The Application of Data Mining Technology in Analysis of the Relation between Majors, Industries and Positions

Xiaoguo Wang and Lin Sun

Electronic Information Engineering Institute, Tongji University,
4800 Caoan Road, Jiading District, Shanghai, China, 201804
xiaoguowang@tongji.edu.cn, sunlin_777@163.com

Abstract. In the context of Prediction System of University Major Setting Research Project, for the machinery manufacturing industry, we study for the association rules model of the relation between majors and positions .We design a set of methods to discover this model, achieve this model with existing data and analyze the practical significance of this model. This association rules model provides a useful exploration to university major settings and employment trend analysis.

Keywords: Association rules, visualization, the relation between majors, industries and positions.

1 Introduction

The correlation between majors, industries and positions has a great significance to the major setting and employment trends. The research on the model of the relation between majors, industries and positions is also an essential part of the Prediction System of University Major Setting Research Project.

The main job of this paper is discovering the association rules model between positions and majors for a specific industry, using association rule mining and visualization technology. The model can provide a scientific guidance to the university major setting and employment trend analysis.

2 General Design

2.1 Data Collection

The scale of China's machinery manufacturing industry has a leading position in the world. In this paper, we focus on this industry and study the correlation between majors and positions. According to the research needs, combined with the actual situation of the research enterprise, the project members and business executives

H. Deng et al. (Eds.): AICI 2011, CCIS 237, pp. 222–229, 2011.
© Springer-Verlag Berlin Heidelberg 2011

design the questionnaire. Then we investigate for the staff in a large machinery manufacturing enterprise and get their education background and current position information as our mining data.

2.2 Data Preprocessing

Data Cleaning. The main task of data cleaning is to eliminate redundant data and noise data, eliminate duplicate or invalid records. Some data records don't have the complete key attribute. That type of records is invalid and we should give these records a further research. On the basis of the research result, we can decide to amend the records or delete them; there are also some records with fuzzy information. That type of records should be given recognition and re-fill.

Data Standardization. The main task of data standardization is to establish the correspondence between Position Classification Dictionary of People's Republic of China and the performance standard of position classification in the enterprise. Based on that correspondence rules, we can design a performance standard of position classification for the Prediction System of University Major Setting Research Project and use it to carry out the standardization of position information. The process is shown in Fig. 1.

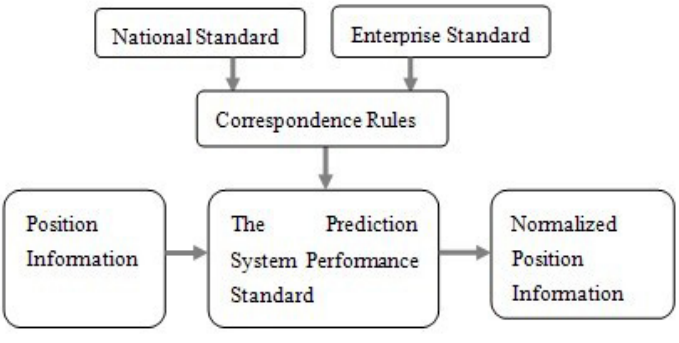

Fig. 1. The process of position information standardization

Data Transformation. Data transformation is an important part of data preprocessing. When the data source can't directly meet the data form that data mining algorithms require, we need to convert them. For example, gender, type of graduate school, educational level and some other attributes need to be converted to Integer type, according to the needs of association rules mining.

2.3 Algorithm Selection

Apriori is a classical algorithm, usually for mining single-dimensional and Boolean association rules in the transaction database. Based on the characteristics of our

mining task, we choose a multi-dimensional improved Apriori algorithm[1], this method is applicable to multi-dimensional association rule mining.

The multi-dimensional improved Apriori algorithm[1] is also separated into two steps. Firstly, search for the frequent item sets. Secondly, deduce the association rules from the frequent item sets. The different part is that we need to quantify the value of the properties before searching for the frequent item sets, using interval instead of the numerical value.

2.4 Visualization

Visualization provides a strong support to association rule mining process. To allow users to participate the mining process, we realize the visualization of mining process and mining results in this paper. Users can choose the properties involved in mining and set the minimum support and minimum confidence, according to their needs.

In summary, we mainly focus on four aspects of the association rules model: data collection, data preprocessing, the association rules model determination, interpretation and analysis for the model. The method flow chart is shown in Fig. 2.

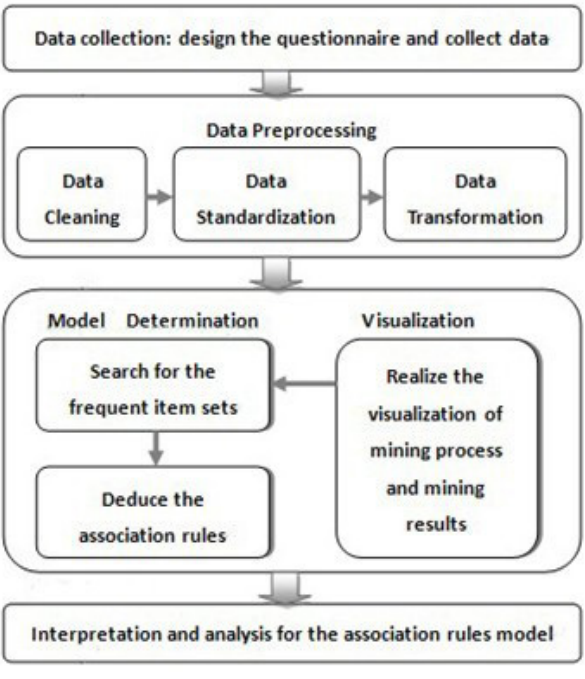

Fig. 2. The method flow chart for discovering the association rules model

3 The Realization

3.1 The Realization of Data Standardization

There are differences between Position Classification Dictionary of People's Republic of China and the performance standard of position classification in the enterprise. The paper designs the correspondence rules for the two standards as below:

- If there is one classification in the enterprise standard which is the same as one classification in the national standard, make a direct conversion from the enterprise one to a national standard one, forming a 1:1 mapping;
- If there is one classification in the enterprise standard which is not same as any classification in the national standard, make a similar conversion from the enterprise one to a national standard one, according to the actual meaning of the classification, forming a 1:1 mapping;
- If there is one classification in the enterprise standard which contents some classifications in the national standard, merge these national classifications together and create a new classification. Then make a conversion from the enterprise one to the new created one, forming a 1:n mapping;
- If there is one classification in the national standard which contents some classifications in the enterprise standard, merge these enterprise classifications together and create a new classification. Then make a conversion from the new created one to the national one, forming a n:1 mapping;
- If there is one classification in the enterprise standard which is not similar to any of the conditions we have mentioned, keep it without any conversion and add it to the performance standard of position classification for the Prediction System as a new classification.

Based on these correspondence rules and Position Classification Dictionary of People's Republic of China, we design a performance standard of position classification for the Prediction System. We convert the twenty-six classifications in the enterprise standard to the fifteen classifications in the system performance standard. The conversion table is shown in Table 1.

Table 1. The performance standard conversion

NO.	The Enterprise Performance Standard	The System Performance Standard
1	Middle-level leaders	The responsible persons
2	General managements	Administrative office staff
3	Financial officer	Economic operators
4	Legal officer	Legal professionals
…	…	…

3.2 The Realization of Data Transformation

After the data collection and data preprocessing, we import the data into the database. Part of properties of the original data in the database is shown in Table 2.

Table 2. The original data in the database

NO.	Gender	Position	Type of graduate school	Educational level	Major	Whether has second majors	…
1	Male	The managements of engineering	985	Full-time undergraduate	Materials science and engineering	No	…
2	Male	The managements of logistics	985	Full-time undergraduate	Applied electronics	No	…
3	Female	The managements of logistics	211	Full-time undergraduate	Ships and ocean engineering	No	…
…	…	…	…	…	…	…	…

We do the data transformation as below:

- Transform the value of gender property into code, 0 is for male, 1 is for female;
- Transform the value of position property into code, according to the system position classification;
- Transform the value of graduate school property into code, 0 is for general universities, 1 is for the key universities of the 985 Project, 2 is for the key universities of the 211 Project(non 985);
- Transform the value of educational level property into code, 0 is for full-time undergraduate, 1 is for adult education.
- Transform the value of major property into code, according to Major Setting Dictionary of People's Republic of China;
- Transform the value of whether has second majors property into code, 0 is for answer no, 1 is for answer yes.

We can see the data after transformation in Table 3.

Table 3. The data after transformation in the database

NO.	Gender	Position	Type of Graduate School	Educational level	Major	Whether has second majors	…
1	0	1	1	0	28	0	…
2	0	6	1	0	42	0	…
3	1	6	2	0	30	0	…
…	…	…	…	…	…	…	…

3.3 The Association Rules Model Determination and Visualization

We apply the multi-dimensional improved Apriori algorithm[1] to association rule mining and realize the visualization of mining process and results. C# is the coding language we use to realize the algorithm and visualization. We can freely choose the properties involved in mining and set the minimum support and minimum confidence. The interface of the visualization is shown in Fig. 3.

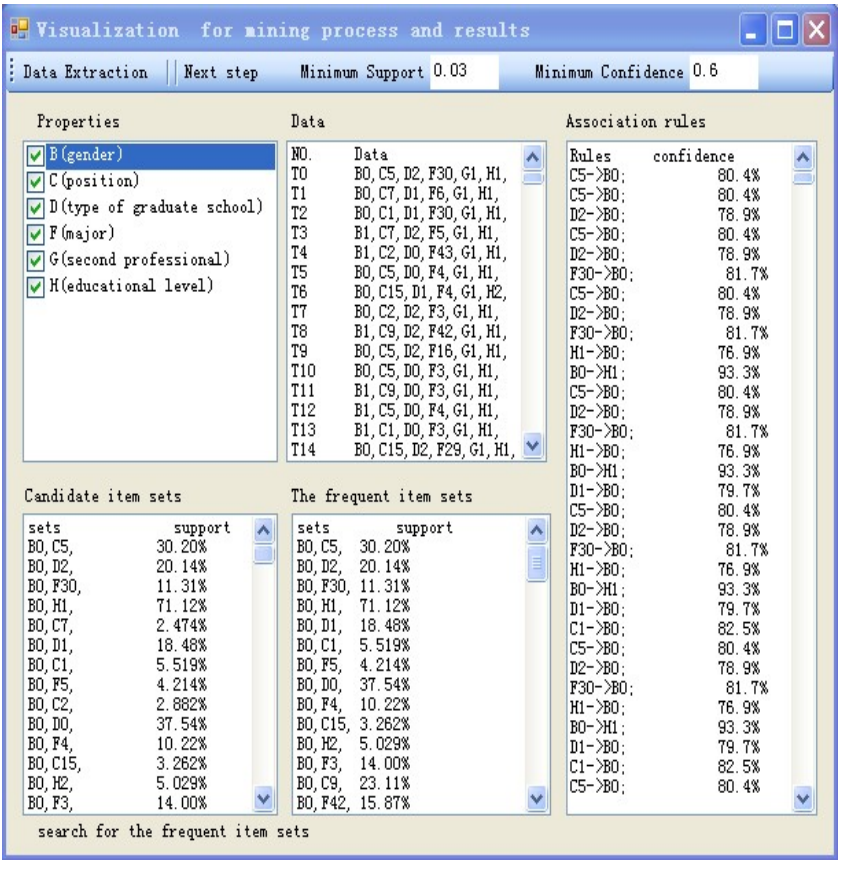

Fig. 3. The interface of visualization for mining process and results

Based on experience, we choose the value of the minimum support as 0.03, and the value of the minimum confidence as 0.6. Association rule mining discovers many association rules. After scientific analysis, we pick up the rules which involved to the five key elements as our interested. The five key elements are gender, type of graduate school, educational level, position and major. The association rules model is shown in Table 4.

Table 4. The assoication rules model

NO.	Conditions	Results	Support	Confidence
1	The responsible persons	Male	5.52%	82.5%
2	The managements of engineering	Male	30.2%	80.5%
3	The managements of engineering	Full-time undergraduate	35.8%	100%
4	The responsible persons	Full-time undergraduate	5.68%	84.9%
5	Civil engineering	The managements of engineering	4.89%	84.5%
6	Marine engineering	The managements of engineering	5.74%	64.9%
7	985, civil engineering	The managements of engineering	3.62%	90.1%
8	Civil engineering, full-time undergraduate	The managements of engineering	5.76%	84.6%
9	985,full-time undergraduate, civil engineering	The managements of engineering	3.59%	90.3%

4 The Analysis

4.1 The Influence of Single Factor

1. In the machinery manufacturing industry, men with a full-time undergraduate degree often take the managerial position.

In the machinery manufacturing industry, we have 82.5% confidence degree for the rule that the responsible persons are male. We have 80.5% confidence degree for the rule that the managements of engineering are male. There is a relatively high rate for a man to be a business leader or a management of engineering in the machinery manufacturing industry. Universities can adjust the male to female ratio of admissions for manufacturing majors to meet the actual needs of the job market.

In the machinery manufacturing industry, we have 84.9% confidence degree for the rule that the responsible persons have full-time undergraduate degree. We have 100% confidence degree for the rule that the managements of engineering have full-time undergraduate degree. Both business leaders and managements of engineering have relatively high educational background.

2. In the machinery manufacturing industry, graduates from civil engineering major and marine engineering major often take the managements of engineering as their position.

In the machinery manufacturing industry, we have 84.5% confidence degree for the rule that graduates from civil engineering major take the managements of engineering as their position. We have 64.9% confidence degree for the rule that graduates from marine engineering major take the managements of engineering as their position.

4.2 The Influence of Factors

In the machinery manufacturing industry, we have 90.3% confidence degree for the rule that civil engineering graduates from the key universities of the 985 Project with full-time undergraduate degree take the managements of engineering as their position. We have 90.1% confidence degree for the rule that civil engineering graduates from the key universities of the 985 Project take this position and 84.6% confidence degree for the rule that civil engineering graduates with full-time undergraduate degree take this position.

As we can see from the rules, there is a high ratio for civil engineering students from the key universities of the 985 Project with full-time undergraduate degree to be managements of engineering in this enterprise. On the one hand, enterprises have a large dependence of graduates who meet the above conditions; On the other hand, students who meet the above conditions should learn more basic knowledge and strengthen their qualities to be the managements of engineering.

5 Conclusion

In this paper, for machinery manufacturing industry, we conduct a useful exploration and research on potential relation between majors and positions and obtain the primary association rules model. With the deepening of research work, this model will be gradually improved and will provide more effective help for the university major setting and employment trend analysis.

References

1. Xiao, B.: A Multi-dimensional association rules algorithm. In: Chongqing Technology and Business University (Natural Science), vol. 22(04), pp. 339–442. Chongqing Technology and Business University Press (2005)
2. Han, J., Kamber, M., Ming, F., Meng, X.: Data Mining Concepts and Techniques. Machinery Industry Press (2007)
3. Sumon, S., Sarawat, A.: Quality data for data mining and data mining for quality data: A constraint base approach in XML. In: International Conference on Future Generation Communication and Networking, pp. 46–49. IEEE Press, Sanya (2008)
4. Jiang, W., Chen, Z.: The view of the major setting for universities from the Needs Education. Modern Education Science 5, 7–10 (2008)
5. Imhoff, C., Galemmo, N., Geiger, J.G., Yu, G., Bao, Y.: Mastering Data Warehouse Design Relational and Dimensional Techniques. Machinery Industry Press (2004)