Volume: 07 Issue: 04 Pages: 2790-2797 (2016) ISSN: 0975-0290

Comparative Study of Improved Association Rules Mining Based on Shopping System

Tang Zhi-hang
School of Computer and Communication, Hunan Institute of Engineering Xiangtan 411104, China
Email:tang106261@126.com

ABSTRACT

Data mining is a process of discovering fascinating designs, new instructions and information from large amount of sales facts in transactional and interpersonal catalogs. Since inventory databases, universal product bar codes and scanners, and other such supply chain management technologies have been around for years, the idea of using data to help manage retail operations is not new. However, more recently, the use of data mining to more thoroughly understand patterns of consumer behavior that affect retail operations has become more prevalent. In order to truly understand consumer behavior though, it is beneficial to understand both what they buy and who they are. With this information, we can go beyond traditional inventory management, and craft a much more personalized shopping experience for you. All organizations that collect, store, and analyze data have a responsibility to protect privacy, to guard against misuse and abuse, and to share data only within the constraints of fairly developed and disclosed policies. It will be able to expand and apply effective marketing strategies and in disease identification frequent patterns are generated to discover the frequently occur diseases in a definite area. The conclusion in all applications is some kind of association rules (AR) that are useful for efficient decision making.

Keywords: Association rule mining, FP growth, decision making

Date of Submission: Feb 02, 2016 Date of Acceptance: Feb 05, 2016

1. INTRODUCTION

Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Although data mining is a relatively new term, the technology is not. Companies have used powerful computers to sift through volumes of supermarket scanner data and analyze market research reports for years. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost.

A typical and widely-used example of association rule mining is Market Basket Analysis. Data are collected using bar-code scanners in supermarkets. Such market basket databases consist of a large number of transaction records. Each record lists all items bought by a customer on a single purchase transaction. Managers would be interested to know if certain groups of items are consistently purchased together. They could use this data for adjusting store layouts (placing items optimally with respect to each other), for cross-selling, for promotions, for catalogue design and to identify customer segments based on buying patterns.

Since inventory databases, universal product bar codes and scanners, and other such supply chain management technologies have been around for years, the idea of using data to help manage retail operations is not new. However, more recently, the use of data mining to more thoroughly understand patterns of consumer behavior that affect retail operations has become more prevalent. In order to truly understand consumer behavior though, it is beneficial to understand both what they buy and who they are. Thus, in the past decade or so, we have seen an increase in the implementation of customer loyalty programs. You have probably seen these programs, and may even participate in them yourself. Generally, if you participate in such programs, you are given some form of reward, either a lower price on items in the store, or 'points' redeemable toward some future good or service. Airlines have been in the business of using such programs to encourage customer loyalty for many years, with grocery and other retail establishments adapting the concept to their operations more recently. But consider what you give when signing up for these programs. In order to receive the card which you subsequently use to gain the added benefit, you fill out a form. On this form, you give your name, gender, address, phone number, birth date, and perhaps any number of other personal characteristics.

2. RELATED WORK

Apriori algorithm has some limitation in spite of being very simple [1]. The major advantages of FP-Growth

Volume: 07 Issue: 04 Pages: 2790-2797 (2016) ISSN: 0975-0290

algorithm is that it uses compact data structure and eliminates repeated database scan FP-growth is faster than other association mining algorithms and is also faster than tree researching. According to availability of determine "Which groups or sets of items are customer's likely to quality services is vital for the well-being of the economy [2]. Market basket circles are covering all major aspects of the service analysis which may be performed on the retail data of customer transactions. According to in situations with a large number of frequent patterns, long patterns, or quite low minimum support thresholds [3], an Apriori like algorithm may suffer from the following two nontrivial costs i.e.it is costly to handle a huge number of candidate sets and it is tedious to repeatedly scan the database and check a large set of candidates by pattern matching, which is especially true for mining long patterns. This is the inherent cost of candidate generation, no matter what implementation technique is applied.

2.1FP GROWTH ALGORITHM

The FP-Growth Algorithm is an alternative way to find frequent itemsets without using candidate generations, thus improving performance. For so much it uses a divide-and-conquer strategy. The core of this method is the usage of a special data structure named frequent-pattern tree (FP-tree), which retains the item set association information. In simple words, this algorithm works as follows: first it compresses the input database creating an FP-tree instance to represent frequent items. After this first step it divides the compressed database into a set of conditional databases, each one associated with one frequent pattern. Finally, each such database is mined separately. Using this strategy, the FP-Growth reduces the search costs looking for short patterns recursively and then concatenating them in the long frequent patterns, offering good selectivity. In large databases, it's not possible to hold the FP-tree in the main memory. A strategy to cope with this problem is to firstly partition the database into a set of smaller databases (called projected databases), and then construct an FP-tree from each of these smaller databases.

The frequent-pattern tree (FP-tree) is a compact structure that stores quantitative information about frequent patterns in a database.

Han defines the FP-tree as the tree structure defined below:

One root labeled as "null" with a set of item-prefix sub trees as children, and a frequent-item-header table;

Each node in the item-prefix sub tree consists of three fields:

Item-name: registers which item is represented by the node;

Count: the number of transactions represented by the portion of the path reaching the node;

Node-link: links to the next node in the FP-tree carrying the same item-name, or null if there is none.

Each entry in the frequent-item-header table consists of two fields:

Item-name: as the same to the node;

Head of node-link: a pointer to the first node in the FP-tree carrying the item-name.

FP-tree construction

Input: A transaction database DB and a minimum support threshold?

Output: FP-tree, the frequent-pattern tree of DB.

Method: The FP-tree is constructed as follows.

Scan the transaction database DB once. Collect F, the set of frequent items, and the support of each frequent item. Sort F in support-descending order as FList, the list of frequent items.

Create the root of an FP-tree, T, and label it as "null". For each transaction Trans in DB do the following:

Select the frequent items in Trans and sort them according to the order of FList. Let the sorted frequent-item list in Trans be [p | P], where p is the first element and P is the remaining list. Call insert tree ([p | P], T).

The function insert tree ([p | P], T) is performed as follows. If T has a child N such that N.item-name = p.item-name, then increment N 's count by 1; else create a new node N , with its count initialized to 1, its parent link linked to T , and its node-link linked to the nodes with the same item-name via the node-link structure. If P is nonempty, call insert tree (P, N) recursively.

By using this algorithm, the FP-tree is constructed in two scans of the database. The first scan collects and sort the set of frequent items, and the second constructs the FP-Tree.

3. ASSOCIATION RULE MINING

One of the most common applications of ARM is market basket analysis [4] that discovers the relations among the items obtained by customers in the database. The improvement in the information technology allows all the retailers to obtain the daily transaction data at a very low cost [5]. Thus, the large amount of useful data to support the retail management can be extracted from large transactional databases. Data mining (DM) is used to obtain valuable information from large databases [6]. The aim of ARM analysis is to describe the most interesting patterns in an efficient manner [7].ARM analysis (also known as the market basket analysis (MBA)) is method of determining customer obtained patterns by mining association from retailer transactional database.

Algorithm used in market basket analysis (MBA) is apriori algorithm because it is a candidate generation algorithm. It is founded on information that this algorithm uses the preceding knowledge of the regular item set possessions. Apriori procedure pays to an iterative tactic that is recognized as a level wise search in which k-item sets are used to discover (k+1) item sets. Based on this possession, if a set cannot pass the minimum verge than all of its super sets will also fail the test as well. Thus, if an item set is not a recurrent item set, then item set will not

use to create large item set. Apriori procedure is the most recurrently used algorithm among the association rules algorithms that were used at the analysis phase. The problems occur in apriori algorithm are that it scans the databases again and again to check the recurrent item sets and it also generate infrequent item sets. Strong associations have been observed among the purchased item sets group with regard to the purchase behavior of the customers of the retail store. The customer's shopping information analyzed by using the association rules mining with the apriori algorithm. As a result of the analysis, strong and useful association rules were determined between the product groups with regard to understanding what kind of purchase behavior customer's exhibit within a certain shopping visit from both in-category and from different product categories for the specialty store.

3.1 DATA SOURCE

Figure 1, below, depicts a simplified relational model which might realistically be used by a supermarket to gather and store information about customers and the products they buy. It is simplified in that the attributes represented in each of the tables would likely be more numerous in an actual grocery store's database. However, to ensure that complexity of the related entities does not confound the explanation of Association Rules in this chapter, the tables have been simplified.

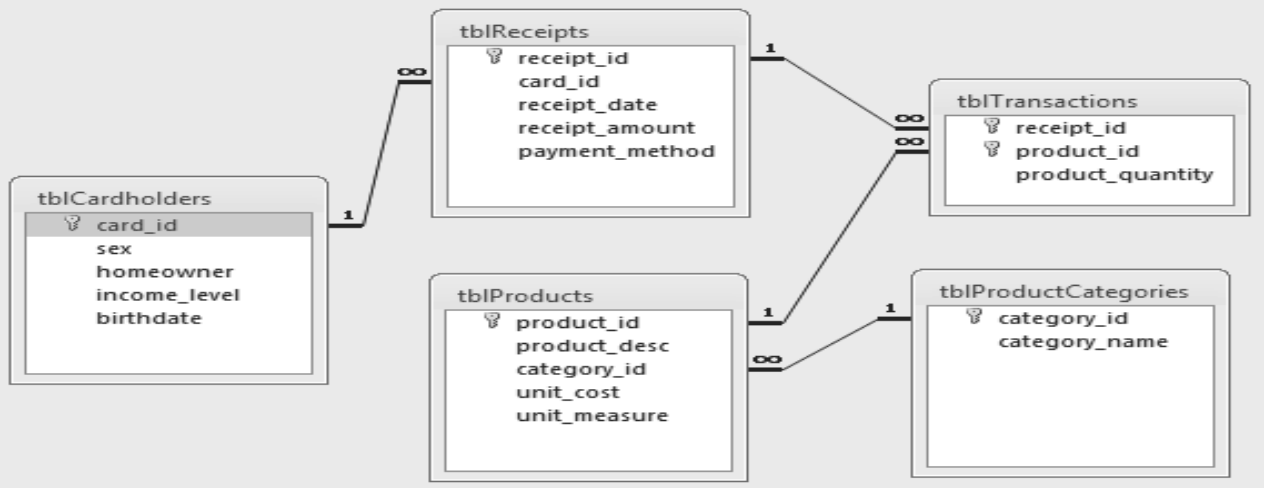

Figure 1 A simplified relational model of supermarket's database.

The datasets used throughout this paper consists of content and collaborative data. Content data was taken from the Supermarkets, Figure 2 depicts the first 19 rows of our previously discussed query, however this query was run on tables containing 108,131 receipts from 10,001 different loyalty card holders, Figure 3 shows the Meta data view.

ExampleSet (108131 examples, 0 special attributes, 12 regular attributes) View Filter (108131 / 108131): all

Row No.	receipt_id	desserts	meats	juices	paper_goods	frozen_foods	snack_foods	canned goo...	beer_wine_...	dairy	breads	produce
1	1	0	1	1	0	1	0	0	0	0	0	1
2	2	1	0	1	1	0	0	0	0	1	0	0
3	3	1	1	1	1	1	0	1	1	1	1	1
4	4	1	1	0	1	1	0	0	0	0	0	1
5	5	0	0	0	0	0	1	0	1	0	0	0
6	6	1	0	1	0	0	0	0	0	0	0	0
7	7	1	0	0	1	0	0	0	0	0	0	0
8	8	0	0	0	0	0	1	0	0	1	0	0
9	9	1	0	0	1	0	0	0	0	0	0	0
10	10	0	1	0	0	0	0	0	0	0	0	1
11	11	0	0	0	0	1	0	0	0	0	0	0
12	12	1	0	0	1	1	0	0	0	1	0	0
13	13	1	0	0	0	0	0	0	0	0	1	0
14	14	0	1	0	1	1	1	0	1	0	0	0
15	15	1	0	0	1	0	0	0	0	0	0	0
16	16	0	1	0	0	0	0	0	0	0	0	0
17	17	0	0	1	0	1	0	0	1	1	0	0
18	18	0	1	0	0	1	1	0	1	0	0	0
19	19	1	0	0	0	1	0	1	1	0	0	1

Figure 2 Query results from an expanded dataset

Volume: 07 Issue: 04 Pages: 2790-2797 (2016) ISSN: 0975-0290

ExampleSet (108131 examples, 0 special attributes, 12 regular attributes)

Role	Name	Type
regular	receipt_id	integer
regular	desserts	binominal
regular	meats	binominal
regular	juices	binominal
regular	paper_goods	binominal
regular	frozen_foods	binominal
regular	snack_foods	binominal
regular	canned goods	binominal
regular	beer_wine_spirits	binominal
regular	dairy	binominal
regular	breads	binominal
regular	produce	binominal

Figure 3 The Meta data view

3.2 PROCESS OF ASSOCIATION MINING

Figure 4 depicts a basic operator workflow. Running the model on the entire dataset. If there are hundreds of thousands or millions of observations in your dataset, the model may take some time to run. Tuning the model on a smaller sample can save time during development, and then once you are satisfied with your model, you can remove the sample operator and run the model on the entire dataset.

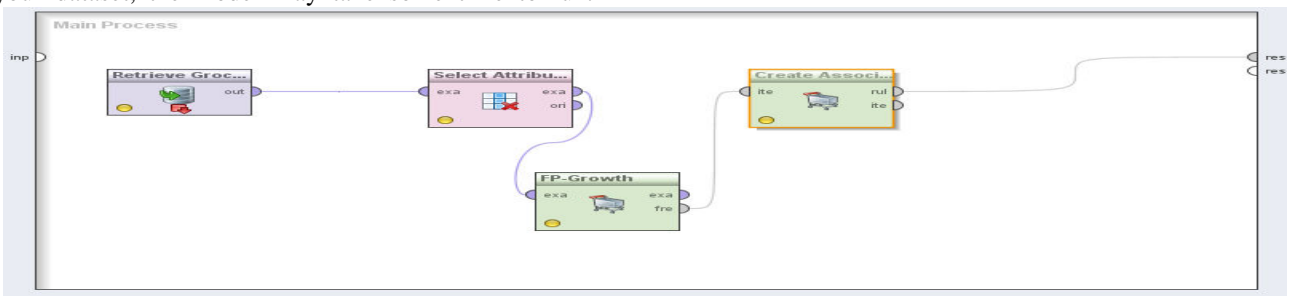

Figure 4 A basic Association rule mining operator workflow

Once any inconsistencies or other required transformations have been handled, we can move on to applying modeling operators to our data. The first modeling operator needed for association rules is FP-Growth (found in the Modeling folder). When min support=0.75 and min support=0.5 Comparative Study depicted in Figure5, calculates the frequent item sets found in the data. Effectively, it goes through and identifies the frequency of all possible combinations of products that were purchased. These might be pairs, triplets, or even larger combinations of items. The thresholds used to determine whether or not items are matches can be modified using the tools on the right-hand side of the screen.

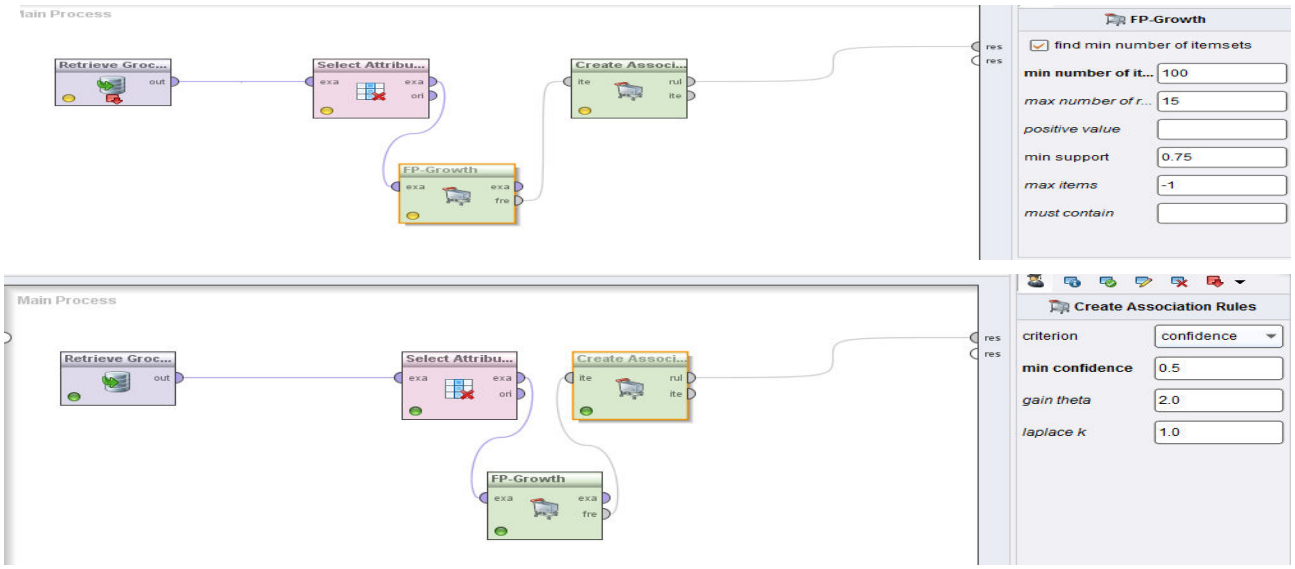

Figure 5 Comparative Study of FP-Growth to our data mining process.

Volume: 07 Issue: 04 Pages: 2790-2797 (2016) ISSN: 0975-0290

At this point, we can run our model and look at the item sets that were found, but we cannot necessarily see the strength of the associations. Figure 6 and Figure 7 show the results of running the model with just the FP-Growth operator in place. Note that the frequency port is connected to the result set port.

As we can see, the operator found frequencies for most items individually, and began to find frequencies between items as well. Although the screen capture does not show all 32 item sets that were found, it if did, you would be able to see that the final set found contains four products that appear to be associated with one another: juices, meats, frozen foods, and produce. There are a number of three-product combinations, and even more two-product sets. The Support attribute seen in Figure 6 indicates the number of observations in the dataset where the single or paired attributes was found; in other words, out of the 108,131

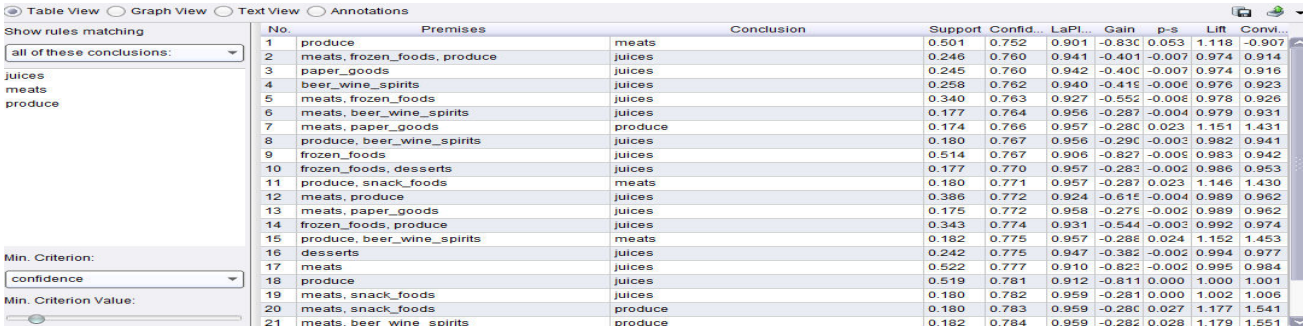

Figure 6 Item sets generated by FP-Growth

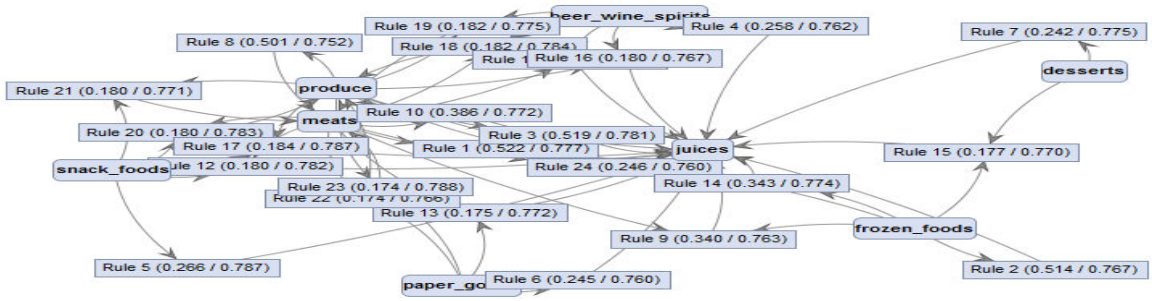

Figure 7 Graph View of Association Rules in Results Perspective.

Figure 8, Figure 9 and Figure 10 show the juices of Association Rules, meats of Association Rules and produce Of Association Rules.

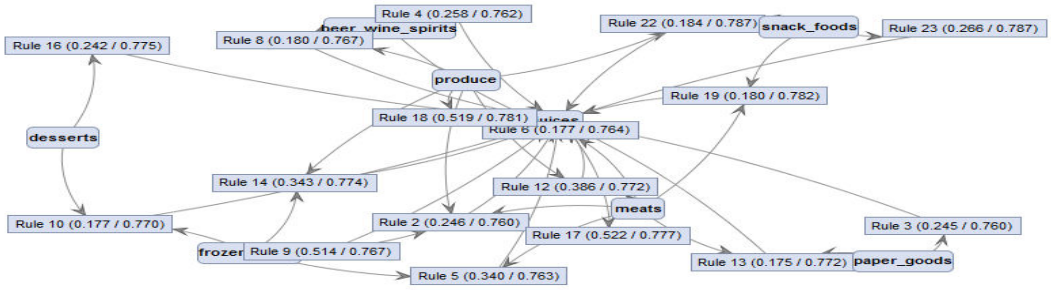

Figure 8 Juices of Association Rules

Volume: 07 Issue: 04 Pages: 2790-2797 (2016) ISSN: 0975-0290

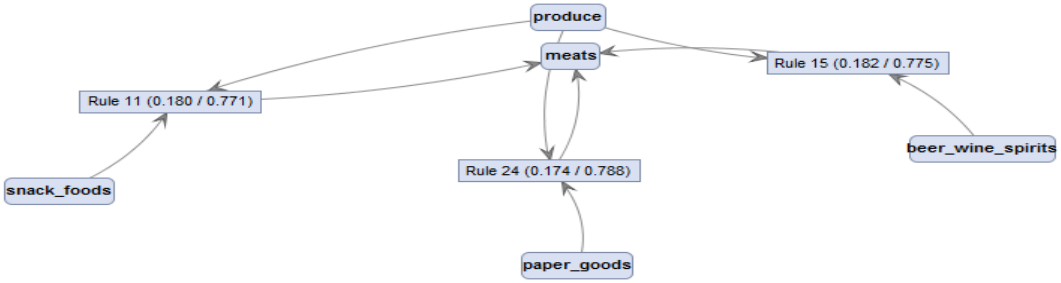

Figure 9 Means of Association Rules

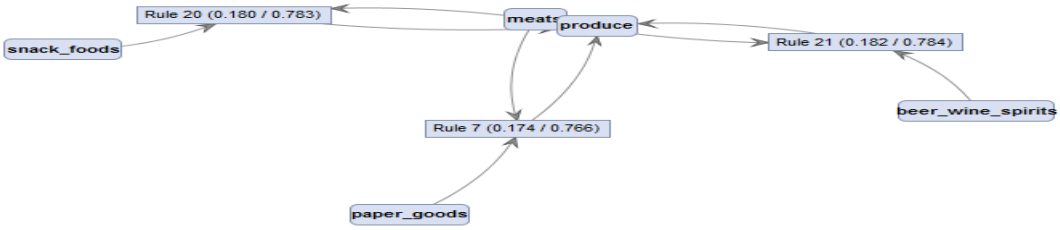

Figure 10 Produce of Association Rules

When min support=0.5, Association Rules are as follows:

[meats] --> [juices, frozen_foods] (confidence: 0.506)

[frozen_foods] --> [juices, meats] (confidence: 0.508)

[frozen_foods] --> [juices, produce] (confidence: 0.512)

[produce] --> [juices, frozen_foods] (confidence: 0.516)

[beer_wine_spirits] --> [juices, meats] (confidence: 0.523)

[beer_wine_spirits] --> [juices, produce] (confidence: 0.532)

[snack_foods] --> [juices, meats] (confidence: 0.533)

[snack_foods] --> [meats, produce] (confidence: 0.534)

[beer_wine_spirits] --> [meats, produce] (confidence: 0.537)

[paper_goods] --> [meats, produce] (confidence: 0.539)

[paper_goods] --> [juices, meats] (confidence: 0.543)

[snack_foods] --> [juices, produce] (confidence: 0.545)

[meats, frozen_foods] --> [juices, produce] (confidence: 0.551)

[frozen_foods, produce] --> [juices, meats] (confidence: 0.554)

[beer_wine_spirits] --> [snack_foods] (confidence: 0.563)

[snack_foods] --> [beer_wine_spirits] (confidence: 0.564)

[desserts] --> [juices, frozen_foods] (confidence: 0.567)

[meats] --> [juices, produce] (confidence: 0.574)

[produce] --> [juices, meats] (confidence: 0.580)

[juices, meats, produce] --> [frozen_foods] (confidence: 0.637)

[meats, produce] --> [frozen_foods] (confidence: 0.646)

[juices, meats] --> [frozen_foods] (confidence: 0.652)

Volume: 07 Issue: 04 Pages: 2790-2797 (2016) ISSN: 0975-0290

[juices] --> [frozen_foods] (confidence: 0.659)

[juices, produce] --> [frozen_foods] (confidence: 0.661)

[frozen_foods] --> [produce] (confidence: 0.661)

[juices, frozen_foods] --> [meats] (confidence: 0.662)

[meats] --> [frozen_foods] (confidence: 0.663)

[frozen_foods] --> [meats] (confidence: 0.665)

[juices] --> [produce] (confidence: 0.666)

[produce] --> [frozen_foods] (confidence: 0.667)

[desserts] --> [meats] (confidence: 0.667)

[juices, frozen_foods] --> [produce] (confidence: 0.668)

[juices] --> [meats] (confidence: 0.670)

[juices, snack_foods] --> [meats] (confidence: 0.677)

[desserts] --> [produce] (confidence: 0.680)

[snack_foods] --> [meats] (confidence: 0.682)

When min support=0.75,Association Rules are as follows:

[produce] --> [meats] (confidence: 0.752)

[meats, frozen_foods, produce] --> [juices] (confidence: 0.760)

[paper_goods] --> [juices] (confidence: 0.760)

[beer_wine_spirits] --> [juices] (confidence: 0.762)

[meats, frozen_foods] --> [juices] (confidence: 0.763)

[meats, beer_wine_spirits] --> [juices] (confidence: 0.764)

[meats, paper_goods] --> [produce] (confidence: 0.766)

[produce, beer_wine_spirits] --> [juices] (confidence: 0.767)

[frozen_foods] --> [juices] (confidence: 0.767)

[frozen_foods, desserts] --> [juices] (confidence: 0.770)

[produce, snack_foods] --> [meats] (confidence: 0.771)

[meats, produce] --> [juices] (confidence: 0.772)

[meats, paper_goods] --> [juices] (confidence: 0.772)

[frozen_foods, produce] --> [juices] (confidence: 0.774)

[produce, beer_wine_spirits] --> [meats] (confidence: 0.775)

[desserts] --> [juices] (confidence: 0.775)

[meats] --> [juices] (confidence: 0.777)

[produce] --> [juices] (confidence: 0.781)

[meats, snack_foods] --> [juices] (confidence: 0.782)

[meats, snack_foods] --> [produce] (confidence: 0.783)

[meats, beer_wine_spirits] --> [produce] (confidence: 0.784)

[produce, snack_foods] --> [juices] (confidence: 0.787)

[snack_foods] --> [juices] (confidence: 0.787)

[produce, paper_goods] --> [meats] (confidence: 0.788)

4 CONCLUSIONS

We have found that juice products are relatively strongly connected to essentially every other product category in our grocery store, but what can we do with this information? Perhaps we already know, through daily experience, that we sell a lot of juice products, in which case this particular data mining model is of little help to us. But perhaps we might not have realized, without this model, just how pervasive juice products are in our product sales. As grocery store managers, we may begin to design product promotions which pair juice products with other strongly associated products in order to boost sales. We may go back and lower our confidence percentage a bit more, to see if other product categories emerge as the next most common conclusions (e.g., frozen foods and produce both have associations above 70% confidence). Or we may decide that we need more detail about specifically what juice products are frequently sold with other items, so we may choose to go back to the data extraction and preparation phase to group juice products into more specific attributes across our 108,131 receipts, to see if we might find even better clarity about what products our customers are most frequently buying together.

ACKNOWLEDGEMENTS:

A Project Supported by Scientific Research Fund of Hunan Provincial Education Department（15A043）

Volume: 07 Issue: 04 Pages: 2790-2797 (2016) ISSN: 0975-0290

REFERENCES:

[1] B.Santhosh Kumar and K.V.Rukmani.Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms,Int. J. of Advanced Networking and Applications,6(2010),400-404.

[2] Ankur Mehay, Dr. Kawaljeet Singh, and Dr. Neeraj Sharma. AnalyzeMarket Basket Data using FP-growth and Apriori Algorithm, International Journal on Recent and Innovation Trends in Computing and Communication, (2013),693-696.

[3] JIAWEI HAN, JIAN PEI, YIWEN YIN and RUNYING MAO.Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach, Data Mining and Knowledge Discovery, 8(2004), 53–87.

[4] Hahsler, M., Grün, B., & Hornik, K. arules: a computational environment for mining association rules and frequent item sets. Journal of Statistical Software, 14(2005), 1–25.

[5] Tan, P. N., Steinbach, M., & Kumar, V. *Introduction to data mining*. Reading: Addison-Wesley, 2005

[6] Chen, M., & Lin, C.. A data mining approach to product assortment and shelf space allocation. Expert Systems with Applications, 32 (2007), 976-986.

[7] Tan, P., Steinbach. & Kumar, V. Introduction to data mining. Boston: Pearson Education, 2006

Authors

< Zhi-hang Tang >, <1974-08-08>, <hunan, China>
Current position, ***Doctor of Hunan Institute of Engineering***
University studies: ***control theory and control engineering in donghua University***
Scientific interest: intelligent decision and knowledge management
Publications <number or main>: *30 Papers*
Experience: Zhihang TANG *was born in Shaoyang, China, in 1974. He earned the M.S. degrees in control theory and control engineering from zhejiang University of techonlogy, in 2003 and Ph.D. from donghua University China in 2009. At the same time ,he is a teacher in department of computer and communication, Hunan Institute of Engineering(Xiangtan, China) from 2003.Chaired the 49th China Postdoctoral Science Foundation grant, presided over science and technology projects in Hunan Province in 2010, presided over the Education Department of Hunan Province in 2010 Outstanding Youth Project, as the first author more than 30 papers were published.His current research interests include intelligent decision and knowledge management.*