



A Prediction and Visual Analysis Method for Graduation Destination of Undergraduates Based on LambdaMART Model

Yi Chen, Beijing Technology and Business University, China*

 <https://orcid.org/0000-0002-4141-0554>

Xiaoran Sun, Beijing Technology and Business University, China

 <https://orcid.org/0000-0003-0157-1975>

Wenqiang Wei, Beijing Technology and Business University, China

Yu Dong, University of Technology Sydney, Australia

Christy Jie Liang, University of Technology Sydney, Australia

ABSTRACT

Predicting graduation destination can help students determine their learning goals in advance, help faculty optimize curriculum and provide career guidance for students. In this paper, the authors first propose a prediction algorithm for graduation destination of undergraduates based on LambdaMART, called PGDU_LM, which uses Spearman correlation coefficient to analyze the correlation between subjects and graduate destinations and extract characteristic subjects, and uses LambdaMART ranking model to calculate students' propensity scores in different graduate destinations. Second, a visual analysis method for students' course grades and graduation destinations is designed to support users to analyze student data from multiple dimensions. Finally, a prediction and visual analysis system for graduation destination of undergraduates, PGDUvis, is designed and implemented. A case study and user evaluation on this system was conducted using the academic data of students from five majors who graduated from a university during 2016-2020, and the results illustrate the effectiveness of this method.

KEYWORDS

Academic Data Set, Course Grades, Graduate Destination, LambdaMART Ranking Model, PGDU_LM, PGDUvis, Prediction, Spearman Correlation Coefficient, Undergraduates, Visual Analysis

INTRODUCTION

Graduation is an important turning point in a student's life. After graduation, one typically has five destinations, including domestic graduate school, overseas study, employment, freelance work, and unemployment. The likelihood of embarking on a particular destination is influenced by various factors that include gender, birthplace, academic performance, and personal strengths (Zhang et al., 2021). Using

DOI: 10.4018/IJICTE.315010

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

an individual's circumstances to predict a student's graduation destinations is important; for students, it helps them clarify their future direction and make reasonable study plans in advance; for teachers, they can optimize their curriculum and provide career guidance to students (Peng, 2020; Liu, 2021).

Generally, the length of undergraduate study is four years, or eight semesters, and about 60 subjects are required. Each subject has credit requirements and assessment results. Each undergraduate accumulates a large amount of data from enrollment to graduation, which includes basic information (including gender, birthdate, birthplace, political affiliation, and major), course grades, graduation destinations (including destination type and industry type), etc. This information constitutes a huge undergraduate academic data set. How to efficiently and deeply analyze these data, find out the distribution characteristics of course grades and graduation destinations, explore the factors that affect students' graduation destinations, and then accurately predict students' graduation destinations is quite challenging.

Most of the existing analysis methods are based on questionnaires, statistical analysis and other methods. Although better results have been achieved, it is difficult to analyze the data in depth in all aspects (Tawafak et al., 2019). The visual analytic techniques that have been developed in recent years map complex data into easily perceivable graphical, symbolic, color-coded and other representations. They are also supplemented with interactive means to enhance people's ability and efficiency in analyzing data so as to quickly discover the features and patterns hidden within the data and provide new ideas for intuitive in-depth analysis of academic data (Alkhalil et al., 2021).

Most existing data mining methods (such as association rule mining, decision trees and classification) and machine learning methods (such as logistic regression) have been used for undergraduate graduation prediction (Gulzar et al., 2019; Jaiswal et al., 2019). This has achieved better results, but the prediction accuracy still needs to be improved. A new machine learning model called LambdaMART that was proposed in recent years can solve the ranking problem directly and effectively without performing classification or regression. This provides a new solution for accurately predicting the graduation destination of undergraduate students.

In this paper, we propose a prediction and visual analysis method for graduation destination of undergraduates based on LambdaMART, called PGDU_LM, and a system called PGDUvis that helps students and teachers deeply analyze the correlation between each subject and the graduation destination, as well as predict the graduation destination based on course grades. The main contributions of this paper can be summarized as follows:

1. A prediction algorithm for the graduation destination of an undergraduate, called PGDU_LM, is proposed; the algorithm uses Spearman correlation coefficient to analyze the correlation between subjects and graduation destinations, extracts characteristic subjects, and uses the LambdaMART ranking model to calculate students' propensity scores in different graduation destinations and then predicts their graduation destinations. The prediction accuracy can reach 86%.
2. A visual analysis method of the course grades and graduation destinations of students is designed. It supports users to analyze student data from multiple dimensions, such as students' course grades, birthplace, graduation destination, industry type, gender, etc., and explore the factors that influence the graduation destination of students.
3. A prediction and visual analysis system for graduate destination of undergraduates called PGDUvis is designed and implemented. It supports users to interactively analyze the academic data of graduates and predict the graduation destination of undergraduates.

RELATED WORK

Correlation Analysis Method

An important feature of the era of big data is the large volume and high dimensionality of data. How to quickly mine the correlation between data from massive and high-dimensional data is an important

issue (Liang et al., 2016). The correlation coefficient is an important indicator to measure the strength of the correlation between variables. Sahar et al. (2019) studied 291 different online courses offered at Georgia Gwinnett College in the summer of 2019 to analyze the correlation between online forum participation rate and MOOC performance. Xiao et al. (2021) studied full-time college students in Xiasha Higher Education Park, Hangzhou, China, and used a quantitative research method, employing SPSS software to analyze the influence of family factors on college students' career choices in depth. Wang et al. (2020) studied the correlation between college students' personality traits and employment intention choices, using a personality trait scale and career orientation scale to survey college students and analyze the correlation between the results. Song et al. (2018) studied the correlation between learning behavior indicators and course grades in the online learning platform, which was obtained from the database of the Beihang online judging platform and used difference analysis, correlation analysis, and comparison analysis to analyze the experimental data. However, the above studies are all about the correlation between course grades and learning behaviors and the correlation between graduate destination and family background, personality characteristics etc. They do not explore the correlation between subjects and graduation destination. Therefore, this paper uses Spearman's correlation coefficient to analyze the correlation between each subject and graduation destination, because the data used in this paper not only have continuous variables but also categorical variables, and some of the data do not show normal distribution. Spearman's correlation coefficient does not require the data to meet the two characteristics of normal distribution and continuous variables, so Spearman's correlation coefficient is chosen for analysis.

Methodology of Prediction

At present, in the present research on the graduate destination prediction of college graduates, the main methods include some machine learning methods such as association rule mining, decision trees, nearest neighbor classification and clustering. For example, He et al. (2021) proposed a combined decision tree and random forest method to predict employment through five main factors: academic achievement, scholarship, graduation qualification, family status (whether poor or not) and association members. Tang et al. (2017, Aug) used employment information of TCM graduates to study employment-influencing factors based on C4.5 algorithm and further used random forest algorithm to improve employment prediction accuracy. Liu et al. (2016) used a database of basic information from college graduates, such as grades and employment, and used the ID3 decision tree algorithm to identify the main factors influencing the employment destination of graduates. Zhou et al. (2020) analyzed graduate employment data based on C4.5 algorithm and achieved high accuracy and practical prediction through cross-validation. Usta et al. (2021) analyzed a rich set of features used in educational search and then used domain knowledge to construct query-related LTR models specifically for certain courses or educational levels. However, most studies have mainly focused on the influencing factors of employment, fast querying course grades, and using other factors to predict employment, and no in-depth research has been conducted to predict graduation destination based on multiple course grades. Therefore, this paper proposes to combine the LambdaMART ranking model with Spearman's correlation coefficient to achieve fast course finding with a strong correlation with graduation destination and fusing multiple course features to predict graduation destination.

Visual Analysis Method

In recent years, visual analytics has become an important method and effective tool for data analysis and scientific decision-making (Schloss et al., 2019), and it has also begun to be applied to the analysis and exploration of educational data (Ji et al., 2021; Vieira et al., 2018). Especially for the analysis of student performance and graduate destination data, Puri et al. (2020) proposed a visual analytics system, RankBooster, to achieve an effective analysis of rankings with a scenario analysis view, showing the situation of different ranking scenarios: a relationship view, visualizing the impact of each attribute on different indicators; and a competition view, comparing the rankings of

universities and their competitors. Gratzl et al. (2013) designed a visual analysis system called Lineup that supports the ranking of top universities at home and abroad by multiple attributes, explores each attribute affecting university ranking by using stacked bar charts, and connects university rankings at different time periods with a line to analyze the change in ranking over time. Chen et al. (2019) proposed an analytical approach to explore the correlation between students' different employment categories and academic performance; this approach utilized a box plot matrix to represent the distribution of compulsory course grades combined with radar plots to show the results of various elective courses and derived some characteristics of course grades and course selection behavior of different groups of students. Ji et al. (2018) proposed an interactive visual analysis system, MVCAS, to display and explore various correlations between general and specialized courses from different levels and perspectives. Most of the above studies have explored correlations between subjects and visual analysis of rankings among different universities. However, studies on visual analytic methods to predict graduation destinations through multiple course grades are relatively rare. In view of this, this paper combines visual analytic methods to provide an in-depth analysis of the correlation between subjects and graduation destinations, information about graduates, and the prediction results of graduation destinations.

BACKGROUND AND RESEARCH PIPELINE

Spearman Correlation Coefficient

Spearman's correlation coefficient is a statistical method used to evaluate the correlation between two variables (Wang, 2017; Zhao et al., 2021; Song et al., 2020), and this paper takes graduate academic data as the sample. The sample size is small, and the joint distribution between variables does not satisfy the normal distribution. Spearman's correlation as a nonparametric statistic has the advantages of low sensitivity to outliers and that the data does not need to satisfy the assumption of normality (Li, 2004), which can meet the characteristics of the data in this study. Therefore, Spearman's correlation in this paper is used to represent the correlation between each subject and each graduation destination, as shown in Equation (1):

$$\rho = 1 - \frac{6 \sum_{i=1}^m d_i^2}{m(m^2 - 1)} \quad (1)$$

The closer the correlation coefficient is to 1 or -1, the more positive or negative the relationship between them. The total number of student records is represented by m , and d_i is the difference between the rankings of the two variables X_i and Y_i in the i -th student record. The number 6 indicates a parameter (Wang, 1997). The calculation process is as follows: X_i (the grade of a course in the i -th student record); Y_i (a graduate destination in the i -th record); sorting X_i in m records; count the number of five graduate destinations in m student records; sorting Y_i by the number of different students in each of the five graduation destinations; record the result after sorting Y_i as the current position in students rank list of a graduate destination, and then calculate the difference between X_i and Y_i ; and finally, the correlation coefficients between each subject and each graduation destination are obtained.

LambdaMART Model Description

The LambdaMART model is a ranking learning model that consists of two parts: LambdaRank and MART (Wu et al., 2010; Burges, 2010). MART is an iterative decision tree model (Hexin et al.,

2021), which consists of multiple decision trees, and the conclusions of all trees are accumulated to make the final result. The Lambda in LambdaRank is the gradient calculated in the MART model and represents the direction and strength for the next iteration of sorting optimization (Wu et al., 2016). The specific steps of the model are as follows:

LambdaMART Model

Variable definitions in the model: Number of regression trees N , number of training samples m , number of leaves per regression tree L , learning rate η , i and h indicates each sample, j indicates each graduate destination.

```

1:  for j=1 to 5 do                                // Each sample is cycled five
                                                times through the five
                                                graduate destinations to
                                                obtain the propensity score
                                                for each graduate
                                                destination. (1: employment,
                                                2: master's degree, 3:
                                                abroad, 4: freelance, 5:
                                                unemployed)

2:      for i=1 to m do
3:           $F_0(x_{ij})=0$                         // Initialize the function  $F_0(x_{ij})=0$ .
4:      end for
5:      for k=1 to N do                            // Iteratively generate N
                                                regression trees.

6:          for i=1 to m do
7:              for h=i+1 to m do
8:                   $\lambda_{ij} = \sum_{(label(i)>label(h))} \lambda_{ihj} - \sum_{(label(i)<label(h))} \lambda_{ihj}$  // The training of each tree
                                                iterates through all the
                                                training data, calculates
                                                the  $\lambda_{ihj}$  for each sample pair
                                                after swapping positions, and
                                                then calculates the  $\lambda_{ij}$  for
                                                each sample (Qu et al., 2019).
                                                 $label(i)$  denotes the value of
                                                the  $i$ -th sample annotation.

9:              end for
10:              $w_{ij} = \frac{\partial \lambda_{ij}}{\partial F_{k-1}(x_{ij})}$  // Calculate the derivative  $w_{ij}$ 
                                                for each  $\lambda_{ij}$ .

11:          end for
12:           $\{R_{lk}\}_{l=1}^L$  // Divide the tree nodes with a
                                                minimum mean square error
                                                and create a regression tree
                                                 $R_{lk}$  with the number of leaf
                                                nodes  $L$ .

```

```

13:      
$$\gamma_{lk} = \frac{\sum_{x_{ij} \in R_{lk}} \lambda_{ij}}{\sum_{x_{ij} \in R_{lk}} w_{ij}}$$
 // Newton's method calculates
                                                the predicted value of each
                                                leaf node in the regression
                                                tree, where  $\gamma_{lk}$  represents
                                                the score of the  $l$ -th leaf
                                                node of the  $k$ -th regression tree.

14:      
$$F_k(x_{ij}) = F_{k-1}(x_{ij}) + \eta \sum_{l=1}^L \gamma_{lk}(x_{ij} \in R_{lk})$$
 // Update the function by
                                                adding the currently learned
                                                regression tree to the
                                                existing model and updating
                                                the original prediction with
                                                the learning rate  $\eta$ 
                                                (Jepkoech et al., 2021).

15:      
$$Lscore\_ij = F_k(x_{ij})$$
 // Each sample corresponds to
                                                the propensity score of five
                                                graduate destinations, with
                                                values ranging from  $[-1,1]$ 
                                                where positive and negative
                                                indicate the direction of
                                                increase or decrease of the
                                                sample in the process of ranking.

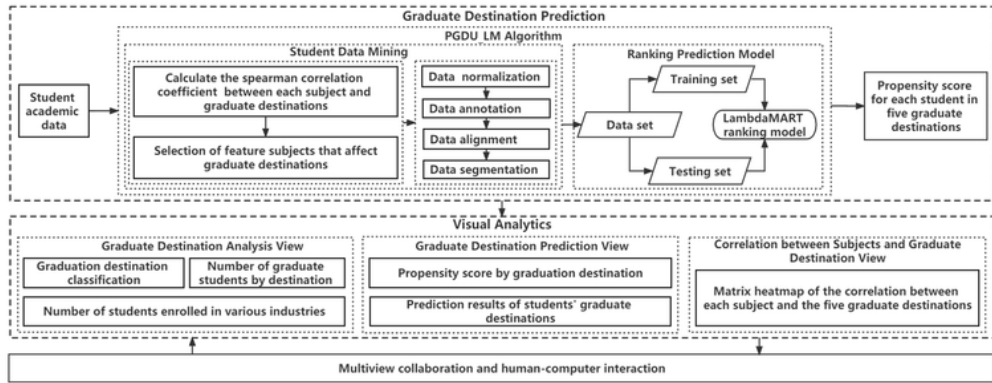
16:      end for
17:      end for
    
```

In this paper, the ranking learning method is applied for the first time to solve the problem of predicting the graduation destination of undergraduates. Specifically, the LambdaMART model is used to calculate students' propensity scores in different graduation destinations to achieve prediction.

The Pipeline of Our Method

The pipeline of our method is shown in Figure 1. First, the academic dataset of students—in which the courses of the five majors of computer science, software engineering, information management, information engineering, and automation add up to about 280 subjects in total—is preprocessed. The names of students are renumbered with 26 letters. Graduation destinations are divided into five aspects after consulting the employment guidance teachers, namely: master's degrees, abroad, employment, freelance, and unemployed, according to the National Economic Classification of Industries (NECI) standard, which classifies a wide range of industry types. Then, this paper uses the PGDU_LM algorithm to calculate the Spearman correlation coefficients between subjects and graduation destinations, extracts subjects related to graduation destinations, constructs students' subjects features, and performs data normalization, annotation division, and format conversion. After that, the LambdaMART ranking model is used to calculate the propensity scores of students' course grades and different graduation destinations, respectively, and the graduation destinations are predicted by the ranking of the scores. Finally, it is designed and implemented an undergraduate graduation destination prediction and visual analysis system, which provides reference to help students and teachers predict students' graduation destinations and analyze the information of graduates and the correlation between each subject and graduation destination through multi-view collaboration and human-computer interaction.

Figure 1. The pipeline of a prediction and visual analysis method for graduation destination of undergraduates based on LambdaMART model



PGDU_LM ALGORITHM

Feature Subjects Construction of Student Graduate Destination

The Spearman correlation coefficients of a total of 280 subjects in five majors and five graduation destinations were calculated separately according to equation (1), and the mean value of the five correlation coefficients of each subject was taken to measure the relevance of the subject to the graduation destination. The 77 subjects with mean values of correlation coefficients between [0.5,1] were extracted to construct student subject characteristics data to demonstrate some of the characteristics. (See Table 1) Figure 4(C) depicts a matrix heatmap of the correlation between each subject and graduation destination.

Table 1. The feature subjects affecting graduate destination (Partial)

Majors	Subject Category	Feature Subjects	Grade range
<ul style="list-style-type: none"> • Computer Science • Software Engineering • Information Management • Information Engineering • Automation 	Required professional subjects	C programming language	Integer [0,100]
		Database principles and applications	
		Software Requirements Engineering	
	Professional elective subjects	WEB system front-end technology	Integer [0,100]
		Automatic identification technology	
		Computer Graphics	
	Public foundation subjects	Higher Mathematics	Integer [0,100]
		University Physics	
		Linear Algebra	
	Practical sessions	Graduation Internship	Integer [0,100]
		Professional Internship	
	English Proficiency	CET4	Integer [0,710]
		CET6	

Data Normalization

The features of each course grade have different magnitudes among them, and this situation affects the results of data analysis. If the features are not normalized, training the model will add more time to find the optimal values. Normalizing each student's grades for each course can speed up the training and also make the final weights controlled within a certain range. In this paper, we will use min-max normalization to linearly transform the students' original course grades and map the data to the [0,1] interval, as shown in Equation (2):

$$Grades_{norm} = \frac{grades - grades_{min}}{grades_{max} - grades_{min}} \quad (2)$$

where $grades_{max}$ and $grades_{min}$ denote the maximum and minimum values of all students' grades in a course, respectively, $grades$ mean the original value of students' grades in a course, and $Grades_{norm}$ means the normalized value of students' grades in a course.

Data Annotation

In this paper, students' grade point average (GPA) is divided into five bands: 4.00~5.00, 3.00~4.00, 2.00~3.00, 1.00~2.00, and 0.00~1.00. The number of students in each band in each graduate destination is counted, and the degree of correlation between the band in which the student's GPA is located and the different graduate destinations is marked according to the number of students. There are five levels of correlation: 0, 1, 2, 3, and 4. The higher the level, the more relevant a certain score is to a certain graduate destination. For example, the number of students with a GPA between 4.00 and 5.00 is fifteen; between 3.00 and 4.00 is twenty-five; and between 2.00 and 3.00 is thirty for a major whose graduate destination is employment. Therefore, students with GPAs between 2.00 and 3.00 are labeled with a grade of 4 for the degree of relevance to employment, between 3.00 and 4.00 with a grade of 3, and so on.

Data Format

The student academic data is converted into a format commonly used in ranking learning models to facilitate the training and testing of the model. The specific format is as follows, and the fields are described in Table 2.

Table 2. The definition of each field in student record

Field	Value range	Definition	Note
Label	0,1,2,3,4	Classification in each graduate destination	Marking of the degree of correlation between the grade band of the student's GPA and the different graduate destinations.
id	3-digit integer	Major and graduate destination	For example, 201, indicating that this student is a software engineering major graduate destination for employment.
c[feature subject no.]	Integer [1,77]	The feature subject no.	The numbering of extracted subject features.
[grade]	Floating-point [0,1]	Course grade	Course grade after data normalization.
Name	Character string	Student name	Set as a student name.

Line: [Line No.], Label: [label], id: [abc], c [feature subject no.]: [grade], Name: [student name]
 Eg. Line:1 Label:4 qid:202 c1:1.000 c2:0.993,, c77:0.356, Name: AAA

Line represents a data record, and each record occupies a line number. Through the analysis, processing, and integration of data, the final data set has 1560 structured data. The id: [abc] of each data is represented by three digits; a corresponds to the major (1: computer science, 2: software engineering, 3: information management, 4: information engineering, 5: automation), b is represented by the number 0, and c indicates the number of graduate destinations (1: employment, 2: master's degree, 3: abroad, 4: freelance, 5: unemployed), and Figure 2 shows the partially formatted data.

LambdaMART Ranking Model for Prediction

The normalized student subject characteristics are used as input to the LambdaMART ranking model, and a scoring function $F(x)$ with different graduation destination propensities through multiple course grades is trained. When new sample data is input, it is scored using the $F(x)$ function, and the propensity score $Lscore_{ij}$ for each student's five graduation destinations is output. The higher the score, the more likely it is that the student will graduate. The specific process is shown in Figure 3, where the student named WQJ has the highest employment propensity score, so the predicted future choice of this student after graduation is employment.

PGDU_LM Algorithm Performance Evaluation

Evaluation Indicator

In this paper, the information retrieval metric $NDCG@T$ is used and mapped to the assessment model (Liu, 2010), where T represents the top T students in the prediction list in alphabetical order by name. $DCG@T$ is calculated as (3):

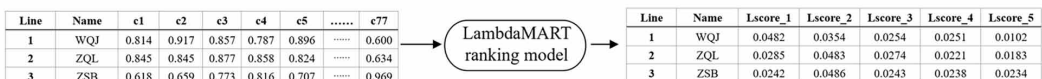
$$DCG@T = \sum_{i=1}^T \frac{2^i - 1}{\log_2(i + 1)} \tag{3}$$

where l_i denotes the label value of each student record, i indicates the location of the current student record. The value of the whole equation is proportional to l_i . Therefore, the higher the label level is, the higher the value of $DCG@T$. The value of DCG after normalization is NDCG:

Figure 2. The formatted student record (Partial)

```
Line:1,Label:3,id:101,c1:0.000,c2:0.814,c3:0.917,c4:0.857,c5:0.000,c6:0.737,c7:0.787,c8:0.896,c9:0.603,.....c75:0.000,c76:0.000,c77:0.000,Name: ABK
Line:2,Label:3,id:102,c1:0.000,c2:0.793,c3:0.927,c4:0.867,c5:0.000,c6:0.808,c7:0.888,c8:0.742,c9:0.571,.....c75:0.000,c76:0.000,c77:0.000,Name: ALD
Line:3,Label:3,id:103,c1:0.000,c2:0.845,c3:0.845,c4:0.877,c5:0.000,c6:0.858,c7:0.898,c8:0.000,c9:0.809,.....c75:0.000,c76:0.000,c77:0.000,Name: AWX
Line:4,Label:2,id:101,c1:0.000,c2:0.773,c3:0.793,c4:0.000,c5:0.000,c6:0.848,c7:0.707,c8:0.824,c9:0.555,.....c75:0.000,c76:0.000,c77:0.000,Name: ABE
Line:5,Label:2,id:101,c1:0.000,c2:0.835,c3:0.927,c4:0.857,c5:0.000, c6:0.606,c7:0.828,c8:0.000,c9:0.634,.....c75:0.000,c76:0.000,c77:0.000,Name: AHY
Line:6,Label:2,id:101,c1:0.000,c2:0.742,c3:0.804,c4:0.816,c5:0.000,c6:0.818,c7:0.838,c8:0.969,c9:0.634,..... c75:0.000,c76:0.000,c77:0.000,Name: AQW
Line:7,Label:2,id:101,c1:0.000,c2:0.835,c3:0.783,c4:0.836,c5:0.000,c6:0.838,c7:0.848,c8:0.876,c9:0.634,.....c75:0.000,c76:0.000,c77:0.000,Name: APO
Line:8,Label:1,id:104,c1:0.000,c2:0.659,c3:0.773,c4:0.704,c5:0.000,c6:0.606,c7:0.606,c8:0.453,c9:0.000,.....c75:0.000,c76:0.000,c77:0.000,Name: AMT
Line:9,Label:1,id:101,c1:0.000,c2:0.618,c3:0.804,c4:0.714,c5:0.000,c6:0.606,c7:0.606,c8:0.752,c9:0.412,.....c75:0.000,c76:0.000,c77:0.000,Name: BXR
Line:10,Label:0,id:105,c1:0.000,c2:0.618,c3:0.671,c4:0.683,c5:0.000,c6:0.646,c7:0.606,c8:0.453,c9:0.412,.....c75:0.000,c76:0.000,c77:0.000,Name: WMY
Line:11,Label:4,id:503,c1:0.000,c2:0.000,c3:0.000,c4:0.000,c5:0.000,c6:0.000,c7:0.948,c8:0.000,c9:0.000,.....c75:0.000,c76:0.000,c77:0.000,Name: ZCT
Line:12,Label:4,id:302,c1:0.000,c2:0.969,c3:0.969,c4:0.929,c5:0.000,c6:0.939,c7:0.000,c8:0.765,c9:0.000,.....c75:0.000,c76:0.000,c77:0.000,Name: ZQL
```

Figure 3. The process of predicting students' graduation destinations based on LambdaMART model



$$NDCG @ T = \frac{DCG @ T}{\max DCG @ T} \quad (4)$$

The $NDCG@T$ in this paper can be interpreted as the quality of the ranking of the top T students in the prediction list according to the prediction scores of different graduate destinations.

Comparison Experiment

In this paper, PGDU_LM is compared with Random Forests algorithm for experiments, using academic data of 936 students as the training set and 624 students as the test set (He et al., 2021). The evaluation results of the NDCG metrics show that the PGDU_LM algorithm is better than the Random Forests algorithm in predicting graduation destinations through multiple course grades. The value of $NDCG@50$ is close to 0.9, indicating that students ranked in the top 50 have higher accuracy in predicting their graduation destination by ranking the propensity scores of each students' five graduation destinations, as shown in Table 3.

PGDUVIS SYSTEM

Design Requirements

Currently, the main tools used for statistical analysis of course grades and graduate destinations are Excel and SPSS (Zhao et al., 2021), which are excellent for recording, querying, and general statistics but still do not meet the needs when the scale of data increases, when multidimensional in-depth analysis of data is performed, and when graduate destinations are predicted. Through research with experts and managers in the field of education, the following main needs have been condensed from two main perspectives: students and teachers:

- R1:** It can statistically analyze the graduate destinations of students of different majors, their birthplace, and the distribution of employment industry types. It helps teachers adjust the enrollment plan and optimize the training program for students.
- R2:** Based on the course grades of existing students in different majors, it predicts students' graduation destinations, helps students plan their future graduation directions, and helps teachers provide targeted career guidance to students.
- R3:** Analyze the correlation between each subject's different majors and graduation destinations and help teachers optimize the training program and adjust the curriculum system according to the degree of influence of each subject on different graduation destinations.

Visual Design

The PGDUvis system consists of three main views: (1) the graduate destination analysis view which includes the graduate destination classification, the number of graduate students by destination and the number of students enrolled in various industries three sub-views, (2) the graduate destination prediction view which includes the propensity score by graduation destination and the prediction

Table 3. The comparison results on $NDCG@Ts$

Algorithm	$NDCG@20$	$NDCG@30$	$NDCG@40$	$NDCG@50$
PGDU_LM	0.80	0.82	0.85	0.86
Random Forests	0.79	0.80	0.77	0.81

results of students' graduate destination two sub-views, and (3) correlation between subjects and graduate destination view. The system interface is shown in Figure 4.

The Graduate Destination Analysis View

The view consists of circle packing and two bar charts, as shown in Figure 4 (A₁). There are four layers in the graduate destination classification view. Each layer shows the birthplace, graduate destination, employment industry type, and student gender. The size of the circle in each layer indicates the number of students, which can be analyzed from multiple aspects and explore the factors affecting graduate destination. (Chen et al., 2022). Figure 4 (A₂) and Figure 4 (A₃) depict the number of students from various countries of origin, as well as the distribution of students in each employment industry type (Nguyen et al., 2018).

The Graduate Destination Prediction View

This view consists of two parts: a table view and a bubble diagram, as shown in Figure 4(B₁). The table view analysis uses the PGDU_LM algorithm to calculate the propensity scores for the five graduation destinations based on students' course grades, denoted as *Lscore_{ij}*. The comparative analysis of the *Lscore_{ij}* scores predicted the students' graduation destination (Blazevic et al., 2021) with seven columns in the table, each row representing one student, the first column representing the student's name, the second column the major, and the remaining five columns the propensity scores for employment, master's degree, abroad, freelance, and unemployed. As shown in Figure

Figure 4. The interface of the PGDUvis system (A) The graduate destination analysis view, showing the graduate destination classification in A₁, including the distribution of graduates' birthplace, graduate destination, employment industry type, and gender in order; the number of graduate students by destination in A₂, and the number of students enrolled in various industries in A₃. (B) The graduate destination prediction view, showing the PGDU_LM algorithm used to calculate propensity scores for five graduation destinations based on students' course grades, denoted as *Lscore_{ij}*, and to analyze students' graduate destinations in B₁, the prediction results of students' graduate destinations in B₂. (C) Correlation between subjects and graduate destination view, showing and analyzing the influence of each subject on graduate destinations



4(B₂), the bubble chart shows more visually the propensity score of each student in each graduation destination and thus analyzes the predicted results of graduation destinations (Yu, 2021). Where each color represents a student, and when the number of students increases, different colors are randomly generated in the view to represent different students. The horizontal coordinates indicate the five graduation destinations, and the vertical coordinates correspond to the student's name. The size of the propensity score in the table view is mapped to the size of the circle in the bubble diagram. If the bubble for a particular graduation destination is larger, it indicates that the student is more likely to choose that graduation destination (Fallon et al., 2019).

The Correlation Between Subjects and Graduate Destination View

This view maps the Spearman correlation coefficients between the extracted subjects and graduation destinations into a matrix heatmap by the PGDU_LM algorithm (Chen et al., 2020), which contains 77 relevant courses in five majors. In Figure 4(C), the correlation between the subjects of each major and the graduation destination is shown separately. The darker the color of the matrix heatmap indicates, the greater the influence of the subject on the graduation destination, and the more relevant it is. The view can analyze the correlation between each subject and the graduation destination in two dimensions: the overall correlation between different graduation destinations and all subjects in a horizontal comparison; and the degree of influence between each subject on different graduation destinations in a vertical analysis.

Interaction Design

The system uses interactive tools, such as filtering, highlighting, and correlation, to help students and instructors explore the factors that affect graduation destinations, predict and analyze the graduation destinations of non-graduating students, and analyze the correlation between each subject and graduation destinations. First, select a major in the system navigation bar "Major" and analyze the information of graduates of that major in the circle packing in figure 4 (A₁) layer by layer. Placing the mouse in any of the circles on a certain level will highlight the information about the birthplace, the five graduation destinations, the types of industries worked in, and the number of students. When a circle is clicked, the view will zoom in and show only the details of the clicked circle. At the same time, in Figure 4(A₂) (A₃), the circle packing are linked with two bar charts to show the distribution of the number of students in the selected region by industry type and by the five graduation destinations.

Secondly, in the "Upload a file" navigation bar in the upper right corner of the system interface, it can upload the course grades CSV file of students who have not graduated from a specific class in a specific major to predict where they will graduate. In the top left corner of the system interface, it is possible to filter and view students' majors in the "Major" navigation bar. In the table view figure 4(B₁), drag the slider to compare and analyze the propensity scores of the five graduation destinations of the uploaded students, and pull the slider of the bubble chart figure 4(B₂) to show the predicted results of the students' graduation destinations visually.

Finally, by selecting a major through the "Major" navigation bar and hovering over the grid of the subject you want to analyze in the matrix heatmap figure 4(C), the hover box will show the full name of the subject and the correlation coefficient with the different graduation destinations. Thus, the correlation between the course and the five graduation destinations will be analyzed.

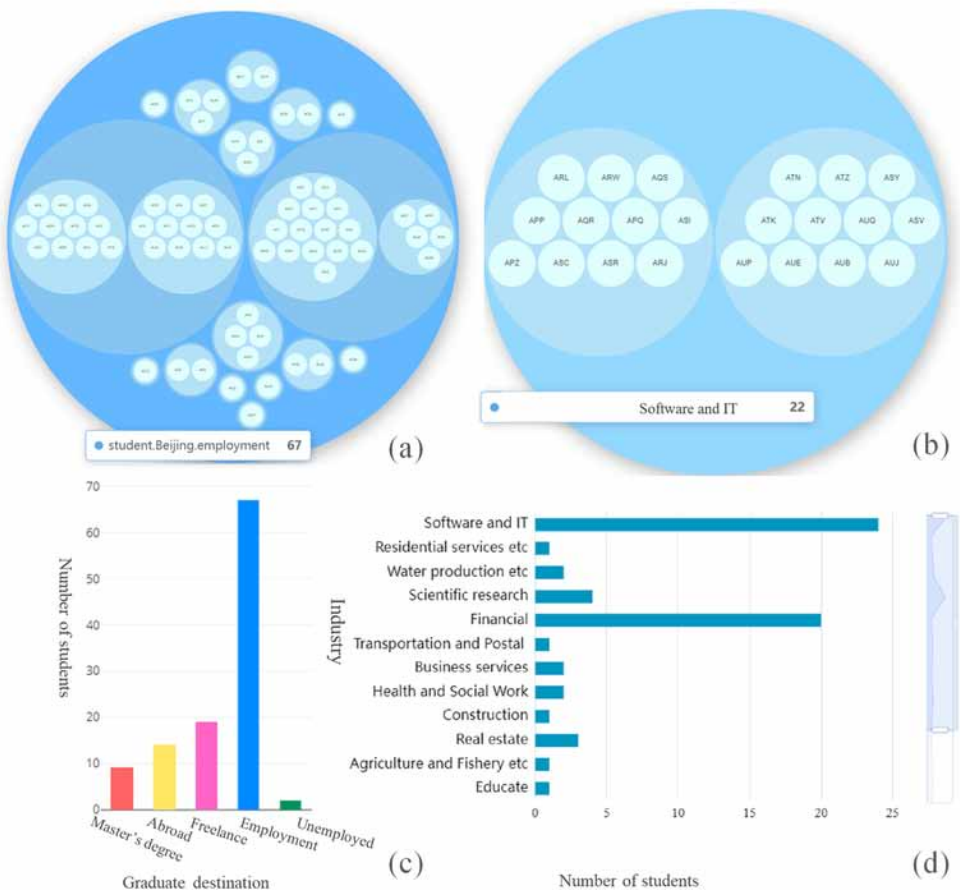
Case Study

Taking the academic data of 1560 graduates of five majors in computer science and technology, software engineering, information management, information engineering, and automation who graduated from a university in the five years from 2016 to 2020, 30 students and 10 faculty members were invited to conduct a case study and user assessment on PGDUvis.

The Analysis of Graduate Destination Information

The users can analyze the information on the graduate destinations of computer science and technology students, as shown in Figure 4 (A₁₁). As shown in Figure 5(a) (c), the number of students in Beijing whose graduation destination is employment is the highest compared with the other three graduation destinations, indicating that there are more employment opportunities in the region and most students prefer to work after graduation. As shown in Figure 5(b)(d), pulling the slider to browse the distribution of the number of students by industry type, students in the region generally choose software and IT services and finance, among which the number of female students choosing the finance industry is greater than the number of male students, which indicates that the industry is more suitable for female students. The two industries currently have good employment prospects for students in this major and are attractive, providing a basis for teachers to help students who have not yet graduated to understand job requirements in advance, and teachers can consider recruiting more students from the Beijing area in the subsequent enrollment plan (R1).

Figure 5. The graduate destinations analysis of students from Beijing. (a) Showing the number of students in employment, (b) Displaying the number of students in the Software and IT service industry, (c) Showing the number of students in five graduation destinations, (d) Displaying the number of students choosing each industry type



Finally, the users can also analyze the correlation between the graduation destinations of students in this major and CET6, as shown in Figure 7. The color shade analysis of the matrix heatmap shows that CET6 has a stronger correlation with a master’s degree and going abroad than the other two graduation destinations. Therefore, students who want to go abroad or get a master’s degree need to be proficient in English, and teachers need to strengthen the cultivation of students’ English ability by increasing English-related subjects, such as speaking practice, writing practice, etc. (R3).

System Evaluation

To understand the effectiveness of the system, the 40 users were invited to evaluate the system based on several analytical tasks. The users were required to complete the listed tasks separately and score the corresponding visualization views out of 10. The mean value of the scores of the 40 users was taken as the assessment result of subjective satisfaction, as shown in Table 4. Satisfaction with the graduate destination analysis view and the graduate destination prediction view is generally high. The users believe that the system can provide detailed analysis of the distribution of graduates’ graduate destinations, birthplace, and the distribution of employment industry types, which can help teachers adjust their enrollment plans and predict graduate destinations through students’ comprehensive scores of multiple subjects, which can provide meaningful employment guidance for students to plan their future graduate directions. However, compared with other views, the satisfaction level of the correlation between subjects and the graduate destination view is lower. The users suggested that the design of this view is rather simple and hoped to increase the innovation in the view design. In conclusion, the 40 users were satisfied with the overall design of the system, demonstrating its effectiveness. Due to the small number of users, there are some limitations in the system assessment. In the future, the validity of the system can be verified by increasing the number of students and teachers from different colleges and majors to conduct large-scale assessments.

CONCLUSION

In this paper, we first propose a prediction algorithm based on the LambdaMART model, PGDU_LM, which supports the prediction of students’ graduation destination according to their course grades. The

Table 4. Survey results of user satisfaction on each view

Num	Questions	Views	Satisfaction Score
1	The number of students choosing automation to see the number of students going abroad and for a master’s degree.	Graduate Destination Analysis View	8.84
2	An analysis of the distribution of students employed in software engineering who choose different types of industries.		
3	Selecting computer science students to analyze the predicted results of the students’ graduation destinations.	Graduate Destination Prediction View	8.62
4	Uploading the course grades of a junior student and predicting graduate destinations.		
5	Selecting the major ‘software engineering’ and analyzing the correlation between the subject ‘C programming’ and the graduate destination ‘master’s degree’.	Correlation between Subjects and Graduate Destination View	7.65
6	Selecting the major ‘information management’ and analyzing the correlation between the subject ‘data structure’ and the graduate destination ‘employment’.		

prediction accuracy can reach 86%. Then, a visual analysis method is designed for students' course performance and graduation destination, which supports users to analyze factors affecting graduation destination from multiple dimensions, such as birthplace, industry type, academic performance, and student gender, and explore the correlation between various subjects and graduation destination. Finally, a graduate destination prediction and visual analysis system PGDUvis is designed and implemented, which provides efficient and convenient visual analysis tools for students to plan their future careers and make learning plans, for teachers to adjust training programs, optimize curriculum settings, and provide career guidance to students.

However, there are still some limitations in this method. First, because of the complexity of the factors that affect the graduation destination of students, only the main factors such as gender, birthplace and course performance are considered at present. In the future, personal family background, economic conditions, social situation and other factors will also be included. Second, although the accuracy of the PGDU_LM algorithm is high, the time complexity of the algorithm's calculation is also relatively high, and we will further optimize the PGDU_LM algorithm in the future to thereby improving the calculation speed.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (No. 61972010) and the 2022 Postgraduate Research Capability Improvement Program Project.

COMPETING INTERESTS

All authors of this article declare there are no competing interest.

FUNDING AGENCY

This research was supported by the National Natural Science Foundation of China [No. 61972010].

REFERENCES

- Alkhalil, A., Abdallah, M. A. E., Alogali, A., & Aljaloud, A. (2021). Applying big data analytics in higher education: A systematic mapping study. *International Journal of Information and Communication Technology Education*, 30(1), 79–89. doi:10.4018/IJICTE.20210701.oa3
- Blazevic, M., Sina, L. B., Burkhardt, D., Siegel, M., & Nazemi, K. (2021). Visual analytics and similarity search-interest-based similarity search in scientific data. *Proceedings of the International Conference on Information Visualization*, 211-217. doi:10.1109/IV53921.2021.00041
- Burges, C. J. C. (2010). From ranknet to lambdarank to lambdamart: An overview. *Learning*. <https://www.researchgate.net/publication/228936665>
- Chen, Y., Li, X., Wang, X., Hu, Y., & Yang, L. (2019). Relevance analysis and visualization of students' employment and their courses achievement. *Journal of Physics: Conference Series*, 1345(2), 022033. Advance online publication. doi:10.1088/1742-6596/1345/2/022033
- Chen, Y., Lv, C., Li, Y., Chen, W., & Ma, K.-L. (2020). Ordered matrix representation supporting the visual analysis of associated data. *Science China. Information Sciences*, 63(8), 236–238. doi:10.1007/s11432-019-2647-3
- Chen, Y., Zhang, Q., Guan, Z., Zhao, Y., & Chen, W. (2022). GEMvis: A visual analysis method for the comparison and refinement of graph embedding models. *The Visual Computer*, 38(9-10), 3449–3462. doi:10.1007/s00371-022-02548-5
- Fallon, J., & Crouse, R. J. (2019). VAMD: Visual analytics for multimodal Data. *2019 IEEE 9th Symposium on Large Data Analysis and Visualization (LDAV)*, 93-94. doi:10.1109/LDAV48142.2019.8944264
- Gratzl, S., Lex, A., Gehlenborg, N., Pfister, H., & Streit, M. (2013). LineUp: Visual analysis of multi-attribute rankings. *IEEE Transactions on Visualization and Computer Graphics*, 19(12), 2277–2286. doi:10.1109/TVCG.2013.173 PMID:24051794
- Gulzar, Z., Raj, L. A., & Leema, A. A. (2019). Ontology supported hybrid recommender system with threshold based nearest neighbourhood approach. *International Journal of Information and Communication Technology Education*, 15(2), 85–107. doi:10.4018/IJICTE.2019040106
- He, Q., Li, X., & Sun, Y. (2021). Company ranking prediction based on network big data. *Journal of the Institution of Electronics and Telecommunication Engineers*, 1–12. Advance online publication. doi:10.1080/03772063.2021.1986144
- He, S., Li, X., & Chen, J. (2021). Application of data mining in predicting college graduates employment. *2021 4th Intelligence Conference on Artificial Intelligence and Big Data (ICAIBD)*, 65-69. doi:10.1109/ICAIBD51990.2021.9459039
- Hexin, L., Yang, X., & Dai, G. (2021). A learning to rank approach for pharmacist assignment. *2021 IEEE 13th International Conference on Computer Research and Development (ICCRD)*, 104-108. doi:10.1109/ICCRD51685.2021.9386461
- Jaiswal, G., Sharma, A., & Yadav, S. K. (2019). Analytical approach for predicting dropouts in higher education. *International Journal of Information and Communication Technology Education*, 15(3), 89–102. doi:10.4018/IJICTE.2019070107
- Jepkoech, J., Mugo, D. M., Kenduiyo, B. K., & Too, E. C. (2021). The effect of adaptive learning rate on the accuracy of neural networks. *International Journal of Advanced Computer Science and Applications*, 12(8), 736–751. doi:10.14569/IJACSA.2021.0120885
- Ji, L., Gao, F., Huang, K., & Chen, Z. (2018). 面向多主体的大学课程成绩相关性可视探索与分析 [Visual exploration and analysis of multi-subject correlation of student performance in college courses]. *Journal of Computer-Aided Design & Computer Graphics*, 30(1), 44–56. doi:10.3724/SP.J.1089.2018.16924
- Ji, L., Yuan, Y., & Gao, F. (2021). Multi-level and multi-perspective visual correlation analysis between general courses and program courses. *The Visual Computer*, 37(3), 477–495. doi:10.1007/s00371-020-01818-4
- Li, W., Xu, S., Zheng, J., & Zhao, G. (2004). Target curvature driven fairing algorithm for planar cubic B-spline curves. *Computer Aided Geometric Design*, 21(5), 499–513. doi:10.1016/j.cagd.2004.03.004

- Liang, J., Feng, C., & Song, P. (2016). 大数据相关分析综述 [A survey on correlation analysis of big data]. *Chinese Journal of Computers*, 39(1), 1–18. doi:10.11897/SP.J.1016.2016.00001
- Liu, H. (2021). 扩招20年中国高校毕业生就业测量与就业变化 [Employment measurement and employment changes of Chinese college graduates in 20 years of higher education enrollment expansion]. *Higher Education Exploration*, 2, 121–128.
- Liu, T. Y. (2010). Learning to rank for information retrieval. *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. doi:10.1145/1835449.1835676
- Liu, Z., & Zhao, Z. G. (2016). 数据挖掘技术在大学生就业分析中的实证研究 [Analysis and calculation of high school graduate student based on data mining]. *Journal of Shenyang Normal University (Natural Science Edition)*, 34, 105–108. doi: j. issn.1673-5862.2016.01.02410.3969
- Nguyen, H., & Rosen, P. (2018). DSPCP: A data scalable approach for identifying relationships in parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 24(3), 1301–1315. doi:10.1109/TVCG.2017.2661309 PMID:28166499
- Peng, Z., Lu, G., & Li, L. (2020). 大学毕业生就业质量的影响因素及路径分析 [Research on graduates employment quality: Influence factors and path analysis]. *China Higher Education Research*, 1, 57–64. doi:10.16298/j.cnki.1004-3667.2020.01.09
- Puri, A., Ku, B. K., Wang, Y., & Qu, H. (2020). RankBooster: Visual analysis of ranking predictions. *Eurovis: Eurographics/IEEE Symposium on Visualization*, 175–179. doi:10.2312/evs.20201068
- Qu, B., Bai, Y., Cai, D., & Chen, J. (2019). 基于LambdaMART算法的微信公众号排序[Ranking WeChat official account based on LambdaMART]. *Journal of Chinese information Processing*, 33(12).
- Schloss, K. B., Gramazio, C. C., Silverman, A. T., Parker, M. L., & Wang, A. S. (2019). Mapping color to meaning in colormap data visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 25(1), 810–819. doi:10.1109/TVCG.2018.2865147 PMID:30188827
- Song, H. Y., & Park, S. (2020). An analysis of correlation between personality and visiting place using Spearman's Rank correlation coefficient. *Transactions on Internet and Information Systems (Seoul)*, 14(5), 1951–1966. doi:10.3837/tiis.2020.05.005
- Song, Y., Yang, J., Fei, X., Ma, X., & Ma, D. (2018). How does college students' online learning behavior impact their academic performance. *13th International Conference on Computer Science & Education (ICCSE)*, 1–5. doi:10.1109/ICCSE.2018.8468726
- Tang, Y., & Wang, P. (2017, August). 基于C4.5和随机森林算法的中医药院校毕业生就业预测应用研究 [Study on employment forecasting of graduates of traditional Chinese medicine based on C4.5 and random forest algorithm]. *China Medical Herald*, 14, 166–169.
- Tawafak, R. M., Romli, A. M., & Alsinani, M. J. (2019). Student assessment feedback effectiveness model for enhancing teaching method and developing academic performance. *International Journal of Information and Communication Technology Education*, 15(3), 75–88. doi:10.4018/IJICTE.2019070106
- Usta, A., Altıngöve, I. S., Özcan, R., & Ulusoy, O. (2021). Learning to rank for educational search engines. *IEEE Transactions on Learning Technologies*, 14(2), 211–225. doi:10.1109/TLT.2021.3075196
- Vieira, C., Parsons, P., & Byrd, V. (2018). Visual learning analytics of educational data: A systematic literature review and research agenda. *Computers & Education*, 122, 119–135. doi:10.1016/j.compedu.2018.03.018
- Voghoei, S., Hashemi Tonekaboni, N., Yazdansepa, D., & Arabnia, H. R. (2019). University online courses: Correlation between students' participation rate and academic performance. *International Conference on Computational Science and Computational Intelligence (CSCI)*, 772–777. doi:10.1109/CSCI49370.2019.00147
- Wang, C. (1997). 计算斯皮尔曼系数公式的证明 [Proof of the formula for calculating the Spearman coefficient]. *Journal of Yanan University*, 1, 73–75, 77.
- Wang, T. (2017). 基于Spearman秩相关系数的红外弱小目标检测 [Infrared small target detection based on Spearman Rank correlation Coefficient]. *Journal of Science Technology and Engineering*, 17(2), 234–238.

- Wang, X., Wu, W., & Fan, R. (2020). Study on the influence of college student's personality traits on employment intention choice. *International Conference on Big Data and Informatization Education (ICBDIE)*, 148-152. doi:10.1109/ICBDIE50010.2020.00041
- Wu, O., You, Q., Mao, X., Xia, F., Yuan, F., & Hu, W. (2016, July). Listwise learning to rank by exploring structure of objects. *IEEE Transactions on Knowledge and Data Engineering*, 28(7), 1934–1939. doi:10.1109/TKDE.2016.2535214
- Wu, Q., Burges, C. J. C., Svore, K. M., & Gao, J. (2010). Adapting boosting for information retrieval measures. *Information Retrieval*, 13(3), 254–270. doi:10.1007/s10791-009-9112-1
- Xiao, P., Zhu, H., Wang, W., & Huo, R. (2021). Research on the influence of family factors on college students' career choice based on SPSS software. *3rd International Conference on Internet Technology and Educational Informatization (ITEI)*, 185-188. doi:10.1109/ITEI5021.2021.00050
- Yu, B. (2021). Visual analysis of English curriculum research based on knowledge graph. *2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*, 592-595. doi:10.1109/CEI52496.2021.9574599
- Zhang, F., Zhang, X., Tang, Z., & Song, X. (2021). Evaluation and prognostics of the higher education based on neural network and AHP-PLS structural equations. *Data Science. ICPCSEE 2021. Communications in Computer and Information Science*, 1452. doi:10.1007/978-981-16-5943-0_39
- Zhao, X., Chen, C., & Li, Y. (2021). Implementation of online teaching behavior analysis system. *Communications in Computer and Information Science*, 1452. doi:10.1007/978-981-16-5943-0_32
- Zhao, Y., Zhao, A., He, H., Xia, Z., & Zhou, Y. (2021). Correlation analysis of weather factors and outage duration. *Proceedings of 2021 IEEE 4th International Electrical and Energy Conference (CIEEC 2021)*, 1-6. doi:10.1109/CIEEC50170.2021.9510213
- Zhou, F., Xue, L., Yan, Z., & Wen, Y. (2020). Research on college graduates employment prediction model based on C4. 5 algorithm. *Journal of Physics: Conference Series*, 1453(1), 1–6. doi:10.1088/1742-6596/1453/1/012033

Yi Chen is a professor in School of Computer Science and Engineering, and director of Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing Technology and Business University, China. She received her PhD degree in Computer Application Technology from Beijing Institute of Technology in 2002. Her research interests include visualization, machine learning and visual analytics. She has published more than 50 papers in academic journals, such as Science China Information Sciences, The Visual Computer, JVLC and JOV. She actively served several conferences, such as IEEE PacificVis, ChinaVis, ChinaVR.

Xiaoran Sun is currently pursuing a master's degree in School of Computer Science and Engineering, Beijing Technology and Business University. Her research interests include educational visualization and visual analytics.

Wenqiang Wei is currently pursuing a master's degree in School of Computer Science and Engineering, Beijing Technology and Business University. His research interests include educational visualization and visual analytics.

Yu Dong is a Ph.D. student at the University of Technology Sydney since 2018 to now. He received his master's degree from Beijing Technology and Business University in 2017. His research interests include information visualization and visual analytics.

Christy Jie Liang leads Data Visualization Research Lab in Visualization Institute at the University of Technology Sydney, whose research interests focus on Data visualization and Visual Analytics. She has contributed substantially to the fundamental visualization research, interactive spatial and temporal Visualization and Visual Analytic techniques for large scale Data. Her research has been widely applied to the domains, including finance, food safety, bio-medical, smart city, and social media.