International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Study of Optimization in Fragmented Item-sets for Business Intelligence

Rajesh V. Argiddi[1], Bhagyashri U. Kale [2]

[1]Assistant Professor Department of Computer Science & Engineering, Walchand Institute of Technology Solapur, India

[2]Department of Computer Science & Engineering, Walchand Institute of Technology Solapur, India

Abstract: ***Association Rule is one of the techniques in the process data mining problems and it might be the most researched one. Some of them have developed fragment rule mining which is based on associations among large data set. Discovering item sets is the key point in fragment rule mining. Major challenge in developing fragment rule mining algorithms is the extremely large number of rules generated which makes the algorithms inefficient and makes it difficult for the end users to cope up with the generated rules. In this research, we concentrate on optimization of fragmented items sets generated from fragment rule mining. We proposed an innovative approach to find optimized association rules within inter-transaction of fragment mining. Design of this method represented in this paper which gives idea of fragmented item sets generated from fragment rule mining on which optimization is performed. This deals mainly with reducing the time and space complexity required for processing the data using fragment mining & generate strong rules using genetic algorithm.. This also reduces the width of sliding window for large data as compared to FITI because of fragmented attributes. Genetic algorithm heuristic is mainly used to generate useful solutions to optimization and search problems. In previous research many have proposed genetic approach for mining interesting association rules from large dataset. In this paper we propose knowledge based method which provides the major advancement, integrating a genetic algorithm in fragment rule mining to obtain effective rules that potentially be used for business intelligent applications.***

Keywords: Association Rule, Business Intelligence, Fragment Mining, Genetic Algorithm

1. Introduction

An informal definition of Knowledge Discovery in Databases (KDD) is to find useful and interesting patterns in data. Data mining is one of the tasks of KDD and is defined as a method to find a part of data which has interesting common features. Most of data mining methods that have been proposed to achieve the task try to find interesting patterns in database. Data mining and knowledge discovery in databases (or KDD) is used as synonyms for each other but data mining makes use of algorithm to find out patterns in the knowledge discovery process. The KDD process generally involves a following processing steps, namely, data selection, feature selection, transformation, Data mining, Presentation and evaluation as shown in Figure1.

Classification is used to analyses the relationship between attributes and classes of objects in transaction table. Clustering is used to identify the classes where they also viewed as groups for set of objects whose classes are unknown. Association refers to discovery of associative relationship among objects [2]. Various techniques, such as statistical analysis, machine Learning, information theory and association rule mining have been used for extraction of knowledge in the literature. Our methodology is based on association rule mining. Traditional association rule mining algorithms focus on association rules among item sets within a transaction. This classical association rule expresses the associations among items within the same transaction, thus we call it intratransactional association rule

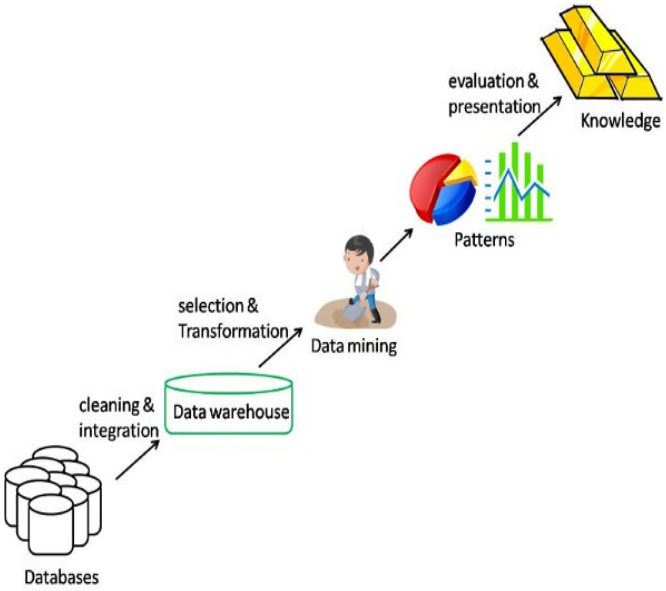

Figure 1: KDD process

Most of work has been carried out on Intra-transaction association rule mining. Intertransaction association indicates association among different transactions [3]. Work related to Inter-transaction association mining was proposed in 2000 and has a broad range of applications, though its basic idea extended from intratransaction association mining. [4]

Financial institutions such as stock markets produce huge datasets that build a foundation for approaching these enormously complex and dynamic problems with data mining tools. Association rule [9] is a technique to detect the hidden facts in large dataset and draw interferences on how subsets of items influence the presence of other subsets. Association mining mostly fits best to direct customer oriented

businesses, the rules generated from this technique is suitable for gaining knowledge .Association rule mining aims to find strong relation between attributes. Apriori algorithm best for single attribute rule generation, Apriori requires more time as the number of transactions gets increased. After this FITI (First Intra-transaction then Inter-transaction) algorithm was introduced, but the drawback of FITI is , its efficiency decreases as the number of transactions increases. To overcome this WangZhong [1] proposed a technique Granule based mining, which allows group of transactions based on common features of the transactions. Further, Prof.R.V.Argiddi [2] had used this approach granule based mining as fragment based mining ,in which they elaborated the work by evaluating a single attribute behavior based on group of attributes and also validated their predictions.

The original problem addressed by Fragment rule mining was to find a correlation among sales of different products from the analysis of a large set of data & generates large number of rules. But users are not interested in all association rules, they are just concerned about the associations among condition attributes and decision attributes. To overcome this we introduced genetic approach for optimization of rules generated from fragment based mining [3]. Genetic algorithm is a family of computational models based on principles of evolution and natural selection .These algorithms convert the problem in a specific domain into a model by using a chromosome-like data structure and evolve the chromosomes using selection, crossover and mutation operators. The range of the applications that can make use of genetic algorithm is quite broad [10].

2. Background

This research integrates issues from the research field of Data Mining, Business Intelligence, Genetic Algorithm, and Association Rule Mining. The following subsections include a brief overview of these topics and their relation to the newly proposed methodology.

2.1 Data Mining for Business Intelligence

Data mining, if done in right direction, can provide an organization a way to optimize its processing of business data. Now days, in market new data mining companies are springing up to the challenge of providing this service. Though data mining is improving the interaction between a business organization using data warehouse and its customers, there are many companies that are trying to vertically integrate to offer the best services to broad markets. This is done by focusing on a particular market like IT industry and trying to understand the types of information collected by industries in that sector. Data mining is then the process of extracting out valid and yet previously unknown & relevant information from large databases and using it to make critical business decisions Data warehouse or exploratory data analysis with large and complex database brings together the wealth of knowledge and research in statistics and machine learning for the task of discovering new patterns of knowledge in very large datasets. The phrase knowledge discovery in databases (Figure1) refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process.

2.2 Association rule mining

Association rule mining [11] is a technique to detect the hidden facts in large dataset and draw interferences on how subsets of items influence the presence of other subsets. Association mining mostly fits best to direct customer oriented businesses, the rules generated from this technique is suitable for gaining knowledge. Association rule mining aims to find strong relation between attributes. All frequent generalized patterns are not very efficient because a portion of the frequent patterns are redundant in the association rule mining. This is why this algorithm produces some redundant rule along with the interesting rule. This drawback can be overcome with the help of genetic algorithm.

2.3 Genetic Algorithm

Genetic algorithm is a part of computational models based on thesis of evolution and natural selection. These algorithms convert the large space problem from any domain into a model by using a chromosome and evolve that chromosome using selection, crossover and mutation operators. In computer security software's , it is mainly used for finding optimum solutions to a specific problem.[4] The implementation of the fragment mining algorithms tends to lead to a very large number of association rules. In most of the case is that they will be confusing and we will have to examine each one of them very carefully to determine if it is of any interest to us. So, This paper tends to mine reasonable trading rules using genetic algorithms future. Genetic algorithm is problem solving heuristics that performs the process of natural evolution. Unlike artificial neural networks (ANNs), designed to function like neurons in the brain, these algorithms utilize the concepts of natural selection to determine the best solution for a problem. As a result, GA is commonly used as optimizers that adjust parameters to minimize or maximize some feedback measure, which can then be used independently.

3. Related Work

The problem of discovering association rules was first introduced in [5] and an algorithm called AIS was proposed for mining association rules. For last fifteen years many algorithms for rule mining have been proposed. Lu et al. [11] first proposed the concept of inter association mining and contributed E-Apriori and EHApriori algorithms to this area. To improve the performance .Moreover, Tung et al. [7] recently proposed the FITI (First Intratransaction Then Intertransaction) algorithm. In FITI, if the average size of the transactions is very large, the extended transactions should be very long. It generates many extra combinations of items because the set of extended items is much larger than the set of items. Thus, this method is very slow if the average size of the transactions is large.

To overcome this, Wanzhong Yang also proposed one innovative technique to process the stock data named

Granule mining technique, which reduces the width of the transaction data and generates the association rules. [1]R.V.Argiddi has proposed fragment based mining which deals mainly with reducing the time and space complexity involved in processing the data in association rule mining technique[2].As in granule mining, fragment based approach fragments the data sets into fragments for processing thereby reducing the input size of data sets fed to the algorithm. In contrast to granule mining, in fragment based mining the condition and decision attributes are summed for obtaining generalized association rules. Kannika Nirai Vaani M, E Ramaraj has now proposed new approach to generate association rules i.e pointing the faster generation of frequent item sets, so that to offer interesting rules in an effective and optimized manner with the help of Genetic Algorithm[4][5].

4. Proposed Methodology

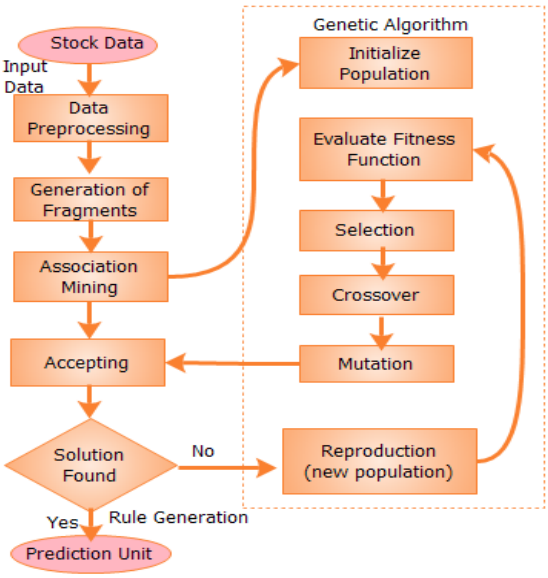

Figure 2: Design of Proposed Methodology

We propose data mining approach using genetic algorithms (GA) to solve the knowledge acquisition problems that are inherent in constructing and maintaining rule-based applications for business. Although there are an infinite number of possible rules by which we could trade, but only a few of them would have made us a profit if we had been following them. This study intends to find good sets of rules which would have made the most money over a certain historical period. Figure 2 represents the design of proposed method. It consists of two major parts: Fragment Rule Mining & Genetic algorithm on covering sets formed after fragmentation.

a) Fragmented rule mining:

In Fragmented rule mining, consider IT stock market dataset for future prediction here we first differentiate the companies based on small and large scale, grouping is done based on the capitalization of the company. This files are stored in the form of excel sheet at the back end. Once we select the dataset as shown in Table 1. Let, A1, A2, A3, A4 are the share market companies B1, B2, B3 are the share investors who interested to invest in market. Data is extracted from the excel sheet and we can apply the fragment mining [6] on this data and generate the rules for prediction. Steps in fragment rule mining are as follows:

1. First, Convert the data into 1's and 0's, this is done by performing operation such as Transaction1=Transaction2 - Transaction1, and if transaction1>0 we put 1, if transaction1<0 we put -1, otherwise 0. Table 2. Shows the converted table.

Table 1: Selected dataset

ID	Date	A1	A2	A3	A4	B1	B2	B3
0	02/01/2008	15.7	17.78	31.75	29.9	18.45	7.7	2.82
1	03/01/2008	15.84	17.96	31.89	30.09	18.88	7.84	2.9
2	04/01/2008	15.71	18.0	32.08	29.92	18.8	7.75	2.86
3	07/01/2008	15.5	17.85	31.88	29.9	19.03	7.75	2.86
4	08/01/2008	15.35	17.69	31.8	29.95	19.11	7.78	2.84
5	09/01/2008	15.15	17.25	31.34	29.98	18.88	7.68	2.76
6	10/01/2008	14.97	17.2	31.0	29.50	18.97	7.61	2.72
7	02/01/2008	15.15	17.35	31.1	29.72	19.0	7.82	2.8
8	02/01/2008	15.17	17.19	31.3	30.3	18.9	7.83	2.8
.....								
18	29/01/2008	15.46	17.31	33.52	32.1	19.35	7.37	2.9
19	30/01/2008	15.7	17.71	34.15	32.75	19.15	7.5	2.86

Table 2: Converted dataset

ID	A1	A2	A3	A4	B1	B2	B3
0	1	1	1	1	1	1	1
1	-1	1	1	-1	-1	-1	-1
2	-1	-1	-1	-1	1	0	0
3	-1	-1	-1	1	1	1	-1
4	-1	-1	-1	1	-1	-1	-1
5	-1	-1	-1	-1	1	-1	-1
6	1	1	1	1	1	1	1
7	0	-1	1	1	-1	-1	0
8	1	-1	-1	1	-1	-1	1

18	1	-1	1	-1	1	-1	-1
19	1	1	-1	1	-1	1	-1

2. Divide dataset attributes in to fragments depending on scale .in case of IT market, small scale companies are consider as condition fragments & large scale companies are under decision fragments. Table3. shows the 2-tier fragments of dataset. While assign $a_{k,1}$ is for 1,$a_{k,2}$ is for 0 & $a_{k,3}$ is for -1 (k=1,2,3,4) from converted table

Table 3: Two-tier structure of Fragments

	Condition Fragments				Decision Fragments		
ID	A1	A2	A3	A4	B1	B2	B3
0	$a_{1,1}$	$a_{2,1}$	$a_{3,1}$	$a_{4,1}$	$b_{1,1}$	$b_{2,1}$	$b_{3,3}$
1	$a_{1,3}$	$a_{2,1}$	$a_{3,1}$	$a_{4,3}$	$b_{1,3}$	$b_{2,3}$	$b_{3,3}$
2	$a_{1,3}$	$a_{2,3}$	$a_{3,3}$	$a_{4,3}$	$b_{1,1}$	$b_{2,3}$	$b_{3,3}$
3	$a_{1,3}$	$a_{2,3}$	$a_{3,3}$	$a_{4,1}$	$b_{1,1}$	$b_{2,1}$	$b_{3,1}$
4	$a_{1,3}$	$a_{2,3}$	$a_{3,3}$	$a_{4,1}$	$b_{1,3}$	$b_{2,3}$	$b_{3,2}$
5	$a_{1,3}$	$a_{2,3}$	$a_{3,3}$	$a_{4,3}$	$b_{1,3}$	$b_{2,3}$	$b_{3,1}$
6	$a_{1,1}$	$a_{2,1}$	$a_{3,1}$	$a_{4,1}$	$b_{1,1}$	$b_{2,3}$	$b_{3,1}$
7	$a_{1,2}$	$a_{2,3}$	$a_{3,1}$	$a_{4,1}$	$b_{1,3}$	$b_{2,3}$	$b_{3,1}$
8	$a_{1,1}$	$a_{2,3}$	$a_{3,3}$	$a_{4,1}$	$b_{1,3}$	$b_{2,3}$	$b_{3,3}$

15	$a_{1,1}$	$a_{2,3}$	$a_{3,3}$	$a_{4,1}$	$b_{1,1}$	$b_{2,3}$	$b_{3,3}$
16	$a_{1,1}$	$a_{2,3}$	$a_{3,1}$	$a_{4,3}$	$b_{1,3}$	$b_{2,1}$	$b_{3,3}$

3. After that group the transactions of condition fragments based on similar rows and form the covering set as shown in below Table.

Table 4: Condition fragments

	Condition Granules					
ID	A1	A2	A3	A4	N	Covering Set
1	$a_{1,1}$	$a_{2,1}$	$a_{3,1}$	$a_{4,1}$	2	(1,7)
2	$a_{1,3}$	$a_{2,1}$	$a_{3,1}$	$a_{4,3}$	4	(2,10,12,14)
3	$a_{1,3}$	$a_{2,3}$	$a_{3,3}$	$a_{4,3}$	2	(3,6)
4	$a_{1,3}$	$a_{2,3}$	$a_{3,3}$	$a_{4,1}$	2	(4,5)
5	$a_{1,2}$	$a_{2,3}$	$a_{3,1}$	$a_{4,1}$	1	(8)
6	$a_{1,1}$	$a_{2,3}$	$a_{3,3}$	$a_{4,1}$	2	(9,15)
7	$a_{1,1}$	$a_{2,1}$	$a_{3,1}$	$a_{4,3}$	1	(11)
8	$a_{1,3}$	$a_{2,3}$	$a_{3,1}$	$a_{4,1}$	1	(13)
9	$a_{1,1}$	$a_{2,3}$	$a_{3,1}$	$a_{4,3}$	1	(16)

4. Next, Consider decision attributes, here perform aggregation of all the Large scale companies and find the minimum and maximum range of these companies, this is because the stock market have very fluctuating data.

5. Further we will find the positive and negative gains of decision attributes based on window size (suppose w=4) i .e inter transaction association is done as shown in Table 5

Table 5: Aggregation of decision fragments

ID	Date	SUM	99.7% of SUM	100.3% of SUM	Delta SUM
1	03/01/2008	29.62	29.53	29.70	1
2	04/01/2008	29.41	W1		-
3	07/01/2008	29.64			-
4	08/01/2008	29.73			-

7	11/01/2008	29.62	29.53	29.70	-1
8	14/01/2008	29.53	W7		
9	15/01/2008	29.16			
10	16/01/2008	29.18			

18	29/01/2008	29.62	29.53	29.70	
19	30/01/2008	29.51	29.44	29.59	

Now, Combine both the tables of large and small scale company as shown in Table 6. Finally after completing all the above procedure the rules will be generated on the basis of minimum support & confidence.

Table 6: Intertransaction association in fragment mining

	Condition Granules				Decision Granules				
ID	A1	A2	A3	A4	Delta SUM=1	Delta SUM=0	Delta SUM=-1	N	Covering Set
1	$a_{1,1}$	$a_{2,1}$	$a_{3,1}$	$a_{4,1}$	1		1	2	(1,7)
2	$a_{1,3}$	$a_{2,1}$	$a_{3,1}$	$a_{4,3}$			1	4	(2,10,12,14)
3	$a_{1,3}$	$a_{2,3}$	$a_{3,3}$	$a_{4,3}$	1		1	2	(3,6)
4	$a_{1,3}$	$a_{2,3}$	$a_{3,3}$	$a_{4,1}$	1			2	(4,5)
5	$a_{1,2}$	$a_{2,3}$	$a_{3,1}$	$a_{4,1}$	1			1	(8)
6	$a_{1,1}$	$a_{2,3}$	$a_{3,3}$	$a_{4,1}$	2	1	1	2	(9,15)
7	$a_{1,1}$	$a_{2,1}$	$a_{3,1}$	$a_{4,3}$	1			1	(11)
8	$a_{1,3}$	$a_{2,3}$	$a_{3,1}$	$a_{4,1}$	1		1	1	(13)
9	$a_{1,1}$	$a_{2,3}$	$a_{3,1}$	$a_{4,3}$	2			1	(16)

Support The rule X ⇒ Y holds with support s if s% of transactions in data contains X ∪ Y. Rules that have s greater than a user-specified support is said to have minimum support.

Confidence The rule X ⇒ Y holds with confidence c if c% of the transactions in data that contain X as well as Y. Rules that have c greater than a user-specified confidence is said to have minimum confidence.

Before generation of rules it gives frequent pattern set. i.e covering set (Fragments). As shown in figure 8. The genetic algorithm has been applied on the generated fragmented item sets to generate the effective & interesting rules containing positive gain.

b) Genetic Algorithm:
A genetic algorithm [9] is a type of searching algorithm. It searches the solution space for an optimal solution to a problem. The algorithm creates a population of possible solutions to the problem and lets them solve over multiple generations to find better and better solution. Algorithm starts with a set of selected chromosomes called population. Here, Table 6. Data is given as set of population.

1. Begin: Generate random population of suitable solutions for the given fragmented item sets.(chromosomes)

2. Fitness Function: Deciding fitness function is crucial part of genetic .Calculate the fitness function f(x) of each chromosome x in the population.

f(x)= Support(x)/MinSupport

3. New population: Reproduce population by repeating following steps until the new population is complete

Selection
New Population is selected from the initial population to be parents to crossover. According to Darwin's evolution theory the best ones should survive and create new offspring. There are many methods how to select the best chromosomes on the basis of fitness function. Ex. Tournament selection, Rank selection, Roulette wheel selection, Boltzmann selection etc.

Crossover
After encoding into binary string, we can move towards crossover. Crossover selects genes from parent chromosomes and creates a new child. There are different ways like single point and multipoint crossover.

Chromosome 1 11011 | 00100110110
Chromosome 2 11011 | 11000011110

Child 1 11011 | 11000011110
Child 2 11011 | 00100110110

Mutation
After a crossover, mutation is carried out. Mutation is used to maintain genetic diversity from one generation of a population to the next by involving random small changes. This is to avoid getting all solutions in population into a local optimal of solved problem. Mutation replaces randomly the new offspring. For binary encoding we can switch some randomly chosen bits from 0 to 1 or from 1 to 0. Mutation is carried out in following way:

Chromosome 1 1101100100110110
Chromosome 2 1101111000011110

Accepting: Assign new child to reproduce population .

4. Reproduction: Use new generated population for a further run of algorithm.
5. Terminating condition*:* If the end condition is satisfied, stop, and return the set of rules in current population, else go to step 2.

5. Conclusion & Future Trend

The aim of this paper is to improve the performance of the fragment rule mining algorithm that mines fragmented rules by presenting fast and scalable algorithm for discovering effective rules in large databases. For this we presented data mining approach for Fragmented item-sets with genetic algorithm. Association mining methods, are usually accurate, but have very large and meaningless results. Genetic algorithms on the other hand provide a robust and effective approach to explore large search space. In recent years lots of work has been carried out using genetic algorithm for mining association rules. This paper studies the existing work on application of Genetic algorithm in mining association rules and uses it as business intelligence tool. Future work includes applying our proposed approach to real data like Indian IT stock market, retail sales transaction and medical transactions to confirm the experimental results in the business area.

References

[1] Wanzhong Yang, "Granule Based Knowledge Representation for Intra and Inter Transaction Association Mining", Queensland University of Technology, July 2009
[2] R.V Argiddi,S.SApte " study of association rule mining in fragmented item-sets for prediction of transactions outcome in stock trading systems" IJCET-2012.
[3] KANNIKA NIRAI VAANI M, E RAMARAJ "AN INTEGRATED APPROACH TO DERIVE EFFECTIVE RULES FROM ASSOCIATION RULE MINING USING GENETIC ALGORITHM" IEEE2013 INTERNATIONAL CONFERENCE.
[4] Kannika Nirai Vaani M, E Ramaraj" E-Rules: An Enhanced Approach to Derive Disjunctive and useful Rules from Association Rule Mining without Candidate Item Generation" IJCA-2013.
[5] R. Agrawal, T. Imielinski, and A. Swami. "Mining association rules between sets of items in large databases". In Proceedings of the ACM SIGMOD International Conference on Management of Data (ACM SIGMOD '93), pages 207216, Washington, USA, May 1993.
[6] R.V Argiddi, S.SApte " Future Trend Prediction of Indian IT Stock Market using Association Rule Mining of Transaction data" IJCA-2012
[7] Anandhavalli M, Suraj Kumar Sudhanshu, Ayush Kumar and Ghose M.K. *"Optimized association rule mining using genetic algorithm"*, Advances in Information Mining, ISSN: 0975–3265, Volume 1, Issue 2, 2009
[8] Prashant S. Chavan, Prof. Dr. Shrishail. T. Patil" Parameters for Stock Market Prediction" IJCTA | Mar-Apr 2013 Vol 4 (2),337-340
[9] Kalyanmoy Deb, "Introduction to Genetic Algorithms", Kanpur Genetic Laboratory (Kangal), Depart of Mechanical Engineering, IIIT Kanpur 2005.
[10] Nikhil Jain,Vishal Sharma,Mahesh Malviya "Reduction of Negative and Positive Association Rule Mining and Maintain Superiority of Rule Using Modified Genetic Algorithm" International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Volume-2 Number-4 Issue-6 December-2012
[11] Lu,H.,Han,J.andFeng "Beyond intratransaction association analysis: mining multidimensional intertransaction association rules", *ACM Transactions on Information Systems*, 18(4), pp.423 - 454, 2000