Data Mining Considering the Instances of Item-Sets

Aman Raj
Information Technology
Institute of Engineering and Management, Kolkata, West Bengal, India

Pratik Singh
Computer Science, Institute of Engineering and Management, Kolkata, West Bengal, India

Debdutta Chatterjee
Information Technology,
Institute of Engineering and Management, Kolkata, West Bengal, India

ABSTRACT
In the field of computer science, data mining is the process that attempts to discover patterns in large data sets. However it deals mostly with the relationship between two or more item objects. For example A to B, where 'A' and 'B' are the item objects. But in the real life scenario not only the relationship between item objects is important, but the relationship of their frequency of occurrence is also the matter of a prime concern. The instances of two or more data items also may be correlated with each other. For example the relation between A and 2B.Where 'A' and 'B' are the data items and '2B' represents two instances of the B type of data items. This paper provides a new approach to find the occurrence dependent data patterns by conventional approaches and also compare the some inter related concepts.

Keywords
Data Items, Instances of data items, Data patterns, Occurrence dependency.

1. INTRODUCTION
Data mining is a field of computer science, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. [1] The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.[1,3]

One basic operation of data mining is finding the frequent item set in the given data using different algorithm like apriori algorithm , Frequent pattern growth etc. are the vital operations, for example the Market Basket Analysis we get the frequent item sets using these algorithms.[3]

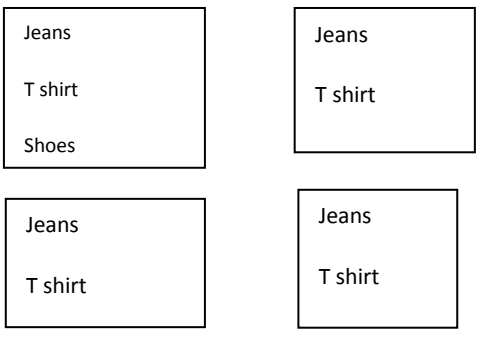

Fig.1: Items in the Baskets of four customers

1.1 Market Basket Analysis:

The discovery of interesting correlation relationships among huge amounts of business transaction records can help in many business decision-making processes, such as catalogue design, cross-marketing, and customer shopping behaviour analysis, market basket analysis process analyzes customer buying habits by finding associations between the different items that customers place in their "shopping baskets" depicted in the figure 1 where in four baskets the purchases of four different customers has been shown. The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers. For instance, if customers are buying milk, how likely are they to also buy bread (and what kind of bread) on the same trip to the supermarket? Such information can lead to increased sales by helping retailers do selective marketing and plan their shelf space. .But in the new approach, this paper considers the fact that a customer tends to buy items a specific number of items depending upon the other purchasing. For example a customer will like to buy one keyboard and one mouse with one laptop rather than buying ten mouses and one laptop in most of the cases. This paper discusses this matter as a important problem and proposes solutions with traditional approaches. Here we get the frequency pattern like jeans, T shirt are frequent, but it neglects the relations like- how many T shirts per jeans are the frequent item set. The Supporting Fact behind this approach is that the People rarely buy one computer many copies of a single operating system. So that item set can be considered as an example of an item set with rare Combination of items. Though the item set itself is frequent in traditional approach. In figure 2 this fact is shown.[2,3]

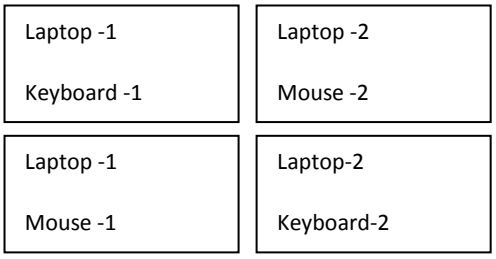

Fig.2: Items in the Baskets of four customers (New approach)

2. SOLUTION OF THE PROBLEM USING THE 'APRIORI' ALGORITHM
A priori is a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994 for mining frequent item sets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent item set properties. In this approach the paper first solves the problem using a priori algorithm. [3]

TID	A	B	C	D	E
1	1	0	2	1	1
2	1	3	2	1	0
3	2	3	3	0	0
4	0	3	2	3	1
5	1	1	2	1	1

Fig.3: Transaction with TID and Items (problem set)

For proceeding the table presented above can be taken into consideration. The table shows the items purchased in each transaction with transaction ID (TID) 1, 2 ..5 and items a, b,c, d, e and f. Here the noticeable fact is that the corresponding value of the item purchased, which refers to the number of items purchased, are not like conventional a priori problem. In the case of a priori the number is either one or zero .Now for proceeding the minimum support can be taken as 0.60 or 60 %. Note the problem set is not only in binary.

STEP 1-Scan the database for count of each candidate (C1)

INSTANCE OF THE SET	SUPPORT COUNT
{A1}	3
{A2}	1
{B1}	1
{B3}	3
{C2}	4
{C3}	1
{D1}	3
{D3}	1
{E1}	3

Fig.4: instances and support count taking one at a time

STEP 2-Compare candidate support count with minimum support count (L1)

INSTANCE OF THE SET	SUPPORT COUNT
{A1}	3
{B1}	1
{B3}	3
{C2}	4
{D1}	3
{E1}	3

Fig.5: after the prune (support =60%)

STEP 3-Generate (C2) candidates from (L1) and - Scan database for count of each candidate

INSTANCE OF THE SET	SUPPORT COUNT
{A1,B3}	1
{A1,C2}	3
{A1,D1}	3
{A1,E1}	2
{B3,C2}	2
{B3,D1}	1
{B3,E1}	1
{C2,D1}	3
{C2,E1}	3
{D1,E1}	2

Fig.6: instances and support count taking two at a time

STEP 4- Compare candidate support count with minimum support count (L2)

INSTANCE OF THE SET	SUPPORT COUNT
{A1,C2}	3
{A1,D1}	3
{C2,D1}	3
{C2,E1}	3

Fig.7: after the prune (support =60%)

STEP 5- Generate (C3) candidates from (L2) and - Scan database for count of each candidate

INSTANCE OF THE SET	SUPPORT COUNT
{A1,C2,D1}	3
{A1,C2,E1}	2
{C2,D1,E1}	2

Fig.8: instances and support count taking three at a time

STEP 5- Compare candidate support count with minimum support count (L2)

INSTANCE OF THE SET	SUPPORT COUNT
{A1,C2,D1}	3

Fig.9: after the prune (support =60%)

So from this algorithm we get the frequent item-set to be {A1,C2,D1}. The algorithm simply scans all of the transactions in order to count the number of occurrences of each item.

3. SOLUTION OF THE PROBLEM USING THE 'FP GROWTH TREE' ALGORITHM

For the problem in the figure 4 the solution using fp growth tree algorithm is presented here [3,4]

TID	a	b	c	M	p	f	h	i	O
1	1	0	2	1	1	2	0	1	0
2	1	3	2	1	0	2	0	2	3
3	2	3	3	0	0	0	3	0	3
4	0	3	2	3	1	2	3	0	2
5	1	1	2	1	1	2	0	1	2

Fig. 11: Transaction with TID and Items (problem set)

TID	Bought Item	Frequent Items(Ordered items)
1	f2,a1,c2,d2,g2,i1,m1,p1	f2,c2,a1,m1,p1
2	a1,b3,c2,f2,i2,m1,o3	f2,c2,a1,b3,m1
3	b3,f2,h3,i1,o3,c3,a2	f2,b3
4	b3,c2,p1,f1,o2,m3	c2,b3,p1
5	a1,f2,c2,i1,p1,m1,b1,o2	f2,c2,a1,m1,p1

Fig. 12: Transaction with TID and Items in sorted order

After the insertion of (f2,c2,a1,m1,p1)

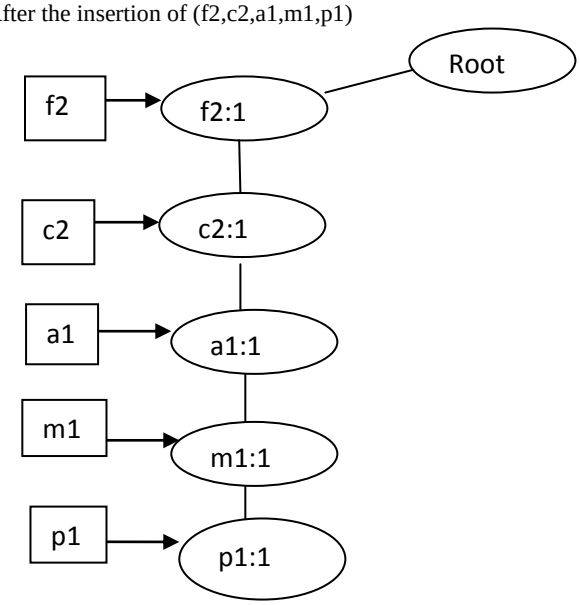

Fig.13:FP tree after intertion of TID-1

After the insertion of(f2,c2,a1,b3,m1)

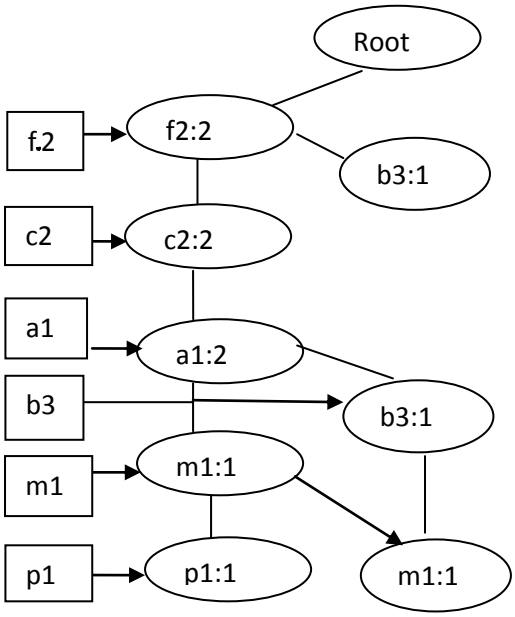

Fig.14:FP tree after intertion of TID-2

After the insertion of(f2,b3)

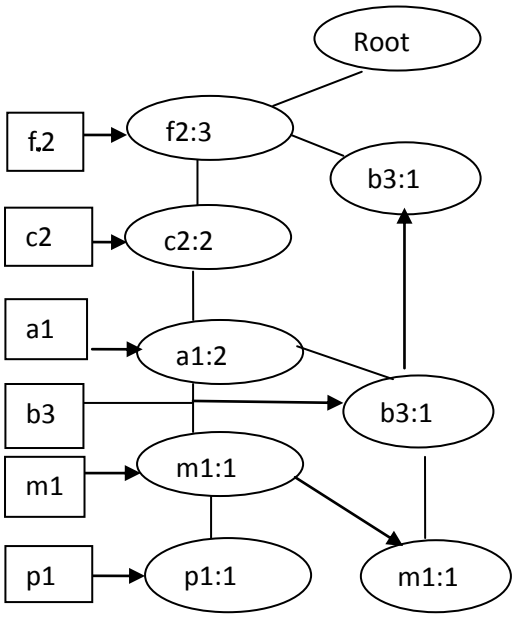

Fig.15: FP tree after intertion of TID-3

After the insertion of(c2,b3,p1)

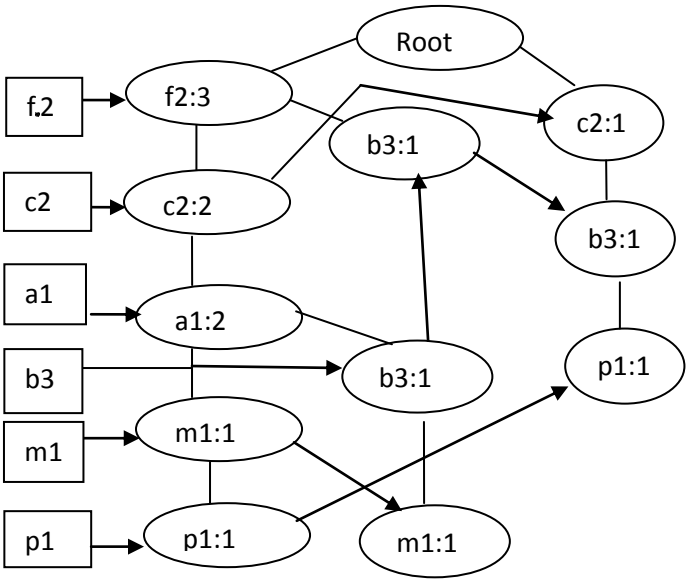

Fig.16:FP tree after intertion of TID-4

After the insertion of(f2,c2,a1,m1,p1)

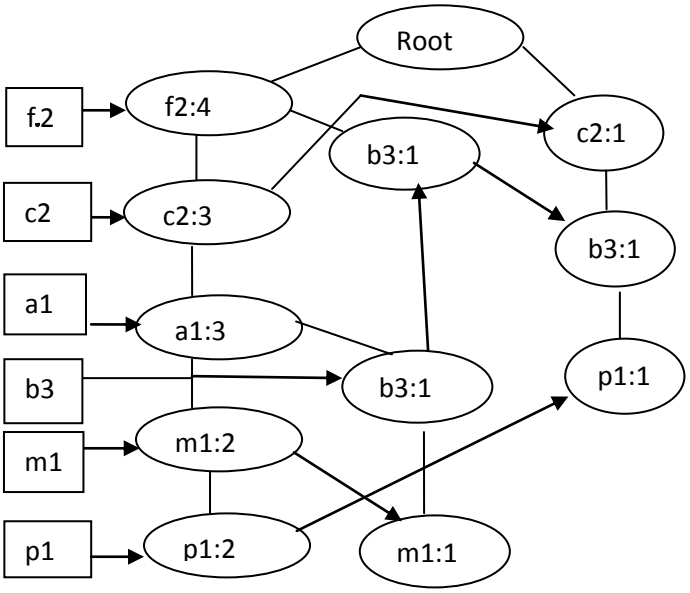

Fig.17:FP tree after intertion of TID-5

Result Set Extraction by the FP Tree: Finding all patterns with 'p1'

Starting from the bottom of the header table
Generate (p1:3)

'p1' exists in paths:

(f2:4, c2:3, a1:3,m1:2, p1:2) and (c2:1, b3:1, p1:1)
process these further Paths with 'p1'

●We got (f2:4, c2:3, a1:3, m1:2, p1:2) and (c2:1, b3:1, p1:1)

●The transactions containing 'p1' have p1.count

●We get (f2:2, c2:2, a1:2, m1:2, p1:2) and (c2:1, b3:1, p1:1)

● Since we know that 'p1' is part of these we can drop 'p1

Conditional Pattern Base (CPB)

● We get paths (p1 dropped):

(f2:2, c2:2, a1:2, m1:2) and (c2:1, b3:1)

●Contains transactions in which 'p1' occurs, To find all frequent patterns containing 'p1' we need to find all frequent patterns in the CPB and add 'p1' to them

●We can do this by constructing a new FP-Tree for the CPB' Finding all patterns with 'p1'

● We again filter away all items < minimum support threshold

– (f2:2, c2:2, a1:2, m1:2), (c2:1, b4:1) => (c2:3)

● We generate (c2p1:3) – Support value is taken from the sub-tree– Frequent patterns thus far: (p1:3, c2p1:3)

4. COMPARISION OFCONVENTIONAL APPROACH TO THE ITEMS-SETS INSTANCE BASED APPROACH

The instance based mining, is not similar to the conventional approach. Though it can be treated as a special problem of the conventional approach, it still contains enough specific applications to be treated carefully like a complete different way of mining. As discussed in this paper in the market basket example, the problem set in this approach is not a set of transactions with the item-sets whose values are in binary 0,1 only. Instead of, limiting the item set count only into binary, this problem broadens the aspect with taking decimal values into account which contains the set of positive integers. The probable applications of this approach are described also.

5. APPLICATIONS

As, we keep on the track of the frequency of occurrence of the item set so the this mining approach will have some many applications in the real life scenario some of are given below

1. Deriving data to help sellers in finding more specific buying pattern.
2. Buyer also can identify the quality of bought products similarly using this approach.
3. Unnecessary inventory stocks can be identified, and ordering can be performed with a co relation between items.
4. Further improvement of this approach can lead to a system to automatically generate specific requirement of certain co related or inter connected items.

6. CONCLUSION

This approach can be used to extract various important, occurrence dependent data in the field of data mining. A new way of thinking can be established in the data mining world. Various new algorithms, dedicated to solve this particular problem can also be thought. The main proceeding and target result should though be to identify the customers buying pattern. This particular approach of thinking is actually more concentrated towards deriving data which are more logical and dedicated towards the goal to achieve more specific selling pattern. Applications in various fields can be visualized as the outcome of implementing this approach will lead to more specific and requirement based data pattern. Though the approach is new it's implementation is easy and all of the traditional approaches can be used as it is shown in this paper. There may be some problem which can arise while adopting the approach, but as clearly they will only be due to complexity in the data handling process. But, today this pattern analysis approach has a wide use and for that reason, those problems can be overlooked to get the result which indeed is more required. Further study and development of this approach fortunately will lead to more specific algorithms deriver to solve tis particular type of problem in a dedicated way. Parallel processing or other such way can lead to less time requirement. As this approach is only a modified problem with huge application, all conventional development and study will remain same relevant to it.

7. REFERENCES

[1] Data mining an overview from database prospective Ming Syan Chen, Jiwaei Han, Philip .s .Yu, IEE transaction on knowledge and data engineering, Vol 8 No 6, December 1996

[2] R. Agarwal and R. Srikant, "Fast algorithms for mining Association rules in large databases", Proc. 20th Int'l Conf. Very large databases, pp. 478-499, sept 1995

[3] Data Mining: Concepts and Techniques, Second Edition Jiwai Han and Micheline Kamber, Morgan Kaufman series.

[4] Pang-Ning Tan, Michael Steinbach, Vipin Kumar: Introduction to Data Mining, Addison-Wesley. Chapter 6: Association Analysis: Basic Concepts and Algorithms.

[5] Rakesh Agrawal, Tomasz Imieliski, Arun Swami, Mining association rules between sets of items in large databases, ACM SIGMOD, May 1993.

[6] Feng Tao, Weighted Association Rule Mining using Weighted Support and Significant framework. ACM SIGKDD, Aug 2003

[7] K. Wang, Y. He and J. Han, Mining Frequent Itemsets Using Support Constraints, VLDB, Sep 2000

[8] Bing Liu, Wynne Hsu, Yiming Ma, Mining Association Rules with Multiple Minimum Supports. ACM SIGKDD, June 1999

[9] C. H. Cai, Ada Wai-Chee Fu, C. H. Cheng, and W. W. Kwong. Mining association rules with weighted items. IDEAS'98, July 1998.

[10] J. Han and Y. Fu, Mining Multiple-Level Association Rules in Large Databases, IEEE TKDE, September/October 1999, pp. 798-805.

[11] Jiawei Han, Jian Pei, Yiwen Yin, Mining frequent patterns without candidate generation, ACM SIGMOD, May 2000.

[12] Guimei Liu, Hongjun Lu, Yabo Xu, Jeffrey Xu Yu: Ascending Frequency Ordered Prefix-tree: Efficient Mining of Frequent Patterns. DASFAA 2003: 65-72.

[13] Jian Pei, Jiawei Han, CLOSET: An Efficient Algorithm for Mining Frequent Closed Itemsets, DMKD, May 2000.

[14] Jianyong Wang, Jiawei Han, Jian Pei, CLOSET+: searching for the best strategies for mining frequent closed itemsets, ACM SIGKDD, Aug 2003.

[15] Zijian Zheng, Real World Performance of Association Rule Algorithms. ACM SIGKDD, 2001.