Application of association rules to determine item sets from large databases

Sambit Kumar Mishra
Associate Professor
Department of Computer Sc.&Engg.
Ajay Binay Institute of Technology, Cuttack

Prof.(Dr.) Srikanta Pattanaik
Professor, S.O.A. University, Bhubaneswar

Prof.(Dr.) Dulu Patnaik
Principal, Government College of Engineering, Bhawanipatna

Abstract

The problem of discovering association rules between items in a large database of sales transactions has been considered in this paper. A new algorithm has been presented in this paper for solving the particular problem which is fundamentally different from the known algorithms. The empirical evaluation shows that the algorithm outperforms the known algorithms, factors ranging from three for small problems to more than an order of magnitude for large problems. The scale up experiments shows that the algorithm scales linearly with the number of transactions.

Key words : Cluster, classification, transaction, association rule, item set, candidate item set, multiple pass.

1. Introduction

Data mining is the process of extracting patterns from data. It searches for unknown patterns in data that can be used to predict future behavior. Basically data mining is a technique not to change the presentation but to discover unknown relationships between the data. It is termed as software that is used to describe data in a different form. It is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified.

Data mining commonly involves four classes of tasks.

(i)Clustering which is the task of discovering groups and structures in the data

that may be similar to another groups and structures of data.

(ii) **Classification** which is the task of generalizing known structure to apply to new data.

(iii)Regression which tries to build or generate a function with the least error.

(iv)**Association rule learning** which searches for relationships between variables.

The problem of finding association rules falls within the purview of database mining. It is also called as knowledge discovery in databases. This work also includes the induction of classification rules , discovery of clausal rules etc.

2. Literature Survey

There are many variants of data mining algorithms that differ in how they check candidate item sets against the database. Mining algorithm in its purest form checks item sets of length for frequency during database pass.

Brin, S et.al.[1] was more eager and continued checking an item set shortly after all its subsets had been determined frequent, rather than waiting until the database pass completes. Savasere, A et.al [2] had identified all frequent-item sets in memory-sized partitions of the database, and then checked those against the entire database during a final pass. Brin, S et.al[1] had considered the same number of candidate item sets where Partition could consider more candidate item sets associated with long patterns. Park et al. [3] had enhanced the mining techniques with a hashing scheme that could identify some candidates which would turn up infrequent if checked against the database. It also used the hashing scheme to re-write a smaller database after each pass in order to reduce the overhead of subsequent passes. Gunopulos et al. [4] had presented a randomized algorithm for identifying maximal frequent item sets in memory-resident databases. Their algorithm worked by iteratively attempting to extend a working pattern. However the randomized version of the algorithm does not guarantee every maximal frequent item set returned is evaluated and found to be efficient to extract long frequent item sets. So it might not be clear how the algorithm presented by Gunopulos et.al[4] would be scaled to disk resident data-sets since each attempt at extending an item set requires a scan over the data.

Zaki et al. [5] had presented the algorithms MaxEclat and MaxClique for identifying maximal frequent item sets. These algorithms were similar to Max-Miner in that they attempted to look ahead and identify long frequent item sets .The important difference was that Max-Miner attempted to look ahead throughout the search, whereas MaxEclat and MaxClique attempted to look ahead only during an initialization phase.

Lin and Kedem [6] had also proposed an algorithm called Pincer-Search for mining long maximal frequent item sets. Like Max-Miner, Pincer-Search attempted to identify long patterns throughout the search. The difference between these algorithms was primarily in the long candidate item sets considered by each Max-Miner that used a simple, polynomial time candidate generation procedure directed by heuristics, while Pincer-Search used an NP-hard reduction phase to ensure no long candidate item set contained any known infrequent item set.

As suggested by Agrawal, R et.al[7] ,finding patterns in databases is the fundamental operation behind several common data-mining tasks including association rule and sequential pattern mining . For the most part, pattern mining algorithms have been developed to operate on databases where the longest patterns are relatively short. This leaves data outside this mold unexplorable using current techniques.

3. Problem Formulation

Algorithms for discovering large item sets make multiple passes over the data. In the first pass, the support of individual items are counted and which of them are large is determined. In each subsequent pass, a set of item sets are found to be large than the previous pass. At the end of pass, which of the candidate item sets are actually large may be determined. While counting candidates of multiple sizes in one pass, instead of counting only candidates of size i in the ith pass, the number of candidates generated from the original database may be counted. This variation may pay off in the later passes when the cost of counting and keeping in memory additional candidates becomes less than the cost of scanning the database.

3.1. Algorithm

```
L= large item sets
$C_i$= database , D
i=2;
while ($L_{i-1}$ !=0)
{
$C_i$= candidate_generate($L_{i-1}$)
i++;
}
for all entries t$\epsilon C_{i-1}$
determine candidate item sets in $C_i$
contained in the transaction with identifier t.TID
Ct= { C ϵ Ci| C-C[i]} ϵ t.set of itemsets ∩ ( C-C[i-1]) ϵ t.set of items};
for all candidates C ϵ Ct do
C.count++;
if( $C_t$ !=0)
$C_i$=$C_i$+<t.TID, $C_t$>;
Li-C.count>=itemsets
```

Table-3.1. Database

TID	Items
100	2 7 9
200	4 5 7
300	7 4 5
400	9 7 2

Table-3.2. Item sets

TID	Set of item sets
100	{2}, {7}, {9}
200	{4},{5},{7}
300	{7},{4},{5}
400	{9}, {7}, {2}

4. Conclusion and future direction

So far the algorithm has been presented and evaluated for mining maximal frequent item sets from large databases. The algorithm applies several new techniques for reducing the space of item sets .The result is orders of magnitude in performance improvements over other algorithms when frequent item sets are long, and more modest and still substantial improvements when frequent item sets are short. Incorporating these constraints into the search may be the only way to achieve tractable completeness at low supports on complex datasets. It may be therefore the extensions for future work to exploit many of the wide variety of interestingness constraints during the search rather than applying them only in a post processing filtering step.

5. References

[1] Brin, S.; Motwani, R.; Ullman, J.; and Tsur, S. 1997. Dynamic Itemset Counting and Implication Rules for Market Basket Data. In Proc. of the 1997 ACM-SIGMOD Conf. on Management of Data, 255-264.

[2] Savasere, A.; Omiecinski, E.; and Navathe, S. 1995. An Efficient Algorithm for Mining Association Rules in Large Databases. In Proc. of the 21st Conf. on Very Large Data-Bases, 432-444.

[3] Park, J. S.; Chen, M.-S.; and Yu, P. S. 1996. An Effective Hash Based Algorithm for Mining Association Rules. In Proc. of the 1995 ACM-SIGMOD Conf. on Management of Data, 175-186.

[4] Gunopulos, G.; Mannila, H.; and Saluja, S. 1997. Discovering All Most Specific Sentences by Randomized Algorithms. In Proc. of the 6th Int’l Conf. on Database Theory, 215-229.

[5] Zaki, M. J.; Parthasarathy, S.; Ogihara, M.; and Li, W. 1997. New Algorithms for Fast Discovery of Association Rules. In Proc. of the Third Int’l Conf. on Knowledge Discovery in Databases and Data Mining, 283-286.

[6] Lin, D.-I and Kedem, Z. M. 1998. Pincer-Search: A New Algorithm for Discovering the Maximum Frequent Set. In Proc. of the Sixth European Conf. on Extending Database Technology.

[7] Agrawal, R.; Imielinski, T.; and Swami, A. 1993. Mining Association Rules between Sets of Items in Large Databases. In Proc. of the 1993 ACM-SIGMOD Conf. on Management of Data, 207-216.