Recruitment big data analysis based on Python crawler Assist professional learning guidance

Zhou Chuanyu, Zhang Qian, Li Na, Xiong Shuyu

***Abstract*— Based on python data extraction technology of ERP, SAP, information security adviser, data analysts use four jobs, collected more than nine hundred jobs information about the fuse, by gathering information, sorting data, using SPSS analysis data, the system analysis of employment requirements of fuse and salary, then found hidden in the data point and the connection between the various data, can be used for students in the future planning in the post on the choice of reference, to professional learning direction, focus, can have clear direction effect. To the professional study of the school guidance work, also can have apparent auxiliary effect.**

***Index Terms*— Data analysis, Major study, Python, SPSS, Specialized guidance.**

I. INTRODUCTION

A. Research background

At present, the employment market for college students is highly competitive. As college students, when they learn professional knowledge and choose the direction of employment, they are not clear about the latest employment situation and requirements, and it is difficult to determine their own development direction. However, nowadays, college students are faced with fierce homogenization competition and devalued diplomas. If they fail to establish their own development direction as early as possible and polish relevant core professional skills, they will face many difficulties in employment. If the data of the job market can be obtained, it can better help college students improve their professional quality, improve their personal ability and have more opportunities.

So we think of using Python data crawl, data analysis and data visualization, and other functions to obtain about employment recruitment, recruitment market personnel employed professional knowledge requirements of big data, by analysis of data to get the job market demand for professional talents with professional knowledge, and the statistics on the market for certain types of professionals required professional knowledge of the general requirements, for college students specialized knowledge learning and the guidance of future employment direction play a supplementary role, and plan in advance.

From a theoretical point of view, information management is a very comprehensive professional, can cultivate management and technology cross "compound talents." However, it is difficult for graduates majoring in information management to find a clear career orientation and direction in the process of choosing a career. Most students choose to do marketing, customer service, administrative and logistics work interests, professional counterparts and career development, which are not related to the major and they do not like. However, the software development, ERP engineer, service outsourcing engineer and other professional positions for which there is a real shortage of talents are rarely qualified. So we have this necessary to understand, as a considerable professional for its employment situation.

We the survey experiment with ERP, SAP, information security consultant, data analyst four jobs data extraction, collected more than nine hundred jobs about the fuse information, the information contained in the city, position, salary, experience, and the minimum degree, job description, company, industry, company nature, number, etc., the analysis of the data, we can by gathering information, sorting data, analyzing data, the system analysis of employment requirements of fuse and salary, then found hidden in the data point and the connection between the various data, in order to help fuses graduates to apply for a job.

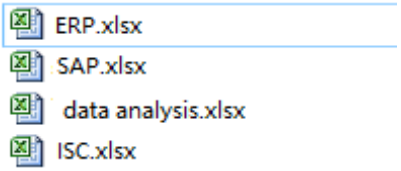
B. Research techniques

Python is a very simple, easy to learn and powerful languages, which can efficiently develop web crawler program, quickly obtain big data, and conduct data mining, statistical analysis and machine learning. Using Python to obtain big data and conduct analysis is the most efficient and effective method.

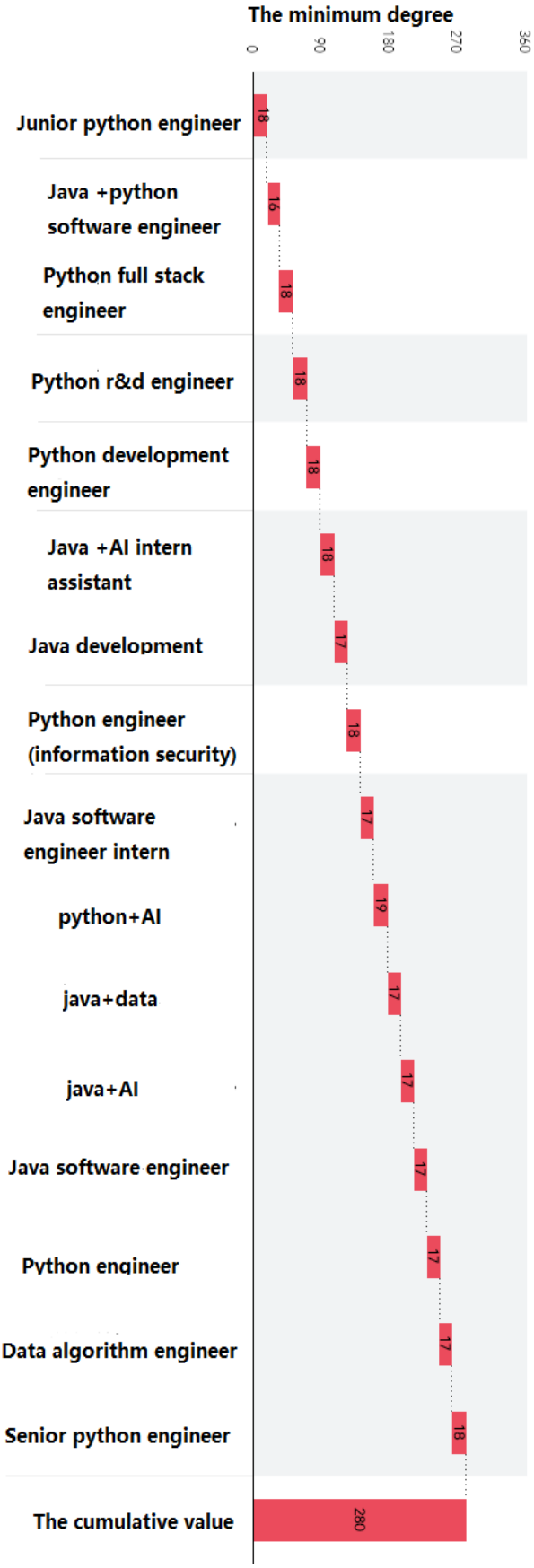
Spss data analysis tool, with complete data input, editing, statistical analysis, reports, graphics production and other functions.136 functions with 11 types. SPSS offers from simple statistical description of statistical analysis to the complicated factors, such as exploratory data analysis, statistical description, contingency table analysis, variance analysis, nonparametric test and multiple regression, survival analysis, analysis of covariance, discriminant analysis, factor analysis, cluster analysis and nonlinear regression, Logistic regression, etc. Data analysis can be visually presented.

II. DATA

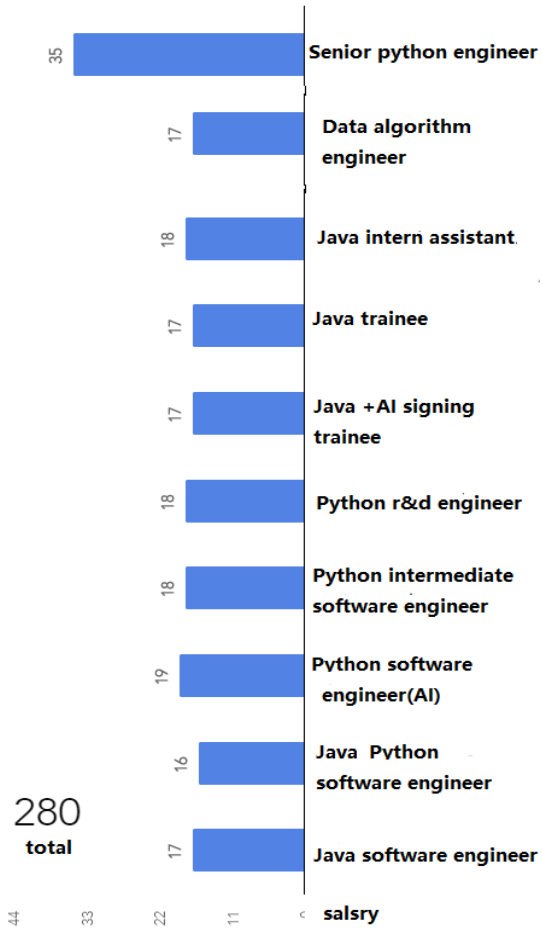
The writing of this article is based on information in ERP, SAP, security adviser, the data analyst four posts above the data of 1000 is used for data analysis, through the crawler technology crawl, with keyword filtering under the condition of ERP information in jobs get 363 article, article of SAP position information get to 174, data analyst position information to the 271 data, the data acquisition of information security adviser to the 205...



In the process of data visualization, due to the large number of elements and large data base, it is not convenient to display here. Here, only the data selected under the condition of Tianjin is taken as an example:



The relationship between position and minimum education



The relationship between position and salary

III. RESEARCH MODELS AND METHODS

A. Research model

Considering that the employment direction of the credit management is too complex and difficult to be counted, we chose ERP, SAP, data analysis and information security to explore the influencing factors of the credit management employment. In order to present our research ideas and model more clearly, we made the model diagram for studying the influence factors of the credit management job as shown in the figure below:

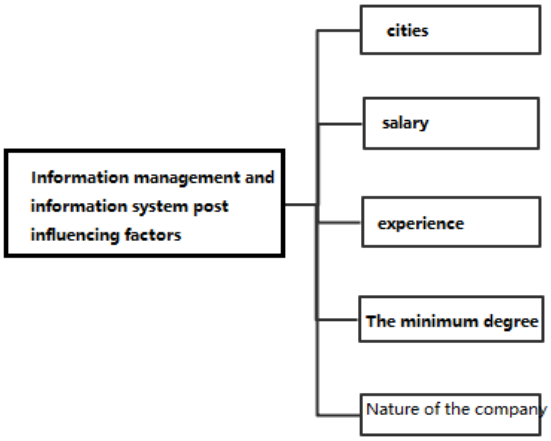


Fig. 1 influence factor model diagram of credit tube employment position

B. The analysis methods and steps used in the data analysis phase of this paper are as follows:

Descriptive statistical analysis is used to make descriptive statistics on the salary and requirements of information management professional jobs. Then, through the sample data collected, the cross-linked table is generated to obtain the correlation of various requirements of different positions for the information management professional jobs. This paper then used multinomial Logistic regression (Logistic regression is a multivariate analysis method to study the relationship between binary or multi-classification observation results and influencing factors) to study the relationship between the requirements of each position and the positions. At the same time, factor analysis method (a kind of data simplification technology, through studying the internal dependence among many variables, to explore the basic structure of the observed data) is used to express its basic data structure with a few hypothetical variables. However, the KMO value of factor analysis is less than 0.5, which reflects the weak relationship between variables, and it is difficult for the principal component to achieve the ideal goal of dimensionality reduction. he extracted factor cumulative variance contribution rate is less than 0.7. Therefore, this problem is not suitable for factor analysis. The system cluster analysis method was used to classify the requirements of different salaries to analyze the degree of similarity of factors affecting the employment position of information management major.

IV. ANALYSIS RESULT

Note: due to the large amount of analysis data and the large table of analysis results, only part is shown here

A. Logistic regression analysis results:

In order to further study the relationship between various influencing factors and posts and the degree of influence, binary multinomial Logistic regression analysis was conducted on posts with SPSS. There are four levels of dependent variables in the construction process, namely ERP, SAP, data analysis and information security. The independent variable is city. Salary, experience, minimum education and company type:

Model fitting information

model	Model fitting criteria	Likelihood ratio test		
	Minus 2 log likelihood values	chi-square	df	Significant level
Only intercept	1.668 e3			
In the end	1.054 e3	614.174	87	000.

Likelihood ratio test

effect	Model fitting criteria	Likelihood ratio test		
	-2 log-likelihood values of the simplified model	chi-square	df	Significant level
intercept	1.054 e3	000.	0	.
experience	1.164 e3	110.878	18	000.
The minimum degree	1.113 e3	59.822	21	000.
Nature of the company	1.168 e3	114.464	30	000.
Pay 2	1.132 e3	78.580	15	000.
City 1	1.119 e3	65.009	3	000.

The chi-square statistic is the difference between the final model and the simplified model in the -2 log-likelihood value. A simplified model is formed by omitting effect from the final model. The null hypothesis is that all parameters of the effect are 0.

A. Since the ellipsis effect does not increase the degree of freedom, the simplified model is equivalent to the final model.

B. unexpected singularity in Hessian matrix. This indicates that some indicator variables should be excluded or some category variables combined.

P value of model fitting information is 0.000, and the model is significant. The significance level of all factors in the likelihood ratio test is high, and all P values are less than 0.05.Parameter estimation reflects the parameters of logistics model, assuming its test results, confidence interval of dominance ratio and other information, which are the main results of multiple regression model. We can get the coefficients, statistics and so on of the variables in the equation from "variables in the equation". Using Wals statistics and concomitant probability, it can be concluded that among all factors as independent variables, the importance of city, >, importance of salary, >, importance of company nature, >, importance of experience, and >, importance of minimum education.

B. Results of cluster analysis

In order to analyze the degree of similarity of factors affecting the employment position of information management major, this paper USES the systematic clustering method to analyze the factors including position, salary, experience, minimum education and city. Next, the results of cluster analysis will be displayed and analyzed

Clustering tables

order	cluster combination		coefficient	First cluster		next
	cluster1	cluster2		cluster1	cluster2	
1	1	3	.220	0	0	3
2	2	4	.134	0	0	3
3	1	2	.053	1	2	4
4	1	5	-.227	3	0	0

The approximate matrix

case	Matrix file input				
	city1	salary2	min degree 1	experience1	jobs
city1	1.000	-.375	-.102	-.084	-.317
salary2	-.375	1.000	.105	-.101	.039
min degree 1	-.102	.105	1.000	.111	.117
experience1	-.084	-.101	.111	1.000	.220
jobs	-.317	.039	.117	.220	1.000

According to the clustering table, when clustering into two categories, jobs and experience are one type, and cities, minimum education and salary are one type. As can be seen from the tree graph, the relationship between post and experience is the strongest, and the relationship between city and minimum education is the least strong. According to the approximate matrix, there is a negative and weak correlation between salary and city, and a positive and weak correlation between position and salary.

C. Analysis results and conclusions

According to the data analysis, we choose fuse professional demand for related of the four jobs in the north of Guangzhou, the largest average salary level in 5000-10000 yuan, is not high to the requirement of work experience, mostly in the following 5 years, the requirement of degree in college and university level, the demand of private enterprises to the related personnel accounts for about half.

The total demand of SAP and ERP in other cities is higher than that in Beijing, Shanghai, Guangzhou and Shenzhen. The total demand of data analysis and information security in Beijing, Shanghai, Guangzhou and Shenzhen is higher, especially the information security is more obvious. Information security and SAP overall wage level is higher than ERP and data analysis. ERP and data analysis require less experience than SAP and information security. SAP and information security require higher degree requirements than ERP and data analysis. Private enterprises have a high demand for ERP, data analysis and information security.

Using logistics regression analysis, it can be obtained that among all factors as independent variables, importance of city, importance of > salary, importance of company nature, importance of > experience, and importance of minimum degree of >.

According to the system clustering, when clustering into two categories, jobs and experience are one type, and cities, minimum education and salary are one type. Jobs are most closely related to experience, and cities are least related to the lowest education. There is a negative and weak correlation between salary and city, and a positive and weak correlation between position and salary.

V. CONCLUSION

In using the python language to obtain the support of a large amount of data, through the cross contingency table analysis, cluster analysis and Logistic regression analysis results show that the position and salary, education, the relationship between different positions can be obviously found that professional conditions and employment benefits, advantages and disadvantages of relations between the strong and weak good relations, can be used for students in the future planning in the post on the choice of reference, to professional learning direction, focus, can have clear direction effect. To the professional study of the school guidance work, also can have apparent auxiliary effect.

ACKNOWLEDGMENT

This research was supported by the Fundamental Research Funds for the Central Universities 2016MS127.

REFERENCES

[1] data spatial information collection method of the subject organization of web crawler of school of environment and surveying and mapping, China university of mining and technology
[2] zhao jiagang, di guangzhi, lv danju, introduction to computer programming -- Python program design preface design, people's posts and telecommunications press, January 2013
[3] Magnus Lie Hetland, Norway, Python basics 2nd edition
[4] zhang wentong, basic course of SPSS statistical analysis
[5] guo zhigang, social statistical analysis method -- SPSS software application
[6] cao studied the applicability of regression analysis conditions based on SPSS weighted regression

Zhou Chuanyu ,student of baoding economic management department of north China electric power university
Zhang Qian professor of baoding economic management department of north China electric power university
Li Na, a student of baoding economic management department of north China electric power university
Xiong Shuyu student of Baoding economic power of north China electric power university