Career Recommendation System Design Based on FP-growth Algorithm

Xuemei Li[1,2,a]*, Maryli F Rosas [2,b]

[a]lxx2677@dlsud.edu.ph, [b]mfrosas@dlsud.edu.ph

[1]College of Electronic and Electrical Engineering, Anhui Sanlian University,HeFei, AnHui, China
[2]College of Science and Computer Studies – Graduate Studies, De La Salle University –Dasmariñas City of Dasmariñas, Cavite, Philippines

Abstract. In recent years, due to the rapid increase of college graduates in China, Chinese college graduates have encountered unprecedented challenges in employment. Therefore, it is necessary to fully understand the needs of enterprises and what factors in the employment process contribute to successful employment. Comprehensive consideration of factors including individual school performance, job requirements, etc. In order to provide suggestions and guidance for undergraduate graduates in employment, this paper collects the employment situation of graduates from the electrical engineering school of a university in the past three years, and analyzes the data based on the association rules of F-growth algorithm. The analysis result is one of the menu design of job recommendation system designed by employment guidance system. The innovation lies in that users do not need to master the algorithm, they can enter their personal information after entering the system, directly through the FP-growth algorithm, get the relevant factors of internship evaluation and the correlation coefficient between various factors, and understand the weight of various attributes in the application. The system also allows graduating students to understand the skills they should acquire and the advantages they should focus on development before entering the workforce. This paper provides some suggestions for students to prepare for employment, and also provides an effective reference for the employment guidance of the majority of colleges and universities.

CCS CONCEPTs: Theory of computation,Theory and algorithms for application domains,Machine learning theory

Key words: association rules FP-growth algorithm correlation coefficient matrix employment guidance system

1. Introduction

In order to achieve the purpose of providing advice to graduates in employment, this paper collects the employment situation of graduates from a college of electrical engineering in the past three years, and analyzes the data based on the FP algorithm. The analysis results are used as one of the menus of the job recommendation part designed by the employment guidance system.

The data mining in this paper includes data correlation analysis, association rules and pattern evaluation.

ITEI 2023, November 24-26, Zhengzhou, People's Republic of China

DOI 10.4108/eai.24-11-2023.2343709

Correlation analysis is to find the degree of correlation between the features of two or more event items. If this event A occurs when another event B also occurs, it is said that event B depends on event A.

Association rules are a form of machine learning. Machine learning is when robots mimic human behavior. The machine itself has a large storage capacity, and it outputs the results through the storage, sorting and summary of various data. It involves many subjects such as linear algebra, statistics, control system and so on.

Association rules are the phenomena that there is a certain relationship or law between two or more items. An introduction to several metrics of association rules, including support, confidence, and promotion.

The degree of support is the probability of the occurrence of one event in the occurrence of all events. For example, S(X)=N (X) /N for the support of item set X when the total event is N, and N (X) represents the support count of X. X→Y represents the probability of events X and Y occurring at the same time. Let's say 20 out of 100 customers buy both a computer and a printer,

The confidence level is the frequency of occurrence when an event Y occurs and contains an event X. This means that 20 out of 50 customers who bought a computer also bought a printer, so the confidence level is 20/50=40%. The expression is shown in equation (1).

$$P(Y \mid X) = P(XY) / P(X) \quad (1)$$

Lift (lift degree) reflects the correlation between A and B in the association rule. Lift degree greater than 1 and higher indicates higher positive correlation; lift degree less than 1 and lower indicates higher negative correlation; lift degree =1 indicates no correlation. Association rules are applied to all kinds of data processing.

In order to improve the data processing speed of association rules, Zhu Anqing, Li Shuai, and Tang Xiaodong applied the parallelization method using FP-growth algorithm on Spark platform [1]. The data scan is completed first, and then time series is introduced. Finally, the nodes of the distributed system bear the load equally to make full use of the functions of each part. They proved to be more efficient.

There is such a problem among suit users: the information of each factor of suit customization is incomplete. Zhao Xin and Wu Tao improved them to solve this problem by using FP-growth algorithm in association rules to mine various data and solve the problem of high internal consumption of resources and low efficiency [2]. They propose an improved association rule and combine it with the k-means algorithm. Finally, the algorithm can save resources and improve efficiency, and it can also get some rules that users are interested in. This is the application of FP-growth algorithm in suit customization enterprises.

Wei Kun, Wang Fang and Huang Shucheng proposed a new improved association rule mining method, which is a frequent pattern mining algorithm MGFP-growth based on FP-growth[3]. The experiments show that the algorithm only scans the database once, which saves the memory space and improves the efficiency of the algorithm.

Liu Cong studied the fast tracking method of close contacts, and used the FP-Growth algorithm to track and accurately locate a specific person. Using relevant data sets for

correlation analysis, the more overlapping paths, the more prominent the advantages of this algorithm [4].

Han Tianpeng, Wang Feng, and Wang Hao proposed an improved frequent pattern mining algorithm on the basis of FP-tree, which uses FP-Growth algorithm to construct data items of incremental database. The FP-tree method has reduced time and complexity and has some advantages compared with FP-Growth [5].

Wang Ying, Gao Qi, Li Tingyu et al. sorted out and mined the data characteristics collected after the product sale. First, the fever data set is constructed, and then the FP-growth algorithm is used to process the service data. At the same time, the algorithm is optimized, and a new improved algorithm is proposed. Experiments show that the algorithm is highly efficient and effective association rules are obtained [6].

In order to analyze landslide disaster situation from data, Zhu Honghu, Wang Jia, Li Houzhi et al. mined and analyzed the data based on clustering method and association rules. Taking Xinpu landslide in the Three Gorges reservoir area of the Yangtze River as an example, they analyzed the displacement rate of very large landslides under the influence of reservoir water level fluctuation and rainfall, and the experimental results showed that the method was effective and reliable. It is of positive significance to the data analysis of causes of landslide disasters [7].

Wulandari, C.P. , Chao, O., Wang, H. (2019)use mutual information measures to discretize a dataset of stroke examinations from Taiwan medical center. In order to simplify the discrete form and the quality of generating rules, interval merging method is proposed. Finally, the prior-rarity method is used to generate rare association rules with relatively low support. In addition, the contents of the rule item set were filtered, and the relative risk values of stroke occurrence were analyzed based on the extracted literature. The results show that the mutual information discretization is superior to the traditional discretization method in supporting better extraction of quantity and quality measurement, which can be used for further analysis. In addition, understanding unusual rule patterns from rare association rules may provide physicians with potentially new and unusual insights and improved awareness of stroke screening results[8].

Paraskevas Koukaras, Christos Tjortjis , Dimitrios Rousidis, They extracted knowledge about the public attitude of the global crisis, used the COVID-19 pandemic as a case, analyzed all the relevant cases from February to August 2020, and put forward the idea of visualization technology to form a theme, using association rules. discover frequent words and generate rules to infer users' attitudes[9].

These analysis methods use association rule analysis algorithms, some of which optimize the model and improve the algorithm. The innovation of the employment recommendation system based on association rule model: this paper combines the FP-growth algorithm with the career advice system of college students, so that users can directly use it, combining theory and application. This paper designs an important part of the career advice system for college students based on the FP-growth model. With the development of machine learning, the application of association rules is involved in various fields, especially in the education industry to analyze the performance and behavior of students and the correlation between the scores of students in various subjects. This paper try to find out the factors that affect students'

employment and understand the weight ratio of each factor, so that students can carry out purposeful learning according to the requirements of the desired position and achieve the requirements of the desired position. This paper provides some suggestions on how to accurately promote students in job hunting, and provides an effective reference for the employment guidance of the education cause of the majority of universities.

2 . Principle of correlation algorithm of association rules

The problem of this paper is that finding different project portfolios is a very time consuming and computationally heavy task, so we need some reasonable search methods to find frequent project sets in a reasonable time. The algorithms of association rules generally include Apriori algorithm and FP-growth algorithm.

2.1 Apriori Algorithm

Apriori is a commonly used data analysis algorithm to find frequent items in a data set. By understanding the frequent items in a dataset, you can understand the basic characteristics of the dataset. The rule of the algorithm is that we first set a support level, then sort the set of candidate items for an item, and eliminate all candidate items below this support. The second candidate set is then sorted, eliminating all secondary candidate sets below this support until the K-1 candidate set is selected above this support. It works as follows: 1. Find the data set. 2. Determine the included feature items in the data set. The set result is a binary value, expressed as 0/1. 3. Perform the first scan to obtain the number of occurrences of an item in the dataset. Support is calculated separately for each item as a member of the candidate based on minimum support. 4. Assuming a minimum support value, scan and filter the small support according to the contained items of a subitem. 5. With constant support, repeat step 4. Until no new candidate sets can be made and frequent term results are obtained.

Its characteristic is that each step of the association rules is carried out on the basis of the previous step, high credibility, simple, not high requirements. However, it is scanned for each item set without any omissions, which may require more time and a large amount of computation.

2.2 FP-growth algorithm

The steps of FP-growth algorithm are as follows: (1) FP-tree is established. (2) Mining frequent item sets from FP tree. FP-growth takes two steps, less time than Apriori. The FP-growth method is used to start mining from the last item in the header table, thus reducing the complexity of the algorithm [10].

3. Establish FP-growth algorithm model

FP-Growth only needs to scan the database twice. The first scan of the transaction database gets the frequent 1 item set. The second scan creates an FP-Tree tree. Therefore, the calculation process is simplified.

3.1 Data Sources

The data comes from the electronic file questionnaire survey of nearly three undergraduate graduates from the electrical engineering school of a university, and the career-related factors include whether the individual is a student leader, academic background, job recruitment rate, academic performance, hands-on skills, certificates and training. In the past three years, there were 243 training sets and 110 test sets. After obtaining the data of these factors related to career selection as feature vectors, this paper first preprocesses the data. Then, different parameters were used for data analysis to obtain different coefficient matrices. Finally, FP-growth algorithm was used to obtain different rules and discuss different situations.

3.2 Model Construction

Firstly, a correlation matrix is constructed in order to obtain the correlation of each feature. The correlation matrix is shown in Table 1 below.

Table 1. Correlation matrix

attribute	A	B	C	D	E	F	G	H
A	1	0.109	-0.018	0.087	0.039	0.080	-0.017	0.164
B	0.109	1	0.371	0.357	0.283	0.284	0.138	0.627
C	-0.018	0.371	1	0.354	0.269	0.304	0.160	0.678
D	0.087	0.357	0.354	1	0.229	0.261	0.102	0.662
E	0.039	0.283	0.269	0.229	1	-0.006	0.053	0.484
F	0.080	0.284	0.304	0.261	-0.006	1	0.048	0.430
G	-0.017	0.138	0.160	0.102	0.053	0.048	1	0.215
H	0.164	0.627	0.678	0.662	0.484	0.430	0.215	1

Where A: Job recruitment rate, that is, the number of people required for the job divided by the total number of applicants; B: students' grades or school performance; C: students' skills, such as some software operations, AutoCAD, etc. D: Certificates obtained, such as Computer Level 2 and English CET-4 and CET-6. E: According to the average salary of students in previous years, the salary offered by each position is divided into two categories. F: Job description, that is, the development space of the job and the room for personal advancement. G: Personal personality, that is, the personality tends to like communication or tends to dislike communication. H: Final internship performance, internship evaluation. The internship evaluation is below or above average for students in the same year. The correlation coefficient between internship evaluation and each factor is shown in Figure 1 below.

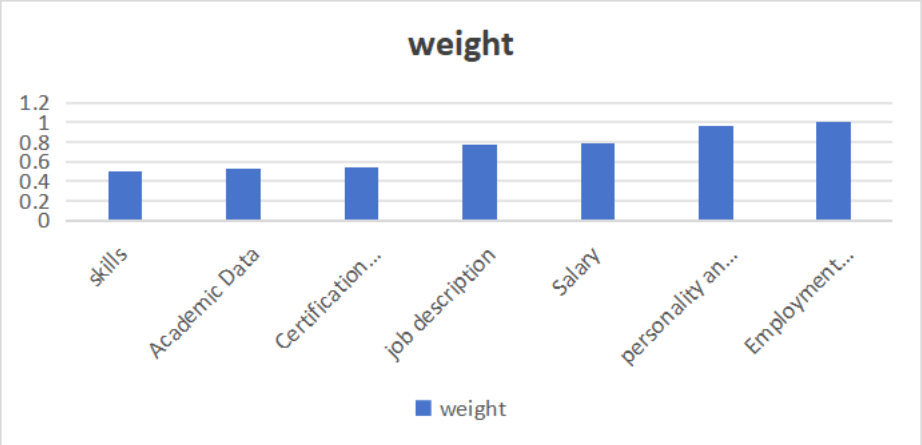

Fig. 1. Correlation coefficients between internship evaluation and various factors

We can see from figure 1 that the internship situation of students has a lot to do with the recruitment rate. For example, only two people are recruited for a position, but there are 500 applicants, so the recruitment is relatively low. In addition, from figure 1, we can see that the prospect of personality and career is valued by the students, which also affects the internship to a great extent. The description of salary and job is related to the internship results to a certain extent.

After cleaning and processing the data, we use FP-growth algorithm to find some useful association rules. The results of some association rules are shown in Table 2 below.

Table 2. Association rules obtained by FP-growth algorithm

premise	conclusion	Support degree	Confidence degree	Lift degree
Salary, academic performance	Internship evaluation	0.329	0.915	0.977
Job description, academic performance	Internship evaluation	0.337	0.924	0.980
Training certificate, salary	Internship evaluation	0.362	0.929	0.980
Salary, skills	Internship evaluation	0.334	0.952	0.988
Academic performance, skills	Internship evaluation	0.337	0.984	0.996
Training certificates, skills	Internship evaluation	0.373	0.993	0.998

From the table we can see some meaningful rules. If the students' academic performance is good and they are satisfied with the salary, we can achieve a good internship evaluation with a confidence level of 0.915. The second rule is that if the student performs well academically, then the company's job description is also satisfactory to the student. If the confidence level is equal to 0.924, the mutual satisfaction between students and enterprises can also be achieved. Other rules are shown in Table 2 above.

The results show that performing well or excelling in the profession requires students to master some professional skills, obtain relevant professional training certificates and good academic performance

4. Results and discussions

The job recommendation system is designed by applying FP-growth algorithm to complete the job recommendation. The system is based on the pycharm platform, which is used to support professional Web development under the Django framework. This paper applies FP-growth algorithm for data analysis and processing, and then tries to find some association rules, so that students can understand what factors are relatively related to the success of internship, and schools and teachers can understand what aspects to cultivate students' quality, so as to improve the adaptability of the transition from school to society. The purpose is to provide some suggestions or references to universities and relevant staff. Finally, this FP-grownth algorithm is applied to the design of college student employment suggestion system.

Project :

Natural Science Foundation of Anhui Province: Evaluation Research of Intelligent Energy System Based on Machine Learning (2022AH051993)

Natural Science Foundation of Anhui Province: Research on optimal control of isolated island microgrid system based on renewable energy(2022AH051991)

Key Project of Natural Science Foundation of Anhui Sanlian University: Design and Research of grid-connected converters for Wind Power Generation (KJZD2023004)

References

[1]Zhu Anqing, Li Shuai, Tang Xiaodong. Parallel FP-growth association rule mining method in Spark platform [J], Computer Science, 2020,vol.47,NO. (12) : 139-144, http://www.Jsjkx.com

[2]Zhao Xin, Wu Tao, Improved FP-growth Fusion K-means algorithm for suit customization collocation method [J], Computer Systems Applications, Volume 31, Issue 6, 2022:368−375, http://www.c-s-a.org.cn

[3]Wei Kun, Wang Fang, Huang Shucheng, Improved Frequent Pattern Mining Algorithm [J], Computer and Digital Engineering, No. 11, 2021:2175-2180.

[4]Liu Cong, Research on Fast Tracking Technology of close Contacts based on "Pruning + Parallel" FP-Growth algorithm [J], Modern Information Technology, Vol. 7, No. 2:36-41

[5]Han Tianpeng, Wang Feng, Wang Hao, Construction of batch incremental FP-tree based on FP-Growth algorithm [J], Journal of Jiaming University (Natural Science),2017, vol. 35, No. 8:21-25

[6]WANG Ying, Gao Qi, Li Tingyu, Zhang Le, After-sales Service Data Mining based on Improved FP-growth Algorithm [J], Modern Manufacturing Engineering, Issue 6, 2021:31-37

[7]Zhu Honghu, Wang Jia, Li Houzhi, et al. Research on association rules of massive landslide deformation in Three Gorges Reservoir Area based on data mining [J]. Journal of Engineering Geology, 30 (5) : 1517-1527

[8]Wulandari, C.P. , Chao, O., Wang, H. (2019) . Applying mutual information for discretization to support the discovery of rare-unusual association rule in cerebrovascular examination dataset. Expert

Systems With Applications118(2019)52-64, Contents lists available at ScienceDirect Expert Systems With Applications journal homepage: www.elsevier.com/locate/eswa
[9]Paraskevas Koukaras, Christos Tjortjis , Dimitrios Rousidis , Mining association rules from COVID-19 related twitter data to discover word patterns, topics and inferences, Information Systems 109 (2022) 102054 Contents lists available at ScienceDirect Information Systems journal homepage: www.elsevier.com/locate/is
[10]Zhou Ting, Zhang Wei, Zhang Zehong, discover word patterns, topics and inferences, Mapping clustering algorithm based on association rules [J]. Microelectronics and Computers, 2006, Volume 23, number 3:26-33