Relevance Analysis and Visualization of Students’ Employment and Their Courses Achievement

Yi Chen *, Xiaoling Li, Xianfa Wang, Yang Hu and Ligong Yang*

School of Computer and Information Engineering, Beijing Technology and Business University, Beijing, 100048, China

*Corresponding author’s e-mail: chenyi@th.btbu.edu.cn, yanglg@btbu.edu.cn

Abstract. Exploring the relationship between college students’ employment and their academic performance is of great significance for optimizing talents training programs and curriculum system, assisting students’ academic planning and employment guidance. In this paper, we present an analysis method for exploring correlation between students’ employment and their academic achievements. Firstly, we divide the students into several groups according to the types of their employment. Then we use box plot matrix to express the distribution of compulsory course achievements for assisting the comparative analysis among different groups of students. In addition, we use the TF-IDF method based on distance weighting to calculate the preference of each group of students quantitatively on various elective courses, and use radar chart to display the results for assisting the analysis of the students’ behaviour in course selection. The method is applied to analyse the course achievement and employment data of students majoring in Computer Science at Beijing Technology and Business University, and some characteristics of the course achievement and course selection behaviour of the students in different groups are obtained. The analysis results show that the method is effective.

1. Introduction
The fundamental task of a university is to cultivate talents, and the quality of talent cultivation in a university is mainly determined by the quality of employment and academic performance of its students. Students’ employment and academic performance seem to be independent of each other, but in fact, they are interrelated.

In the past decades, researchers have made extensive research on the factors affecting students’ employment and academic performance based on student data, and have achieved many results [2-4]. However, there are few researches about the relationship between student employment and student achievement. And most of the existing results are based on statistical analysis methods for quantitative and qualitative analysis, which is applicable to a certain scale of sample analysis, it is difficult to achieve the analysis of large-scale data sets [1]. Visualization and visual analysis techniques emerged in recent years provide a new means for analysing large scale of dataset.

In this paper, based on statistics and visualization methods, we will analysis the relationship between students’ employment and academic performance, and explore the distribution of students’ academic achievements and the tendency of course selection behaviour in different employment categories. It can provide reference for optimizing talents training program and curriculum system, assisting students’ academic planning and employment guidance. The contributions of this paper are as follows.
(1) We present a visualization method based on box plot matrix, which can effectively express the distribution of students' compulsory course scores in different employment groups, and help users to compare and analyse the distribution of students' performance in different employment groups.

(2) We also present a TF-IDF-based method for analysing students' preference in selecting courses, which can help teachers effectively analyse the tendency of students' choosing courses in different employment groups.

(3) The above methods are applied to analyse the achievements and employment of the students majoring in computer science in Beijing Technology and Business University. The characteristics of academic performance and selective behaviour of each employment group are obtained.

2. Employment grouping and classification of courses

The student dataset used in this paper includes the employment information and course achievements of 532 undergraduate students majoring in computer science and software engineering at Beijing Technology and Business University. The employment information data includes each student’s employment unit, employment group etc. We divide the students into several groups based on the different type of employment [8]. The course achievements data includes each student’s achievements of all compulsory, selective courses, College English Test Band 4 (CET4) and College English Test Band 6 (CET6) during their four-year study. We divide the grades of the course achievements into three levels: A (80-100), B (60-79), and C (<60); and divide the CET4 and CET6 scores into four levels: A (>=550), B (425-549), C (<425) and D (missing test). Compulsory courses and selective courses are further classified into several classifications respectively according to their contents [3]. The employment grouping and classification of courses are shown in Table 1.

| Group Name  | The details of each group                                      |
|------------|----------------------------------------------------------------|
| Employment groups | Study abroad (going abroad), Domestic study (going to graduate school), IT companies, Research institutes, and Unemployed.        |
| Compulsory Courses | Specialized Courses (Computer Organization, Data Structure, Computer Network, Operating Systems, etc.), English Courses (College English Comprehensive Course, Basic Courses (Advanced Mathematics, Linear Algebra, University Physics, etc.), College English Speaking Course, etc.), Quality Education Courses (Moral Accomplishment, Physical Education, etc.), CET4 and CET6. |
| Selective Courses | Computer courses, Science courses, Engineering courses, Literature courses, Art courses, Physical education courses, Law courses, Economic management courses, Competition courses, Language courses, Engineering practice courses, Liberal arts practice courses, Failing classes (giving up the elective courses). |

3. Analysis and visualization method for the achievements of compulsory courses

3.1. Constructing the distribution matrix of compulsory course achievements

This paper analyses the characteristics of students' abilities in different employment groups by statistics of the grade distribution of their compulsory courses. According to 5 employment groups and 6 compulsory courses classifications, the distribution matrix \( S = (S_{ij})_{5 \times 6} \) of all students' achievement is constructed, in which \( S_{ij} \) denotes the grade distribution matrix of students' achievement in classification j of compulsory courses in i\(^{th}\) employment group, \( S_{ij} \) is shown in formula (1).
\[ S_{ij} = \begin{bmatrix} s_{t1} & s_{t2} & \cdots & s_{tn} \\ s_{p1} & s_{p2} & \cdots & s_{pn} \end{bmatrix} \] (1)

\[ S_{31} = \begin{bmatrix} 50\% & 50\% & 0\% \\ 83\% & 17\% & 0\% \\ 67\% & 33\% & 0\% \\ 67\% & 17\% & 17\% \end{bmatrix} \] (2)

Where \( s_{p1} = \frac{Q_A}{Q}, s_{p2} = \frac{Q_B}{Q}, s_{p3} = \frac{Q_C}{Q} \), \( Q = Q_A + Q_B + Q_C \). \( Q_A, Q_B, Q_C \) indicate the number of A, B and C courses for a student respectively, and \( n \) is the total number of students in this group. Each column of matrix \( S_{ij} \) represents the distribution of A, B and C grades in the employment group.

For example, suppose that we need to construct the achievement distribution matrix of IT enterprise group (i = 3) students' specialized courses (j = 1). There are four students in this group, specialized courses includes six courses, and each student's achievement distribution is shown in Table 2. The matrix \( S_{31} \) is obtained from Table 2, as shown in Formula 2.

Table 2. Number of Achievements of A, B and C in Students' Specialized Courses

|       | \( Q_A \) | \( Q_B \) | \( Q_C \) |
|-------|-----------|-----------|-----------|
| Student1 | 3         | 3         | 0         |
| Student2 | 5         | 1         | 0         |
| Student3 | 4         | 2         | 0         |
| Student4 | 4         | 1         | 1         |

3.2. Visualization of achievement distribution of compulsory courses based on box plot matrix

According to the result of the compulsory course grade processing in section 3.1, matrix \( S_{ij} \) is shown by box plot and pie charts as shown in Figure 1. In box plot, the abscissa represents achievement of the grade, and the ordinate represents the proportion of the grade in a certain course category [7]. The upper-most edge line and the lower-most edge line of the box plot represent the maximum and the minimum boundary. The length of the middle box represents the distribution of the students' grades. The upper, middle and lower lines of the box represent the quartile, median and three-quarters of the students' grades, respectively. The pie chart in the upper right corner shows the overall distribution of different grades in this course classification. The colour mapping of pie chart corresponds to the colour of box plot. In summary, box plot and pie chart shown in Figure 1 can express grade distribution of compulsory courses effectively.

The distribution proportion of students' CET4 and CET6 grade is shown by pie chart. The box plot and pie chart are embedded into the matrix to get the box plot matrix as shown in Figure 2. The box plot
matrix shows the grade distribution of different course classifications of students in the same employment group horizontally and the grade distribution of different employment groups in the same course classification vertically.

4. Analysis and visualization method of tendency to selective courses

4.1. Tendency analysis method based on TF-IDF

Selective courses are chosen by students according to the professional teaching plan [6]. They have a certain degree of autonomy and selectivity [4]. This paper classifies elective courses into 13 classifications, as shown in Table 1. TF-IDF algorithm is a method for text mining. It evaluates the importance of a word in a document set based on distance weight. We apply this method to calculate the tendency of the elective classes of students in different employment groups, and mine the correlation between the employment groups and the elective courses [5].

![Box plot matrix](image1)

Figure 2. Box plot matrix. The first line represents the classifications of courses: specialized courses, English courses, basic courses, and quality education courses; the first column represents different employment groups, namely: domestic graduate school, study abroad, IT companies, research institutes and unemployed. A, B, C and D represent grade of course achievements.

![Radar Chart](image2)

Figure 3. Radar Chart of Label Association Index for Selective Courses. Each dimension of the radar chart represents the classifications of course the student prefers to choose. ‘failure’ represents abandon selective courses.

Based on the TF-IDF method, the tendency index $F$ of each group of students for each classification of elective course is calculated. The specific steps are as follows:

**Step1**: Constructing Selective Course Matrix

According to the specific course selection of students in each employment group, the times of each student chooses different course classifications is counted, and the student course elective matrix $T_e =$
\((t_{uv})_{n \times m}\) is obtained, as shown in equation (3). Where \(e \in (1,5)\) represents five employment groups, \(n\) is the number of students in this group, \(m\) represents the number of selective course classifications, in this data set, \(m = 13\). \(t_{uv}\) refers to the number of selective courses in classification \(v\) by student \(u\).

\[
T_e = s_t \begin{bmatrix}
t_{u1} & \ldots & t_{uv} & \ldots & t_{um} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
t_{u1} & \ldots & t_{uv} & \ldots & t_{nm}
\end{bmatrix} \quad (3)
\]

\[
w_u = \left(\frac{1}{d_u} \sum_{r=1}^{m} \frac{1}{d_r}\right) \quad (4)
\]

\[
t_{uv}' = w_ut_u \quad (5)
\]

\[
f_v = \sum_{u=1}^{n} t_{uv}' / \sum_{v=1}^{m} \sum_{u=1}^{n} t_{uv}' \quad (6)
\]

**Step2:** Calculating Selective Courses Frequency Based on Distance Weight

Get the student's course selection vector \(t_u\) from step 1, \(t_u = [t_{u1}, \ldots, t_{uv}, \ldots, t_{um}]\), \((u = 1, 2, \ldots, n)\), and \(t_{uv}\) denotes the number of elective courses classification \(v\) selected by student \(u\). Then calculating the distance between all the students in the employment group, the distance matrix \(D_e = (d_{ur})_{n \times n}\) is obtained by the Euclidean distance formula, where \(d_{ur}\) is the distance between student \(u\) and student \(r\). The distance matrix \(D\) is used to calculate the average distance between student \(u\) and other students in the employment group: \(\bar{d}_u = \frac{1}{n} \sum_{r=1}^{n} d_{ur}\). The course selection weight \(w_u\) of student \(u\) to the analysis of overall course selection tendency of the employment group is calculated by the formula (4). Then \(W_e = [w_1, w_2, \ldots, w_u, \ldots, w_n]^T\) is the weight vector of students in this employment group. The weighted course classification matrix \(T_e' = (t_{uv}')_{n \times m}\) is obtained by formula (5).

In the employment group, the weighted frequency of course classification \(v\) is \(f_v\), which is calculated by formula (6). When the number of an elective course classification in the employment group is 0, this classification value \(f_v\) is 0

**Step3:** Calculating the Reverse Frequency \(rf\)

By summarizing all the students, the course classification matrix \(T_{all} = (t_{uv})_{h \times m}\) is obtained, in which \(h\) is the number of all the students and \(m\) is the number of elective course classification. In this data set \(h=532, m=13\), and then the inverse frequency index of the classification \(v\) of elective courses is calculated by formula (7).

\[
rf_v = \sum_{u=1}^{n} \sum_{v=1}^{m} t_{uv} / (\sum_{u=1}^{n} t_{uv} + 1) \quad (7)
\]

\[
F_v = f_v \times rf_v \quad (8)
\]

**Step4:** Calculating the Tendency Index of Elective Courses

Formula (8) can be used to obtain the tendency index \(F_v\) of elective course classification \(v\) in the employment group. Using the above method, the tendency index of the students in the group is calculated, that is \(F = [F_1, F_2, \ldots, F_p, \ldots, F_m]\).

For example, suppose we need to calculate the tendency index of elective courses for students in IT enterprise group (\(e=3\)) and there are four students in this group and 13 classifications of elective courses. The times of elective course of the students in this group is shown in Table 3. The elective matrix \(T\) of
this group of students is $T_3 = \begin{bmatrix} 3 & 2 & \ldots & 1 \\ 3 & 1 & \ldots & 2 \\ 1 & 3 & \ldots & 2 \\ 2 & 0 & \ldots & 4 \end{bmatrix}_{4 \times 13}$. According the above steps, the tendency index of all the students in IT enterprise group is $F = \{0.2, 0.1, \ldots, 0.3\}$.

| Table 3 Number of elective course for each classification for IT enterprise group students |
|---------------------------------|----------------|----------------|
|                                 | Computer courses | Science courses | Engineering courses |
| student1                       | 3               | 2              | 1               |
| student2                       | 3               | 1              | 2               |
| student3                       | 1               | 3              | 2               |
| student4                       | 2               | 0              | 4               |

4.2. Radar chart for expressing the tendency of elective courses
Calculate the tendency index $F$ of each employment group for the 13 elective course classifications according to the method described in 4.1, and obtain a matric of $5 \times 13$. The matric is visualized using a radar chart, where each dimension of the radar chart represents an elective course classification, and five lines with different colours show the tendency of elective course classification, as shown in Figure 3.

5. Discussion
The experimental results of this paper shows five employment groups: domestic graduate school, study abroad, IT companies, research institutes and unemployed. The data of students are analysed from selective courses and compulsory courses.

5.1. Grade distribution of compulsory courses
We analysed the grade distribution of compulsory courses in different employment groups. As shown in Figure 2, each row of the matrix can analyse the distribution of grades of different course classifications in the same employment group and the columns of the matrix can compare the distribution of grades of the same course classification in different employment groups. For example, in the first line of the matrix, the pie chart in each square reflects that the grade A of the graduate group has a proportion of 80% or more in the classifications of specification courses, basic courses, and quality education courses. Through each box in the box plot, the analysis of the distribution of A, B and C grades of each course classifications can be clear. For example, in the first column, the ratio of grade A in the first four employment groups are higher, and the ratio of obtaining the C grade is lower.

5.2. Tendency analysis of selective courses classification
Analysing the selective behaviour of students in different employment groups has the following conclusions:

The comprehensive analysis (as shown in Figure 3) result is that there is no obvious tendency for students in various employment groups to choose courses, which indicates that most students do not consider long-term career planning when choosing courses. However, the domestic graduate students have obvious tendency when selecting course, as shown in Figure 3. For example, such students are more inclined to select the competition courses, participating more actively in competition courses during the undergraduate period, to lay a solid foundation for future academic study. At the same time, unemployed students have abandoned elective courses many times. This behaviour reflects the lack of patience in learning process of such students.

6. Conclusion
This paper presents a method of correlation analysis between students' employment and their academic performance to provide suggestions on learning and employment for students in schools.
Acknowledgments
This work is supported by National key R&D program of China (Grant No. 2018YFC1603602), National Natural Science Foundation of China (61972010), and Basic Research Project of the Ministry of Science and Technology (Grant No. 2015FY111200). The authors would like to thank the CISAT2019 conference, which provided the exchange platform for our research.

References
[1] Ren L, Du Y, Ma S, Zhang X, Dai G. (2014) Visual Analytics Towards Big Data [J]. Journal of Software. 25(9):1909-1936.
[2] Bendinelli A. et al. (2012) Visualization of longitudinal student data [J]. Physical Review Special Topics - Physics Education Research, 8(2): 1554-9178
[3] Sheng Y. et al. (2017) Research on University Public Elective Course Investigation and Curriculum Optimization [J]. Journal of Changchun Education Institute, 33(6):1671-6531.
[4] Nonis S. et al. (2006) Academic Performance of College Students: Influence of Time Spent Studying and Working [J]. Journal of Education for Business, 81(3):151-159.
[5] Yahav I, et al. (2018) Comments Mining With TF-IDF: The Inherent Bias and Its Removal [J]. IEEE Transactions on Knowledge and Data Engineering, 31(3):437-450.
[6] Chen Y, et al. (2002) Design and Implementation of Priority-based Course Scheduling Algorithm [J]. Journal of Beijing Technology and Business University (Natural Science Edition), 20(2):32-35.
[7] Chen Y, Du X, Yuan X. (2017) Ordered small multiple treemaps for visualizing time-varying hierarchical pesticide residue data [J]. The Visual Computer, 33(6-8):1073-1084.
[8] Chen Y, Zhang X, Feng Y, et al. (2015) Sunburst with ordered nodes based on hierarchical clustering: a visual analyzing method for associated hierarchical pesticide residue data [J]. Journal of Visualization, 18(2):237-254.