Research of College English Test Knowledge Points Association Analysis and Implementation of Table of Association Rules

Tao Wu[1, a], Dongyang Liu[1], Hui Zhang[1] and Min Wu[1]

[1] Center of Modern Educational Technology, University of Science and Technology of China, Hefei 230026, China;

[a]jacksonwu1992@163.com

Keywords: knowledge points association analysis, data mining.

Abstract. In order to better predict the potential weaknesses of user knowledge points, this paper firstly carried on knowledge points association analysis by a large amount of data collected in the practical application. Secondly, we did the data mining process including data collection, data preprocessing, the actual mining, result analysis and so on, and we built the table of the knowledge points association rules which had a large increase in many aspects. Finally, we gave an algorithm to find potential weak knowledge points.

1. Technical Basis

At present, there are various teaching and predicting methods aiming at The College English Test, but most predictions and knowledge points association analysis are based on the teacher's experience, which are somewhat subjective and biased[1] [2]. Students can scarcely grasp the association of knowledge points when studying and taking the test for teachers or guidance books mostly make association analysis of knowledge points based on their own experience.

This paper carried on knowledge points association analysis based on a large amount of preprocessed data collected in the practical application，and then we produced a relatively complete and reliable table of knowledge point association rules. More importantly, we have created an algorithm to find potential weak knowledge points based on the table of the knowledge points association rules, which give a more effective guidance to learners based on their grasp of knowledge points

1.1 Knowledge Point.

From the view of content, knowledge point is a relatively logical collection of partial knowledge in one field. Learning a course is actually learning a series of knowledge points of this course. In this paper, the knowledge in College English is a collection of knowledge involved in objective questions in the College English test, which are the basic components of the teaching and learning of English knowledge.

1.2 Data Mining and Association Analysis

Data mining is the process of finding valuable information in large scale data.

The four main tasks of data mining include prediction modeling, association analysis, clustering analysis, anomaly detection [3]. Among them, the association analysis is an important research work in this paper. Association analysis is used to discover the interesting connection hidden in large data sets [4]. The minimum support degree of association rules is used to rule out the less frequently rules, and the confidence is to estimate the conditional probability of rule A→B.

1.3 The Apriori Algorithm

The core of Apriori algorithm is based on the recursive algorithm of the two stage frequent item set, which is one of the most influential mining algorithms for mining Boolean association rules.

2. The Set of Table of Knowledge Points Association Rules

2.1 Data Collection

In order to improve the accuracy of the results of association rules, the project team decided to use the College English diagnostic system in one high school on the line. Students of this school were in the middle level of the college entrance examination, and they are of great enthusiasm and demand to pass The College English Test, so the use of the system will be more frequent, which will be easy to get potential weak points of knowledge. After a year of use, a total of 4590 students used the system, and most students completed more than two papers.

2.2 Data preprocessing.

System design can't avoid the system error caused by the running of system. Firstly, we used data cleaning to correct data.

During the running of system, some students may have wrong operation or not serious practice, system operation may also produce error, and database sometimes recorded error information. So, it is necessary to clean up the sample data. After cleaning up the data, the data of 4406 system users are retained.

2.3 Mining Process

2.3.1 Process continuous data

The original data show that the correct rate of the knowledge points of the students are continuous data directly derived from the database, whose range is 0 to 1. Therefore, the first step is to set the line value to convert continuous data into Boolean data of True and False. Analysis shows that the average correct rate of all 18 knowledge points is 43.12%, among which 14 knowledge points has a correct rate of 30%-60%. Therefore, we decided to make a compromise between the average correct rate and the passing score, and set the line value to 0.5. The data ranging from 0 to 0.5 are represented by False, and the data ranging from 0.5 to 1 are represented by True.

2.3.2 Data hierarchy

From the above, we can see that the sample number of different knowledge points is different, which leads to the difference of proportion of the total number of different knowledge points. Therefore, in order to mine association rules of each knowledge points, the analysis of rules should be carried out according to different samples. Observation shows that there is a regular distribution of four hierarchies including more than 90%, about 75%, 40%-50%, below 10%, so it has four hierarchies based on the proportion of the sample quantity of each knowledge point. Due to limited space, we only make a presentation on the knowledge point of the highest proportion of each hierarchy, as is shown in Table2.1.

Table 2.1 Data hierarchy

Hierarchy	Total Number	Name of Knowledge Point	Sample Quantity	Proportion
First	4406	1.2 Synonym	4406	100.00%
Second	3334	3.3 Non-Finite Verb	3325	99.73%
Third	2969	3.6 Application of Parallel Structure	2197	74.00%
Fourth	282	1.5 Understand Relationship of Text	281	99.65%
No Analysis		3.4 Prepositions and Numerals	1	No Analysis

2.4 Parameter setting

2.4.1 Analysis of the high accuracy

When analyzing the association rules between the high correctness rates of knowledge points, we set the maximum number of former of 3, the minimum support degree of 30% and minimum confidence of 90% through experiment. So we got the number of association rules for the first hierarchy of 50, second hierarchy of 82, third hierarchy of 0 and fourth hierarchy of 0. It can be seen that the number of association rules of the third hierarchy and the fourth hierarchy is 0, so we will discuss the reasons below.

Mentioned in the previous article, the proportion of the third hierarchy is lower than the two hierarchies above, so we reduce the minimum support to 20%, and then we have 46 association rules. The correct rate of two knowledge points of the fourth hierarchy is the lowest and we have a small number of samples of this hierarchy, so there is not enough data of high correct rate to support a sufficient number of frequent item sets.

2.4.2 Analysis of the low accuracy

When analyzing the association rules between the low correctness rates of knowledge points, we set the maximum number of former of 3, the minimum support degree of 40% and minimum confidence of 90% through experiment. So we get the number of association rules for the first hierarchy of 77, second hierarchy of 51, third hierarchy of 18 and fourth hierarchy of 1442. It can be seen that the number of association rules of the third hierarchy is small and the number of the fourth hierarchy is large, so we will discuss the reasons below.

Mentioned in the previous article, the proportion of the third hierarchy is lower than the two hierarchies above, so we reduce the minimum support to 35%, and then we have 66 association rules. For the fourth hierarchy, two knowledge points of this hierarchy have a large amount of data below the line value, so this will produce a large number of frequent item sets. Therefore, we increased the minimum support degree to 75% and minimum confidence to 98% through experiment, and then we get 77 association rules.

2.5 Mining results

We can get the final numbers of association rules as shown in Table 2.2 through the above association analysis and the specific mining analysis.

Table 2.2 Numbers of Association Rules

Total Setting	Hierarchy		Minimum Support	Minimum Confidence	Number	Total	
Line = 0.5；Maximum Number of Former of 5；	Analysis of the High Accuracy	First	90%	30%	53	189	477
		Second	90%	30%	88		
		Third	90%	20%	48		
		Fourth	90%	30%	0		
	Analysis of the Low Accuracy	First	90%	40%	82	288	
		Second	90%	40%	54		
		Third	90%	35%	89		
		Fourth	98%	75%	63		

Through the above association analysis, finally 477 association rules can be produced. Due to limited space, we only make a presentation on association rules of the highest confidence of each hierarchy, as shown in Table 2.3.

Table 2.3 Results of knowledge point Association Rules

Hierarchy		Latter	Former	Support	Confidence
Analysis of the High Accuracy	First	3.5	2.1 and 2.7 and 2.4	31.82	96.576
	Second	3.5	2.1 and 3.2 and 2.7	31.434	95.992
	Third	1.2	1.3 and 3.6	20.108	96.817
	Fourth	无			
Analysis of the Low Accuracy	First	1.3	1.4 and 1.2 and 1.1	40.604	97.596
	Second	2.1	3.2 and 1.3 and 2.2	40.792	98.897
	Third	2.1	2.3 and 1.2 and 1.3 and 2.2	35.231	98.948
	Fourth	1.1	2.6 and 2.2	79.433	100

3. The algorithm to find potential weaknesses of knowledge point based on the association rules table

3.1 background and design ideas

Based on a large amount of data mining, we have produced a relatively complete and reliable table of knowledge point association rules. Therefore, this paper will propose an algorithm to find potential weaknesses of knowledge point based on association rules, which give guidance to learners based on their grasp of knowledge points and association rules table.

3.2 Specific steps and flow chart

According to the background and design ideas, the specific steps are as follows:

（1）Read correct rate of knowledge point from database, store them to List R in order of their ID, if some correct rate is null, then it is initialized to 0;

（2）Get the maximum 5 items from List R, store them into List A;

（3）Read the association rules Table L of high accuracy from database;

（4）Read association rule Li from Table L in loop， conduct the following steps from（5）to（7）;

（5）Read former part Xj of each association rule Li, set Flag which is used to judge whether all the former part is in List A to 0;

（6）Make a judgment whether former part Xj is in List A, if existed, then go to judge next former part Xj +1;if not existed, set Flat to 1,break the loop, go back to step（4）and go to next rule Li +1;

（7）After break the loop in step（6）, judge whether Flag is equal to 0, if so ,it means the match succeed, store association rule Li into Table B, go to next rule Li +1;

（8）Exclude rules whose latter part is in List A;

（9）Retain the rule of the highest confidence whose latter part is same in List B;

（10）Get the minimum 5 items from List R, store them into List B;

（11）Read the association rules Table K of low accuracy from database;

（12） Read association rule Ki from Table K in loop， conduct the following steps from（13）to（15）;

（13）Read former part Yj of each association rule Ki, set Flag which is used to judge whether all the former part is in List C to 0;

（14）Make a judgment whether former part Yj is in List C, if existed, then go to judge next former part Yj +1;if not existed, set Flat to 1,break the loop, go back to step（13）and go to next rule Ki +1;

（15）After break the loop in step（14）, judge whether Flag is equal to 0, if so ,it means the match succeed, store association rule Ki into Table D, go to next former part Ki +1;

（16）Exclude rules whose latter part is in List C;

（17）Retain rule of the lowest confidence whose latter part is same in List D;

（18）Delete the rule whose latter part is same in List B and D;

The flow chart is Fig 3.1.

3.3 Experimental verification

In order to evaluate whether the algorithm can effectively improve user's knowledge point learning, the project team designed and implemented an experiment to verify it by comparing the scoring rate of knowledge point.

We randomly selected 20 students in a university and divided them into two groups, and each group had 10 students, so that the initial average correct rate of each group is almost same. The first group used the original algorithm to take tests while the second group used the above algorithm. Results shows that the original correct rates of two groups are 42.96% and 43.12%, and then the correct rates are 46.25% and 48.66% after doing the experiment. We can see that the correct rate of each group is both increased, but the second group have increased more than the first. So that we can make a conclusion that the new algorithm is better than the original one to improve the score rate of knowledge points.

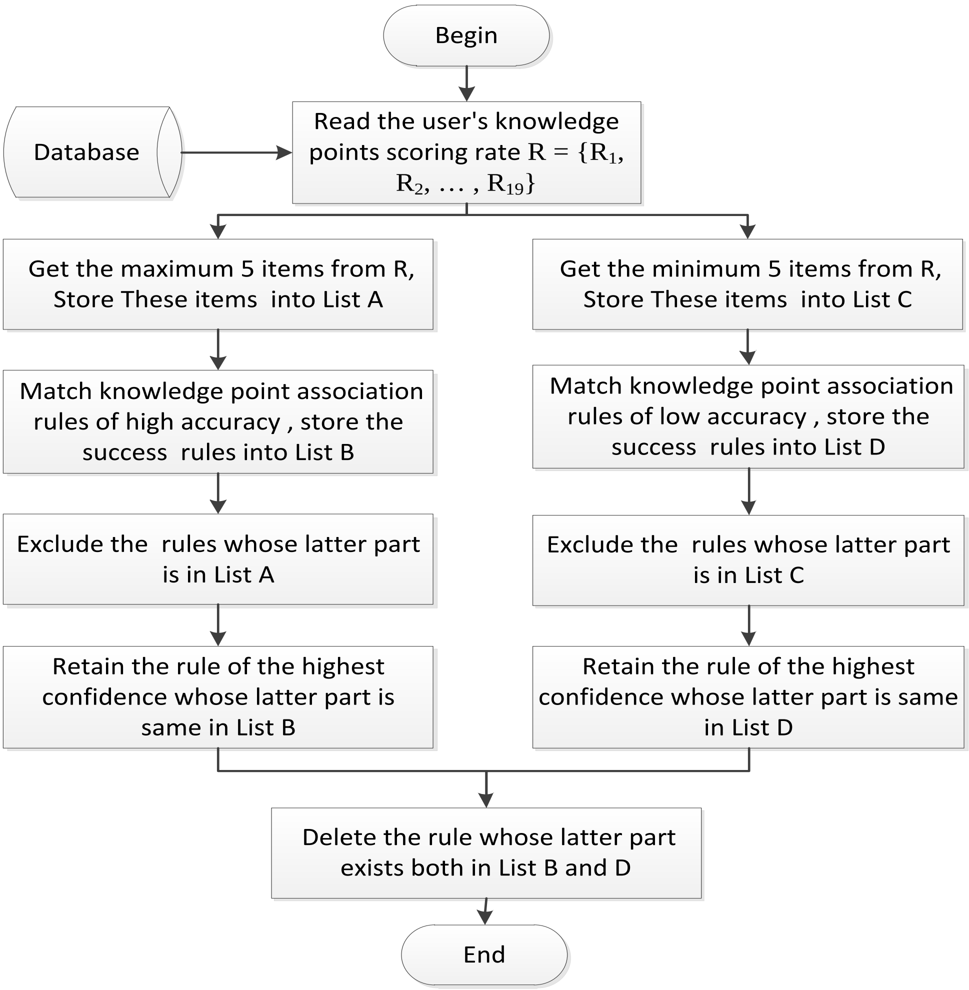
Begin
Database
Read the user's knowledge points scoring rate $R = \{R_1, R_2, \dots, R_{19}\}$
Get the maximum 5 items from R, Store These items into List A
Get the minimum 5 items from R, Store These items into List C
Match knowledge point association rules of high accuracy , store the success rules into List B
Match knowledge point association rules of low accuracy , store the success rules into List D
Exclude the rules whose latter part is in List A
Exclude the rules whose latter part is in List C
Retain the rule of the highest confidence whose latter part is same in List B
Retain the rule of the highest confidence whose latter part is same in List D
Delete the rule whose latter part exists both in List B and D
End

Fig. 3.1 the process to find potential weaknesses of knowledge point

4. Conclusion

In this paper, the knowledge point association rules table is produced by association analysis, and 477 association rules has an increase in minimum support and confidence, which greatly improves the reliability and survivability. More importantly, we propose an algorithm to find potential weaknesses of knowledge point based on association rules. The algorithm takes knowledge point association and learners' grasp of knowledge into consideration, which is more likely to find potential weaknesses of knowledge point and gives guidance to learners based on their grasp of knowledge points.

References

[1]. Yu Xue,Yi Zhuang,Tianquan Ni,Siru Ni,Xuezhi Wen. Self-adaptive learning based discrete differential evolution algorithm for solving CJWTA problem[J]. Journal of Systems Engineering and Electronics,2014,01:59-68.

[2]. Jun Pan,Mingcen Jiang,Mingzhong Wang. The Application of Adaptive Learning Rate Optimization BP Neural Network Method in Water Quality Evaluation[A].ESIAT 2012[C].Information Engineering Research Institute,USA:,2012:6.

[3]. Jianhua Fan,Deyi Li. An Overview of Data Mining and Knowledge Discovery[J]. Journal of Computer Science and Technology,1998,04:348-368.

[4]. Caiping Cai,Wenxue Ye,Tianzhen Zhang,Wangzhen Guo. Association analysis of fiber quality traits and exploration of elite alleles in Upland cotton cultivars/accessions(Gossypium hirsutum L.)[J]. Journal of Integrative Plant Biology,2014,01:51-62.