The Data Mining of the Human Resources Data Warehouse in University Based on Association Rule

Zhang Danping
Nanchang Hangkong University, Nanchang, China
Email: zhangdp9968@163.com

Deng Jin
Nanchang Hangkong University, Nanchang, China

Abstract—**Based on an actual dataset of college human resources, we analyzed the data warehouse technologies and combined them with the pracical work. The snowflake structure has been described about a data warehouse of university human resource and the data warehouse has been constructed. This paper reduces and categorizes features by explorative data analysis. When the data warehouse had built and applied on human resources management in university, we studied mining technologies and processes based on association rule. Association rules show the relationship among teaching, research and social practices. The results of this study can be well explained and have some management suggestions on human resource management in university.**

Index Terms—**Human Resources Management, Association Rule, Feature Reduction, Data Mining**

I. Introduction

In the modern information society, the establishment of highly qualified faculty has become the core work of human resource management in university. The human resources of university is inherent the knowledge, skills, attitudes, experience and innovative ideas about the college full-time teachers. These teachers in university demonstrated the particularity of the following: having high levels of schooling , pursuing the unique style of teaching and research preferences, seeking to meet mental demand, pursuing the relaxed academic environment, eagering to participate in various learning and training. Meanwhile, there are a lot of human resource management problems in the university of our province ,such as severely environmental constraints , the low quality of management, the lack of information communication, training mechanism is not perfect, the real income did not meet their psychological expectations ,etc. To respond to these many problems, teaching staff is essential to universities in our province because they can train a group after group of outstanding talents. Managers of universities must seize the faculty building, on the one hand to grasp the number of teachers , on the other to exert the teacher's enthusiasm and creativity.

The word "information" will be the core of word in human resources management. At present, the personnel departments of colleges and universities make use of management information system has accumulated a large amount of data. Now these data on the existing resources are short of an effective organization, collation and extraction, it is not only difficult to conduct an objective analysis, but also not to know what to make of in decision support. Based on the workload and work factors so much complicated and important in college human resources management, there is an urgent need to achieve the information management of scientific and systematic. Since the emergence of data warehouse technology and development, data integration and transformation can be greatly simplified, complex data can be effectively organized. Through statistics and analysis with fair and objective, we can quickly and correctly identify the model of implicit and accurate grasp of future developments, improve the efficiency and capacity utilization.

With the fierce competition, in order to obtain institutions of higher learning by leaps and bounds, the university must make full use of advanced information technology. The personnel department must establish human resource data warehouse. Data mining analysis to support decision-making will become a trend.

II. Design Human Resource Data Warehouse

Data warehouse is an object-oriented, integrated, nonvolatile, and time variable data set, it is used in decision support for management [1]. In this paper we used an actual university human resource management project as its background to show a relatively design and implementation process of the complete data warehouse. The theme is the teaching and scientific performance evaluation of university teachers, and the all data source about this theme is from the personnel management database system.

doi:10.4304/jcp.6.1.139-146

First of all, data warehouse can be seen as a special DBMS, we developed a platform as shown in Fig. 1 include:

According to the definition of W. H. Inmon, A Data Warehouse process should contain the following five parts [1]:

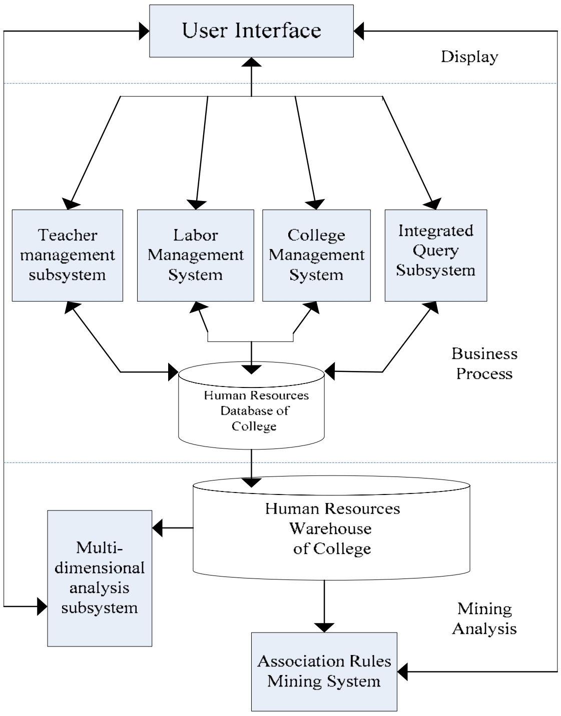

User Interface
Display
Teacher management subsystem
Labor Management System
College Management System
Integrated Query Subsystem
Business Process
Human Resources Database of College
Human Resources Warehouse of College
Multi-dimensional analysis subsystem
Mining Analysis
Association Rules Mining System

Figure 1. Structure of Human Resources Data Warehouse in university.

(a).Data pretreatment: for example, remove, etc.

(b).Data conversion, and for your choice stage mining way.

(c). All sorts of model, the mining scheme.

(d). Take on the results.

(e).To explain and use the Results.

In fact, the establishment process about the human resources data warehouse in university is the traditional database-centric structure transfer from operating systems to a data warehouse architecture as the central process. To achieve the changes in architecture, the first thing to do is understood the system which already has, which is the basis for the establishment of data warehouse. After understanding of the data system we can analysis about user needs so as to determine the theme of data mining system.For a large-scale data warehouse systems, it involved in many business systems and its function are complex .Spiral development methodology used by the target will be divided into a large number of the implementation phase, divide and rule [2]. Association rule mining subsystem is one of relatively simple and minor problems.

In our research, a structure of human resources data warehouse in university has built. From the user interface, there is four subsystem has displayed: teacher management subsystem, labor management subsystem, college management subsystem and integrated query subsystem. There is a human resource database processed a large amount of data that come from these subsystem. Then all data of this relational database would be exeracted, transformed and loaded so as to create a relational data warehouse. This is the human resource data warehouse in unversity. Underlying this warehouse, we can do mining analysis with the help through association rules mining system and multidimensional analysis system.

III. Pretreatment and Conversion of the Data in Human Researches Data Warehouse

The paper adopts human resource data sets from a university of science in Nanchang, Jiangxi, which is a real data set (we had hidden personal information as name etc.).

A. Normalization of "basic informati-on" data set

Data set records the staff's personal information, statistics, professional and technical titles in our experience, etc. Data set reflects the basic personal information database of employees and has hundreds of characteristics, all distribution in 10 database table. According to the expert proposal, we selected 24 characteristics of representative, including: departments, nationality, ethnic, date of birth, titles, title start time, degree, education level, graduated colleges, professional qualification, the position, the type of party, etc.

Using relevant statistical tools, we had a basical statistics and analysis to "basic information" data set. For example: the youngest age is 25, the biggest is 68. Graduated colleges involved 174 universities. Of which there are 286 is own university graduates, accounting for 8.4%. Both overseas and in HongKong and Macao University of 48, accounting for 6.4%. Last 411 is graduated from other domestic university, accounting for 55.2%. A total of eight political parties parties, the Communist Party of China which has 491 members, accounting for 57.2 percent and so on.

B. The principles of data collation in data set

Some characteristics which has not significant correlation with the theme were Deleted. such as "department", "nationality", "nation", "cadres appointed date", etc.

To delete the underlying concept (described in more details) features, to retain more high-level and abstracted characteristics. Such as to delete "Title Series" , "jobs", "title of the beginning of time", which is described the contents of the "title name", "title-level" two characteristics about professional and technical positions can be related to.

To delete the characteristics which definition were confused, or to clear its value. Such as "education level" in the classification is not accurate enough clear and be deleted. "Graduate" not only describe the characteristics of the professional staff to learn, but also to describe the staff’s the highest degree . we described it clearly as the highest degree.

C. *The data processing methods of data set*

Characteristics of the date should turned into a numerical variable. Such as "age", "school age", "service", respectively, said the characteristics of the original "Date of Birth", "employment date" ,"work date" etc.

The numerical characteristics should turned into discrete variables. The criteria of discretization is that the value of each characteristic which was discreted contains discretization of employees. At the same time discretization process on the demarcation point discretization is not sensitive[5]. According to this criteria, the distribution of mapping about the characteristics of the numerical value should be summarized in graphics by the bottom classification, and the data should be discretizated. Such as "age" variables can be separated into 3:25-40, 41-50 and 51-68. Similarly, a "Schoolage", "service" and other characteristics can also be returned to the discrete variables.

The classification variable should be reduction. Such as we reducted the value of “degree” into “undergraduate, master's degree or doctorate” three categories, "Party" reducted to " party members , democratic parties and the masses" three types. Some features may have different characteristics of the results for variable value reduction , such as " graduation institutions " can be reducted "own school, other universities, both overseas and in HongKongand Macao University of" three types, or "own school, theother 21 institutions, foreign and University of HongKongand Macao, and other "categories, or "own school , 985 institutions, both overseas and in Hong Kong and Macao University, the other "categories.

After more than data processing, the data set contains 14 independent features. Some of which retain the characteristics of the different characteristics of expression (the value of type, classification type) to apply to different data mining algorithms. Reservations 14 independent features include: date of birth (age), the employment date (school age), the work of date (seniority), gender, title name, title-level, degree, graduation time, graduated from institutions, professional, job level, job, level names, and parties.

IV. Construction of the Human Resource Data Warehouse in University

A. *The Data Definition*

The data definition is definited the data structure, includes the following several parts:

(a)Define the logic and storage mode of the data warehouse. For example: tables, indexes, views, stored procedures, physical storage mode.

(b)Define the data source it will be extracted from. It is generally relational database, and distributed in different locations or in storage in DBMS.

(c)Define some rules, the algorithm and model. We canextract and process of data from the source data the sourcedata will be converted to structure data to the data warehouse filled in a data structure.

C. *Data types*

In the SQL Server, the data values of a platform datasetare arranged in a matrix table structure. table column was called avariable, it has several types of variables as follows:

(a)Numerical type: int, smallint, long, real, numeric, money, float.

(b)Character type: char, varchar, text.

(c)Date type: datetime, smalldatetime.

(d)Binary type: binary, varbinary, bit, image.

In ASCII character maximum length not exceed 200 deposit. Data collection of matrix structure observation of each variable value must exist. If a data value is missing, the system will automatically fill a missing value.

C. *Build Data Sets*

We used SQL Server 's Query Analyzer tools. The concrete steps are as follows:

(a)Open the SQL Server's Query Analyzer tool.

(b)Choose the database table that will be run .We can Compile the programe sentence with CREATE in order to produce the table, then the statements can be run.

Of course we also can success in this way：

(a) Open SQL Server.

(b) Select the database in the Enterprise Manager.

(c) Click the right mouse button to choose New Table attribute, enter the New Table window.

(d) Operate it step-by-step.

D. *Modify data sets*

If the user would like to modify an already existing data set, he can directly enter the Query Analyzer tool in SQL Server. He can choice the database is running to rewrite the database use the sentences about CREATE (first to DELETE this table) and then run it.

Of course we can select database in the SQL Server Enterprise and then click this table directly into the definition window to correct the table properties.

E. *Browse Data Sets*

There are four methods of data browsing can be carried out here:

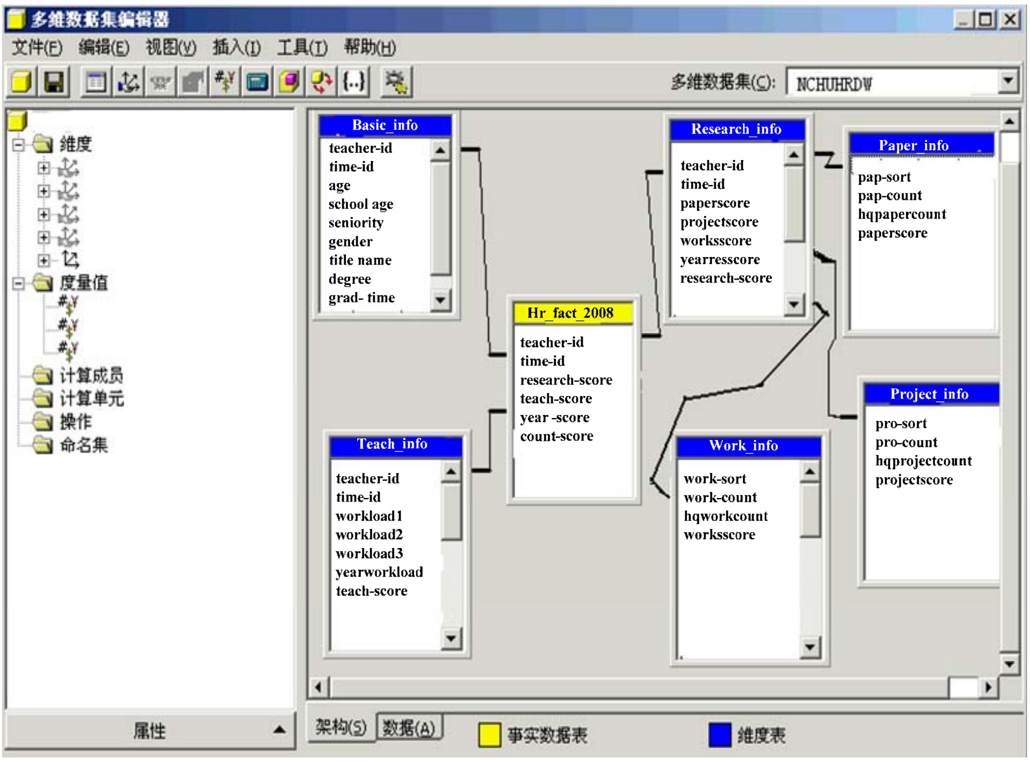
The first one is equipped with PowerBuilder in the data warehouse management workstation, then we can use Power Builder in the Database window to browsing.

The second is in the SQL Server's Query Analyzer tool to prepare SQL statement (also can use the system stored procedure), and then run the corresponding results can be obtained.

The third is in the SQL Server's Enterprise Manager by selecting the database directly in the browser.

The fourth is indirectly browse from safeguard module in platform directory.

This chart(Fig. 2) is shown the snowflake structure about the data warehouse of university human resources. It is set up by a fact table and some dimension tables that is related of the fact table. In the center, the fact table surrounded by all the dimension tables similar as the snowflake structure. In Fig.2, there are three dimension tables underlying the HR fact: Basic_info table, Research_info table and Teach_info table. At the same time, As a fact table, Research info table had three dimension tables too. They are Paper_info table, Work_info table and Project_info table.



多维数据集(C): NCHUHRDW

Basic_info
teacher-id
time-id
age
school age
seniority
gender
title name
degree
grad- time

Hr_fact_2008
teacher-id
time-id
research-score
teach-score
year -score
count-score

Teach_info
teacher-id
time-id
workload1
workload2
workload3
yearworkload
teach-score

Research_info
teacher-id
time-id
paperscore
projectscore
worksscore
yearresscore
research-score

Paper_info
pap-sort
pap-count
hqpapercount
paperscore

Work_info
work-sort
work-count
hqworkcount
worksscore

Project_info
pro-sort
pro-count
hqprojectcount
projectscore

🟨 事实数据表 🟦 維度表

Figure 2. The snowflake structure about the data warehouse of university human resources is realized in SQLserver2005.

V. THE RELATION OF DATA MINING AND DATA WAREHOUSE

Usually, the data object be excavated must be extracted from the data warehouse or data marts. Shown as Fig. 3.

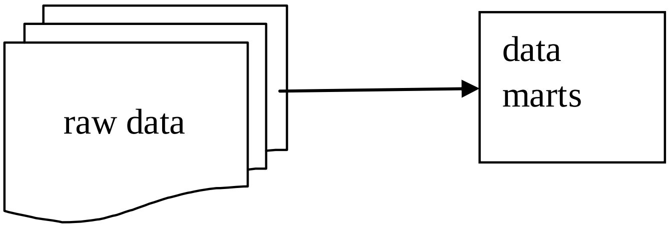


Figure 3. extracted the data sets that data mining required from the Data warehouse.

Because the two processes which tidied up the information data mining needed and tidied up the information data warehouse needed are similituded, the data that you want to dig has been a part of the data warehouse, it will be very convenient. If information was converted into the data warehouse needed, and it doesn't need to make further consolidation in excavation, at the same time it avoid for data maintenance and operation. Because the database be digged up is not usually a subset of the physical data warehouse, it is a subset of logic, the use of the database management system must be the resources type that supports data mining. Otherwise, it is better to handle the database that dig in as an independent data warehouse database.

The work to establish a data warehouse is arduous, it is a original data collection to solution the data integrity issues and diversity. The warehouse will be a vast system that uses database management system, sometimes it takes many years and a great deal of money to finish. From simple application to think, we can extracted a single database from one or more database s it were run and support transaction processing. Then we dig it! (Fig. 4), the new database is called a data mart.

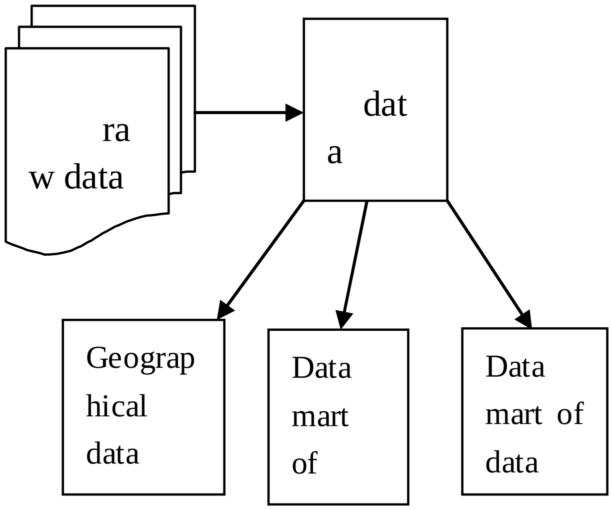


Figure 4. extracted the data sets that data mining required from real-time database.

VI. Mining Association Rules

Association analysis is a kind of commonly used data mining tasks. It used to describe the models that strongly correlated the characteristics of data. The pattern was found usually in the form of implication rules. Due to the size of search space is the index, so the goal of association analysis is from an effective way to extract the most interesting model. At early stage, the associated model mainly used for the discovery of retail transaction data analysis, carried out the placing of goods more rational, and ultimately increase sales, so sometimes the method is referred as the "basket analysis" [3].

The definition of association rules described as follows: The definition of association rules described as follows: If I= {i1, i2,i3,...im} is composed of a data collection by the m different components, in which the element is called items, a collection of items called itemsets, including the k-item itemsets known as the k itemsets, given a services (transactions) D, that is, the transaction database, each of which services (transactions) T is the data of a subset of I, that is, $T \subseteq I$, T has a unique identifier TID; If and only if $X \subseteq T$, we said set of transaction T contains X; then association rules are as the implicate form "X → Y". which, $X \subseteq I$, $Y \subseteq I$, $X \cap Y = \Phi$, it means that the X to be met ,the Y should be met. Association rules in transaction databases set up with support s (support) and a confidence level c (confindence) [4].

A. Analysis of classic Apriori algorithm for association rule mining

Apriori algorithm is the first influential algorithm which come through frequent itemsets mining Boolean association rules, and many other association rule mining algorithms all use Apriori algorithm as the core of the basic idea [5].

Apriori algorithm use a priori knowledge of frequent itemsets, through a breadth-first search strategy, using a layer-by-layer iterative search method to find all the frequent itemsets, and then accordance with the confidence level of the frequent itemsets to generate rules one by one[6]. Its pioneering use of the pruning based on the support technology, and control the exponential growth about a further set of options. the following is the process that Apriori algorithm from frequent itemsets:

```
1: k = 1
2: F: k =$\{i \mid i \in I \wedge \sigma(\{i\}) \geq N \times minsup)$ { Found all the frequent 1 - itemsets }
3: repeat
4: k=K+1
5: Ck= apiroir-gen(Fk-1) { Have a further set of options }
6: for every affairs $t \in T$, do
7: Ct=subset(Ck,t) { Identification of all the candidates belonging to t }
8: for $c \in Ct$,do
9: o(c)= o(c)+1
10 :  end for
11 : end for
12: Fk= $\{c \mid C \in Ck \wedge \sigma(c) \geq N \times minsup)\}$ { Extraction of frequent k-itemsets }
13 :until F k= $\Phi$
14 : Result= $\cup$F k
```

In which, I means the transaction database, Fk means frequent k-itemsets, Ck, means the candidate k-itemsets, σ(c) that contains the itemset c the number of records in, N means the total number of records of database I, minsup means minimum support.

The algorithm initial determine the degree of support for each item through single-pass scan the data sets. After this step, it has been a set of all frequent 1 a collection of F, (steps 1 and 2).

Next, the algorithm will use the last iteration found the frequent (k-1) set of new candidate k-itemsets (step 5). Candidate sets generated using the apriori-gen function.

In order to count the support option, the algorithm needs to scan data sets (step 6-10). When we finished calculation of support options about the candidate set, the algorithm will delete all the candidate set which support less than minsup (step 12). When there is no new frequent itemsets generated, that is, for the empty set Ft, the algorithm steps finished (step 13).

The core of association rule mining algorithm is to find frequent itemsets, this step determines ensemble performance of the association rule mining algorithm. In order to solution the problem, we often need to scan the database many times. A lot of time will be spent scanning the database and I / O operation[7]. Therefore, how to quickly and efficiently find all frequent itemsets of various association rule mining algorithms is the main issue, but also measured the various advantages and disadvantages of association rule mining algorithm standards.

The algorithm uses SQL technology directly generate frequent itemsets, and the results saved to the temporary table, omit the Apriori algorithm steps to generate the candidate sets. Such as generating title (title), degree (degree) are combined of 2 – itemsets. The SELECT statement forms can be written as follows:

```
SELECT title,degree,count(*)  FROM  basic_info
GROUP BY title,degree
HAVING count(*)>=11_mincount
```

Which is to be excavated basic_info table, 11_mincount is a minimum number of record which contented with the minimum support.

The efficiency of the frequent itemsets generated in this way is determined by the implementation of the number of SQL statements. That may be the number of combinations of different attributes from frequent itemsets. The algorithm draws on the connection thought of Apriori algorithm, connects the property portfolios of operation, and avoids the situation recurring combinations. Overall, this method makes full use of the relational database and data warehouse for minning, saves the time of calculating candidate set , reduces the algorithm's dependence on memory.

B. Realize Association Rule Mining Algorithm

Use in the procedure PowerBuilder8 .0, first of all the user entered information , we can access properties to be excavation, minimum support, minimum confidence and minimum impact. We can turn this minimum support contented the conditions into the smallest number of records, and save these properties will be excavated to is_dfxl array. Then we created the temporary tables to storage frequent itemsets, combined each property names in is_dfxl array, for a condition of divide into groups, used dynamic SQL statement to generate the frequent itemsets, and saved the results to a temporary table in the corresponding ,at last rules generated for the frequent itemsets in turn.

After generated rules , the no.1 combination of rules, can be filtrated rules that contented the conditions by calculating the confidence level affect. Through the operation of connect, we can use the rule that the latter number is k-1 in turn to generate the rule that that the number is k and content the requirements, of which 25≦k≦ll_p-1. ll_p is the current number of requent itemsets. Finally, the conversion rules for the way users can understand the output.

C. Rule Mining

According to the Uniform Code, we established code table for each of these attributes. The data were inserted into the corresponding code table.

We entered the property name to be excavated, the first time the properties of association rules mining for: age, degree, title, and then we entered each threshold. In this Mining, the minimum support is set to 0.1, the minimum confidence is set to 0.5, the smallest impact as 1.25 degrees. The results of the first excavation are shown in Table I.

Conducted for the second time, the properties of association rules mining is: gender, age, degree, title. In this excavation, we set the minimum support of 0.15, Set the minimum confidence level 0.6, minimum impact as 1.2 degrees. Mining the outcome of the second is shown in Table II.

TABLE I. THE FIRST PART OF ASSOCIATION RULE MINING AND SUPPORT, THE CONFIDENCE LEVEL

NO.	rule	support degree	confidence level	impact degree
1	Degree = Bachelor, Age = middle-aged teachers	0.13	0.64	1.26
2	Title = Intermediate, Age = young teachers	0.28	0.62	1.98
3	Title = Associate Professor, Age = middle-aged teachers	0.26	0.65	1.28
4	Title = Professor, Degree = Ph.D.	0.13	0.59	1.50
5	Degree = Ph.D., Title = Professor, Age = middle-aged teachers	0.21	0.62	1.53
6	Age=young teachers, degree = Master, title= Intermediate	0.10	0.63	2.42

TABLE II. THE SECOND PART OF ASSOCIATION RULE MINING AND SUPPORT, THE CONFIDENCE LEVEL

NO.	rule	support degree	confidence level	impact degree
1	Age = young teachers ⇨ Gender = female	0.20	0.64	1.37
2	degree = Master ⇨ Gender = female	0.19	0.61	1.30
3	Degree = Ph.D.⇨ Gender =men	0.25	0.67	1.26
4	Title = Intermediate⇨ Gender = female	0.21	0.63	1.34
5	Title = Professor ⇨ Gender =men	0.18	0.84	1.57
6	Age = middle-aged teachers，Degree = Ph.D.⇨ Gender =men	0.16	0.78	1.46

D. The interpretation of association rule

During the first excavation, from the rules of bachelor → middle-aged teachers, 13% support degree, 64% confidence level, the impact degree is 1.26, we can see that the the middle-aged bachelor degree teachers account for 13%, and 64% confidence level showed that the teachers which degree is bachelor, accounting for 64% are middle-aged teachers.

From the rules young teacher → Intermediate title, 21% support degree, 68% confidence level, the impact degree is 1.98, we can see that the titles for the Intermediate youth accounted for 21% of teachers, and 68 % confidence level indicated that the young teachers, with Intermediate title 68% of the total titles.

Analysis from such two rules, we can see the structure of teacher education where have bigger room for improvement: the degree of some middle-aged teachers is lower,the professional title of the young teachers is generally not high.

Mining in the second, a gender attribute are added. Through the following four rules: professor→ men, 18% support, 84% confidence level, the impact degree is 1.57 ; Intermediate → women, 21% support, 63% confidence level, the impact degree is 1.34.

Dr→ men, 25 % support, 67% confidence level, the impact degree is 1.26; master's degree→ women, 19% support, 61% confidence level, the impact degree is 1.30, we can see the degree of female teachers and the professional title of the level needs to be improved.

E. The revelation of association rule

Through the generation of association rules mining ,we analysed and explained the rules for building the future teachers, we have received the following revelation.

First, the approach of construction of university teachers have the main two ---- training and introduction. from the association rules on the above analysis we can found: If we introduce talent, we should as far as possible introduce high-degree or high-titles young teachers; In the development of talent we should be targeted at different types of training teachers to take a different approach.

Second, these association rules can be used to help the departments concerned to take measures to improve teachers' graduate-level, improve teachers' title structure, such as the introduction of high-degree personnel, or school staff have been sent for further training, and we can encourage teachers to take the initiative study or to pursue a degree. Thereby we can enhance the overall quality of teachers.

From this example, we felt the method used for mining association rules can excavate correlation in large amounts of data that is no possible to be found in surface .These rules will help decision-makers to make further reasonable adjustments for the construction of teaching faculty.

VII. CONCLUSIONS

This paper applies mining association rules to universities for human resources system, and analyses the relationship from the basic information of teachers. Human resources data warehouse in university has been pooled and integrated the basic information of the teachers year by year. In this paper, we used the data in fact tables of one year, applied the betterment association rule mining algorithms, mining the relationship among the basic properties and their mutual influence, such as teacher's age, title, degree, gender, etc. We found the teaching faculty has a number of problems in the age structure, degree structure, teacher's titles. Then we provided reference and basis for a more rational organization of the use of teachers to provide a measure of resources, so as to achieve the optimization of teaching faculty.

REFERENCES

[1] William H.Inmon, Building the Data Warehourse, Fourth Edition, John wiley & sons. 2006.8,21-27.

[2] Han J,et al. Generalization-based data mining in object-oriented databases using an object-cube model.Data and Knowledge Engineering.2003,Vol.25:55-97

[3] S. Papadimitriou, A. Brockwell, C. Faloutsos. Adaptive, unsupervised steam mining. VLDB Journal.2004.1,3(3):222-239.

[4] E. Omiecinski. Alternative Interest Measures for Mining Associations in Databases. IEEE Trans. On Knowledge and Data Engineering,2003.1，5(1):57-69.

[5] S. Shekhar, C. T. Lu, P. Zhang. A Unified Approach to Detecting Spatioutliers. Geoinformatica, 2003.7,7(2):13-16.

[6] Liu Peng, A unified strategy of feature selection, LNCS,2006(4093):457-464

[7] Ian H., Frank E., Data Mining : Principles Machine Learning Tools and Techniques(Second Edition),Morgan Kaufmann,2005

Zhang Danping was born in Shanghai in 1967, graduated from Jiangxi Normal University at 1988. She obtained a master's degree in management science and engineering.

The main areas of her research is data mining, information management and information system design. She is a professor of economics and management College in Nanchang Hangkong University, engages in the teaching and scientific research, head of Information Management and Information System professional. She has presided over more than ten topics such as provincial and ministerial level. In 2000 she won the title of outstanding teachers in Nanchang City. In 2004, she published a textbook as editor. In recent years, she published nearly 20 academic theses, including 8 papers published in the core journals, EI retrieve 1, ISTP retrieval 1.

Professor Zhang is an expert in Jiangxi Province Logistics Association.

Deng Jin was born in Nanchang in 1966, graduated from Beijing University of Aeronautics & Astronautics at 1988., She obtained a master's degree in information management.

The main areas of her research is human resource management, information management and information system design.etc. She is a professor of Nanchang Hangkong University, and she is HR management of Nanchang Hangkong University. she published 2 works and more than 30 academic theses.