DISCOVERING NEW TECHNIQUES IN ASSOCIATION RULE MINING

David ZEMAN, Doctoral Degree Programme (1)
Dept. of Information Systems, FIT, BUT
E-mail: zemand@fit.vutbr.cz

Supervised by: Dr. Jaroslav Zendulka

ABSTRACT

This paper shows some new procedures and ideas in data mining. We try to find new ways, how to improve and speed up the process of extracting associations from databases. We want to show specific problems of mining association rules in multimedia databases, trying to discover and describe weak points of the whole process, propose new algorithm and introduce chance of using clustering analysis to accelerate the data mining. Endeavour of this project is further growth of efficiency and preparation of suitable conditions for interesting association discovery.

1 INTRODUCTION

Process of data mining becomes very often subject of exploratory articles and studies, up to present often connected only with alphanumeric databases. Because of expansion of the internet and massive distribution of hardware with high performance, multimedia databases come to scene. Image and sound galleries become very popular, while processing of weather and satellite pictures are already very important. Data mining is unavoidable part of these systems and creates very important role of the process.

Multimedia databases have own specific problems which need to be solved. Mainly it is capaciousness of databases and heterogeneousness of data structure. That is why new algorithms and methods have to be found.

2 ASSOCIATION RULE

Data mining process can bring different kinds of interesting information as evolution analysis or deviation analysis, characteristic, classification or association rules. And just association rules are the main topic of this project. They allow us to obtain interesting information about relations among objects in the set based on statistical data.

2.1 PROFILE OF ASSOCIATION RULE

Association rule has a form A⇒ B. We can interpret the rule as: "If set contains element A, then set contains also element B". Use of association rules is most spread in business sphere, where it can have meaning: "If customer will buy commodity of type A, then he will with certain probability buy also commodity of type B". We can find the same utilization in image databases with meaning: "If picture contains object A, then it contains with certain probability also object B".

More formally we can define association rule as follows. Let I = { i_1, i_2, … , i_m} is set of literals called items. Let D be the set of transactions, where each transaction T is set of items, where T ⊆ I. Let A be set of items. We say, that transaction T contains A if and only if A ⊆ T. Association rule is implication of the form A⇒ B, where A ⊂ T, B ⊂ T a A ∩ B = ∅. For use of association rules in image databases the last condition is removed and replaced by not so strong condition B ⊄ A ∧ B ≠ A. That is caused by possible multiple occurrence of arbitrary item in the database.

For our purpose this definition is sufficient. Formal definition is in [1, 3, 4]. Now we have to familiarize with conception of support and confidence.

We know that the association rules need not to be accepted for every item in database. That's why we use two other parameters, which are important for assessment, how the rule is frequent and strong.

- Support – support of the rule A⇒ B is *s*, if *s%* of transactions in database contains A ∪ B.
- Confidence – rule A⇒ B is accepted with confidence *c*, if *c%* of transactions in database, which contain item A, contain also item B. Then we can count confidence of association rule as :

$$c(A \Rightarrow B) = \frac{s(A \Rightarrow B)}{s(A)}$$

These two parameters tell us, how often we can find the rule in database and how is strong. The basic task then is to find all rules, rules which support and confidence is higher then minimal support and minimal confidence determined in advance. The rules are called strong rules.

The whole process of finding association rules can be divided in two parts:

- First step is retrieval of frequent sets of items (itemsets). Frequent itemsets are called *large*. That means we choose from all generated sets those, which support is higher then user's defined minimal support. All other set are signed as *small*.
- Second step is process of generating association rules from large itemsets. We generate all possible rules from large itemsets and then pick up those, which meets user's minimal confidence. We can say, that rule AB ⇒ CD is accepted, if :

$$\frac{s(ABCD)}{s(AB)} = c(AB \Rightarrow CD) \geq \min c$$

Where *s* is support of the rule, *c* is confidence of the rule, *min c* is minimal confidence entered by user. The rule will surely have minimum support because is generated from large itemset.

2.2 ASSOCIATION RULES IN IMAGE DATABASES

Connection of association rule mining and image databases brings some problems, which we did not meet when we work with transactional databases. Mining from image databases needs adjustment of the structure of the database. Classical approach is to transform database to transactional and use actual advantages. Then every image is regarded as transaction. Transaction is represented by identifier and items representing objects in image. But this approach has also many disadvantages. Transaction does not look at attributes, which can be important for image mining as shape, spatial relationship and texture type. With this attributes we can generate many more interesting rules. More can be found in [4].

We can also see that very important is repetition of objects in image. Transactional databases do not care about quantity, since it is not interesting information. In image databases we have to solve this problem. Also we guess that spatial relationship is one of the attributes we can not leave out of consideration. In transactional databases nothing similar exists. Here spatial predicates as “above”, “next to”, “inside” can be fundamental. Especially in combination with color, shape and other attributes new inquired rules can be created.

3 METHODS AND ALGORITHMS FOR DATA MINING

3.1 APRIORI

Groundwork for numerous set of algorithms is Apriori algorithm. It ignores any spatial relationship or multiple occurrences of objects. Algorithm is very simple and if we have already image database in the form of transactional, then is also fast, but to that correspond results. Algorithm works only with one attribute not considering relations to other predicates. From this reason generated set of possible association rules is very small and many interesting algorithms are not generated at all. It is not subject of this work to explain detailed structure of this algorithm. Principle of this algorithm can be found in literature [3].

3.2 MAXOCCUR

Many algorithms try to improve process considering different approach to database using auxiliary structures or special memory procedures. We can discover many problems, which are connected with special type of data. Such a problem can be fore mentioned multiple occurrence. Algorithm MaxOccur try to solve this problem using auxiliary data structure, where the multiple occurrence of each object is stored. Algorithm was introduced in [4]. We will mention modified version of this algorithm in section (4.1). Obtaining information about multiple occurrences and storing it for latter use is neither time nor space consuming. From the view of complexity is this adjustment acceptable. Although we have to find and store information about multiple occurrences, result that the set of possible rules is much larger is very interesting for us.

4 NEW APPROACH

We can go further and try again to widen the generated set of possible association rules. Very interesting idea is to add new items to the candidate itemset. These candidates can arise from a subset of original candidate set with some joint feature.

For example let's have a set of blue circle, red circle, blue triangle. We can generate new objects such as circle without specification of color by joining blue circle with red circle. On the other side we can generate blue object without specification of shape by joining blue circle with blue triangle. From this extended candidate itemset can be again generated more association rules.

Simple solution of this problem is to generate all possible combinations of the objects and subsequently eliminate all those, which are not interesting for further processing. But we have to realize that real databases are larger and larger and such generation of possible combination would be very difficult and time consuming.

4.1 MAXOCCUR WITH CLUSTERING

We propose some solution which is based on clustering analysis which can form new candidate objects to candidate itemset.

This algorithm is based on MaxOccur algorithm. Main difference is extended generating of candidate set. This set is enriched by new items. We want these new items to be generated by cluster analysis. Basic idea is that we put current objects from database as input of cluster analysis, and we receive clusters of "similar" objects. Clustering is gathering objects into clusters based on maximum similarity among objects in the same cluster and maximum dissimilarity among objects from different clusters. That means we can consider these new clusters as new objects, which aroused from original objects based on idea of similar features. Then new objects presenting each cluster can be easily added to the set of candidate items. After this step algorithm continues commonly as MaxOccur algorithm. We used the simplest way for counting of support that is just sum of previous support. We know that more suitable and smart algorithm for support adjusting must be developed.

Solving this problem with cluster analysis seems to be very suitable, because we can work directly with several or all attributes of each object. We can take advantage of idea using geometric representation. Every item in database is represented by object in n-dimensional space. We can then measure similarity as geometric distance among objects. Each attribute of objects in database represents one dimension. While human eyes are very good at cluster recognition in up to three dimensional space, cluster algorithms have no problems with multidimensional space. That means we can work with all features of objects together.

We do not know yet, if we can use directly any of already well-studied and described algorithms for clustering, but Density-based methods seem to be very suitable, because these methods can work with clusters of arbitrary shape. Hierarchical methods can be very effective when searching for clusters in multidimensional space. Probably the best solution is some combination of different types of methods.

Using cluster analysis also brings some problems, which need to be solved. At first clustering analysis certainly brings delay to the whole process. Further lot of clustering algorithms need some input parameters. In our case is certainly hard to assess these inputs or leave this setting on user. Other problem, which is typical for clustering process, is interpretation of results. In the worst case cluster algorithm gives us only list of clusters and we need to interpret these results and transform into objects, which we add to candidate set.

4.2 NOTE

During the working on related topic and considering the use of clustering new idea

comes to play. We can use clustering analysis on a different place in our algorithm. We said that procedure of generating association rules can be divided in two parts. First ensure generating of large itemsets, second one generating association rules from these sets. First step is very important for reducing the complexity of the whole process and discard all sets which are not important for further generation of the rules. That greatly reduces size of data, which we are working with.

Our idea is to go further and once again try to reduce the size of data by reducing the number of sets. It is just proposal and needs more attention. Principle is in using clustering analysis on the set of transactions. Why? Because after clustering the whole set of transactions, we receive set of clusters. Each cluster can be then represented by one transaction. It is plain, that this transaction has common features of the whole cluster and can be suitable replacement for the whole cluster. Great advantage is that all objects in certain distance from center of cluster can be considered as noise and we can remove them since these objects can't affect generating of association rules because of not meeting minimal support.

5 CONCLUSION

We proposed new procedure, how to enlarge set of association rules that we can obtain from given database. This alternate of algorithm needs to be implemented and tested to see its possibilities of use in real applications. Second introduced idea tries to improve process by reducing the set of data we are working with. This is again only suggestion and certainly meets series of problems. But solving them we think can bring so needed time and space savings.

Our task is further study of cluster analysis. We try to find concrete type of clustering algorithm, which can be ideal for our purposes, and consider carefully use of clustering in different sections of data mining. Connection of association rule problem and clustering can bring further improvement and allow us obtaining more information from our data set.

REFERENCES

[1] Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules, San Jose, IBM Almaden Research Center, 1994

[2] Han, J., Kamber, M.: Data Mining: Concepts and Techniques, Morgan Kaufrnann Publisher, 2000

[3] Kotásek, P.: Knowledge discovery from databases – association rules [semestral project], Brno University of Technology, Brno, Czech Republic 1997

[4] Zaiane, O. R., Han, J., Zhu, H.: Mining Recurent Items in Multimedia with Progressive Resolution Refinement, Proc. 2000 Int. Conf. on Data Engineering (ICDE'OO), San Diego, CA, March 2000

[5] Zaiane, O. R., Han, J., Li, Z., Chee, S. H., Chiang, J. Y.: MultiMediaMiner: A System Prototype for MultiMedia Data Mining, In: Proc. 1998 ACM-SIGMOD Conf. on Management of Data, (system demo), Seattle, Washington, June 1998, pp. 581–583