SUGGESTIONS TO EXTRACT ASSOCIATION RULES WITH MULTIDIMENSIONAL DATABASE+

Hilal Hadi Salih* Soukaena Hassan Hashem**
Shaimaa Akram Hassan***

Abstract:

Data mining derives its name from the similarities between searching for valuable information in a large database and mining rocks for a vein of valuable ore. The term data mining is actually misnomer, since mining for gold in rocks is usually called "gold mining" and not "rock mining", thus by analogy, data mining should have been called "knowledge mining" instead. The more general terms such as Knowledge Discovery in Databases (KDD) describe a more complete process.

This research concentrates on one particular aspect to extract the association rules from multidimensional databases by the following:

It is proposed to deal with multidimensional database to extract the association rules that is done by dividing the multidimensional database into two databases the first one consists of TIDs and the items, while the second one consists of TIDs and dimensions. Then extracting the frequent itemsets for each one separately, the frequent itemsets for the first one is extracted by the traditional apriori algorithm but for the second one a new algorithm has been proposed for that purpose. The two obtained sets will be combined into one set. Finally applying the association rules generation algorithm to get the final rules of the multidimensional database.
Keywords: multidimensional database, association rules algorithm.

المستخلص

تم اشتقاق مصطلح تعدين البيانات من التشابه الحاصل بين البحث عن المعلومات القيمة في قواعد البيانات الكبيرة والبحث في المعدن الخام. مصطلح تعدين البيانات هو مصطلح خاطىء لان البحث عن الذهب في الصخور يسمى تنقيب الذهب وليس تنقيب الصخور. اذا تنقيب البيانات يجب ان يصطلح عليها تنقيب المعرفة والمصطلح الأشمل هو اكتشاف المعرفة في قواعد البيانات.

في هذا البحث تم التركيز على استخلاص قوانين الأرتباط في قواعد البيانات متعددة الابعاد من خلال:
تجزأة قاعدة البيانات متعددة الأبعاد الى قاعدتين الأولى تحوي الارقام التعريفيه والأبعاد والثانية تحوي الارقام التعريفيه والعناصر ومن ثم استخلاص مجموعة العناصر المتكررة لكل جزأ على حدة. في الجزأ الأول تم استخدام الطرق التقليدية، اما في الجزأ الثاني فقد تم اقتراح خوارزمية جديدة لهذا الغرض. وأخيرا تم دمج النتائج من المرحلتين وتوليد قوانين ارتباط نهائية.

1- Introduction:

+ Received on 28/11/2007 , Accepted on 21/7/2008

* Prof/ University of Technology/ Baghdad-Iraq

** Lecturer /University of Technology/ Baghdad-Iraq

*** Technical Trainer /Medical- Technical Insititue- Mansur / Baghdad-Iraq

Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships that can be hidden among vast amount of data. From these patterns and relationships, businesses and organizations can make valid predictions about future trends in all areas of business. Data mining tools can answer business questions that had traditionally been very time consuming to resolve. These tools find hidden patterns and predictive information that experts might have missed because they laid outside of their expectations.

Data mining can be divided in two sub groups: discovery and exploration. Discovery in the sense that meaningful patterns are uncovered in data and distinguished properly and exploration means that these meaningful patterns are used to create useful applications for data modeling purposes. The process of discovering knowledge in data is referred to as knowledge discovery. It is when data is being worked on from a conventional data store not a data warehouse or an online transaction system. Knowledge discovery can be performed in two ways either manually or through an automated process.

This process of knowledge discovery is finding a connection between data and facts. It can, in fact, be said to be a relationship between prior facts and subsequent facts. The prior fact is called the antecedent and the subsequent fact is known as the consequent. There should be no assumption of an underlying cause and effect connection between the antecedent and consequent. Through the process of data mining, it should always be kept in mind that data is not facts, in order to have knowledge and facts one needs data. Knowledge is the main connection between antecedent facts and consequent facts. This knowledge plays the main role between antecedent and consequent facts and takes the form of several shapes in data mining. Knowledge can be in the form if-then-else rules that are known as Knowledge Based Expert Systems (KBES) or it can be in form of a complex mathematical transformation. There are four main techniques used in data mining: *Auto-clustering, link analysis, visualization, and the rule induction.* These four techniques are used in the creation of knowledge information based on data from the database [1, 2, 3].

2- Apriori Association Rule:

The efficient discovery of association rules has been a major focus in the data mining research community. Many algorithms and approaches have been proposed to deal with the discovery of different types of association rules discovered from a variety of databases. However, typically, the databases relied upon are alphanumerical and often transaction-based. The problem of discovering association rules is to find relationships between the existence of an object (or characteristic) and the existence of other objects (or characteristics) in a large repetitive collection. Such a repetitive collection can be a set of transactions for example, also known as the market basket. Typically, association rules are found from sets of transactions, each transaction being a different assortment of items, like in a shopping store ({milk, bread, etc}). Association rules would give the probability that some items appear with others based on the processed transactions, for example milk →bread [50%], meaning that there is a probability 0.5 that bread is bought when milk is bought. Essentially, the problem consists of finding items that frequently appear together, known as frequent or large itemsets [4].

The problem is stated as follows, see Table 1, Let $I = \{i1, i2, ...im\}$ be a set of literals, called items. Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. A unique identifier *TID* is given to each transaction. A transaction T is said to contain X, a set of items in I, if $X \subseteq T$. An *association rule* is an implication of the form "$X \Rightarrow Y$", where $X \subseteq I$, $Y \subseteq I$, and $X \cap Y = \varnothing$. The rule $X \Rightarrow Y$ has a *support s* in the transaction set D is s% of the transactions in D contain $X \cup Y$. In other words, the support of the rule is the probability that X and Y hold together among all the possible presented cases. It is said that the rule $X \Rightarrow Y$ holds in the transaction set D with *confidence c* if c% of transactions in D that

contain X also contain Y . In other words, the confidence of the rule is the conditional probability that the consequent Y is true under the condition of the antecedent X. The problem of discovering all association rules from a set of transactions D consists of generating the rules that have a *support* and *confidence* greater that given thresholds. These rules are called *strong rules [4, 5]*.

The basic algorithm (Apriori) is shown below:

```
APPIOPI(data, minsup, mincon):
 1  tcount ← length(data)
 2  items ← LISTUNIQUEITEMS(data)
 3  icount ← length(items)
 4  scount ← max(tcount * minsup /100, 1)

 5  f1 ← FREQUENCIES(items, data)
 6  REMOVEINFREQUENT(f1, scount)
 7  if not f1: return NIL

 8  for k ← 2 to icount:
 9          candidates ← GENERATECANDIDATES(fk−1)
10          if not candidates: break
11          fk ← FREQUENCIES(candidates, data)
12          REMOVEINFREQUENT(fk, scount)
13          if not fk: break
14  return BUILDASSOCIATIONS(f, mincon)
```

```
GENERATECANDIDATES(fk):
1  candidates ← NIL
2  for u ∈ fk, v ∈ fk, u < v:
3          if u1:-1 = v1:-1:
4                  c ← u ∪ v-1
5                  s ← SUBSETS(c)
6                  if length(filter(λx : x ∈ fk, s)) = length(s):
7                          candidates.append(c)
8  return candidates
```

```
BUILDASSOCIATIONS(f, mincon):
 1  rules ← NIL
 2  for k ∈ fis, itemset ∈ f_k, i ← 1 to length(itemset):
 3          for c ∈ COMBINATIONS(itemset, i):
 4                  lhs ← c
 5                  rhs ← NIL
 6                  for k ∈ itemset: if k ∉ c: rhs.append(k)
 7                  confidence ← 100 × SUPPORT(itemset, f) / SUPPORT(lhs, f)
 8                  if confidence > mincon:
 9                          support ← SUPPORT(itemset, f)
10                          rules.append(lhs → rhs(support, confidence))
11  return rules
```

3- The Proposed System:

To get a clear understanding of the present study, it will be organized to the following points:

First step: Take the multidimensional database, see Table (1).

Table (1) multidimensional database has two dimensions with items.

TID	D1	D2	items
1	A	1	ABC
2	A	1	AB
3	B	1	AB
4	B	3	ABC

Second step: divide the multidimensional database in Table (1) into two databases, the first one contain the TID and the two dimensions while the second one contains the TID and items, see Table (2).

Table (2): the first and second databases.

TID	D1	D2
1	A	1
2	A	1
3	B	1
4	B	3

TID	items
1	ABC
2	AB
3	AB
4	ABC

Third step: apply the traditional apriori association rule on the second database to extract the frequent itemset. These frequent itemsets are {A, B, C, AB, AC, BC, ABC} since the minimum support was assumed to be 2 TIDs.

Fourth step: apply the proposed algorithm to extract the frequent itemsets from the first database.

1- Take the first dimension D1 and find the frequent items in it, for example the threshold of the minimum support is equal to 2 TIDs, two frequent items(a,*) and (b, *) can be found because the value a and b appears two times; the value * for the other dimension means that it is not relevant. Repeat the operation on the second dimension to find the frequent item (*, 1) only because the value 1 appears three times; value * for the other dimension means that it is not relevant.
2- Now take the first dimension with concern of the second dimension the resulting frequents items are: (a, 1) and (b, 1). Since there are no more dimensions in above example, the search will be terminated.
3- If there are more than two dimensions then the same procedure will be applied, taking each dimension alone, then taking each dimension related with the others.
4-

Fifth step: the general frequent itemsets for the multidimensional database will be as in the follow:

{A, B, C, AB, AC, BC, ABC, (a, *), (b, *), (*, 1), (a, 1), (b, 1)}

Sixth step: The association rules will be generated using apriori algorithms according to the frequent itemsets displayed in the previous step.

4- The Implementation:

To explain the proposed system and to prove its quality, it will be implemented using visual basic programming language as follows:

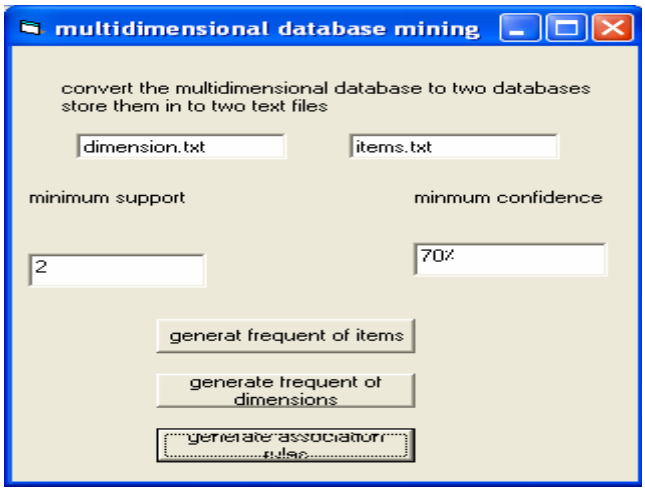

Figure (1): the main window for the proposed system.

In Figure (1) the multidimensional database will be divided into two databases and each one will be stored in a notepad file. The name of these two files will be entered in the first two textboxes. In the other two textboxes the value of the minimum support and minimum confidence will be enteredand when

-The first command is activated; the frequent itemsets of the database contain the TIDs and items will be found and displayed, see Figure (2).

-The second command is activated; the frequent itemsets of the database contain the TIDs and dimensions will be found and displayed, see Figure (3).

-The third command is activated; the association rules of the two types of frequent itemsets will be generated and displayed, see Figure (4).

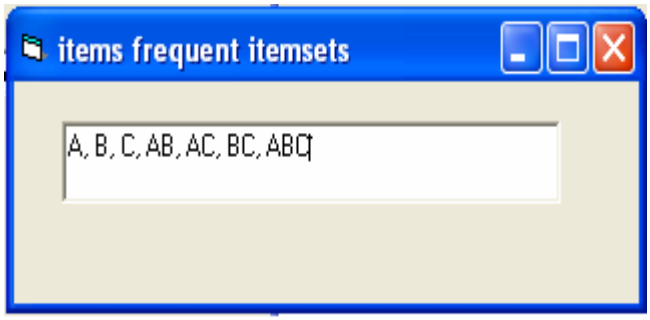

Figure (2): the main window for generating frequent itemsets of TIDs and items

.

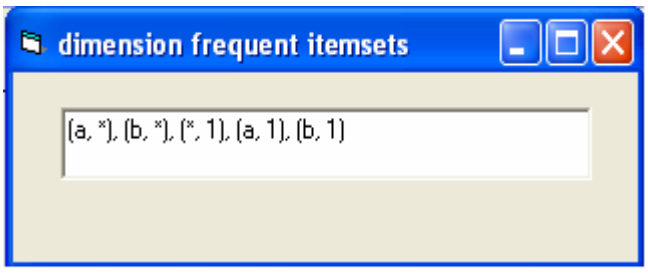

Figure (3): the main window for generating frequent itemsets of TIDs and dimensions.

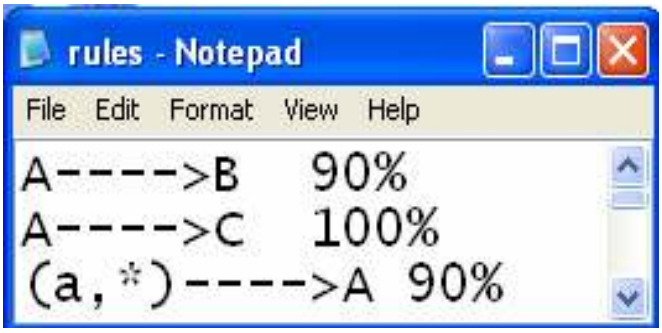

Figure (4): the main window for generating the association rules.

5- Conclusions:

From the present study, the following conclusions can be summarized:

1) Direct mining of a multidimensional database is not efficient because it has inconsisten values of items and dimensions.

2) For more efficient results, it is suggested to divide the multidimensional database into two databases the first one contains the TIDs and items and the other contains the TIDs and dimensions.
3) The frequent itemsets of the TIDs and items database will be found by applying the traditional apriori algorithms. To treat the second part of the database the proposed algorithm is used.
4) To strengthen the proposed system of mining, the resulting frequent itemsets are combined together to generate the association rule for complete database. This unifies the multidimensional in one block.

6- References:

1) M. S. Chen, J. Han, and P. S. Yu. "Data mining: An overview from a database perspective". IEEE Trans. Knowledge and Data Engineering, 8:866-883, 1996.
2) U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. "Advances in Knowledge Discovery and Data Mining". AAAI/MIT Press, 1996.
3) J. Han and M. Kamber. "Data Mining: Concepts and Techniques". Morgan Kaufmann, 2000.
4) Mitra S., and Ahharya T., "Data Mining Multimedia, Soft Computing, and Bioinformatics", John Wiley and Sons, Inc., 2003.
5) Mohammadian M., "Intelligent Agent for DM and Information Retrieval", Idea Group Publisher, 2004.