The Improved Apriori Algorithm was Applied in the System of Elective Courses in Colleges and Universities and Analysis of Carbon Emissions

She Wei[1,a] ,Tan Yuyin[2,b], Xie Huijuan[3,c]

[1] Hainan University, 570228, Haikou, China

[2] Hainan University, 570228, Haikou, China

[3]Hainan College of Economics and Business, 571127, Haikou, China

[a]694394619@qq.com, [b]26473428@qq.com, [c]360088369@qq.com

Keywords: Data Mining, Apriori Algorithm, Professional Elective System, low-carbon

Abstract. Apriori algorithm is one of classical algorithms in data mining, can mine frequent item sets what the association rules needed. A classic application is shopping analysis in the supermarket, such as "beer" and "diaper". Aiming at the system of professional elective courses in the colleges and universities, this paper proposes a modified Apriori algorithm that can analyze what the combination of each professional elective course that influenced the employment can provide basis for decision making for the college professional elective course system and provide some suggestions for the employment of students. At the same time, the association rules can also provide the analysis of carbon emissions in the environmental protection, and provide decisions to the low-carbon for China's government.

Introduction

Data mining was also translated into two different meanings. It is one of steps in the Knowledge-Discovery in Databases (KDD). Data mining generally refers to searching for hidden information by the algorithm from a great deal data. Data mining usually achieves the above objectives through many methods, such as the Statistics, Online Analysis Processing Information Retrieval, Machine Learning, Expert System (depending on the old rules of thumb) and Pattern Recognition. In the data mining model, association rules model is a wide application. The concept of association rule (a simple and useful rule) was proposed by Agrawal, Imielinski, Swami. The association rule reflects the unknown association relationship among each data item in the database, discovering frequent item sets is the core technology of the association rules in data mining. Apriori algorithm was not only an widely used algorithm in the data mining but also was inefficient because of scanning the database. This paper proposes an improved Apriori algorithm that could analyze for the combination of college professional elective course and employment, so as to find out the professional elective association between course combination and employment, the results can provide some suggestion for professional elective course system and employment in the colleges and universities.

Classic Apriori Algorithm

Apriori algorithm that was a breadth first algorithm based on repeatedly scanning the database to find all the frequent item sets, each scan only considered the same length of all items [1]. Apriori algorithm generated frequent item sets with searching layer by layer and iterative method [2]. First, scanning the transaction database D, 1- frequent item sets L1 would be found in database D by the user minimum support degree, then 2- candidate item sets C2 would be generated by L1 connecting operations, 2- frequent item sets L2 would be found from C2 by rescanning the transaction database D, and so on, it would not be end until no more k- frequent item sets or the candidate sets were empty.

Improved Apriori Algorithms

The common and improved Apriori algorithm as follows: width first algorithm, depth first algorithm, the data set partitioning algorithm, sampling algorithm, incremental updating algorithm and parallel algorithm for mining [3]. the FP-growth algorithm that was one of improved depth first algorithms in the Apriori algorithm according to professional selective courses system in university and the data of employment could analyze and process the transaction database and find the frequent item sets.

The implementation of FP-growth algorithm:

Step 1: input transaction database D, and set to the minimum support value min_sup.

Step 2: 1- frequent item sets G was found according to the first scanning the transaction database D, the frequent item table L1 including item-name domain and pointer domain that was the first node pointing to FP-tree and had the same item-name. Pointer could not point to the child nodes of FP-tree root node, because it didn't have a prefix when it was a suffix in the inverse the traversal of FP-tree. In other words, 1-frequent item sets mostly had no significance.

Step 3: FP-tree was generated by the second scanning the transaction database D. Create FP-tree roots and set null, item-nameS of each transaction T in the transaction database D were added to FP-tree in sequence. Each of FP-tree nodes beside root node included three domains: item-name, count and link.

The specific operation as follows:

First, scanning the first transaction T1, the items in the transaction would be arranged in the order by L1, the first frequent item i1 as the child node of root node was inserted into FP-tree, the item-name and the count was separately set i1 and 1; the second frequent item i2 as the child node of the node i1 was inserted into FP-tree, the item-name and the count was separately set i2 and 2; the third frequent item i3 as the child node of node i2 was inserted into FP-tree, the item-name and the count was separately set i3 and 1, the pointer of i3 in the L1 pointed to the node, in this transaction other frequent items were inserted into FP-tree in turn.

Second, the items in the transaction would be arranged in the order by L1, the first frequent item j1 was inserted into FP-tree. First, if the children N1 and j1 had the same name among the root nodes in the FP-tree, the value of N1 was added 1; second, a new node was created as the child node of root node, and the item-name and the count was separately set j1 and 1. The second frequent item j1 was inserted into FP-tree, In the first case, if the children N2 and j2 in j1 had the same name in the FP-tree, the value of N2 was added 1; when the second case happened in j1, a new node would be created as the child node of j1, and the item-name and the count were separately set j2 and 1. If the pointer in the corresponding J2 in L1 was null, the pointer was pointed to this node, or the value of the pointer was saved in the link of this node, and then the pointer would point to the node. Other frequent items in the transaction were inserted into the FP-tree as the node j2 did.

Last, scanning the other affairs T_K, referencing to the method of scanning the second transaction T2. FP-tree would be structured after scanning the whole transaction.

Step 4: mining the FP-tree by the bottom-up way. The branch structural mode base would be found out by count domain and link domain in the table L1, FP-tree would be constructed and the frequent patterns were generated [4].

The FP-growth algorithm did not generate candidate item sets, scanned the transaction database only twice, greatly reduced frequency of classical Apriori algorithm that scanning the transaction database, so avoided to produce a mass of candidate item sets, effectively improve the operation efficiency. At the same time, because the number of similar professional elective professional courses system was considerably less than the number of item that Apriori algorithm was applied in other fields, and the number of courses that allowed students to take was limited, so the long-branch in the FP-tree number was effectively avoided.

The Results of Testing Algorithms

In theory, the efficiency of the FP-growth algorithm was more effective than classic Apriori algorithm. The two algorithms were compared with java language in the experiment. The experimental environment as follows: CPU was the processor (R) Pentium (R)CPU G2010 @ 2.80GHz, memory was 4GB (3.41GB was enough), the operating system was Windows 7ultimate, 32 bit. They were tested using the Eclipse software. This paper provided 7228 affairs, including 112 attributes, the support degree of them respectively was 0.1, 0.2, 0.3, 0.4, and 0.5, and the test results showed the efficiency of the FP-growth algorithm was more effective than classic Apriori algorithm, especially under the condition of small supporting degree. The results were shown in fig 1.

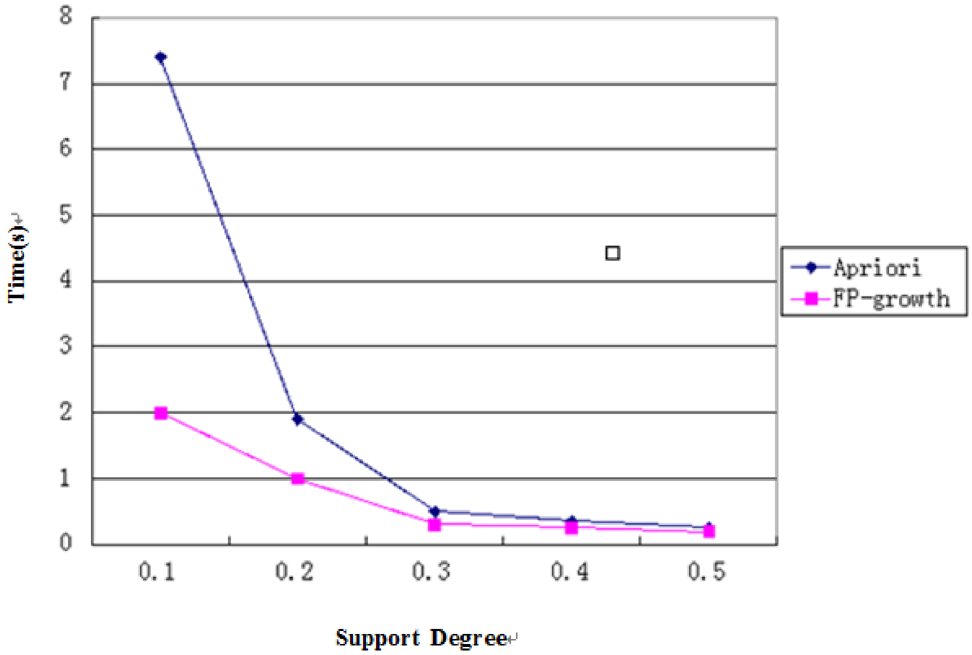

Fig 1 Experiments and Results

The Process of Application

The paper mainly used the association rules, the hidden and potential regular pattern and the relevant item sets could be found by the mining data of employment and professional elective courses before the graduation, could be good for the system of the professional elective courses and employment. In this paper the improved Apriori algorithm can also be applied in the prediction and analysis of carbon emissions, the association rules those were found can predict how to increase or decrease the carbon in environmental protection, and provide valid decision-makings for the nation development.

Data Collection, Data arrangement and Collation and Establishing a Transaction Database

Data Collection. At present, the university has managed the employment by using the information management graduates, some colleges and universities have input the employment information into the database, but there are a large number of universities without the database electronic information system. Max Institute is good for collecting the information of tracking evaluation of employed, students and employer [5]. In this paper, the employment data for the students who had graduated for half or half a year was received by questionnaire that asked students to write whether they had been employed, whether they were professional counterparts, how they got jobs, whether they were satisfied with jobs and the prospects of their occupation

Data Arrangement. The electronic questionnaire received should be rearranged, including excluding the unfit data and merging the courses with the different names but the same essence into a specialty elective course.

The Establishment of the Transaction Database. The transaction database with employment and professional selective courses table would be generated by connecting the table inputted information of the electronic questionnaire and graduate list in the dean's office.

The Main Association Rules

The valid data, that was from related accounting graduates between half a year and 5 years, was tested. The condition was as follows: they were employed, generally interested in the specialty, employed professionally, were candidates for the posts, employment satisfaction and career prospects were ignored. The supporting degree was set 0.15, the length of discovered rule was 3, and the discovered rule is 6, as shown in Table1.

Table1 Association Rule and Support

Serial Number	Association Rule	Support
1	Financial Applied Writing- Public Relation and Etiquette - Accounting Information Application	29%
2	International Trade-International Accounting - Financial Engineering	23%
3	Public Relation and Etiquette - Marketing – Public Management	18%
4	Financial Applied Writing- International Trade- Capital Management Practice	16%
5	Financial Applied Writing- Public Relation and Etiquette - Tax Planning	15%
6	Finance- Management Consulting- Public Management	15%

The courses on practical skills were good for employment, which could be seen from the results of the association rules, and the employment would not be less affected by the strong theoretical course. The association rules can promote the investment on the related course in college and university, and recommended students electing related courses.

Summary

This paper introduced the Apriori algorithm and FP-growth algorithm in the data mining, and combined the association rules application to the elective course system and employment. Frequent item sets could provide the decision basis for the professional elective course system, and provide some suggestions for the employment. There is important to employment that is becoming increasingly difficult. The improved Apriori algorithm can also be applied in the prediction and analysis of carbon emissions, the association rules can also provide available decisions for decreasing 40-50% carbon the nation development.

The Project: The Statistical Accounting System Research And the Ability to Construction on Climate in Hainan Province,2013112

References

[1] Luo,K.&He,C.W. The Extraction Algorithm of Improved Association Rules Based On Apriori [J]. Computer and Digital Engineering,2006,(2): 48-49.

[2]Yang,J.F.&Liu,F. A new Improved Apriori Algorithm [J].Microcomputer and Application,2010,(1): 55-57.

[3] Yang,Q. Improvement of of Algorithm of [J].Computer Knowledge and Technology,2013,(9): 2037-2039.

[4] Wang,A.P.&Wang,Z.F.Use of Association Rules for Collapsing Algorithm in the Data Mining [J]. Technology and Development of the Computer,2010,(4): 105-108.

[5]Jiang,S.Y,&Li,X.Theory and Practice of the Data Mining[M].Beijing:Electronic Industry Press,2011:30-50.