Analysis of Association Rule in IT Job Requirement

-Example of 1111 HR Service Provider

Wen-Hsing Kao, Jason C. Hung, Wei-Jen Cheng , and Chi-Ching Lin
Department of Information Technology
The Overseas Chinese Institute of Technology
No:100, Chiao Kwang Rd., Taichung 407, Taiwan, R.O.C.
Email: star@ocit.edu.tw, jhung@ocit.edu.tw, s9618209@ocit.edu.tw

Abstract *- As the evolution of the Internet application continues, people now are enjoying the convenience and user-friendly interface of career engagement from the human resource service providers. In this research, we are proudly having the opportunity to cooperate with 1111.com.tw, one of the most famous HR service providers in Taiwan, and to mine its valuable database. The main technology of data mining discussed in this paper is association rules and the purpose of this research is to find out the associated rules in employees' or job seekers' computer skills and the IT job requirements. Therefore, we can provide the real trend of the IT job requirements and give better directions to students and the people who are seeking jobs.*

Keywords: Data Mining, Association Rules, Human Resource, Dependency Network.

1. Introduction

Data Mining is widely used in many fields nowadays and it becomes the common sense to find out some undiscovered knowledge form a well-formatted and well-organized database. By using this, we can always learn the rules of how decisions were made and even predict the future behavior of an object. Association rule mining has become one of the most popular methods of data mining. In this project, we focus on human resource field and try to discover the potential valuable knowledge of the reaction between employees and employers through the platform. In order to present more reliable and accurate result, we start from filtering out the unwanted and incomplete data as for pre-processing of the database. Secondly, we analyze the job requirement for IT field by using association rule mining and obtain the relationship of job requirements and computer skills. Among the data of the relation, we figure out the useful support, and dependency network. Finally, we expect that the result can reflect the real world job trend and provide references for schools, students, and job seekers from this research.

2. Literature Review

2.1. Introduction of Data Mining

Basically, data mining can be explained as knowledge discovery in database. In other words, it can be defined as extracting interesting knowledge from lots of data in a large database which could be an online database or a data warehouse. According to different requirements, people can use different mining models for their specific researches. There are three types of the models including association rules, grouping, and classifying while we focused on association rules.

2.2. Association Rules

Association rules data mining is one of the most popular and well researched methods which is to find out interesting and valuable knowledge that is implicit in large databases. It was first announced by Agrawal in 1994 and the most famous algorithm is Apriori (Agrawal, Imielinski, and Swami, 1993) [1]. The main function of the algorithm was to figure out the relation in each transaction record of the whole business database. Meanwhile, its candidate was based on support and confidence. For example, we can find out rules in the selling database of a computer shop like following: 80% of the customers purchased printer cartages will also buy printing paper. 80% of the customers purchased desktop computers will also buy monitors.

The definition of association rule mining is: Let I equals a set of item selling in a shop. In the database, each transaction is including a unique transaction ID and a group of purchased items while the group of items is called an itemset. Assume that X is an itemset and all of the items in X are included in a transaction called T then we can say that T support X. The support count of an itemset X is defined as the transaction amount of the supported itemset X while the support of an itemset X is defined as the ratio of the itemset X transaction amount to the total transaction amount.

The form of association rules is:

$X \rightarrow Y$ [support, confidence] where X and Y are both represented as itemsets plus we call X as antecedent and Y as consequent.

The support of rule $X \rightarrow Y$ is defined as the support of the itemset $X \cup Y$ and the confidence of rule $X \rightarrow Y$ is defined as the ratio of the transaction amount that meets with both antecedent and consequent to the transaction amount that meets with antecedent.

$$Confidence(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)}$$

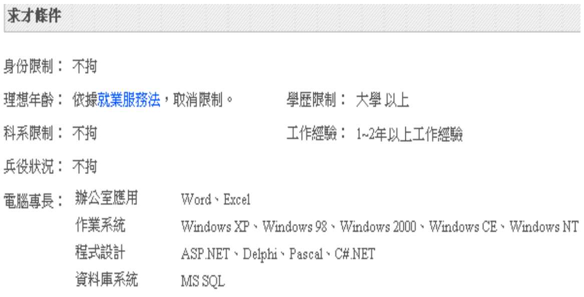
The support and confidence we mentioned above are the evaluating standard of the association rules and they are used for the judgment of the rules' satisfaction. If they are set too high, we will not be able to get rules or lose some important rules whereas if they are set too low, there might be many messy and unreliable rules coming out. How to set the values is depended on the experience of the people who try to analyze the data so the main target of the algorithm is to find out the rules that are satisfied to the minimum support and the minimum confidence. And it can be divided into two steps:

1. To discover all frequent itemsets.
2. To produce the associated rules by using the frequent itemsets.

3. Method

3.1. Motivation and purpose

Currently, the HR service provider is listing all of the job requirements as shown below.



求才條件

身份限制： 不拘

理想年齡： 依據就業服務法，取消限制。　學歷限制： 大學 以上

科系限制： 不拘　工作經驗： 1~2年以上工作經驗

兵役狀況： 不拘

電腦專長： 辦公室應用　Word、Excel
作業系統　Windows XP、Windows 98、Windows 2000、Windows CE、Windows NT
程式設計　ASP.NET、Delphi、Pascal、C#.NET
資料庫系統　MS SQL

Figure 1. 1111 job requirement interface

In order to promote the effectiveness of job matching and help job seekers to get proper careers, there should be some information showing what kind of skills they need to have in certain fields so that they can learn or improve the skills and increase the opportunities of obtaining a suitable job. On the other hand, employers can shorten the time of finding suitable people to fit their needs.

To schools, teachers can teach certain subjects and train students to fit the nearby industrial area companies' needs while to students, they can always know what are the popular skills in certain careers that they are heading forward to. The achievement is to connect employers, job seekers, schools and students and to provide a better communication among them.

3.2. Procedure

The mining procedure is shown below and the descriptions of its 6 steps are followed.

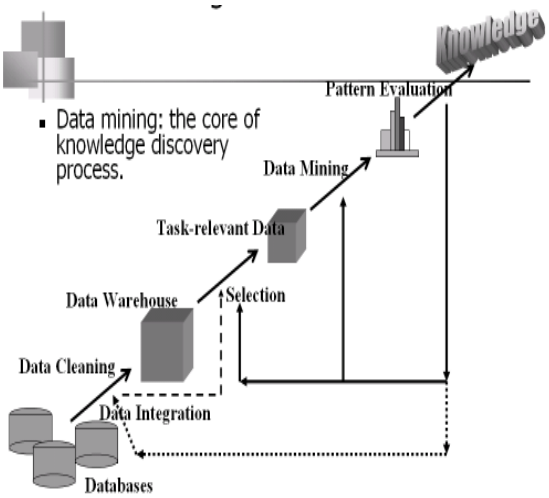


Figure 2. Data mining procedure

3.2.1. Data Collection. Just like other applications, we need to collect data and store the information in a database for pre-processing.

3.2.2. Data Pre-Processing. In order to obtain higher quality of the mining result, we need to filter the raw data and get rid of unwanted and inaccurate data records. Usually, it takes 80% of whole processing time to do data integration, data processing, data conversion, and data simplifying in this step.

3.2.3. Data Warehouse. Integrated data, detailed data, and historical data are stored for On-Line Analytical Processing, multidimensional structuring, and statistical processing.

3.2.4. Data Mining. Mining potential, precise, and useful knowledge is the purpose of this step. First of all, we use tools and figure out the relation in every job requirement transaction to create a model. Then, we evaluate the accuracy, quality, explainability, healthiness, and expansibility of the model. Finally, we discover the usage and application of the association rules.

3.2.5. Pattern Evaluation. We evaluate the discovered knowledge and filter out the useless one so the rest of the information will be valuable.

3.2.6. Result Demonstration. It is not easy for people to understand the complex mined result so we need to make the interesting knowledge and the useful information demonstrated as a chart or a table.

4. Practice and Analysis of Association Rule

4.1. Practice Procedure

The objects of online human resource service provider are mainly from the people use the Internet so it differs from traditional job seeking way and the job requirements also different. Therefore, the purpose of this research is to analyze the relations in computer related skills and job requirements.

4.1.1. Data Collection. The main database contains the data that was collected in July, 2008 is provided by 1111.com.tw and it includes totally 9076 data records in the table of jobs, employers, and job seekers.

4.1.2. Data Pre-Processing. We filtered out the necessary data and then converted the data format plus selected the necessary columns for data pre-processing.

1. The filtering code was developed by Microsoft Visual C# 2008 and there were 1462 records of filtered data.

2. According to the input requirement of Microsoft SQL Server 2005 association rule model, we converted the raw data from horizontal listing to vertical listing for data mining.

3. The data columns related to computer skills were selected and then the data in those columns were joined and encoded.

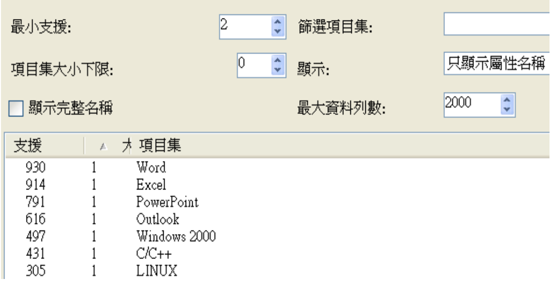
4. Some inaccurate and unwanted data like null value were filtered out in order to get a precise analyzed result.

4.2. Practice Result and Analysis

Under the circumstances of getting all required data and devices, we set the computer skills for IT job requirement as the execution properties and we received 1462 records of filtered data. We used the model of SQL Server 2005 association rule to calculate the filtered data and finally obtained the results as following.

.

1. Support count analysis: as the Figure 3 showing below, we can see that the order of the basic skills for IT job requirement is Word, Excel, and PowerPoint. Besides, the abilities of operating Windows 2000, LINUX and coding in C/C++ are also necessary to advanced skills.



最小支援: 2

篩選項目集:

項目集大小下限: 0

顯示: 只顯示屬性名稱

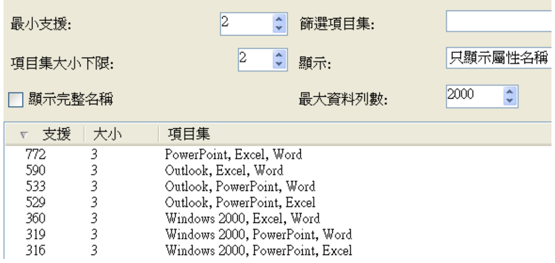
□ 顯示完整名稱

最大資料列數: 2000

支援	大	項目集
930	1	Word
914	1	Excel
791	1	PowerPoint
616	1	Outlook
497	1	Windows 2000
431	1	C/C++
305	1	LINUX

Figure 3. Association rule itemset 1

2. Job requirement association analysis: as the Figure 4 showing below, we know that the most basic skills of IT job requirement is the ability of operating Microsoft Office and Windows 2000.



最小支援: 2

篩選項目集:

項目集大小下限: 2

顯示: 只顯示屬性名稱

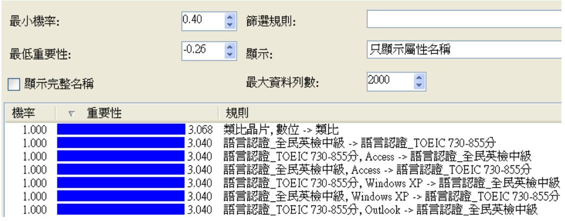
□ 顯示完整名稱

最大資料列數: 2000

支援	大小	項目集
772	3	PowerPoint, Excel, Word
590	3	Outlook, Excel, Word
533	3	Outlook, PowerPoint, Word
529	3	Outlook, PowerPoint, Excel
360	3	Windows 2000, Excel, Word
319	3	Windows 2000, PowerPoint, Word
316	3	Windows 2000, PowerPoint, Excel

Figure 4. Association rule itemset 2

3. Primary and secondary analysis: as the Figure 5 showing below, we can see the result of association rule produced by Apriori algorithm. The result is telling that people with PC hardware repairing skill will be more competitive if they have certificates of English language plus the abilities of operating Access, Windows XP, and Outlook.



最小機率: 0.40
篩選規則:
最低重要性: -0.26
顯示: 只顯示屬性名稱
□ 顯示完整名稱
最大資料列數: 2000

機率	重要性	規則
1.000	3.068	類比晶片, 數位 -> 類比
1.000	3.040	語言認證_全民英檢中級 -> 語言認證_TOEIC 730-855分
1.000	3.040	語言認證_TOEIC 730-855分, Access -> 語言認證_全民英檢中級
1.000	3.040	語言認證_全民英檢中級, Access -> 語言認證_TOEIC 730-855分
1.000	3.040	語言認證_TOEIC 730-855分, Windows XP -> 語言認證_全民英檢中級
1.000	3.040	語言認證_全民英檢中級, Windows XP -> 語言認證_TOEIC 730-855分
1.000	3.040	語言認證_TOEIC 730-855分, Outlook -> 語言認證_全民英檢中級

Figure 5. Association rule result

4. Dependency network analysis: as the Figure 5 showing below, we can see the relations of items represented as the lines and the more lines an item has the more relations it builds. It also presents the importance of the items

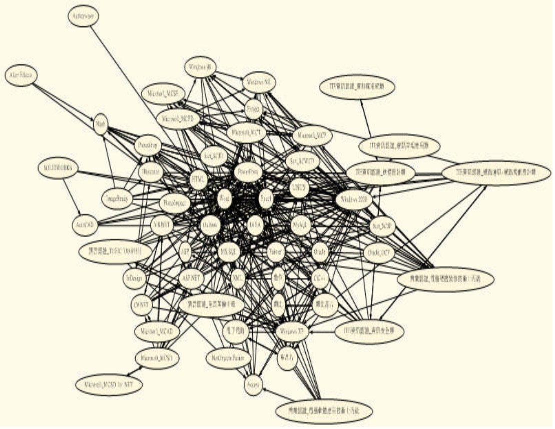


Figure 5. Dependency network

5. Conclusions

In this paper, we used the model of association rule mining and analyzed the special rules and relations of IT job requirements in the human resource database. Therefore, we got the conclusion describing below.

The most basic job requirement for a computer software programmer is the ability of C/C++. Thus, if a job seeker also has the certificates of SCJP and SCWCD, he or she will be more competitive and easier to win a programming job.

To the people search for a job related to database administration, the skill of operating MS SQL server along with the Microsoft MCPD, MCT, and MCP certificates is most wanted. It also tells that the Microsoft SQL server is taking the database market in Taiwan.

HTML has become a must and associated with the ITE (Information Technology Expert) certificates of information security and network communication to the job related to web design. This means a web designer has to take care not only the visual design but also the network security.

To be a graphic designer, one will need the skill of Photoshop basically. With the skill of Illustrator, InDesign, and Windows XP will be a plus. We also find out that MySQL is required for some jobs and this means a graphic designer with database administrating ability will be more competitive in this field.

According to the result above, schools can arrange courses in a practical and effective way while students can prepare for their future needs in an accurate and competitive way.

The direction of further research is to use other columns in the database and associate with other data mining technology so that we can continue follow up this subject and discover more interesting knowledge. The ideas will be like following:

1. Discover the association rules among working experience, job location, business size, and job requirement.

2. Extend the working model and technology to related position or occupations.

3. Continue the analysis of job requirement association rule and compare the results of certain period in order to obtain the trend of human resource.

References

[1] Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules.IBM Research Report RJ9839, IBM Almaden Research Center.

[2] Berry, M. & Linoff, G. S. (2000). Data Mining: Concepts and Techniques. John Wiley and Sons, INC.

[3] Fayyad, (1996) U.M., "Data Mining and Knowledge Discovery: Making Sense out of Data," IEEE Expert, Vol.11, No.5, pp.20-25.

[4] Frawley, W.J., Paitetsky-Shapiro, G. and Matheus. C. J. (1991), "Knowledge Discovery in Databases: An Overview," Knowledge Discovery in Databases, California, AAAI/MIT Press, pp.1-30

[5] Grupe, F.H. and Owrang(1995), M. M., "Database Mining Discovering New Knowledge And Cooperative Advantage," Information Systems Management, Vol.12, NO.4, pp26-31.

[6] J. Han and M. Kamber (2001), Data Mining: Concepts and Techniques,Morgan Kaufmann Publishers.

[7] Mintzberg, H.,(1973), The Nature of Managerial Work NY: Harper & Row.

[8] Quinn, R. E., and Camerson, K., 1988, Paradox and Transformation: Toward a Framework of

Change in Organization and Management, Cambrige, Mass: Ballinger.

[9] Robbins, S.P., (1993), Organizational behavior, Englewood Cliffs, NJ: Prentice-Hall.

[10] Usama Frayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth(1996), knowledge Discovery and Data Mining: Towards a Unifying Framework, Proceeding of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, August., AAAI Press.

[11] Von der Embse(1973), T.J. "Choosing a management development program: a decision model." Personnel Journal, 907-912.