any constraint of the form u(s) >= minimum utility defined, where u(s) is utility value of item set denoting threshold defined [14].

Total utility of the item set can be finding simply by multiplying the utility value of individual items in the item set with the corresponding frequencies of the items of item set in the transactions that includes the particular item set. Utility mining approach measure significance of an item set from two dimensions which are support value of item set i.e., frequency and the semantic significance of the item set measured by user.

In our previous example let us take the utility of an item set as unit profit associated with sale of that item set then defining the utility threshold as min(utility)= 500, the item sets that are most earning are Z, XZ, YZ and XYZ in which only two frequent item sets. The item sets which are not frequent and still they are giving more profit are YZ and XYZ. The item set Y is such that its one unit sold can earn more profit than one unit sold from X and Z.

Furthermore, it may be noted that the relation between item sets in the case of frequency satisfies anti-monotonicity, i.e., if XY is not frequent then any superset of XY (e.g., XYZ) would not be frequent. X is not a high-utility item set but XYZ is. It shows that the algorithms for frequent pattern mining cannot be used directly for mining high utility patterns [5, 6, 7].

C. High Utility Itemset Mining

The problem of frequent itemset mining has been redefined as the problem of high-utility itemset mining to overcome certain limitations. Within this problem, a transaction database contains transactions that take into account purchase amounts, as well as the unit income of each item.

Table 4: Transaction Table with Quantities

Transactions	Items
T-0	a(1), b(5),c(1),d(3),e(1)
T-1	b(4), c(3),d(3),e(1)
T-2	a(1), c(1),d(1)
T-3	a(2), c(6),e(2)
T-4	b(2), c(2),e(1)

Table 5: Items with Associated Profit

Item	Unit Profit (Rs.)
A	5
B	2
C	1
D	2
E	3

Let's take an example of transaction T-3 which shows that customer has purchased two units of item 'a', six units of item 'c' and two units of item 'e'. Now the above table shows the unit profit of each item. It indicates the unit profit of items 'a', 'b', 'c', 'd' and 'e' are rupees 5, 2, 1, 2 and 3 respectively. For example, this means that every unit of "a" being sold would produce a profit of rupees 5.

High-utility itemset mining problem is to find the itemsets (group of items) that produce a high profit in a database when they are sold together. The user has to provide a value for a threshold called "minutil" (the minimum utility threshold). A mining algorithm for high-utility itemsets produces all the sets of high-utility items that are the itemsets which generate at least "minutil" profit. Let's take an example and set the "minutil" to rupees 25. The result of a high utility itemset mining algorithm would be the following. Some of the high utility itemsets mine by algorithm is

{a, c}: 28(Rs.) {a, c, e}: 31(Rs.)
{a, b, c, d, e}: 25(Rs.) {b, c}: 28(Rs.)
{b, c, d}: 34 (Rs.) {b, c, d, e}: 40 (Rs.}
{b, c, e}: 37(Rs.) {b, d}: 30(Rs.)
{b, d, e}: 36(Rs.) {b, e}: 31 (Rs.)
{c, e}: 27(Rs.)

High-Utility Itemsets

Now let's take an example of itemset {b,d} which is a high utility itemset here because it has a utility of rupees 30 (produce a profit of 30 rupees),which is greater than the minutil threshold (set to rupees 25 by user).Normally the utility of an itemset from a transaction is computed by

 ISBN: 978-93-91355-11-1

multiplying the quantity of each item with their unit profit. Let us explain it with an example, the profit of set {a,e} in transaction T-0 of database is 1*5+1*3=8(Rs).Same way the profit of {a, e} in Transaction T-3 is 2*5+2*3=16(Rs). Now sum the utilities of itemset in the whole database from the transactions where it appears. For example, total utility of set {a, e} is computed as 8+16=24(Rs) because it is present only in the transaction T-0 and T-3.

D. Why the problem of high-utility itemset mining is interesting?

For the following reasons the topic of high utility itemset mining is very interesting. Firstly, discovering itemsets that generate a high profit in customer transactions may be more interesting from a practical perspective than those that are purchased frequently. Secondly, it is more challenging to mine the high-utility itemsets from the transaction database.

In frequent itemset mining, there is a well-known property of the frequency (support) of itemsets stating that given a set of items, all of its supersets must have a lower or equal support. This is also referred to as the "Apriori property" or "anti-monotonicity" property and is very useful in pruning the search space, since if an itemset is not frequent, then we know that all its supersets are also infrequent and can be pruned. In high-utility itemset mining there is no such property [19, 20]. Therefore, given an itemset its supersets may have higher, lower or same utility. For example, in the previous example, the utility of itemsets {a}, {a,e} and {a,b,c} are respectively 20 Rs, 24 Rs and 16 Rs. We won't go into the details but a key idea is to use upper limits on the use of itemsets to preserve the anti-monotonicity property so that the search space can be pruned.

E. Diverse Pattern Mining

In this section, we will study the concept of diverse frequent patterns and diverse rank to rank the frequent patterns based on the item category given in a database. Diverse-Frequent Patterns Concept Given a frequent pattern obtained at some "minSup" from a transactional database, its diversity is based on the category of items within it. Diversity of pattern is considered to be low if all the items of it belongs to same category, [21, 22] but if items are from different categories, then diversity is relatively high. An item can be from different levels of categories. Taking an example, the item 'charger' belongs to electronics and house hold categories. The notion of concept hierarchy can be considered to find the category of an item. Concept hierarchy is basically a tree in which the items are arranged in a hierarchical manner [23]. We assume that all the leaf levels present the items of a frequent pattern in the concept hierarchy. Starting from leaf level a mapping process is used from lower to higher level in the hierarchy and finally all the items are merged to top level item called root.

For example, in the figure above it can be observed that items battery, charger and mobile mapped to a higher-level item electronics and in same way items chair, table and bed mapped to higher level furniture. And finally, the items mapped to higher level root. This hierarchy can contain more levels and always leaf level contains the items in the tree. The frequent patterns can be given rank depending upon the categories of the items for example.

Table 6. Frequent Patterns with Rank

Frequent Patterns	Rank
{chair, table, bed}	3
{chair, table, mobile}	2
{chair, mobile}	1

Using the concept hierarchy given in Figure, we explain the notion of diverse patterns [18]. Let us take example of frequent patterns given in the table above at min support 0.25. Here the rank assigned to the frequent patterns based on the categories of the items they belong in the table. The reason behind assigning such rank numbers is, we found that items belonging to different categories are more interesting than the frequent patterns belonging to few or same category. For example, frequent pattern {chair, table, bed} all items belong to same category furniture, so this pattern is of less interest to user. In pattern {chair, table, mobile} all items are not from same category, items chair and table have common parent furniture but item mobile is from different category. Thus, it is more interesting. And in pattern {chair, mobile} both the items are from different category and have common

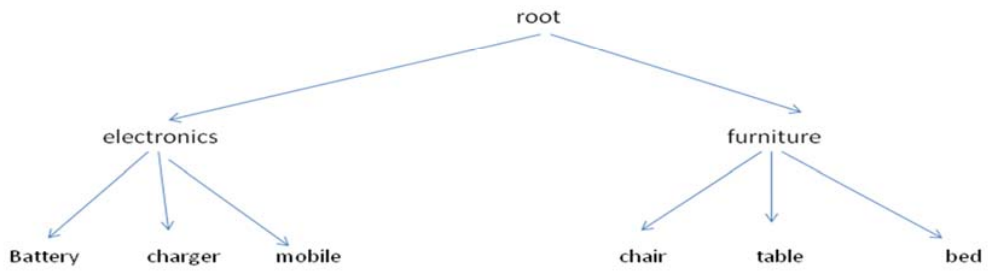

Fig. 1: Concept Hierarchy

parent root which is at level 0. And this is more interesting pattern in all. Thus, we can say that pattern having items with less common parent are more interesting in database. Now, given two common patterns of the same length, different merging behavior can be realized if we observe how a given frequent pattern at the level of the leaf is moving towards the patterns of the higher level. That is, one pattern may easily merge into a few higher-level items within a few levels, and the other pattern may merge into a few higher-level items by crossing more levels. To differentiate the patterns in terms of diversity, the notion of diversity rank [18] is given. Diverse Rank [22], to rank the frequent patterns based on the items' categories within it. We have studied the algorithm called diverseFP [18] to mine diverse-frequent patterns based on the diverse rank.

III. Comparison

Above study of the mining algorithm shows that all the techniques of mining are relevant for discovering the patterns useful for business decisions. Simple Frequent pattern technique produce itemsets which are bought frequently by the customers and thus help the business vendor present such items together on same shelf in the store. But all frequent itemsets are not giving good profits to business and this is a limitation of the technique. Utility itemset mining attach a new measure and removes limitation of simple frequent pattern mining [20]. This new measure with all items is termed as utility of the item. This technique filters the itemset on the basis of the minimum utility value defined. Thus, produce itemsets giving high profits only. On the other hand, diverse pattern gives totally a different measure which check the category of items purchased together. It finds that items of an itemset if belongs to different category should be counted as more diverse and can be more profitable for business. The new measure of diversity is attached with itemsets and search space is pruned on the basis of minimum diversity defined.

IV. Conclusion

This Study of all methods in our paper found that the frequent pattern discovery methods that discover itemsets with high frequency in transaction database is not enough for business profits. It may result in producing such frequent itemsets which are less profitable than the non-frequent itemsets. High utility itemset mining method is of more interest which discover itemsets with high utility (profit). Itemsets having larger value of utility are pruned from search space by comparing with minimum utility defined and itemsets with lower utility are discarded. Thus, this method makes it more interesting. Another measure of diversity also found more interesting as it not only considers items from same category but found itemsets which contains items from different categories and produce more profitable items. So diverse pattern mining is more interesting. All the existing algorithm of diverse pattern mining are multi phases which discover frequent itemsets in first phase and prune search space in next phase. All the existing methods are very lengthy and complex that uses diverse rank also produce the mining results in number of phases. As a future work we have planned to design some new algorithm that will find and prune the diverse pattern in single phase and more efficient.

References

[01] R. Agrawal, T. Imielinski, A. Swami, 1993, mining association rules between sets of items in large databases, in: proceedings of the ACM SIGMOD International Conference on Management of data, pp. 207-216

[02] K. Ali, S. Manganaris, R.Srikant , Partial classification using association rules, in: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining , Newport Beach, California, 1997, pp. 115-118

[03] C.F. Ahmed, S.K. Tanbeer, Jeong Byeong-Soo, Lee Young-Koo, Efficient tree structures for high utility pattern mining in incremental databases, in: IEEE Transactions on Knowledge and Data Engineering 21(12) (2009)

[04] B.Barber, H.J Hamilton , Extracting share frequency item sets with infrequent subsets, Data Mining and Knowledge Discovery 7(2) (2003)153-185.

[05] C.H. Cai, A.W.C Fu, C.H.Cheng , w.W.Kwong, Mining association rules with weighted items,in:Proceedings of IEEE International Database Engineering and Applications Symposium, Cardiff, United kingdom, 1998, pp.68-77

[06] Chan, Q.Yang,Y.DShen, Mining high utility item sets, in:Proceedings of the 3rd IEEE International Conference on Data Mining , Melbourne , Florida, 2003, pp.19-26

[07] S. Lu, H. Hu,f. Li, Mining weighted association rules, Intelligent Data Analysis 5(3)(2001) 211-225

[08] S Laxman, P S Sastry, A survey of temporal data mining, Sadhana, Vol. 31, part 2, April 2006, pp. 173-198.

[09] H.Mannila , H.Toivonen, Levelwise search and borders of theories in knowledge discovery, Data Mining and Knowedge Discovery 1(3)(1997) 241-258

[10] J.Pillai, O.P.Vyas ,Overview of item set utility mining and its applications , in: International Journal of Computer Applications (0975-8887), Volume 5-No.11(August 2010)

[11] V.Podpecan , N. Lavrac, I. Kononenkom, A fast algorithm for mining utility-frequent item sets,in workshop on Constraint-Based Mining and Learning at ECML/PKDD, 2007, pp. 9-20.

[12] C.Ramaraju , N.Savarimuthu,A conditional tree based novel algorithm for high utility item set mining,International Conference on Recent Trends in Information Technology (ICRTIT) 2011, pp. 701-706

ISBN: 978-93-91355-11-1

[13] H.Yao, H.J.Hamilton, C.J.Butz, A foundation approach to mining item set utilities from databases,in:Proceedings of the Third SIAM International Conference on Data Mining, Orlando, Florida, 2004, pp.482-486

[14] H.Yao,H.J.Hamilton ,Mining item set utilities from transaction databases, in Data and Knowledge Engineering 59(2006) pp.603-626

[15] Cheng-Wei Wu, Souleymane Zida1, Vincent S. Tseng, FHM: Faster High-Utility Item set Mining using Estimated Utility Co-occurrence Pruning, 21st International Symposium for intelligent systems, (ISMIS 2014), Springer LNAI pp 83-92.

[16] Wei Wu, Bai-EnShie, and Philip S. Yu, UP-Growth: An Efficient Algorithm for High Utility Item set Mining, KDD'10, July 25–28, 2010, Washington, DC, USA.

[17] Jieh-shanYeh, Yu-chiang Li, Chin-chen Chang, Two-phase algorithms for a novel utility-frequent model, in: Emerging Trends in Knowledge Discovery and Data Mining, Lecture notes in Computer Science, Vol. 4819/2007, pp. 433-444.

[18] R. Uday Kiran, P.Krishna Reddy, Discovering diverse-frequent patterns in transactional databases, International Conference COMAD-2011, December 19-21 2011.

[19] Dawar S, Goyal V (2015) UP-Hist tree: An efficient data structure for mining high utility patterns from transaction databases. In: Proceedings of the 19th international database engineering & applications symposium. ACM, New York, pp 56–61.

[20] Dawar S, Goyal V, Bera D (2017) A hybrid framework for mining high-utility itemsets in a sparse transaction database. Appl Intell 47(3):809–827.

[21] Swamy MK, Reddy PK (2015) Improving diversity performance of association rule-based recommender systems. In: Database and expert systems applications. Springer, New York, pp 499–508.

[22] Wu D, Luo D, Jensen CS, Huang JZ (2019) Efficiently mining maximal diverse frequent itemsets. In: International conference on database systems for advanced applications. Springer, New York, pp 191–207.

[23] Zaki MJ (2000) Scalable algorithms for association mining. IEEE Trans Knowl Data Eng 12(3):372–390.

ISBN: 978-93-91355-11-1