INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 18, Issue: 3, Month: June, Year: 2023
Article Number: 5045, https://doi.org/10.15837/ijccc.2023.3.5045

CCC Publications

AGORA
UNIVERSITY PRESS

WOMDI-Apriori Data Mining Algorithm for Clustered Indicators Analysis of Specialty Groups in Higher Vocational Colleges

Fei Gao, Jing Yang*, Yang Yang, Xiaojing Yuan

Fei Gao
Non-governmental Higher Education Institute of China, Zhejiang Shuren College
Hangzhou 310015, China
gaofei@zjsru.edu.cn

Jing Yang*
Hebei Women and Children Activity Center
Shijiazhuang 050081, China
*Corresponding author: pljinger@126.com
*ORCID: 0000-0003-3216-9444.

Yang Yang
School of Transportation Science and Engineering, Beihang University
Beijing 100191, China
yangphd@buaa.edu.cn

Xiaojing Yuan
School of Traffic and Transportation, Beijing Jiaotong University
Beijing 100044, China
xjyuan@bjtu.edu.cn

Abstract

The cluster effect of specialty groups plays an important role in the development of Higher Vocational Colleges. The purpose of this research is to scientifically explore the interaction mechanism of specialty groups clustering indexes in higher vocational colleges, quantitatively analyze the correlation of these indexes, and explore reasonable measures to promote the specialty groups clustering effect in higher vocational colleges. Firstly, data denoising and field screening were carried out on the original data, and then the variables were clustered and divided into LHS (Left Hand Side) and RHS (Right Hand Side). Then, an improved multi-dimensional interactive Apriori association rule mining algorithm considering index weights and orientation constraints was proposed. The improved Apriori algorithm and the traditional Apriori algorithm were applied to mine the structured data sets. The results show that the improved WOMDI-Apriori algorithm in this study improves the accuracy by 79.96% compared with the traditional Apriori algorithm. The results indicate that, when the indicators of brand, key and characteristic majors at or above the provincial level, proportion of full-time teachers with double qualifications, and the number of internship students accepted by cooperative enterprises are at a low level, the number of projects and satisfaction proportion of employers with graduates would be negatively affected; The major category of equipment manufacturing is subjected to various factors coupling, which may lead to different graduates' counterpart employment rate; for association rules where the successor of the mining results is dominated by negative results, measures should be taken to avoid or reduce the possibility of their occurrence as much as possible. For association rules in which the successors of the mining results are dominated by positive results, measures should be taken to facilitate the occurrence of these frequent item sets whenever possible. The framework proposed in this research can provide theoretical guidance for analyzing operating characteristics and promoting the positive effects of specialty groups in higher vocational colleges.

Keywords: WOMDI-Apriori data mining algorithm, clustered indicators analysis, specialty groups, higher vocational colleges.

1 Introduction

"Double high plan" is another major project after the demonstration, backbone, high-quality higher vocational college construction plan in China, in which high-level college and high-level specialty (specialty group) is essentially a symbiotic relationship. In the limited resources of the fierce competition environment, higher vocational colleges, will inevitably focus on strengthening the construction of advantageous specialty groups, choose the cluster development mode, and the growth of specialty groups is indeed conducive to breaking through bottlenecks such as scattered resources, indistinctive features and insufficient linkage to improve the adaptability of talents, carry out technological innovation and focus on regional needs.

Specialty groups are a collection of majors or directions with common foundations, complementary advantages and resource sharing, and are committed to changing the division between majors and promoting cross-border cooperation in knowledge flow. Although not all countries have the terminology of specialty group, the research on interdisciplinary education can still provide important reference. Camilleri et al. pointed out that European professional higher education, as an applied education, requires to break the curriculum boundary and run through professional experience to promote knowledge integration [1]. Norton et al. proposed that American community colleges entrust multidisciplinary courses to achieve education content integration[2]. Scholars have carried out relevant research on the connotation definition, existing problems and development paths of specialty groups. Around the definition of specialty groups, the different interpretations contain the similarity theory, the joint force theory and the common theory, which are the most representative. In recent years, various views have shown a trend of convergence, such as Gu Yong'an [3]. Specialty groups also face a series of obstacles in the process of formation, operation and exertion of influence. Zeng Xianwen and Zhang Shu analyzed the role of specialty groups with the help of human capital value measurement models [4]. Zeng Xianwen and Yan Meng introduced concepts such as class density, group intensity, and concentration of specialty groups to quantitatively analyze specialty groups [5]. Around the existing problems, scholars have elaborated on the optimization path from different angles, such as Zong Cheng, he has pointed out that recombine talent training models, innovate the formation of teacher teams, and jointly build a sharing training base [6]. Zhao Mengcheng believes that it is necessary to establish specialty group faculty, inter-specialty team, curriculum sharing mechanism, and information sharing platform to achieve micro-organizational changes [7].

With the development of computer technology, the acquisition of big data has become possible [8]. The research and solution of problems in the social science field also increasingly rely on the use of big data methods. For example, the genetic algorithm is used to optimize production scheduling problem[9].The intelligent mechanism is established to realize the coordination and subsequent response of emergency resources [10]. Text data mining is used to analyze the relationship between innovation and development in economic field[11]. Machine learning and data mining tools are often able to solve some problems that cannot be efficiently discovered by traditional means. For example, The economic lot-size problem is explained through machine learning to optimize resource allocation[12]. With the help of machine learning and educational data mining, students' performance can be predicted based on video learning systems[13]. Data mining technology is a data analysis method that extracts implicit and potentially valuable laws for decision making from a large amount of data, and its process is user-oriented and knowledge discovery-oriented data analysis process, while association rule analysis is an important branch of data mining technology. In the work of traffic accident causation analysis and risk identification, association rule analysis is a common research method for data mining of this problem, and the Apriori association rule mining algorithm is one of the main methods of association rule analysis [14]; the association rule mining algorithm was first proposed by Agrawal et al. when they analyzed the problem of market shopping baskets, and the algorithm can accurately and effectively mining the correlation between two or more factors, but it cannot quantify the importance of a single factor in the association rule, and the computational effort will increase exponentially when there are more data items, the computational efficiency decreases, and it is easy to ignore rare data [15]. In the work of association rule mining for professional clustering effect, the factors are essentially treated with equal weights, which cannot reflect the degree of influence of different indicators on professional clustering effect and easily ignore the factors that need to be focused on, and at the same time, it is impossible to filter out the factors with weak influence on the results, which causes a large number of useless operations and affects the efficiency of model calculation [16]. As a result, many researchers began to improve the traditional Apriori algorithm. An intelligent method is used for improving the Apriori algorithm in order to extract frequent itemsets[17]. The improved Apriori algorithm can reduce the time complexity of association rule mining[18].

At present, the academic community has carried out useful exploration in the field of specialty groups, but it has yet to be enriched. On the one hand, the research theme still lacks in-depth analysis of the elements interrelationship and action mechanism within the specialty group, and the internal law of the development of the specialty group needs to be further excavated. On the one hand, the research method is mainly qualitative research, and in general it lacks the strong support of first-hand data and systematic empirical analysis. This study will use WOMDI-Apriori data mining algorithms to analyze cluster indicators for 232 specialty groups of higher vocational colleges, establish a corresponding analysis framework, and explore basic problems such as the essential attributes and operating mechanisms of specialty groups, so as to provide theoretical support for related research, and provide new ideas and perspectives for creatively understanding and solving problems, so as to contribute to the deepening of related research and the development of practice.

2 Data pre-processing

The original sample data set contains a total of 242 base data, each covering 41 field variables, constituting a 242*41 matrix. Before model construction and data mining analysis, the sample structure design needs to be implemented, and the first task is data pre-processing, including data cleaning (denoising), field screening, variable coding and numeralization, and the final available matrix output. The sample structure design process is shown in Figure 1.

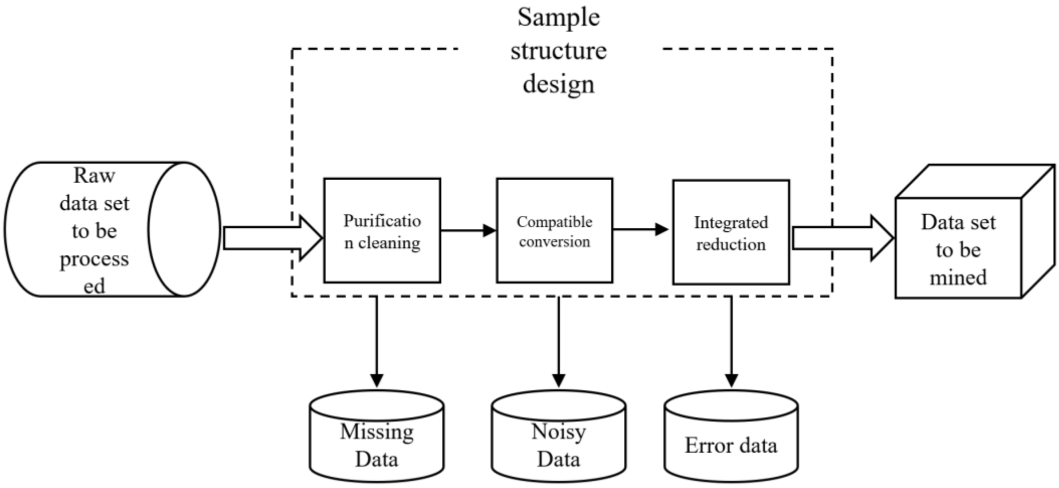


Figure 1: Sample structure design flow

2.1 Data denoising

Data Mining refers to the process of extracting hidden, deep and potentially valuable information from large, fuzzy and noisy big data through purification and denoising, algorithm design and other means. Data mining contains many meanings: (1) the data must be real and valid; (2) it contains the information resources required for data mining; (3) the information mined has the value of utilization; (4) it is not necessary to mine the general knowledge in the massive data, but more to mine specific laws. Data pre-processing is the first step in the process of data mining work, and it is also the most crucial one. Data pre-processing usually accounts for nearly 60% of the entire data mining workload, which is extremely time-consuming. It follows that when data mining is carried out, it is necessary to ensure the reliability and validity of the data at the very beginning, and effectively doing the work of data pre-processing can further improve the quality of the data in the database, effectively compensate for the incompleteness of the research data set, and provide more reliable data information for data mining work.

The most obvious problems of the original data in the record information of specialty groups are incompleteness, inconsistency and noisiness, which can directly cause the waste of computational resources and even lead to the bias of computational results. The main problematic characteristics of the data set are as follows: (1) worthless field variables, (2) incomplete missing data, (3) noisy data, (4) inconsistency and compatible transformation problems, and (5) redundant data and similar mergeable data. The general process of dealing with such data problems in this study includes: purification and cleaning, compatible conversion, and integration and subsumption. Specific examples are as follows:

(1) Merging of redundant data: the value of the "category of backbone major" field of the data of serial number code 226 in the original data is "manufacturing", the "category of backbone major" field of code 155 in the original data takes the value of "equipment", considering that they are actually "equipment manufacturing", therefore, both are modified. The data such as these are processed in a unified manner.

(2) Abnormal data deletion: The 10 data items indicating employment rate of "9625%" and "0%" are deleted.

(3) Data correction: the indicator value of "graduates' counterpart employment rate" is changed from "823" to "82.3"; the field " The value of "Number of full-time teachers (sum)" in a data entry is "150.0018", which is changed to "150"; the field "Number of hours taught by part-time faculty as a percentage of total professional hours in one academic year (%)" was changed to "84.2";. The value of "%" in a data entry is "447", which is replaced by "44.7"; the value of "Funding for horizontal projects (million yuan)" in a data entry is "(+)". The value of "(+)17" in one data is changed to "17"; the value of 3 data in the field "Brand, key and characteristic majors at or above provincial level" is empty and is filled with the value of "0" is filled in.

(4) No value field variable rejection: All values of the "Rank" field are empty, which is a worthless field, so it is rejected here.

2.2 Field Filtering

According to the characteristics of the association rule mining algorithm, the fields with too much dispersion (biased data fields) should be deleted in the design of this data sample structure, and the facing industries of the specialty groups should be the main object of this analysis. The fields that are not considered include "serial number", "school name", "name of specialty group", "city where the school is located", "name of included majors", "name of involved colleges", "name of included major categories", "ranking ".

2.3 Variable coding

(1) Clustering analysis of variable values

In the original data of this study, all fields take values in discrete variable form. The association rule mining algorithm for continuous variables generates memory overflow errors in the computation process, so the input data set needs to be in discrete variable form. However, too discrete variables (e.g., all integers with field values from 1 to 100) can cause the results to have too little support, and valuable association rules can be easily ignored and missed, so the discrete variables need to be clustered to obtain more focused and reliable association rule mining results

(2) Division of LHS (Left Hand Side) and RHS (Right Hand Side)

According to the characteristics of association rule mining, in the data mining process, there are causative and resultant terms, i.e., it is necessary to define the precedence term (LHS) and the successor term (RHS).

Here, the filtered and de-noised data were divided into causal dimensions, and the fields related to "status quo" and "input" were used as LHS, and the causal dimension fields were analyzed by professionalism: industry oriented, including major quantity, number of colleges involved, including major categories, number of majors sharing cooperative enterprises, number of majors sharing employers, number of majors sharing courses, number of majors sharing on campus and off campus training bases, number of majors sharing full-time teachers, brand key and characteristic majors at or above the provincial level, number of majors sharing off-campus part-time teachers, number of full-time students in specialty group , number of full-time teachers, proportion of full-time teachers with double qualifications number of hours taught by part-time faculty as a percentage of total professional hours for one academic year, number of specialty group training bases , average equipment value of students on campus training bases, frequency of use of on-campus training bases for one academic year, total number of cooperative enterprises , total number of courses jointly developed by cooperative enterprises , number of cooperative enterprises support part-time teachers , number of internship students accepted by cooperative enterprises , total value of equipment donated by the cooperative enterprise.

The fields of "consequence" and "output" related factors are used as RHS, and the fields of the result dimension through professional analysis include: the initial employment rate of graduates of the specialty group, graduates' counterpart employment rate, satisfaction proportion of employers with graduates , the number of graduates accepted by off-campus internship training bases , number of students accepted by the cooperative enterprise, total number of employees trained for the enterprise, number of provincial or above teaching achievement awards, number of provincial or above scientific research achievement awards, horizontal project funds, number of invention patent , number of industry standard, number of provincial or above awards for college student, number of provincial or above scientific research projects.

2.4 Sample structure design

Based on the above analysis, the clustering calculation is performed by the LOOK UP function embedded in the database, based on the "if" loop and copying the values of the clustered variables and returning them to the source file, processing all the fields according to the above process work and constructing the final input value matrix. The results of dimensional division and variable value clustering are shown in Table 1 below.

Table 1: Sample structured design data set

Dimension	variable	Value clustering	Value quantity
LHS	Category of backbone major	Finance and trade, electronics and information, equipment manufacturing, medicine and health, civil engineering, culture and art, education and sports, tourism, public management and services, agriculture, forestry, fisheries and animal husbandry, light industry and textiles, food, drugs and grains, transportation, energy, power and materials, and other categories	16
	Including major quantity	3,4,5,6	4

	Including major categories	1,2,3,4	4
	Brand, key and characteristic majors at or above the provincial level	0,1,2,3,4	5
	Number of majors sharing courses	0,1,2,3,4,5,6,7,8,10,11	11
	Number of majors sharing on campus and off campus training bases	1,2,3,4,5,6,7,8,9,10,11,14,17,18,33	15
	Number of majors sharing full-time teachers	$< 10, 11 \sim 50, > 50$	3
	Number of full-time students in specialty group	$< 1000, 1000 \sim 2000, 2000 \sim 3000, > 3000$	4
	Proportion of full-time teachers with double qualifications	$< 50, 50 \sim 70, 70 \sim 90, 90 \sim 100$	4
	Average equipment value of students on campus training bases	$< 1, 1 \sim 10, 10 \sim 100, > 100$	4
	Number of cooperative enterprises support part-time teachers	$< 10, 10 \sim 50, 50 \sim 100, > 100$	4
	Number of internship students accepted by cooperative enterprises	$< 100, 100 \sim 500, > 500$	3
	Total value of equipment donated by the cooperative enterprise	$< 50, 50 \sim 100, 100 \sim 500, 500 \sim 1000, > 1000$	5
RHS	Graduates' counterpart employment rate	$< 50, 50 \sim 70, 70 \sim 90, 90 \sim 100$	4
	Satisfaction proportion of employers with graduates	$50 \sim 90, 90 \sim 100$	3
	Number of students accepted by the cooperative enterprise	$< 100, 100 \sim 300, > 300$	3
	Number of provincial or above teaching achievement awards	0,1,2,3,4,5	6
	Number of provincial or above scientific research achievement awards	0,1,2	3
	Horizontal project funds	$< 10, 10 \sim 100, > 100$	3
	Number of invention patent	$0 \sim 5, 5 \sim 10, > 10$	3
	Number of industry standard	0,1,2,3,4,5,7,10	8
	Number of provincial or above awards for college students	$< 50, 50 \sim 100, > 100$	3

Number of provincial or above scientific research projects	$< 5, 5 \sim 10, > 10$	3
Total number of employees trained for the enterprise	$< 100, 100 \sim 1000, 1000 \sim 10000, > 10000$	4

3 Algorithm design and modeling

3.1 Association rules mining

Association rule mining is one of the main technologies of data mining, and it is also the most common form of mining patterns in unsupervised learning systems. Association rule mining is to mine valuable knowledge from a large amount of data to describe the relationship between data items [19].

A complete association rule can be expressed as the implication form of "$x => y$". X is the leading term, also known as the cause layer, y is the subsequent term, also known as the result layer. The association rule "$x => y$" is the basic condition that can be seen and established in the study, which meets the requirements of the preset support, confidence and lift. Support, confidence and lift are three important parameters that characterize the association rule, among which:

(1) Support: the number of transactions of itemset x contained in dataset D is called the support number of itemset x, expressed as σx. The support rate (also known as support) of item set X is recorded as: support (x), that is, probability $P(X)$:

$$\text{Support}(x) = \frac{\sigma x}{\{D\}} \times 100\% \tag{1}$$

Where {d} is the number of transactions in dataset D. if support (x) is not less than the preset minimum support threshold (min_support), X is called a frequent itemset, otherwise x is an infrequent itemset.

The support of item set (D) is the support rate of association rule $x => y$, which is essentially the proportion of transactions in D containing $(X \cup Y)$, that is, the frequency of probability p$(X \cup Y)$, as support $(x => y)$:

$$\text{Support}(X => Y) = \frac{|X \bigcup Y|}{|D|} = \text{Support}(X \bigcup Y) = \text{P}(X \bigcup Y) \tag{2}$$

(2) Confidence: two association rules defined on I and D, such as $x => y$, whose confidence means that "the transaction meeting condition x also meets condition Y". The confidence of association rules with $x => y$ is the conditional probability p$(y \mid x)$ of itemset y on the premise of including itemset x, which is recorded as: confidence $(x => y)$.

$$\text{Confidence(X => Y)} = P(Y \mid X) = \frac{P(X,Y)}{P(X)} = \frac{Sup(X => Y)}{Sup(X)} = \frac{Sup(X \cap Y)}{Sup(X)} \tag{3}$$

Where $X \subseteq I$, $Y \subseteq I$, $X \cap Y = \varnothing$.

(3) Lift: the lifting degree is used to characterize the correlation degree of the leading item and the subsequent item. In order to avoid the interference of pseudo strong association rules and prevent invalid association rules from appearing in the final result, the lifting degree index is hereby introduced, and this index is also used as the judgment condition of effective association rules:

$$\text{Lift}(X => Y) = \frac{P(Y \mid X)}{P(Y)} = \frac{Conf(X => Y)}{Sup(Y)} = \frac{Sup(X => Y)}{Sup(X)Sup(Y)} \tag{4}$$

The greater the degree of promotion, the higher the degree of correlation between itemset X and itemset y. Generally, we believe that only association rules with lift greater than 1 are effective strong association rules, and X and y are positively correlated at this time; If the lifting degree lift < 1, it means that X and y have no correlation degree or are mutually exclusive item sets. Such non effective association rules will not be considered in the results.

(4) Minimum support, minimum confidence threshold and strong association rules

In the modeling process, users can specify the minimum support (recorded as min_support) and the minimum confidence (recorded as min_confidence). The former describes the minimum importance that association rules must meet, and the latter specifies the minimum reliability that association rules must meet, and min_support $\in (0,1]$, min_confidence $\in (0,1]$.

Data set D meets the minimum support threshold and the minimum trust threshold on item set I, and the association rules with the lift greater than 1 are called valuable strong association rules.

3.2 Weighting model

In order to eliminate the weight deviation of subjective weighting method and objective weighting method in the process of weight assignment as much as possible, a combined weighting method based on the sum of deviation squares is adopted [20], and the subjective weight obtained by IAHP method and the objective weight obtained by rough set model are integrated and optimized to calculate the actual consideration weight of the corresponding index:

Combine the weight vector $\overline{\omega'}$ obtained by the subjective weighting method and the weight vector ω^* obtained by the objective weighting method to obtain the final reasonable weight vector $\bar{\omega}$. Assuming that w is the optimal weight vector, the deviation between the subjective weight vector w and the objective weight vector w should be minimized [21]. Establish optimization model:

$$\begin{cases} \min & \theta \sum_{j=1}^{n} \left(\overline{\omega'_j} - \bar{\omega}_j\right)^2 + (1-\theta) \sum_{j=1}^{n} \left(\omega_j^* - \bar{\omega}_j\right)^2 \\ \text{s.t} & \begin{cases} \bar{\omega} \geq 0 \\ \sum_{j=1}^{n} \bar{\omega}_j = 1 \end{cases} \end{cases} \tag{5}$$

Where θ denotes the trust degree in the result of subjective weighting; $1-\theta$ denotes the trust degree in the result of objective weighting; $\bar{\omega}_j$ represents the weight of the j th index attribute. There is an optimal solution $\bar{\omega} = \left[\omega^L, \omega^U\right]$ in formula (5) of the optimization model, where $\omega^L = \left(\omega_1^L, \omega_2^L, \omega_3^L \ldots, \omega_n^L\right)$ is the lower bound of the interval number of model solutions, $\omega^U = \left(\omega_1^U, \omega_2^U, \omega_3^U \ldots, \omega_n^U\right)$ is the upper bound of the interval number of model solutions, and $\bar{\omega}$ represents the interval weight vector of the final combination weighting. The steps of the established subjective and objective joint weighting model are as follows:

Step 1: Determine the set of objects to be evaluated and the corresponding index set. Let $\mathbf{X}$ be a collection of objects denoted as $\mathbf{X} = \{x_1, x_2 \ldots, x_n\}$, $\mathbf{A} = \{a_1, a_2, \ldots, a_n\}$ as the index set, $a(x)$ is the value of object $\mathbf{X}$ on attribute $\mathbf{A}$, meanwhile, the index value can be discrete value or continuous value.

Step 2: Interval number feature vector method is used to determine the subjective weight. According to the index comparison scale, interval number judgment matrix $\underline{\mathbf{B}} = \left[\mathbf{B}^L, \mathbf{B}^U\right]$ is determined, and the weight vector $\overline{\omega'} = \left[\omega'^L, \omega'^U\right]$ of each index $a_j (j \leq m)$ is calculated based on IAHP method.

Step 3: Apply rough set theory to get objective attribute weight. Attribute set $\mathbf{C} = \{c_1, c_2, \ldots C_n\}$ is the set of evaluation indicators determined in Step1, and the domain $\mathbf{U} = \{u_1, u_2, u_3 \ldots, u_n\}$ is a set of events with various possible causes at the corresponding time. The value of each traffic crash recorded on the sub-index is regarded as a piece of information of $u_t, u_t = \{c_{1t}, c_{2t} \ldots c_{nt}\}$, and discretize it to establish a discretized twodimensional information table. Then, the weight vector is calculated according to Formula (5).

Step 4: Obtain the optimal combination of weights. Calculate the final weight vector $\bar{\omega} = \left[\omega^L, \omega^U\right]$ according to Formula (5). The weight value is assigned to the field variables of traffic crash analysis, and the relative support, relative confidence, and relative lift of each association rule are determined.

3.3 Construction of the WOMDI-Apriori algorithm

The traditional Accident Tree analysis and FP-Tree algorithm and other methods are convenient for operators to grasp the overall characteristics of the problem, but they can't realize the relevance thinking between the multi-attributes of indicators in each dimension, nor can they effectively quantify and analyze the causes of the negative results. The association rule mining algorithm was first proposed by Agrawal et al. when analyzing the market basket problem. The algorithm can accurately and effectively mine the correlation between two or more factors, but it can't quantify the importance of a single factor in the association rule. In addition, when there are many data items, the amount of calculation will multiply, the computational efficiency will be reduced, and it is easy to ignore rare data [19]. At the same time, the traditional algorithm can't filter out the factors that have a weak influence on the results, leading to a large number of useless operations, affecting the calculation efficiency of the model. Most importantly, because the association rule mining algorithm was first developed for market basket analysis, some scholars did not improve and optimize the algorithm in an orderly way when applying the algorithm to analyze the association rules of other problems. If the traditional Apriori association rule mining algorithm is directly applied to study the problems in the fields of economics or education, a large number of invalid or even incorrect disordered association rules are output in the result [20].

This research improves and optimizes the algorithm from three perspectives: 1) the algorithm is constrained by the form of orderly and directional rule Association, so that the traditional Apriori association rule mining algorithm for the application scope of the shopping basket field can be compatible with the problem of clustered indicators analysis of special groups in higher vocational colleges; 2) Through the subjective and objective weighting model, the index weights of all field variables are calculated, and based on the weight optimization results, the concepts of "relative support", "relative confidence" and "relative improvement" are proposed; 3) Breaking the traditional mining output method of "cause leading item => consequence subsequent item", introducing the idea of multi-dimensional interactive association, not only considering the association relationship of "cause leading item => consequence subsequent item", but also exploring the association rules between dimensions from the perspective of the autocorrelation (leading item) of the dimension part of the cause layer and the internal autocorrelation (subsequent item) of the result dimension.

Figure 2 shows the steps of the multi-dimensional interactive improved Apriori algorithm considering directional constraints proposed in this paper.

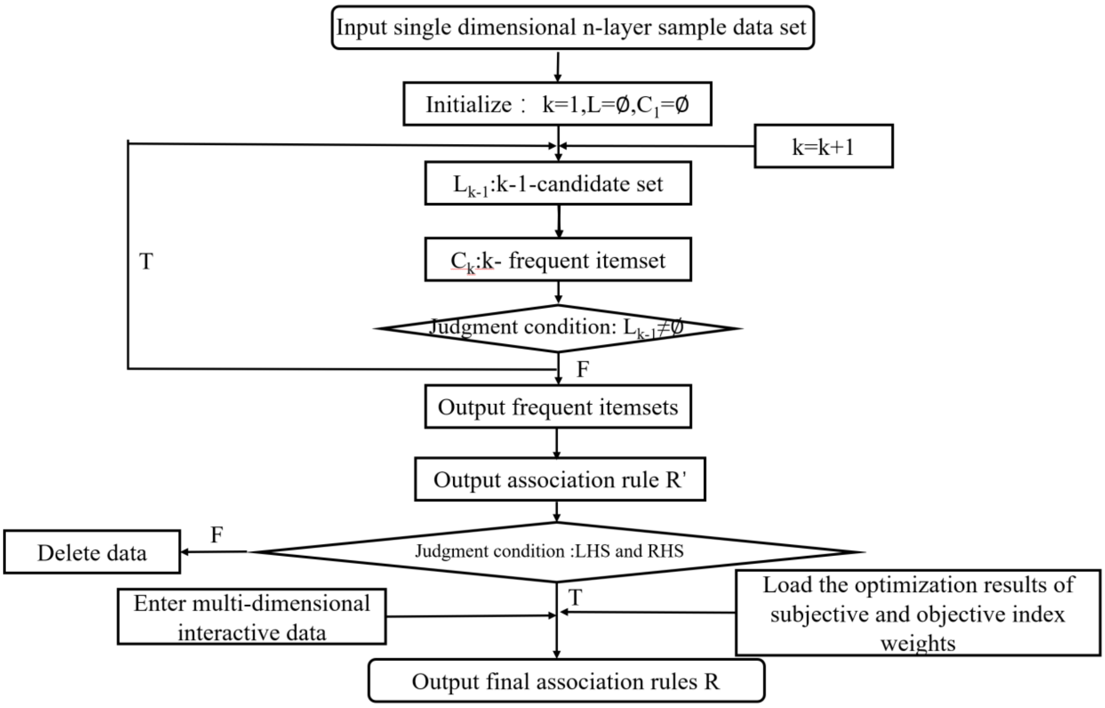

Input single dimensional n-layer sample data set
Initialize： k=1,L=∅,C$_1$=∅
k=k+1
L$_{k-1}$:k-1-candidate set
T
C$_k$:k- frequent itemset
Judgment condition: L$_{k-1}$≠∅
F
Output frequent itemsets
Output association rule R'
F
Delete data
Judgment condition :LHS and RHS
T
Enter multi-dimensional interactive data
Load the optimization results of subjective and objective index weights
Output final association rules R

Figure 2: Professional clustering effect association rule mining process

4 Association rule mining analysis

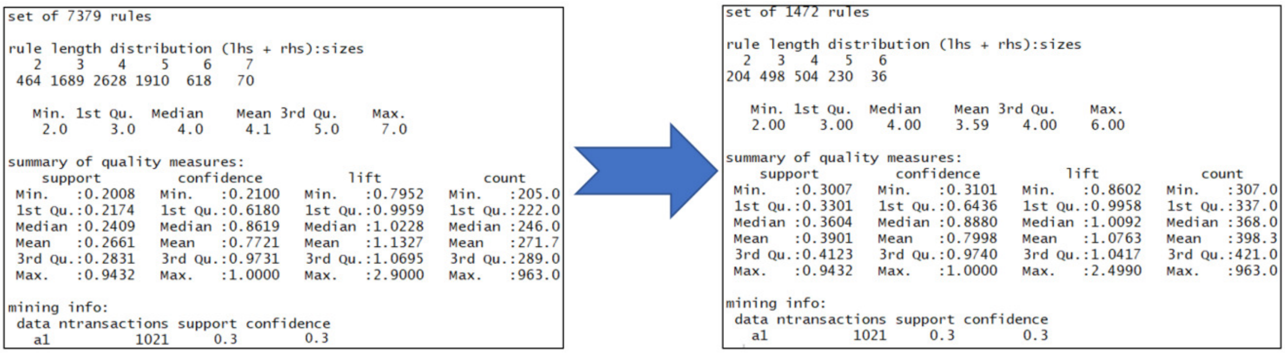
4.1 Parameter calibration

First, the input of data of each dimension is carried out; then, the initial threshold is set, and the minimum support threshold min_sup=0.30 and the minimum confidence threshold min_conf=0.35 are set after continuous manual debugging considering the main features of the specialty group clustering effect in higher vocational colleges; meanwhile, the min_lift threshold of the lift is set to 1, so that the frequent itemsets with no positive association between the preceding and following items can be eliminated.

4.2 Improving the accuracy improvement calibration of the algorithm

The original Apriori association rule mining algorithm and the improved WOMDI-Apriori association rule mining algorithm are applied to the input dataset by calling the "arules" function package in R language [22], and the mining results are summarized in Figure 3.

The above calculation results can be identified: the original Apriori association rule mining algorithm without loaded orientation constraints mines a total of 7379 items, while applying the WOMDI-Apriori model proposed in this paper, under the same threshold setting (min_sup=0.30, min_conf=0.35), only the mining results output only 1472 eligible association rules. In other words, if the algorithm is not optimized and improved, at least 5907 invalid association rules will be generated, and the improved WOMDI-Apriori model improves the computational accuracy by 79.96% over the traditional Aprori model under the conditions of the base data in this paper. It can be seen that the direct application of the traditional Apriori algorithm to such problems will cause some confusion to the results, so the optimization and improvement of the Apriori algorithm in this paper is necessary for such research problems of association rule analysis of professional clusters in higher education schools.



```
set of 7379 rules

rule length distribution (lhs + rhs):sizes
   2    3    4    5    6    7
 464 1689 2628 1910  618   70

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.0     3.0     4.0     4.1     5.0     7.0

summary of quality measures:
    support          confidence          lift             count
 Min.    :0.2008   Min.    :0.2100   Min.    :0.7952   Min.    :205.0
 1st Qu.:0.2174   1st Qu.:0.6180   1st Qu.:0.9959   1st Qu.:222.0
 Median :0.2409   Median :0.8619   Median :1.0228   Median :246.0
 Mean   :0.2661   Mean   :0.7721   Mean   :1.1327   Mean   :271.7
 3rd Qu.:0.2831   3rd Qu.:0.9731   3rd Qu.:1.0695   3rd Qu.:289.0
 Max.   :0.9432   Max.   :1.0000   Max.   :2.9000   Max.   :963.0

mining info:
 data ntransactions support confidence
   a1          1021     0.3        0.3
```

```
set of 1472 rules

rule length distribution (lhs + rhs):sizes
  2   3   4   5   6
204 498 504 230  36

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00    3.00    4.00    3.59    4.00    6.00

summary of quality measures:
    support          confidence          lift             count
 Min.    :0.3007   Min.    :0.3101   Min.    :0.8602   Min.    :307.0
 1st Qu.:0.3301   1st Qu.:0.6436   1st Qu.:0.9958   1st Qu.:337.0
 Median :0.3604   Median :0.8880   Median :1.0092   Median :368.0
 Mean   :0.3901   Mean   :0.7998   Mean   :1.0763   Mean   :398.3
 3rd Qu.:0.4123   3rd Qu.:0.9740   3rd Qu.:1.0417   3rd Qu.:421.0
 Max.   :0.9432   Max.   :1.0000   Max.   :2.4990   Max.   :963.0

mining info:
 data ntransactions support confidence
   a1          1021     0.3        0.3
```

Figure 3: Comparison of original Apriori algorithm and improved Apriori algorithm

Figure 3 can show the percentage of the total frequent item set in the sparse matrix of the mining results. Through the above mining summary information can also be found, after the algorithm improvement and optimization, the rules that contain more items in the mining results are: 204 association rules for those containing 2 items, 498 association rules for those containing 3 items, 504 association rules for those containing 4 items, 230 association rules for those containing 5 items, and 36 association rules for those containing 6 items The median is 4, which means that 50% of the frequent itemsets contain no more than 4 items, and the mean value of 3.59 means that the average number of items contained in all frequent itemsets is 3.59.

5 Results and discussion

5.1 Valuable association rule extraction

The parameter scatter plot of the mining results is plotted by the "plot function" in arulesViz, a visual analysis toolkit for association rules in R language, as shown in Figure 4. The horizontal coordinate in the figure represents the relative support (R-Sup), the vertical coordinate represents the relative confidence (R-Conf), and the color shade is the relative lift (R-Lift) size.

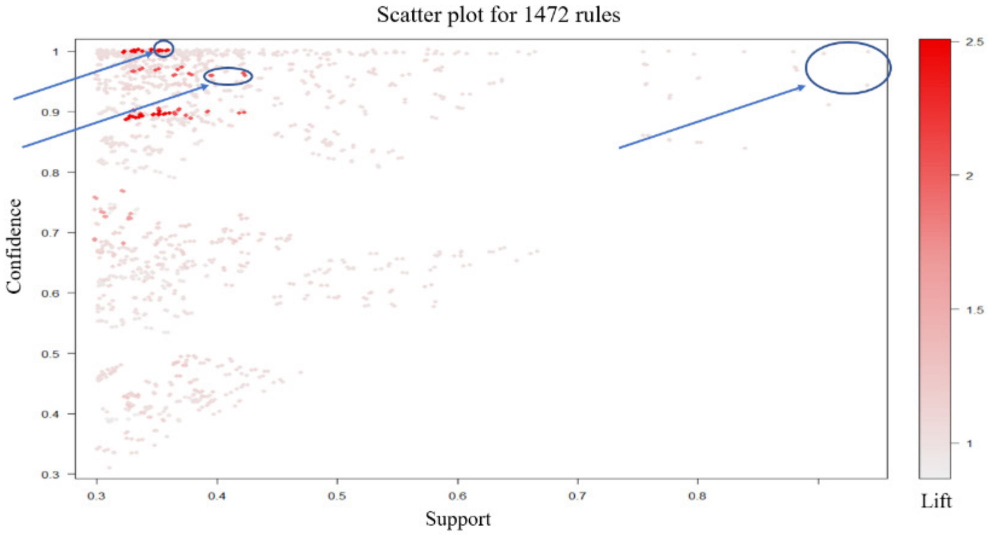


Figure 4: Scatter plot of association rule mining results

The scatter plot of association rules of the mining results in Figure 4 shows that the relative support of some association rules ranges from 0.30 to 0.65, and another part ranges from 0.75 to 0.95, indicating that high-frequency rules and low-frequency rules coexist among all threshold-eligible rules; from the confidence index, the relative confidence of most association rules is between 0.3 and 1, but the region below 0.9 relative confidence has a lighter color, which means that the relative lift of this part of association rules is not enough, and it is possible that the association rules with insufficient degree of association between item sets or even invalid ones; meanwhile, for the lift index, by observing the colors in the scatter plot, about half of the association rules have a relative lift less than 1, indicating that a large proportion of the rules fail to satisfy the constraint of a lift greater than 1, i.e., they are invalid association rules.

On the basis of filtering out the valid association rules (R-lift>1) and focus locking on the regions with high confidence and high support (blue circles in Figure 4), the valuable strong association rules are further extracted.

5.2 Analysis of high support association rule extraction results

In order to extract the valuable strong association rules more precisely, the "inspect" and "sort" functions in R Studio are called, and the 1472 valid association rules obtained by applying WOMDI-Apriori association rule mining algorithm are conditionally sorted according to the relative support (R-Sup) from the largest to the smallest, and the rules with relative lift (R-Lift) less than 1 are eliminated by extension, and the extracted output results are shown in Table 2.

The high support association rule characterized by the relative support ranking corresponds to the higher frequency of frequent item sets, and by analyzing the results in Table 2, the following pattern is summarized:

(1) The relative confidence of the three association rules with the highest ranking of high support is also at a high level, and their relative confidence is greater than 9; indicating that these association rules are all strongly correlated rules. The relative lift is greater than 1 and less than 1.1, which means that these association rules are positively correlated, but the correlation is at a low level.

(2) The lower level indicators including brand, key and characteristic majors at or above the provincial level, proportion of full-time teachers with double qualifications, and number of internship students accepted by cooperative enterprises may basically cause the loss of horizontal and vertical projects, and satisfaction proportion of employers with graduates may be affected negatively as well; On the contrary, the specialty group of medicine and the health in the higher vocational colleges with three brand, key and characteristic majors at or above the provincial level , often have a high number

Table 2: Extraction of high relative support results

rules	R-support	R-confidence	R-lift
{Brand, key and characteristic majors at or above the provincial level= 0, Proportion of full-time teachers with double qualifications=<50, Number of internship students accepted by cooperative enterprises =<100}=> { Satisfaction proportion of employers with graduates = 50∼90, Horizontal project funds =<10, Number of provincial or above scientific research projects =<5}	0.941	0.998	1.01
{ Category of backbone major=Medicine and health, Brand, key and characteristic majors at or above the provincial level= 3, Number of full-time students in specialty group =>3000, Cooperative enterprises support part-time teachers Total =>100, Cooperative enterprises accept internship studentsNumber =>500}=> { Satisfaction proportion of employers with graduates = 90∼100, Number of scientific research projects at or above the provincial level= 5∼10}	0.933	0.905	1.01
{Including major categories= 2, Total value of equipment donated by the cooperative enterprise = 100∼500}=> {Scientific research achievement award at or above the provincial level=1}	0.892	0.973	1.02

of students, cooperative enterprises support part-time teachers, and cooperative enterprises accept internship students. Higher vocational colleges with such conditions are at a high level of satisfaction proportion of employers with graduates , and there are also a considerable number of scientific research projects.

(3) In most situations, if the specialty groups with two major categories, and total value of equipment donated by the cooperative enterprise is between 1 ∼ 5 million RMB, they can often get 1 provincial or above scientific research achievement award.

5.3 Analysis of high-confidence association rule extraction results

In order to extract the valuable strong association rules more precisely, the "inspect" and "sort" functions in R Studio are called, and the 1472 valid association rules obtained by applying WOMDI-Apriori association rule mining algorithm are conditionally sorted according to the relative confidence (R-Conf) from the largest to the smallest, while those rules with relative lift (R-Lift) less than 1 are eliminated by extension, and the extracted output results are shown in Table 3.

The conditional probability of occurrence of high confidence association rules characterizing frequent item sets according to the relative confidence ranking is higher, and the following pattern is summarized by analyzing the results in Table 3.

(1) By observing Table 3, we can find that the association rules with the highest three relative confidence levels do not have high relative support, indicating that the item sets with high conditional probabilities are not necessarily frequent item sets. However, their R-lifts are all greater than 2, suggesting that these high-confidence association rules are strongly correlated frequent item sets.

(2) Also in major category of equipment manufacturing, when the number of full-time students in specialty group is greater than 3000, counterpart employment rate of graduates is instead at a lower level of 50%-70%; on the contrary, when in the case of number of full-time students in specialty group is less than 1000, graduates' counterpart employment rate is higher. Possible explanations are that other conditions have changed, affecting the prior distribution of conditional probabilities, such as brand, key and characteristic majors at or above the provincial level, proportion of full-time teachers

Table 3: Extraction of high relative confidence results

rules	R-support	R-confidence	R-lift
{ Category of backbone major=Equipment manufacturing, Brand, key and characteristic majors at or above the provincial level=3, Number of full-time students in specialty group =>3000, Proportion of full-time teachers with double qualifications=<50}=>{ Satisfaction proportion of employers with graduates =50∼90, Graduates' counterpart employmentrate =50∼70}	0.351	0.998	2.49
{Category of backbone major=Equipment manufacturing, Including major quantity=5, Number of full-time students in specialty group =<1000, Total value of equipment donated by the cooperative enterprise =>1000}=>{ Graduates' counterpart employmentrate=90∼100, Number of provincial or above awards for college student =>100}	0.359	0.996	2.5
{Number of majors sharing on campus and off campus training bases =9, Average equipment value of students on campus training base =10∼100, Total value of equipment donated by the cooperative enterprise =100∼500}=>{Teaching achievement award at or above provincial level=4, Invention patent=5∼10 }	0.361	0.996	2.49

with double qualifications and differences in conditions such as including major quantity.

(3) Medium level input of major sharing on campus and off campus training bases, equipment in the campus training base and equipment donated by the cooperative enterprise can create a medium level of output of teaching achievement awards at or above provincial level and invention patents.

5.4 Recommendations based on association rule mining results

For association rules in Tables 2 and 3 where the successor of the mining results is dominated by negative results, measures should be taken to avoid or reduce the possibility of their occurrence as much as possible. For association rules in which the successors of the mining results in Tables 2 and 3 are dominated by positive results, measures should be taken to facilitate the occurrence of these frequent item sets whenever possible. Based on the above results, the specific measures are as follows:

(1) The high support association rules indicates that such frequent item sets have a high frequency of occurrence, and if the successor item is a negative outcome, the occurrence of the preceding item should be avoided as much as possible to reduce the frequency of negative outcomes; for example, the first and third association rules in Table 2 should take measures to prevent the occurrence of the preceding item to avoid higher education institutions from obtaining lower employers satisfaction , and fewer project and research awards.

(2) The high support association rules indicates that such frequent item sets have a high frequency of occurrence, and if the successor item is a positive outcome, then the frequency of the preceding item should be increased as much as possible to enhance the frequency of the positive outcome; for example, the second association rule in Table 2 should take measures to promote the occurrence of the preceding item so that higher education institutions can obtain higher employers satisfaction and more research projects.

(3) The high confidence association rules indicates that such frequent item sets have a high frequency of occurrence. If the successor is a negative result, once the preceding term occurs, extra attention should be paid at this time; for example, the first association rule in Table 3, once the combination of the frequent set of items of the preceding term matches the situation in the table, extra

attention should be paid to prevent the lower cunterpart employment rate of graduates and employers' satisfaction proportion with graduates.

(4) The high confidence association rules characterize such frequent item sets with high probability of occurrence. If the successor is a positive outcome, it should contribute to the frequent item set coupling condition of the prior as much as possible, e.g., the second association rule in Table 3 should set the coupling condition of the prior as much as possible in order to facilitate higher graduates' counterpart employment rate and more provincial or above awards for college student.

6 Conclusions

The main findings of this research are obtained as follows:

(1) The improved WOMDI-Apriori algorithm proposed in this study has 79.96% higher accuracy than the traditional Apriori algorithm, which will cause some confusion to the results if the traditional association rule mining algorithm is applied directly. Therefore, it is necessary to improve the Apriori algorithm for the problem of association rule analysis of clustering effect of specialty groups in higher vocational colleges.

(2) When brand, key and characteristic majors at or above the provincial level, proportion of full-time teachers with double qualifications, and cooperative enterprises accept internship students' number are at low levels, the number of projects obtained by higher vocational colleges and the satisfaction of employers with graduates are negatively affected. For the major category of equipment manufacturing, the number of full-time students in specialty group of higher vocational colleges does not fully determine the graduates' counterpart employment rate, and the level of this indicator is also related to the coupling effect of other factors.

(3) High support association rules characterize the high frequency of such frequent item sets. If the successor items in the specialty group cluster effect mining results are negative results, the occurrence of the leading items should be avoided as much as possible to reduce the frequency of negative results; if the successor items in the specialty group cluster effect mining results are positive results, the occurrence frequency of the leading items should be increased as much as possible to improve the frequency of positive results. High confidence association rules characterize such frequent item sets with a high probability of occurrence. If the successor in the specialty group cluster effect mining result is a negative result, once the preceding term occurs, extra attention should be paid at this time; if the successor in the specialty group cluster effect mining result is a positive result, the frequent term set coupling condition of the preceding term should be contributed as much as possible.

Further research directions:

(1) In analyzing the clustering effect of specialty groups in higher vocational colleges, the LHS and RHS divisions for each field variable in the original data may be unreasonable and need to be further optimized and adjusted in the future.

(2) In the future, the fields need to be clustered and divided into multiple dimensional field clusters, so that each field may appear in the mining results as a causative factor or a result. For the analysis of the cluster taking values of each field index, further refinement is needed in the future based on the position of statistical quartiles, medians, and means.

Acknowledgments

This work was supported by Humanities and Social Science Planning Fund Project of Chinese Ministry of Education “Research on the Generation Mechanism and Measurement Model of Specialty Groups Agglomeration Effect in Higher Vocational Colleges under the Background of ‘Double high plan’" (21YJA880013).

Author contributions

The authors contributed equally to this work.

Conflict of interest

The authors declare no conflict of interest.

References

[1] Camilleri, A., Delplace, S., Frankowicz, M. et al.(2014) . *Professional Higher Education in Europe Characteristics, Practice Examples and National Differences*, Brussels: European Association of Institutions in Higher Education. 2014.

[2] Grubb , W., Badway, N., Bell, D.,Kraskouskas, E.(1996). *Community College Innovations in Workforce Preparation: Curriculum Integration and Tech-Prep*, Washington, DC: Office of Vocational and Adult Education.1996.

[3] Gu, Y.A.(2016). Applied Undergraduate Specialty Cluster: An Important Breakthrough in the Transformation and Development of Local Universities, *China Higher Education*, 22, 35–38, 2016.

[4] Zeng, X.W., Zhang, S.(2010). On the Construction of Specialty Group in Higher Vocational Colleges — a Qualitative Discussion. *Contemporary Education Science*, 13, 15–18, 2010.

[5] Zeng, X.W., Yan, M.(2010). On the Construction of Specialty Group in Higher Vocational Colleges - based of Quantitative Analysis, *Chinese Vocational and Technical Education*, 18, 33–36, 2010.

[6] Zong, C. (2020). Vocational Colleges and Universities: How to Build and How to Evaluate, *Journal of Vocational Education*, 7, 40–45, 2020.

[7] Zhao, M.C. (2020). On the Nature of Major Clusters Construction of Higher Vocational Colleges and its Organizational Reform Ways on Micro Level, *Research in Educational Development*, 9, 63–70, 2020.

[8] Yang, Y., He, K., Wang, Y.P. et al.(2022). Identification of Dynamic Traffic Crash Risk for Cross-area Freeways Based on Statistical and Machine Learning Methods, *Physica A: Statistical Mechanics and Its Applications*, 595, 127083-,2022.

[9] Xu, W., Sun, H.Y., Awaga, A.L., Yan, Y.,Cui, Y.J. (2022). Optimization Approaches for Solving Production Scheduling Problem: A Brief Overview and a Case Study for Hybrid Flow Shop Using Genetic Algorithms, *Advances in Production Engineering & Management*, 17(1), 45–56, 2022.

[10] Sun, H.Y., Xu, W., Yu, Y.Y., Cai , G.Y.(2022). An Intelligent Mechanism for COVID-19 Emergency Resource Coordination and Follow-Up Response, *Computational Intelligence and Neuroscience*, 1–10.2022.

[11] Cicea, C., Lefteris, T., Marinescu, C., Popa, Șc., Albu, Fc.(2021). Applying Text Mining Technique on Innovation-Development Relationship: A Joint Research Agenda, *Economic Computation And Economic Cybernetics Studies And Research*, 55(1), 5–22, 2021.

[12] Sousa, Junior W.T. de, Montevechi, J.A.B., Miranda, R. de C., Rocha, F., Vilela, F.F.(2019). Economic Lot-Size Using Machine Learning, Parallelism, Metaheuristic and Simulation, *International Journal of Simulation Modelling*, 18(2), 205–216, 2019.

[13] Teoh, C.W., Ho, S.B., Dollmat, K.S. et al. (2022). Predicting Student Performance from Video-Based Learning System: a Case Study, *Informatics and Service Science*, 9(3), 64–7, 2022.

[14] Singh, P.K., Othman, E., Ahmed, R.et al.(2021). Optimized Recommendations by User Profiling Using Apriori Algorithm, *Applied Soft Computing*, C, 107272, 2021.

[15] Redhu, S., Hegde, R.M. (2020). Optimal Relay Node Selection in Time-varying IoT Networks Using Apriori Contact Pattern Information, *Ad hoc networks*, 98(Mar.):102065.1-102065.9.2020.

[16] Yang, Y., Wang, K., Yuan, Z., Liu, D. (2022). Predicting Freeway Traffic Crash Severity Using XGBoost-Bayesian Network Model with Consideration of Features Interaction, *Journal of Advanced Transportation*, 4257865.2022.

[17] Karimtabar, N., Fard, M.J.S.(2022). Finding Frequent Items: Novel Method For Improving Apriori Algorithm, *Computer Science-AGH*, 23(2), 161–177, 2022.

[18] Pan, T. (2021). An Improved Apriori Algorithm for Association Mining Between Physical Fitness Indices of College Students, *International Journal of Emerging Technologies in Learning*, 16(9): 235–246, 2021.

[19] Yang,Y., Tian,N., Wang,Y., Yuan, Z. (2022). A Parallel FP-Growth Mining Algorithm with Load Balancing Constraints for Traffic Crash Data, *International Journal of Computers Communications & Control*, 17(4): 4806.2022.

[20] Yang, Y., Yuan, Z., Meng, R. (2022). Exploring Traffic Crash Occurrence Mechanism toward Cross-Area Freeways via an Improved Data Mining Approach, *Journal of Transportation Engineering Part A Systems*, 148(9): 04022052.2022.

[21] Yang, Y., Yuan, Z., Chen, J., Guo, M. (2017). Assessment of Osculating Value Method Based on Entropy Weight to Transportation Energy Conservation and Emission Reduction, *Environmental Engineering & Management Journal*, 16(10), 2413–2424, 2017.

[22] Narváez-Bandera, I., Suárez-Gómez, D., Isaza, C E. et al. (2022). Multiple Criteria Optimization (MCO): A Gene Selection Deterministic Tool in RStudio, *PLOS ONE*, 17. 2022.

COPE
Member since 2012
JM08090

This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).
https://publicationethics.org/members/international-journal-computers-communications-and-control

Cite this paper as:

Gao, F.; Yang, J.; Yang, Y.; Yuan X.J. (2023). WOMDI-Apriori Data Mining Algorithm for Clustered Indicators Analysis of Specialty Groups in Higher Vocational Colleges, *International Journal of Computers Communications & Control*, 18(3), 5045, 2023.
https://doi.org/10.15837/ijccc.2023.3.5045