Personalized Professional Recommendation System Based on Undergraduate Questionnaires

Jin Wan, Qiang Sun, Xiang Li, Jin Ding,Quanyin Zhu*

Faculty of Computer & Software Engineering, Huaiyin Institute of Technology, Huaian, China
*Corresponding author's email: hyitzqy@126.com

***Abstract*—In order to satisfy the majority of the undergraduate students who desire their major, a personalized professional recommendation system based on students' interests and learning ability is proposed. Data transfer module, Professional recommendation module, and Visual graph analysis module are contained in this system. The basic information such as hobbies, learning abilities are fitted a series of professional characteristics by users passing respectively. Parts of personalized recommendation professional health directions are unfolded through a visual map, which is conducive to his professional selection. At present, the correct rate of Text Segmentation reaches 98% by analyzing users' preferences, the accuracy rate of making professional predictions by using random forest algorithm is up to 97%. The investigation results demonstrate that the application of the proposed system catch the students' mind of personalized professional recommendations.**

Keywords- Hobbies and interests; Machine learning and data Mining; Recommendation algorithm; Visualization.

I. INTRODUCTION

The main force of national professional talents and scientific and technological innovation is college students, and majors are the important cornerstone for undergraduates to achieve novel learning and grow into professional talents [1]. Studies demonstrate that professional interests have an important role in promoting college students' professional adherence, academic performance, learning motivation, and career planning [2]. The diversity of professional schools and colleges, how to help college students choose their own specialties based on their personal interests and learning abilities becomes a more popular research direction. Nowadays, there are varieties of recommended professional software on the Internet. The only source of data for their analysis is SAT scores. However, the test results not only have a large number of professional bases, but also students may not know what disciplines and knowledge they are good at clearly [3]. In order to make professional recommendations more personal to the actual situation, a personalized professional recommendation system is proposed and demonstrated. Personalized professional recommendation refers to the combination of students' interests and hobbies, the learning ability and professional characteristics of different subjects, and making suggestions about specialty selection [4].

Data mining refers to the process of searching for a special relationship between implied knowledge and signals from a large amount of data. There are multiple processing steps in the data mining process, which generally include three phases: data preparation, data mining, result interpretation and evaluation [5]. In the context of establishing associations of data through data mining, machine learning techniques are used to simulating human behavior and machines acquire new knowledge through training of large amounts of data. The study of machine learning originated in the 1950s. At present, China's data mining and machine learning make remarkable progress which include computer chess, voice recognition, automatic car driving and so on [6].

Based on our past work research, A recommendation system is built that combines personalization and professional recommendation to improve the accuracy of professional recommendations and becomes more humanized, aiming at each individual's current situation.

In the second part, we will introduce the architecture of the entire system. The third part introduces the module of the recommendation algorithm. The fourth part introduces the module of data experiment.

II. SYSTEM ARCHITECTURE

User registration module, user test module and display module are contained in the system.

The user registration module is used in collect the user's initial information to ensure the service of the system.

The user test module provides the user with the test and sends the user's test data to the Web server via Http requests.

The display module is used for receive the user's request data and analyzes the data, and finally returns the user's professional recommendation information.

The main innovation of the system is the processing of user data through machine learning algorithm, realizing the recommendation of users.

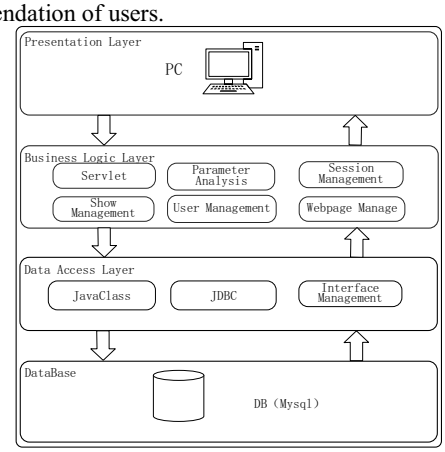

Figure 1. System Architecture

978-1-5386-7445-1/18/$31.00 ©2018 IEEE
DOI 10.1109/DCABES.2018.00045

1) Presentation Layer

The presentation layer consists of a number of web front-end pages, which is tectonic by bootstrap technology, and the user interacts with the system through the presentation layer after authentication.

2) Service Layer

This layer presents test issues and personalized professional recommendation results.

The service layer links the presentation layer and data layer, providing registration function, test function, personalized professional recommendation function, etc.

The Web server receives the client's request, retrieves the data through the interface of the data layer and returns it to the user.

3) Data layer

MySQL mainly stores user's initial metadata information, user's test data, log information and function module information.

III. RECOMMENDATION SYSTEM ALGORITHM

A、Content-based recommendation

In this personalized recommendation system, the recommendation algorithm is the core part of the whole system, which largely determines the merits of this recommendation system performance [7]. Content-based recommendation, collaborative filtering recommendation, knowledge recommendation and hybrid recommendation are contained in the current mainstream recommendation algorithms [8]. Taking into account the personalized features of current recommendations, the recommendation algorithm based on content (CB) is more appropriate. The CB process consists of the following three steps:

1) Select some features for each Item to represent Item.

2) Use the feature data of the user's favorite Item to learn the profile of this user.

3) Provide the user with the most relevant item by comparing the features of the user profile and the candidate item.

B、Space vector model

All the professional sets are D = $\{d_1, d_2, d_3, \ldots\ldots, d_n\}$, and the set of words that appears in all majors is T= $\{t_1, t_2, t_3, \ldots\ldots, t_n\}$. In other words, n majors to deal with, and there are n different words in it, eventually we are going to use a vector to represent a major. For example, the JTH major is represented as D_j = ($w_{1j}, w_{2j}, w_{3j}, \ldots\ldots, w_{nj}$), where w_{1j} represents the weight of the first word t1 in the professional j, and the larger the value, the more important it is [9]. The key is how to calculate the value of each component of D_j for the JTH major. At present, the most frequently used calculation method or the frequently used term frequency-inverse document frequency in information retrieval [10].

In Eq.(1),TF-IDF corresponding to the k in the dictionary in the JTH major is:

$$\text{TF-IDF(tk,dj)=TF(tk,dj)}\log\frac{N}{Nk} \tag{1}$$

In Eq.(2),the weight of the KTH word in the professional j is:

$$w(k,j)=\frac{TF-IDF(tk,dj)}{\sqrt{\sum_{S=1}^{|T|}TF-IDF(ts,dj)^2}} \tag{2}$$

Based on the fact that users give his preferences to some major. So, have to do is to analyze each professional user community of interests, to train the users in a particular professional attribute feature set, the user attributes most related n the item can be returned to the user as recommended [11].

C、Chinese word segmentation

A participle is the process of regrouping a sequence of words into a sequence according to certain specifications. The word segmentation method based on the string matching, the participle method based on understanding and the statistical participle method are contained in Segmentation algorithms [12].

JIEBA participle (a Chinese word segmentation method) supports three participle modes: 1. Precise mode, trying to cut the sentence most precisely, suitable for text analysis; 2. Scan all the words in the sentence. 3. Search engine mode, based on the accurate model, the long word segmentation again, improves recall rat [13].

Table 1. TF-IDF segmentation

Key	Value	Key	Value	Key	Value
Everyday	0.8539	Read	0.4269	Reporter	0.4269
Translations	0.8539	Work	0.4269	Teacher	0.4269
Meet	0.4269	Emails	0.4269	Training	0.4269
Life	0.4269	Documents	0.4269	Education	0.4269
Ability	0.4269	Edit	0.4269	Hours	0.4269
Feeling	0.4269	Time	0.4269	Perform	0.4269

D、Collaborative filtering algorithm

There are mainly item-based recommendation algorithm and user-based recommendation algorithm. The item-based-recommendation algorithm is mainly about the similarity between item and item. The user-based recommendation algorithm related to the similarity between users and users [14]. There are varieties of similarity measures between different samples [15].

1) Euclidean Distance

Compute the distance of two 'n' dimensional vectors a(x11,x12... x1n) and b (x21, x22,... ,x2n) in Eq.(3):

$$d_{12}=\sqrt{\sum_{k=1}^{n}(x_{1k}-x_{2k})^2} \tag{3}$$

2) Manhattan Distance

Compute the distance of two 'n' dimensional vectors a(x11,x12... x1n) and b (x21, x22,... ,x2n) in Eq.(4):

$$d_{12} = \sum_{k=1}^{n} | x_{1k} - x_{2k} | \quad (4)$$

IV. THE EXPERIMENTAL PROCESS

The collected experimental data are from questionnaires filled out by students from various colleges and universities, which included 13 colleges and more than 50 majors, including 5,000 questionnaires. The data includes students' interests, English ability, professional information, subject classification information, participation activities, reading data, news and employment tendency, etc. The method of processing the characteristic data of students according to category data. Figure 2 shows the data structure.

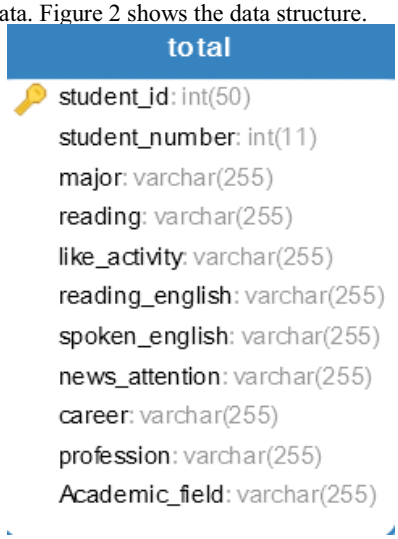

Figure 2. Some of Student Features

A. Data Exploratory

Explore the segmentation effect of processing Chinese data in content-based recommendation system.

Explore the distribution of each feature in the data set, taking the relationship between major features and majors as an example. Figure 3 is the distribution relationship between the major and the sub-division.

1) The result of jieba Chinese word segmentation

Table2. Chinese Word Segmentation Experiment

model	evaluation
Text Analysis(Chinese word segmentation)	0.98

Numeric data and categorical data are contained in data sets, so it is necessary to extract the characteristics of the data. Seven dimensions in students are contained in the original data, which transforms to thirty-six dimensions.

$$STUDENT = \{\{x_1^1, x_1^2, ..., x_1^{36}\}, ..., \{x_j^1, x_j^2, ..., x_j^{36}\}\}$$

Because the data has missing values, the data needs filling or discarded before the test.

We used the median and random forest algorithm to fill the data, and then compared the experimental results.

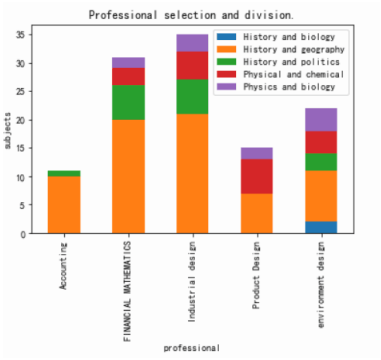

Figure 3. Features Distribution

B. Model Training

Some of the concepts used to evaluate models need to be mastered before model training.

TP: The prediction is positive, actually positive.

TN: The prediction is negative, actually negative.

FP: The prediction is positive, actually negative.

FN: The prediction is negative, actually positive.

Accuracy: The formula is R = TP+TN/TP+TN+FN+FP. The accuracy is based on the prediction results, which represents how many of the positive samples are actually positive samples.

Recall: The formula is TP/ (TP+FN), which counts all "correctly retrieved item (TP)" as the proportion of all" actual retrieved (TP+FP)".

F1 Score: The formula is P*R/2(P+R), P is defined as accuracy, and R is defined as recall. F1 Score is one of the indicators used to measure the accuracy of the second classification model in statistics.

Prediction and Recall measure whether a model is good or bad. F1 Score is a weighted average of model precision and recall rate, with a maximum value of one and a minimum value of zero.

The experiments which are based on traditional machine learning models, such as KNN, LR, Decision-Tree, RandomForestClassifier, GradientBoostingClassifier, Logistic Regression and SVM.

The experimental results of different models show in table 3.

Table 3. Experimental Drawing of Classification Algorithm

model	evaluation		
	precision	recall	f1-score
KNN	0.86	0.86	0.85
DecesionTree	0.97	0.96	0.97
RandomForestClassifier	0.97	0.96	0.97
GradientBoostingClassifier	0.97	0.96	0.97
LogisticRegression	0.96	0.96	0.96
SVM	0.90	0.89	0.89

According to Recall Accuracy, RandomForestClassifier is the final of the final model of the algorithm.

C. Experimental Results

According to the analysis of 109 data, the learning rate graph of the random forest classifier is as follows. As can be seen from the Figure 5, there is no over-fitting or under-fitting of the model.

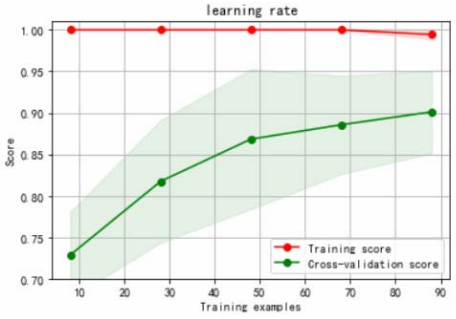

Figure5. Learning Rate Analysis

V. CONCLUSIONS

In the proposed experiment, we adopt the content recommendation model and the RandomForestClassifier model. According to the experimental results, we can see content recommendation model achieved the accuracy of text analysis is 98% and RandomForestClassifier model achieve the accuracy of data classification is 97%. In the process of further research, the precision rate of recommendation is in need of further data mining.

ACKNOWLEDGMENTS

The work in this paper is supported by The National Undergraduate Innovation and Entrepreneurship Training Program (201811049099X) and The Provincial Key Research and Development Program of Jiangsu (BE2015127).

We would like to thank the anonymous reviewers for their constructive comments.

REFERENCES

[1] Ghosh S, Winston L, Panchal N, et al. NotifiVR: Exploring Interruptions and Notifications in Virtual Reality[J]. IEEE Transactions on Visualization & Computer Graphics, 2018, 24(4):1447.

[2] He S, Zhou Z, Farhat F, et al. Discovering Triangles in Portraits for Supporting Photographic Creation[J]. IEEE Transactions on Multimedia, 2017, PP(99):1-1.

[3] Pozueco L, Rionda A, Pañeda A G, et al. Impact of on-board tutoring systems to improve driving efficiency of non-professional drivers[J]. Iet Intelligent Transport Systems, 2017, 11(4):196-202.

[4] Sergis S, Sampson D. Learning Object Recommendations For Teachers Based On Elicited ICT Competence Profiles[J]. IEEE Transactions on Learning Technologies, 2016, 9(1):67-80.

[5] Xu J Y, Wang Y, Barrett M, et al. Personalized Multilayer Daily Life Profiling Through Context Enabled Activity Classification and Motion Reconstruction: An Integrated System Approach[J]. IEEE Journal of Biomedical & Health Informatics, 2016, 20(1):177-188.

[6] Wang zhisheng, li qi, wang jing, et al. Real-time personalized recommendation based on implicit user feedback data flow [J]. Journal of computer science, 2016(1):52-64.

[7] D. Ayata, Y. Yaslan and M. E. Kamasak, "Emotion Based Music Recommendation System Using Wearable Physiological Sensors," in IEEE Transactions on Consumer Electronics, vol. 64, no. 2, pp. 196-203, May 2018.

[8] Han zhongyuan, Yang mu yun, kong lei, et al. Expansion of micro-blog query based on lexical time distribution [J]. Journal of computer science, 2016, 39(10):2031-2044.

[9] P. Castro P. and M. A. Valenzuela, "Space Vector Modeling of a SAG Mill Drive and Evaluation During Mill Shutdowns," in IEEE Transactions on Industry Applications, vol. 52, no. 5, pp. 4442-4453, Sept.-Oct. 2016.

[10] Yin yi, feng Dan, display. A personalized recommendation method based on utility [J]. Journal of computer science, 2017, 40(12).

[11] A. Arkkio and T. P. Holopainen, "Space-Vector Models for Torsional Vibration of Cage Induction Motors," in *IEEE Transactions on Industry Applications*, vol. 52, no. 4, pp. 2988-2995, July-Aug. 2016.

[12] The field adaptability method of Chinese word segmentation model [J]. Journal of computer science, 2015, 38(2):272-281.

[13] M. Zhang, N. Yu and G. Fu, "A Simple and Effective Neural Model for Joint Word Segmentation and POS Tagging," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1528-1538, Sept. 2018.

[14] C. Zhang, M. Yang, J. Lv and W. Yang, "An improved hybrid collaborative filtering algorithm based on tags and time factor," in *Big Data Mining and Analytics*, vol. 1, no. 2, pp. 128-136, June 2018.

[15] Yang qiang, li zhixu, jiang jun, et al. Entity matching based on non-master attribute values [J]. Journal of computer science, 2016, 39(10):2075-2087.