Association Rule Mining for Career Choices Among Fresh Graduates

Leibao Zhang\textsuperscript{1}, Xiaowen Tan\textsuperscript{1}, Shuai Zhang\textsuperscript{2,*}, Wenyu Zhang\textsuperscript{2}

\textsuperscript{1}School of Public Finance and Taxation, Zhejiang University of Finance and Economics, Hangzhou, China
\textsuperscript{2}School of Information, Zhejiang University of Finance and Economics, Hangzhou, China

Email address: labaor@zufe.edu.cn (Leibao Zhang), oohtxw@zufe.edu.cn (Xiaowen Tan), zhangshuai@zufe.edu.cn (Shuai Zhang), wyzhang@e.ntu.edu.sg (Wenyu Zhang)

*Corresponding author

To cite this article:
Leibao Zhang, Xiaowen Tan, Shuai Zhang, Wenyu Zhang. Association Rule Mining for Career Choices Among Fresh Graduates. Applied and Computational Mathematics. Vol. 8, No. 2, 2019, pp. 37-43. doi: 10.11648/j.acm.20190802.13

Received: May 12, 2019; Accepted: July 2, 2019; Published: July 19, 2019

Abstract: Nowadays, an increasing number of colleges have built information systems to manage masses of educational data, but actually most data is in an idle state and fails to create any value. As an efficient data analysis method, association rule mining can precisely make good use of these disordered data and extract useful but latent information from them. In this paper, an example of 228 students, who graduated from the School of Information of Zhejiang University of Finance and Economics, China in 2017, is taken to discover the association rules between their career choices and academic performance using Apriori algorithm. The main purpose of this paper is to offer fresh graduates a reference to future career choices and help teachers guide them in better career planning. The experimental results indicate that the courses students are good at largely affect their career choices, although their overall career scope is not narrow.

Keywords: Data Mining, Association Rule Mining, Apriori Algorithm, Career Choice, Academic Performance

1. Introduction

Recently, with the constant enlargement of college enrollment scale, higher education is gradually stepping into the stage of massification. However, what follows is the sharp increase in both the number of fresh graduates and the employment pressure each year. How to enter a suitable industry and further get an ideal job according to one’s own academic background also becomes a puzzle for most fresh graduates. Against this background, the research on graduate employment attracts more attention [1-3], but few of them make an approach to the relationship between the career choices and individual academic performance through an effective method.

Data mining is a process of extracting potentially useful information from vast amounts of data, and it is also known as educational data mining when it comes to educational-related data [4-5]. Generally, data mining can be realized through the methods of association rule mining, classification, prediction and so on. Among them, association rule mining can be applied to identify frequent patterns and interesting causal relationships that satisfy the predefined minimum values of parameters concerned [6]. In this paper, to explore the relationship between the fresh graduates’ career choices and their individual academic performance, an example of 228 students, who graduated from the School of Information of Zhejiang University of Finance and Economics, China in 2017, is taken as the research object. The Apriori algorithm based on R programming is used to realize the association rule mining.

The rest of this paper is structured as follows. Section 2 summaries the current related works on association rule mining and Apriori algorithm. Section 3 elaborates the approaches to data preprocessing and association rule mining. Section 4 presents the experimental results of association rule mining. Conclusions and directions of future work are discussed in Section 5.

2. Related Work

With the rapid development of data mining technology,
association rule mining has become a quite active research hotspot for its special advantages. Ho et al. [7] combined association rule mining with fuzzy theory to help investors decide the right trade time in financial market. Kumar and Toshniwal [8] used k-means algorithm to group the road accident locations based on accident frequency count, and then applied association rule mining to discover the main factors that affect the road accidents at these locations. Mane and Ghorpade [9] predicted the students' admission tendency for each branch of engineering through association rule mining with pattern growth method. Nahar et al. [10] applied association rule mining to identify the sick and healthy factors that cause heart disease for males and females. To extract KANSEI knowledge, Shi et al. [11] applied association rule mining and rough sets to explore the relationship between form features and KANSEI adjectives. Based on the multi-objective particle swarm optimization algorithm, Tyagi and Bharadwaj [12] adopted support and confidence as two objective functions and then proposed a new approach to obtain direct and indirect association rules with high quality. Wang et al. [13] introduced a weighted fuzzy association rule mining method to discover the alarm sequences, and the method combines fuzzy sets, Apriori algorithm and alarm time sequence metrics together.

The Apriori algorithm [14] popularized in association rule mining has been widely used in various fields. D’Angelo et al. [15] proposed a pervasive computing architecture based on a trust model, in which Apriori algorithm is used to extract the users’ behavioral patterns from their historical data during network interactions. Deng et al. [16] built a causation analysis model for traffic accidents, and the Apriori-Genetic algorithm is used to mine the main influencing factors. To make the mobile e-commerce shopping more convenient and high-quality, Guo et al. [17] proposed an improved Apriori algorithm according to the characteristics of the mobile e-commerce recommendation system. Ilayaraja and Meyyappan [18] used the Apriori algorithm to identify the frequent diseases at given time periods according to a data of patients from different areas. Combining the Apriori algorithm with frequent subgraph tree, Nair and Thomas [19] proposed a two-phase approach to discover the frequent subgraphs in graph datasets. Prasanna and Ezhilmaran [20] used an improved Apriori algorithm, together with a modified Genetic algorithm, to help the investors select better stock rules for their investment.

However, there exist few studies focusing on the relationship between career choices and academic performance through data mining, especially the method of association rule mining. Thus, based on a data set supported by Zhejiang University of Finance and Economics, this paper explores the relationship between fresh graduates’ career choices and their individual academic performance. To be specific, the academic performance here is mainly depicted by students’ course grades. The experimental results are conducive to making students’ career objectives clearer and improving their chance of being employed.

3. Data Preprocessing and Association Rule Mining

This section is made up of two parts. One is a detailed description of data preprocessing, and the other is the elaboration of the Apriori algorithm used for association rule mining.

3.1. Data Preprocessing

The raw data used in this paper includes the employment and academic performance information of 228 fresh graduates from Zhejiang University of Finance and Economics, who graduated in 2017 and respectively majored in Information Management and Information System, Computer Science and Technology, E-Commerce and Software Engineering. Owing to the existence of missing values and noises, it is essential to preprocess the raw data for more desired experimental results.

For the employment information, three different types of these students’ career choices are summarized based on the national economy industry classification and the characteristics of their majors, i.e., Work in Information Industry, Work in Finance or Business, and Work in Other Industries. For academic performance information, except for the deletion of students’ private information like their names and ID numbers, three main steps are taken as follows.

1. The combination of similar attributes. The raw data of academic performance is mainly consisted of students’ course grades and the total number of the courses is 207. Among all courses, some courses are rather similar, such as English (2), English (3) and English (4). Taking these three courses as an example, they are combined into a new attribute named College English through averaging.

2. The removal of unsuitable attributes and records. Although the raw data contains the information of a lot of courses, about 90% of the courses are taken by less than one third of all 228 students. Thus, to make the experimental results more accurate, these courses must firstly be removed. Secondly, courses which gave the similar grades to all students are removed, such as College Chinese, Military Training and so on. And lastly, students with too many missing attribute values are also removed.

3. The transformation of course grades. According to the Implementation Measures of Credit System for Undergraduate Students of Zhejiang University of Finance and Economics, the grade of one course can be given in either hundred-mark system or grade point system, but grade point average (GPA) must be employed to measure students’ overall quality. To satisfy the data type requirement of association rule mining in R programming, five levels listed in Table 1 are used to transform the course grades and GPA. The NA values are labeled as “Unselected” for those students who did not take the course. Besides, although students must attend the make-up exams if they fail the compulsory courses, they can take other courses instead if they fail the optional courses. Thus, the label “Unselected” is also given to the records of cancelling or abandoning the exams belonging to optional courses.
3.2. Association Rule Mining

In this paper, the R programming and Apriori algorithm are used to discover the rules between the selected fresh graduates’ career choices and their academic performance. For the Apriori algorithm, which centers mining the frequent itemsets of Boolean association rules, its realized process can be divided into the following two steps.

First, discover all the frequent itemsets whose support values are not lower than the predefined threshold through iterations.

Second, take advantage of frequent itemsets to generate the rules which satisfy the predefined minimum confidence value.

Thus, two key parameters are included in this algorithm, i.e., support and confidence. Assuming \( Z=\{Z_1, Z_2, \ldots, Z_m\} \) is a set of items, the traditional association rule is an expression as \( X \Rightarrow Y \), in which \( X \subseteq Z \), \( Y \subseteq Z \), and \( X \cap Y = \emptyset \). The support of each rule represents the probability that \( X \) and \( Y \) simultaneously occur in item \( Z \). The confidence of each rule represents the probability that \( Y \) occurs in the items where \( X \) has been included. After a series of tests, the threshold of support and confidence values are set as 0.02 and 0.75 respectively in this paper. Furthermore, lift is also an important parameter which should be considered. The lift of each rule represents the ratio of its confidence and the probability that \( Y \) occurs. When the value of lift is equal to 1, it means that there exists no correlation between \( X \) and \( Y \). Thus, the rule can be kept only when its lift value is higher than 1. The computational formulas of these three parameters are respectively shown in equations 1 to 3, in which \( \text{Number}(X \cap Y) \) represents the number of itemsets containing \( X \) and \( Y \) and \( \text{Number}(Z) \) represents the number of all itemsets [21].

\[
\text{Support}(X \Rightarrow Y) = \frac{\text{Number}(X \cap Y)}{\text{Number}(Z)} \quad (1)
\]

\[
\text{Confidence}(X \Rightarrow Y) = \frac{P(Y|X)}{P(X)} = \frac{\text{Number}(X \cap Y)}{P(X)} \quad (2)
\]

\[
\text{Lift}(X \Rightarrow Y) = \frac{\text{Confidence}(X \Rightarrow Y)}{P(Y)} = \frac{P(Y|X)}{P(Y)} \quad (3)
\]

4. Statistical Analysis and Experimental Results

In this section, a descriptive statistical analysis is firstly given to show the basic information about the processed data set. And then, the experiments results are elaborated to reveal the potential relationship between the selected fresh graduates’ career choices and their academic performance.

4.1. Descriptive Statistical Analysis

In order to better know the used data, the descriptive statistical analysis concerning the career choices and course grades is conducted in this sub-section. As illustrated in Figure 1, students graduated with four different majors have a wide career scope, but most students in each major still choose to work in information industry, which is in accordance with their general academic background. In terms of the remaining two career choices, students majoring in Information Management and Information System tend to work in other industries like public administration, social work and so on. On the other hand, students majoring in E-Commerce have a
higher possibility to work in finance or business than other majors, partly because they have more access to the courses related to this area. Since different learning experiences have effect on students’ career choices to some extent [22].

The overall level of students’ grades in 15 selected courses is measured in quartiles and mean. These two methods usually reflect the central tendency with different features. Table 3 gives the transformed results of the descriptive statistics and the records with all grades kept for each course. The first quartile shows that at least 25% students get poor grades respectively in Database Principle and Analysis, Managerial Operation Research, Mathematics and Accountancy. The values of the median and mean are the same here and the levels are either “Middle” or “Good”. As for the third quartile, in all courses except College English, Database Principle and Analysis, and Accountancy, the students get grades no lower than the level “Good”. It is obvious that students are not good at these three courses relative to the other ones. Thus, improving their academic performance on these courses is of great necessity, especially College English, because English ability has been a key factor for students’ future development now.

4.2. Experimental Results

Figure 2 is the scatter plot that depicts the association rules of students’ career choices. The value of support and confidence are represented by the X axis and Y axis respectively. The points in this figure represent the association rules and the colors correspond to different values of lift. As illustrated in Figure 2, the support values of most association rules are lower than 0.1 and the confidence values of most association rules are higher than 0.75. However, the lift values are mostly between 1.5 and 4, which indicates that the students’ academic performance do have great effect on their career choices. This finding is also supported by Uyar et al. [23], in which academic performance is proven to influence whether students choose to work in accounting.

**Figure 1.** The distribution of career choices for students with four different majors.

**Table 3.** The descriptive statistical results of all courses.

| Course                          | Number | First Quartile | Median | Mean  | Third Quartile |
|--------------------------------|--------|----------------|--------|-------|----------------|
| College English                | 205    | Middle         | Middle | Middle| Middle         |
| M-Commerce                     | 111    | Middle         | Good   | Good  | Good           |
| E-Government                   | 122    | Good           | Good   | Good  | Excellent      |
| Multimedia Technology          | 175    | Middle         | Good   | Good  | Good           |
| Database Curriculum Design     | 205    | Good           | Good   | Good  | Excellent      |
| Database Principle and Analysis| 125    | Pass           | Middle | Middle| Middle         |
| Business Data Analysis         | 88     | Middle         | Good   | Middle| Excellent      |
| Enterprise Resource Planning   | 131    | Good           | Good   | Good  | Good           |
| Managerial Operation Research  | 111    | Pass           | Middle | Middle| Good           |
| Mathematics                    | 125    | Pass           | Middle | Middle| Good           |
| Accountancy                    | 158    | Pass           | Middle | Middle| Middle         |
| Marketing                      | 125    | Middle         | Good   | Good  | Good           |
| Mobile Development Technology  | 140    | Good           | Good   | Good  | Excellent      |
| Management                     | 124    | Middle         | Good   | Good  | Good           |
| Economic Principle             | 157    | Middle         | Middle | Middle| Good           |
| Computer Networks              | 140    | Middle         | Good   | Good  | Good           |

4.2. Experimental Results

Figure 2 is the scatter plot that depicts the association rules of students’ career choices. The value of support and confidence are represented by the X axis and Y axis respectively. The points in this figure represent the association rules and the colors correspond to different values of lift. As illustrated in Figure 2, the support values of most association rules are lower than 0.1 and the confidence values of most association rules are higher than 0.75. However, the lift values are mostly between 1.5 and 4, which indicates that the students’ academic performance do have great effect on their career choices. This finding is also supported by Uyar et al. [23], in which academic performance is proven to influence whether students choose to work in accounting.
Figure 2. Rules of students’ career choices.

Figure 3 is the connection graph of the association rules of students’ career choices. The size of each circle represents the support value of the rule, and the bigger the circle is, the higher the support value is. Meanwhile, the color of each circle represents the lift value of the rule, and the darker the circle is, the higher the lift value is. Owing to the extremely large quantity of the association rules, six rules for each career choice are selected as examples. As illustrated in Figure 3, most rules of students working in information industry have higher support values but lower lift values. On the contrary, the rules of students working in finance or business almost have highest lift values but lowest support values. This may be because most students choose to work in information industry rather than finance or business.

Table 4 gives the partial results of association rules between students’ career choices and academic performance. Take several rules listed as examples, rule 1 shows that male students who get level “Good” in Enterprise Resource Planning, Mobile Development Technology and Computer Networks but level “Middle” in College English have a greater possibility to work in information industries. Sheer interest in one industry tends to have a positive effect on students’ career choices. It has been proven by Edwards and Quinter [22] and Lent et al. [24], that students’ interest is the main motives for their career choices. Thus, the phenomenon revealed by rule 1 can be largely explained that those students are likely to have more interest in information technology and more enthusiasm to learn the related courses well, so that they can further work in information industries more smoothly. Rule 3 shows that female students who get level “Good” in Database Curriculum Design and Mobile Development Technology tend to work in finance or business. Employability skills play a key role in determining students’ career choice. Having a good command of the skills which are exactly sought by employers offers most students bigger chance to be employed [25-26]. From this perspective, this phenomenon can be explained that these two courses make female students equipped with the employability skills to work in financial or business organizations, such as the ability to develop the systems or APPs that are often necessary by many of those organizations against this era of big data. Rule 5 shows that students who get level “Good” in GPA, Mobile Development Technology and Multimedia Technology, and even level “Excellent” in E-Government are more likely to work in other industries. Seeking opportunities to show mastered skills and knowledge is also an influential factor for students’ career choice [22]. The phenomenon revealed by this rule can largely be explained that most of these courses have a wider applicability in other industries like public administration or entertainment, so the students with good GPA want to give full play to their
advantages and further win a promising promotion prospect for themselves if being employed in these industries. In addition, getting a “Good” level in Mobile Development Technology appears in each career choice, which can be reflected in rules 1, 3, 5 and 6. This indicates that Mobile Development Technology represents a relatively practical skill for most students.

| Number | Rules                                                                 | Support  | Confidence | Lift    |
|--------|-----------------------------------------------------------------------|----------|------------|---------|
| 1      | {Gender = Male, College English = Middle, Enterprise Resource Planning = Good, Mobile Development Technology = Good, Computer Networks = Good} => {Career Choice = Work in Information Industry} | 0.024390 | 1          | 2.070707 |
| 2      | (Managerial Operation Research = Pass, Accountancy = Middle, Economic Principle = Middle) => {Career Choice = Work in Information Industry} | 0.048780 | 0.909091  | 1.882461 |
| 3      | {Gender = Female, GPA = Middle, Database Curriculum Design = Good, Mobile Development Technology = Good} => {Career Choice = Work in Finance or Business} | 0.029268 | 0.857143  | 4.086379 |
| 4      | {Gender = Male, M-Commerce = Good, Marketing = Good, Management = Good} => {Career Choice = Work in Finance or Business} | 0.024390 | 0.833333  | 3.972868 |
| 5      | {GPA = Good, Mobile Development Technology = Good, E-Government = Excellent, Multimedia Technology = Good} => {Career Choice = Work in Other Industries} | 0.024390 | 1          | 3.253968 |
| 6      | {Gender = Female, Enterprise Resource Planning = Good, Mobile Development Technology = Good, Management = Good} => {Career Choice = Work in Other Industries} | 0.039024 | 0.888889  | 2.892416 |

5. Conclusion

Nowadays, solving the problem of employment for fresh graduates has become rather hot-debated. In this paper, association rule mining and Apriori algorithm are applied to explore the relationship between students’ career choices and academic performance. Firstly, the raw information of 228 students graduated from the School of Information of Zhejiang University of Finance and Economics in 2017 is collected. Secondly, the raw data is preprocessed and transformed to meet the requirements for the further study. Thirdly, a data set containing 205 students’ career choices and their academic performance is used to conduct the descriptive statistical analysis and association rule mining. The experimental results indicate that the selection among three career choices is largely influenced by students’ grades in different courses, but Mobile Development Technology represents a relatively practical skill for most students owing to its applicability to each career choice. The rules mined can help students learn to understand their future career objectives better based on their academic characteristics. On the other hand, the rules can also help teachers to advise students on their career planning.

However, there still exist some limitations that need to be addressed in future work. For instance, the course information is dispersive that only 15 courses are kept. The limited data set influences the accuracy of experimental results. Moreover, some other data mining technologies or heuristic algorithms can be combined with association rule mining to improve the efficiency and accuracy of the experimental results.

Acknowledgements

This work was supported by a grant from Industry-university Cooperation and Cooperative Education Project of the Ministry of Education, China (No.201802170003, No.201701028082).

References

[1] Gao, L. (2015). Analysis of employment data mining for university student based on WEKA platform. Journal of Applied Science and Engineering Innovation, 2 (4), 130-133.

[2] Mishra, T., Kumar, D., & Gupta, S. (2016). Students’ employability prediction model through data mining. International Journal of Applied Engineering Research, 11 (4), 2275-2282.

[3] González-Romá, V., Gamboa, J. P., & Peiró, J. M. (2018). University graduates’ employability, employment status, and job quality. Journal of Career Development, 45 (2), 132-149.

[4] Chen, M. S., Han, J., & Yu, P. S. (1996). Data mining: An overview from a database perspective. IEEE Transactions on Knowledge and Data Engineering, 8 (6), 866-883.

[5] Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications, 33 (1), 135-146.

[6] Solanki, S. K., & Patel, J. T. (2015). A survey on association rule mining. In Proceedings of the 5th International Conference on Advanced Computing & Communication Technologies, February 21-22, Haryana, India, pp. 212-216.

[7] Ho, G. T., Ip, W. H., Wu, C. H., & Tse, Y. K. (2012). Using a fuzzy association rule mining approach to identify the financial data association. Expert Systems with Applications, 39 (10), 9054-9063.

[8] Kumar, S., & Toshniwal, D. (2016). A data mining approach to characterize road accident locations. Journal of Modern Transportation, 24 (1), 62-72.

[9] Mane, R. V., & Ghorapade, V. R. (2016). Predicting student admission decisions by association rule mining with pattern growth approach. In Proceedings of International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques, December 9-10, Mysuru, India, pp. 202-207.
[10] Nahar, J., Imam, T., Tickle, K. S., & Chen, Y. P. P. (2013). Association rule mining to detect factors which contribute to heart disease in males and females. Expert Systems with Applications, 40 (4), 1086-1093.

[11] Shi, F., Sun, S., & Xu, J. (2012). Employing rough sets and association rule mining in KANSEI knowledge extraction. Information Sciences, 196, 118-128.

[12] Tyagi, S., & Bharadwaj, K. K. (2013). Enhancing collaborative filtering recommendations by utilizing multi-objective particle swarm optimization embedded association rule mining. Swarm and Evolutionary Computation, 13, 1-12.

[13] Wang, J., Li, H., Huang, J., & Su, C. (2016). Association rules mining based analysis of consequential alarm sequences in chemical processes. Journal of Loss Prevention in the Process Industries, 41, 178-185.

[14] Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In Proceedings of the ACM SIGMOD International Conference on Management of Data, May 26-28, Washington D. C., USA, pp. 207-216.

[15] D’Angelo, G., Rampone, S., & Palmieri, F. (2017). Developing a trust model for pervasive computing based on Apriori association rules learning and Bayesian classification. Soft Computing, 21 (21), 6297-6315.

[16] Deng, X., Zeng, D., & Shen, H. (2018). Causation analysis model: Based on AHP and hybrid Apriori-Genetic algorithm. Journal of Intelligent & Fuzzy Systems, 35 (1), 767-778.

[17] Guo, Y., Wang, M., & Li, X. (2017). Application of an improved Apriori algorithm in a mobile e-commerce recommendation system. Industrial Management & Data Systems, 117 (2), 287-303.

[18] Ilayaraja, M., & Meyyappan, T. (2013). Mining medical data to identify frequent diseases using Apriori algorithm. In Proceedings of International Conference on Pattern Recognition, Informatics and Mobile Engineering, February 21-22, Salem, India, pp. 194-199.

[19] Nair, J. J., & Thomas, S. (2017). Improvised Apriori with frequent subgraph tree for extracting frequent subgraphs. Journal of Intelligent & Fuzzy Systems, 32 (4), 3209-3219.

[20] Prasanna, S., & Ezhilmaran, D. (2016). Association rule mining using enhanced Apriori with modified GA for stock prediction. International Journal of Data Mining, Modelling and Management, 8 (2), 195-207.

[21] Agrawal, R., Srikant, R. (1994). Fast algorithms for mining association rules. In Proceedings of the 20th VLDB Conference, September 12-15, Santiago, Chile, pp. 487-499.

[22] Edwards, K., & Quinter, M. (2011). Factors influencing students career choices among secondary school students in Kisumu Municipality, Kenya. Journal of Emerging Trends in Educational Research and Policy Studies, 2 (2), 81-87.

[23] Uyar, A., Güngörmüs, A. H., & Kuzey, C. (2011). Factors affecting students’ career choice in accounting: The case of a Turkish university. American Journal of Business Education, 4 (10), 29-38.

[24] Lent, R. W., Brown, S. D., Talleyrand, R., McPartland, E. B., Davis, T., Chopra, S. B., Alexander, M. S., Suthakaran, V., & Chai, C. M. (2002). Career choice barriers, supports, and coping strategies: College students’ experiences. Journal of Vocational Behavior, 60 (1), 61-72.

[25] Wye, C. K., & Lim, Y. M. (2009). Perception differential between employers and undergraduates on the importance of employability skills. International Education Studies, 2 (1), 95-105.

[26] Bridgstock, R. (2009). The graduate attributes we’ve overlooked: Enhancing graduate employability through career management skills. Higher Education Research & Development, 28 (1), 31-44.