ATLANTIS PRESS
Advances in Computer Science Research, volume 44
3rd International Conference on Wireless Communication and Sensor Network (WCSN 2016)

Application of the Data Mining in the University Human Resource Management

Li Yi
MIS sciences, Business School, Donghua University, China
MIS sciences, Business School, Shanghai Jian Qiao university, China
Shanghai, China
Email: 13585735718@163.com

Yao Wei-Xin
MIS sciences, Business School, Donghua University, China
Shanghai, China
Email: 13585735718@163.com

Abstract-Competitions among different universities are the competition of talents. The content is about how to obtain talents, how to properly allocate talents, how to comprehensively obtain information of human resources and promote the level of human resource management of universities. The paper used cluster analysis and decision tree, etc tools tries to solve problems existed in the university human resource management such as talents division, brain drain by using data mining tools. Through the reference case, the paper uses ten teachers' data as example, extracts information from hidden information and comes to the conclusion on teacher talent division; the conclusion can help the use and allocation for talents in universities and colleges.

Keywords-Data mining; Human Resource; Maximal tree.

I INTRODUCTION

A. *Research Background*

Colleges and universities in the 21st century is no longer a general sense of teaching and scientific research units, has become a national innovation system and the main force of developing the country by relying on science and education, for the talents cultivation, scientific research and social services.

B. *Research Significance*

The traditional method of the university human resource management is to accumulate the surface information from daily management so as to implement decisions, which results in uneven salary allotment, invalid performance appraisal, low satisfaction of staff and then causing brain drain.

Therefore, we need to adopt hidden information from the analysis of the surface data by using the related theories of data mining. We make sure that human resource plays an important role and we will achieve the goal of promoting universities' comprehensive strength.

C. *Problems Existed in the Process of the University Human Resource Management*

University human resource owns the basic characteristics of the human resource, but it also has unique characteristics. They are mainly reflected in the following 3aspects[4]:

①strong self-awareness

②relatively strong achievement motivation

③strong desire of mobility

The inappropriate university human resources allocations will influence the overall efficiency of the unaiversities. Universities have the shortage of scientifically rational long-term planning of the human resource management. The total quantity of human resources and the hierarchical structure are not reasonable. Lacking of the scientifically rational long-term planning of the human resource management, the work of the university human resource management can't provide guarantee to the long-term development of the universities.

II DATA MINING TECHNOLOGY IN THE APPLICATION OF THE UNIVERSITY HUMAN RESOURCES MANAGEMENT

With the increase of university informatization level, universities possess relativelycomplete information management system and adequate basic data of human resources. We can choose, pre-process and change the human resources data and then make data provisions for data mining.

A. *Divide the Talent Type*

In universities, grasping the universities' personnel constitution and types and judging the faculty belong to which kinds of personnel are essential to personnel selection and draw up talents development strategy. The data mining technology enables us to obtain human resources information from several data base of the related workers' working condition and find their relation and mode, and then it reflects the internal personnel constitution objectively.

B. *Pattern Introduction*

By using the mode, we can find the types of the talents and determine the type of the teacher.

In the record of the data warehouse, we can set up unspecified sample collections H0 and name all the unspecified objects as samples, such as h1,h2,…,hn,H={h1,h2,…,hn} are sample collections. In order to get the reasonable sample classification, we should quantify their concrete attributes. The attribute of quantification can be called as sample index. If there are m indexes, we can use m dimensional vector to describe sample, that is:

hi=(hi1,hi2,…,him) (i=1,2,….,n)

As the data we gathered in actual data are not on the closed interval of [0,1], we need to standardize the raw data and seek its average. For example, there are n samples in sample collection;we can get n data for one index k of the sample. h'1k,h'2k,…,h'nk refers to data of the i sample obtained from k indexes. Their average can be calculated according to the Formula 1:

$$h'_{\bar{k}} = \frac{h'_{1k}+h'_{2k}+\cdots+h'_{nk}}{n} = \sum h'_{ik}/n \quad k=1,2,\ldots,m \qquad (1)$$

We can get the standard deviation Sk of raw data according to the Formula 2 .

$$s_k = \sqrt{\frac{\sum_{i=1}^{n}(h'_{ik}-h'_{\bar{k}})^2}{n}} \qquad (2)$$

We can get the standardized value of each data according to the Formula 3.

$$h''_{ik} = \left|\frac{h'_{ik}-h'_{\bar{k}}}{s_k}\right| \qquad (3)$$

If the standardized data h"ik we get is not on the closed intervals of [0,1], we should adopt the following Formula4 of extremism standardization:

$$h_{ik} = \frac{h''_{ik}-h''_{mink}}{h''_{maxk}-h''_{mink}} \qquad (4)$$

h''_{maxk} and h''_{mink} refers to the maximum and the minimum in h"1k, h"2k, …h"nk and h"nk.

Establish a pattern similarity relation R, general form as follows Formula5:

$$R = \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nn} \end{bmatrix} 0 \le r_{ij} \le 1; i = 1,2 \ldots n; j = 1,2 \ldots n \qquad (5)$$

There are several methods to calculate rij. We adopt the minmaxmethod, which is as Formula6.

$$r_{ij} = \sum_{k=1}^{m} \min(h_{ik}, h_{ij}) \Big/ \sum_{k=1}^{m} \max(h_{ik}, h_{ij}) (i,j \le n) \qquad (6)$$

III. EXAMPLE

Choose the annual assessment data base of the faculty of the university and add performance index and evaluation result (39 fields) on the basis of information base of the faculty of the university (30 fields). The added 9 fields assessment indexes are: observing discipline, enterprising spirit and sense of responsibility, organizational coordination ability, knowledge level, innovation ability, human competencies, teaching quality, teaching performance.

Choose target factors and candidates that you need to consider and build independent mining database. Select four targeted factors: teaching quality, knowledge level, productive capacity and teaching performance as show in Table1 and Table2.

TABLE I. THE ORIGINAL DATABASE

Name	Gender	Professional title	Education	..	Assessment index								
					Teaching attitudes			Working ability			Teaching performance		
					Enterprise	…	Follow the rules	The level of knowledge	…	Working ability	Quality of teaching	…	Teaching performance
Bree	Female	Assistant teacher	Bachlor	..	4	…	5	3	…	3	4	…	5
Mike	Male	Lecture	Master	..	5	…	4	5	…	3	4	…	4
Susan	Female	Assistant professor	Master	..	6	…	5	4	…	5	5	…	5
…	…	…	…	..	…	…	…	…	…	…	…	…	…

TABLE II. TEN TARGET INFORMATION OF TEACHERS

Alternative offer	Knowledge Level	Working ability	Teaching quality	Teaching performance
Teacher 1	3	2.5	4	3
Teacher 2	2	3.5	3	2
Teacher 3	1	1.5	4	1
Teacher 4	4	4	4	3
Teacher 5	5	3	3	4
Teacher 6	6	3.5	3	4
Teacher 7	5	2.5	2	3
Teacher 8	4	1.5	5	2
Teacher 9	3	0	4	2
Teacher 10	4	3.5	3	2

The scoring criteria of knowledge level:

very strong 6, strong 5, relatively strong 4, medium 3, generally strong 2, relatively weak1.

The scoring criteria of productive capacity:

very strong 4, strong 3.5, relatively strong 3, not too strong2.5, generally strong 1.5, relatively weak0.

The scoring criteria of teaching quality:

Very good 5, good 4, relatively good 3, generally good 2, not good 1.

The scoring criteria of teaching performance:

very good 4, good 3, relatively good 2, generally good 1.

There are 10 objects that has been divided. Domain of discourse: U={T1,T2,T3,T4,T5,T6,T7,T8,T9,T10}

We can get the average vectors from Formula1: h'k={3.7,2.6,3.5,2.6} k=1,2,3,4

We can get the normal vectors from Formula 2 : sk={1.42,1.17,0.81,0.92} k=1,2,3,4

We can get the standard matrix (a) from Formula 3.

We can get the normalized matrix (b) Formula 4.

$$\begin{vmatrix} 0.49 & 0.04 & 0.62 & 0.44 \\ 1.20 & 0.81 & 0.62 & 0.65 \\ 1.90 & 0.90 & 0.62 & 1.75 \\ 0.21 & 1.24 & 0.62 & 0.44 \\ 0.92 & 0.38 & 0.62 & 1.53 \\ 1.62 & 0.81 & 0.62 & 1.53 \\ 0.92 & 0.04 & 1.86 & 0.44 \\ 0.21 & 0.90 & 1.86 & 0.65 \\ 0.49 & 2.18 & 0.62 & 0.65 \\ 0.21 & 0.81 & 0.62 & 0.65 \end{vmatrix} \quad \begin{vmatrix} 0.17 & 0.00 & 0.00 & 0.00 \\ 0.58 & 0.36 & 0.00 & 0.17 \\ 1.00 & 0.40 & 0.00 & 1.00 \\ 0.00 & 0.56 & 0.00 & 0.00 \\ 0.42 & 0.16 & 0.00 & 0.83 \\ 0.83 & 0.36 & 0.00 & 0.83 \\ 0.42 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.40 & 1.00 & 0.17 \\ 0.17 & 1.00 & 0.00 & 0.17 \\ 0.00 & 0.36 & 0.00 & 0.17 \end{vmatrix}$$

(a) (b)

Set up the fuzzy similar matrices (c) :

$$\begin{vmatrix} 1.00 & & & & & & & & & \\ 0.15 & 1.00 & & & & & & & & \\ 0.07 & 0.46 & 1.00 & & & & & & & \\ 0.00 & 0.24 & 0.16 & 1.00 & & & & & & \\ 0.12 & 0.42 & 0.59 & 0.09 & 1.00 & & & & & \\ 0.08 & 0.55 & 0.84 & 0.16 & 0.70 & 1.00 & & & & \\ 0.12 & 0.20 & 0.18 & 0.00 & 0.17 & 0.14 & 1.00 & & & \\ 0.00 & 0.25 & 0.17 & 0.23 & 0.12 & 0.17 & 0.50 & 1.00 & & \\ 0.13 & 0.40 & 0.25 & 0.42 & 0.22 & 0.26 & 0.07 & 0.24 & 1.00 & \\ 0.00 & 0.48 & 0.22 & 0.49 & 0.20 & 0.26 & 0.00 & 0.34 & 0.40 & 1.00 \end{vmatrix}$$

(c)

A. *Clustering Analysis*

The maximal tree clustering method is adopted,to build a special image and take the classified objects as vertex. When rij≠0, connect the vertex “i” and the vertex “j”, they can form a line. The concrete practice is to draw the vertex “i”first and then connect each vertex according to the sequence of rij from big to small with no loop until all the vertexes are connected. In doing so, we can get the maximal tree. Actually, it is a tree with empowerment. You can get weight from each side, which is rij. As the connection is different, the maximal tree is not the only one.

Then you can get cut set from the maximal tree, that is to cut the edge rij<λ of those weights, λ∈[0,1]. In doing so, you get cut the tree into several subtrees without connection. Even though the maximal tree is not the only one, you can get the same subtree after the cut set, there sub numbers are modes that concluded from the data warehouse.

After analysis, we can get Figure1:

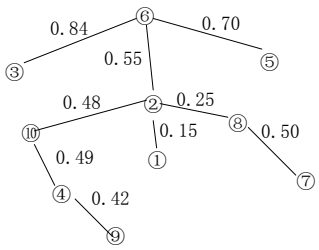

Figure1. Talent classification diagram

Take λ=0.84, divide it into 9 categories: {T3,T6}, the rest belongs to one category.

Take λ=0.48, divide it into 5 categories: {T2,T3,T5,T6},{T4,T10},{T1},{T7,T8},{T9}.

Take λ=0.25, divide it into 3 categories: {T7,T8} (the third category teachers with weak capacities.),{T1} (the second category teachers with weak capacities.),the rest belongs to the first category teachers with strong capacity.

Take λ=0.15, all belongs to one category.

Take λ=0.25, calculate the average index of different modes as shown in Table 3.

TABLE III. EACH MODEL AVERAGE INDEX

Alternative offer	Knowledge Level	Working ability	Teaching quality	Teaching performance
Third classified teacher	0.75	0.50	0.70	0.625
Second classified teacher	0.50	0.625	0.80	0.75
First classified teacher	0.60	0.68	0.69	0.64

B. *Prediction*

For each of the modes, we can seek its average index according to the following Formula7:

$$\mathrm{Mode}_{ij} = \sum h_{ik}/p \quad i=1, 2\ldots, s, j=1, 2,\ldots, m \tag{7}$$

"s" stands for the total modes. "k"stands for the mode (which is the mode in i)is deduced from which of the records in the data warehouse. "p"stands for the total quantities of records deducing from the mode.

The sample X(X1, X2,…,Xn)is n fuzzy subsets of the mode in the domain of discourse X. Compare it with the classification of the modes in the data warehouse and calculate the close degree between them as Formula 8.

$$(X, \mathrm{Mode}_i) = (1/2)[X \cdot \mathrm{Mode}_i + (1 - X \odot \mathrm{Mode}_i)] \tag{8}$$

("·"stands for inner product in fuzzy operation; "Θ"stands for outer product in fuzzy operation.)

According to the principle of selecting the near, that is Formula 9:

$$(X, \mathrm{Mode}_i) = \max(X, \mathrm{Mode}_1), (X, \mathrm{Mode}_2), \ldots, (X, \mathrm{Mode}_s) \tag{9}$$

Determine the mode of which sample it is close to and predict the result from the whole condition of the mode.

For example, examine the information of a new teacher *X*= (4,3,3,2)

For the third category: inner product=0.67, outer product=0.625, close degree=0.523

For the second category: inner product=0.625, outer product=0.67, close degree=0.48

For the first category: inner product=0.68, outer product=0.64, close degree=0.52

We can get that it has the maximum close degree with the third category. The teacher belongs to the category that has a weak comprehensive ability.

IV CONCLUSION

People, money and commodities are three resources of school in the conventional sense. As time progressed, information has become an invisible resource. Data mining technology uses contents hid in information to manage university human resources, which enables human-based management in human resource management to combine with specific data in order to divide personnel types preferably and avoid losses brought by brain drain.

REFERENCE

[1] Ian H.Witten,EibeFrank.DataMining:*Practical Machine Learning Tools and Techniques*,Second Edition[M].NewYork:MorganKaufmann,(2005).

[2] Yu-feng Zhao,Li-yun He,Bao-yan Liu,Jun Li,Feng-yi Li,Rui-li Huo,Xiang-hong Jing. Syndrome classification based on manifold ranking for viral hepatitis[J]. Chinese Journal of Integrative Medicine . (2014)

[3] TAO Dao-qiang1,MA Liang et al.The decision tree algorithm based on classification matrix. Computer Engineering and Design, 2309-2313(2012).

[4] Ye yanfeng,Why brain drain voluntarily. Operators and managers,9(38) (2003).

[5] Xiaoyue Wang,Abdullah Mueen,Hui Ding,Goce Trajcevski,Peter Scheuermann,Eamonn Keogh. Experimental comparison of representation methods and distance measures for time series data. Data Mining and Knowledge Discovery . 2 (2013).