Mining System in HR: A Proposed Model

Abeer Fatima
School of Engineering and IT
Manipal University
Dubai,UAE

Sophia Rahaman
School of Engineering and IT
Manipal University
Dubai,UAE

Abstract— **Today it is a widely accepted fact; the most important resource of an organization is the human resource. One precondition of an efficiently working and profitable organization is to have an adequate human resource strategy. The HR department deals with abundant employee information and it is critical to handle it in the right manner so that it reflects and aids in sound decision making. Data mining gives organizations the opportunity to handle the enormous amount of data with adequate efficiency to distil useful information in order to achieve greater and efficient results. This paper aims to study the role of HR in an education environment, and identify the data mining techniques that can be applied in this field. It also analyzes the benefits reaped specific to employee hiring, performance, training and skills sets for sound managerial decision making. The paper proposes a model for application of data mining in a HR process with respect to talent management, recruitment, manpower planning, and employee engagement to ease and benefit the HR system in an educational environment.**

Keywords- ***education, human resource, talent management, recruitment, manpower planning, data mining, knowledge discovery***

I. INTRODUCTION

The higher education institutions have a unique position in the society. These are areas of particular importance due to the production, dissemination and sharing of knowledge. In addition to these conventional associations between the universities and knowledge, higher education institutions have the unique potential to promote the synthesis and integration of different types of knowledge and to amplify the knowledge application in order to enhance social change. Higher education besides providing new knowledge to those who study it, settles out the learnt knowledge and gives to the future graduates the acquired skills and competencies in the specialty fields, the ability to think in a complex way and valorize superiorly the whole educational baggage which is acquired in social benefit [1].

HR is responsible for a comprehensive set of managerial activities and tasks concerned with developing and maintaining a workforce-human resource. HR aims to facilitate organizational competitiveness; enhance productivity and quality; promote individual growth and development; and complying with legal and social obligation. Besides that, in any organizations, they need to struggle effectively in terms of cost, quality, service or innovation. All these depend on having the right people, with the right skills, deployed in appropriate locations at appropriate points in time [2].

One precondition of an efficiently working and profitable company is that it has to have an adequate human resources strategy. This includes the procurement (recruiting, selecting and launching), management (manpower development, performance appraisal and career planning) and the 'drain' (reducing staff, retirements, etc.) of the human resources. From the cost effectiveness point of view it is equally important that the man-power should be well-skilled and performance-orientated, thereby making the organization profitable. An important point of view that the least possible cost should be incurred for manpower-development and recruiting, to save the costs. One of the important decisions for human resources managers is that these two mutually exclusive factors should be optimized to bring out the maximum profit and performance in an organization [3].

II. MOTIVATION OF RESEARCH

Academic institutions are universities and different forms of educational institutions engaged in higher education management and delivery. Therefore, these institutions are in need of an integrative discipline for studying, research and learning about the knowledge assets - human intellectual capital and technology. In the past decades, especially in least developed countries, these institutions worked in a relatively stable environment, and seemed isolated from much competitive pressure. However, the global environment has changed so drastically that the decision and operation processes of academic institutions have become more volatile and dynamic than ever [4]. Today higher education institutions must also think like businesses because of the increase in competition and globalization. One of the main components in any educational institutions is the faculty of various schools and universities. The faculty is an integral factor that adds value to the development of the institutions and imparts valuable knowledge to the students.

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 university full-time faculty. Due to 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 [1]. Data mining analysis is to support decision-making to aid the human resource department to make efficient and quick decisions. An organization is more likely to be successful if it manages its entire resources well, including its people whose value is enhanced by development [5]. This research paper aims at studying the HR specific to the educational environment and brings out the role of data mining in achieving quality enhanced development in its employees.

III. METHODOLOGY

The study adapts a case based analysis approach for identifying the requirements of HR in an educational environment and the suitable data mining techniques. The methodology is as follows:

- *Step 1: Identify suitable cases*

  Selecting cases for human talent prediction, manpower planning and recruitment

- *Step 2: Case Study Analysis*

  The cases identified are analyzed to understand the applications of data mining in HR. A detailed analysis of the techniques used and their findings were identified.

- *Step 3: Identify Data Mining Techniques*

  Various techniques adapted to the case identified were analyzed and the working and application with respect to each algorithm were studied in detail for a good understanding of its working which also included identifying the preprocessing of data that is required for application of the specific technique.

- Step 4: Analyze application of data mining to the case

  In this step the results of data mining to the various cases have been studied to understand the working, application, benefits and limitations of the selected techniques. The areas and situations for which specific techniques were understood.

- *Step 5: Proposed Framework*

  Develop HR model: A suitable HR process flow has been developed at this stage which would represent the major functions of the HR in an educational environment and the relationships between those functions.

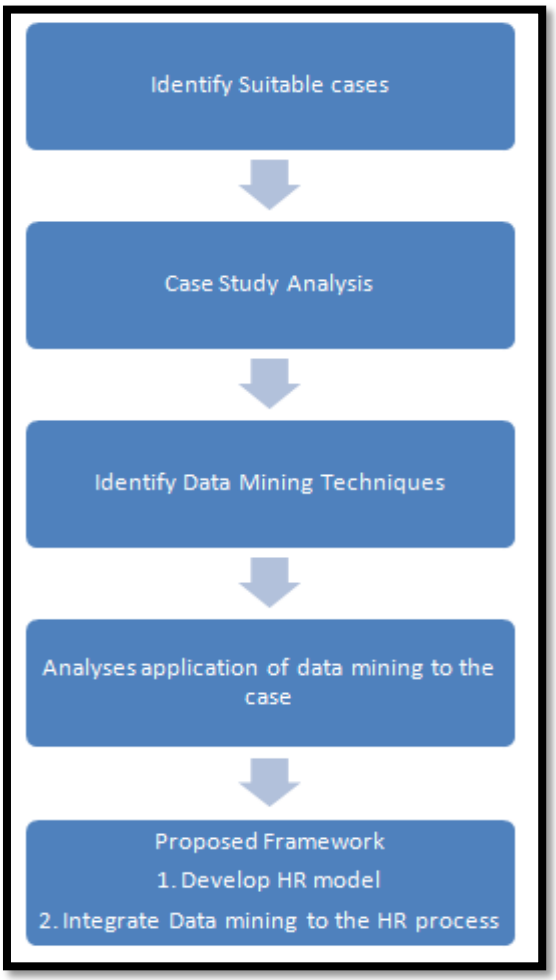

Figure 1. Methodology

IV. DISCUSSION

A. Case Overview

- *Human Talent Prediction using C4.5 Algorithm*

  This study attempts to discover employees' performance patterns from the existing employees' performance data using the decision tree classification techniques. The techniques are chosen based on the common techniques for classification and prediction in data mining. The decision tree is a 'divide-and–conquer' approach from a set of independent instances. The case uses C4.5 classifier which comes from the decision tree. The input variables for the process are performance factors for selected attributes; and the outcome is the employees' performance represented by the status of recommendation for promotion, whether "yes" or "no" as shown in the table. The C4.5 classifier uses two heuristic criterions to rank possible tests: information gained by using attribute selection measure and the default gained ratio that divides information gain by the information provided by the test outcomes. With these criteria, C4.5 classifier can also be used to determine the important or interesting attributes from the dataset. In this study, the important attributes are identified through the number of hits for each of the attributes from generated classification rules. In this study, as can be seen from the result analysis, C4.5 classifier has a great potential for performance prediction. The generated classification rules can be used to predict the performance of an employee whether he/she has potential to be promoted or not, based on his/her performance [6].

- *The Data Mining of the Human Resources Data Warehouse Based on Association Rule*

  This study uses real time university human resource management project as its background to show a relative design and implementation process of the complete data warehouse. The theme was the teaching and scientific performance evaluation of university faculty, and the data source from the personnel management database system. The approach of construction of university faculty is seen as two integral subsystems ---- training and introduction. During the first excavation, from the rules of bachelor → middle-aged faculty, 13% support degree, 64% confidence level, the impact degree is 1.26, they can see that the middle-aged bachelor degree faculty account for 13%, and 64% confidence level showed that the faculty were middle aged bachelor degree holders. Second attribute gender is added to the mining process and from the association rules on the above analysis it was found: If they introduce talent, they should as far as possible introduce high-degree or high-titles young faculty; in the development of talent they should be targeted at different types of training faculty to take a different approach [7].

- *Selection with the Help of Data Mining (Association As Well As Classification Algorithms Implemented)*

  The TESCO Global Inc. makes retailing by selling 70000 different types of products. To carry out this an adequate number of employees with the necessary qualification are needed. Currently about 300 employees work in the TESCO in Miskolc. According to this an effective staff-procurement strategy has been elaborated. In case of a vacancy, only the application materials will be found that fulfil the requirements of qualifications and expertise. With the Yule association coefficient only relations between alternative criterions can be analyzed. With the use of Csuprov association coefficient there is a possibility to analyze the attributes that are not alternative [8].

B. Case Analysis:

The following algorithms have been studied in this paper:

- C4.5 algorithm :

  C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. C4.5 is the most well-known inductive learning algorithm, which can be used to build decision trees as well as prediction rules. C4.5 is a software extension of the basic ID3 algorithm designed by Quinlan, but it can also deal with continuous attributes and null attribute values. To deal with the continuous attributes, they should be discretized first.

  To build a decision tree from training data, C4.5 tree employs a greedy approach that uses an information theoretic measure (gain ratio) as its guide. Choosing an attribute for the root of the tree divides the training instances into subsets corresponding to the values of the attribute. If the entropy of the class labels in the subsets is less than the entropy of the class labels in the full training set, then information has been gained through splitting on the attribute. C4.5 tree chooses the attribute that gains the most information to be at the root of the tree. The algorithm is applied recursively to form sub-trees, terminating when a given subset contains instances of only one class [9].

- K- Nearest neighbor:

  The k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. K-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is

among the simplest of all machine learning algorithms.

Both for classification and regression, it can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor. The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

- Apriori Algorithm:

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis.

Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found [10].

Apriori uses breadth-first search and a Hash tree structure to count candidate item sets efficiently. It generates candidate item sets of length k from item sets of length k-1. Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent k-length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates [11]. The benefits of C4.5 algorithm are handling continuous and discrete, handling training data with missing values, handling attributes with differing costs, pruning trees after creation. The main advantage of k-NN methods is their simplicity and lack of parametric assumptions. In the presence of a large enough training set, these methods perform surprisingly well, especially when each class is characterized by multiple combinations of predictor values. The advantages of apriori algorithm include using large itemset property, easily parallelized and easy to implement. The Apriori algorithm takes advantage of the fact that any subset of a frequent itemset is also a frequent itemset. These benefits make these algorithms suitable for application for HR in higher education institutes.

V. PROPOSED FRAMEWORK

The drawback with a manual system lies in the fact that most decision are taken on basis of human experience and decision which could lead to errors sometimes.

Therefore it is recommended the system should be automated and the decision taken should consider the past information within the organization which is available in abundance.

The proposed system as depicted in Figure 2 the first step would be to accept applications for the job into the information system which would also contain the past employees details and performance. Based on the past performance data and using Classification Data Mining Technique (CDMT) the characteristics of the good and outstanding employees could be compared against the new candidates. The selected new candidates are then associated with the suitable job using the Association Data Mining Technique (ADMT).

Next the performance of the existing and new employees is evaluated and performance scores are stored in the data warehouse. The characteristics of the good and outstanding employees are identified using CDMT. Using the k- nearest neighbor the average and poor employees are filtered having characteristics similar to those of the good and outstanding employees.

Based on the past trainings and development methods of the good and outstanding employees, the average and poor performing employees should be associated with similar training based on their attributes using ADMT.

The best among the good and outstanding performing employees are motivated. In the next 6-12 months the employees are again evaluated on their performance and the cycle continues.

This cyclic process would enhance and regulate the functioning of the HR department and therefore help improve and develop the working culture within an educational environment.

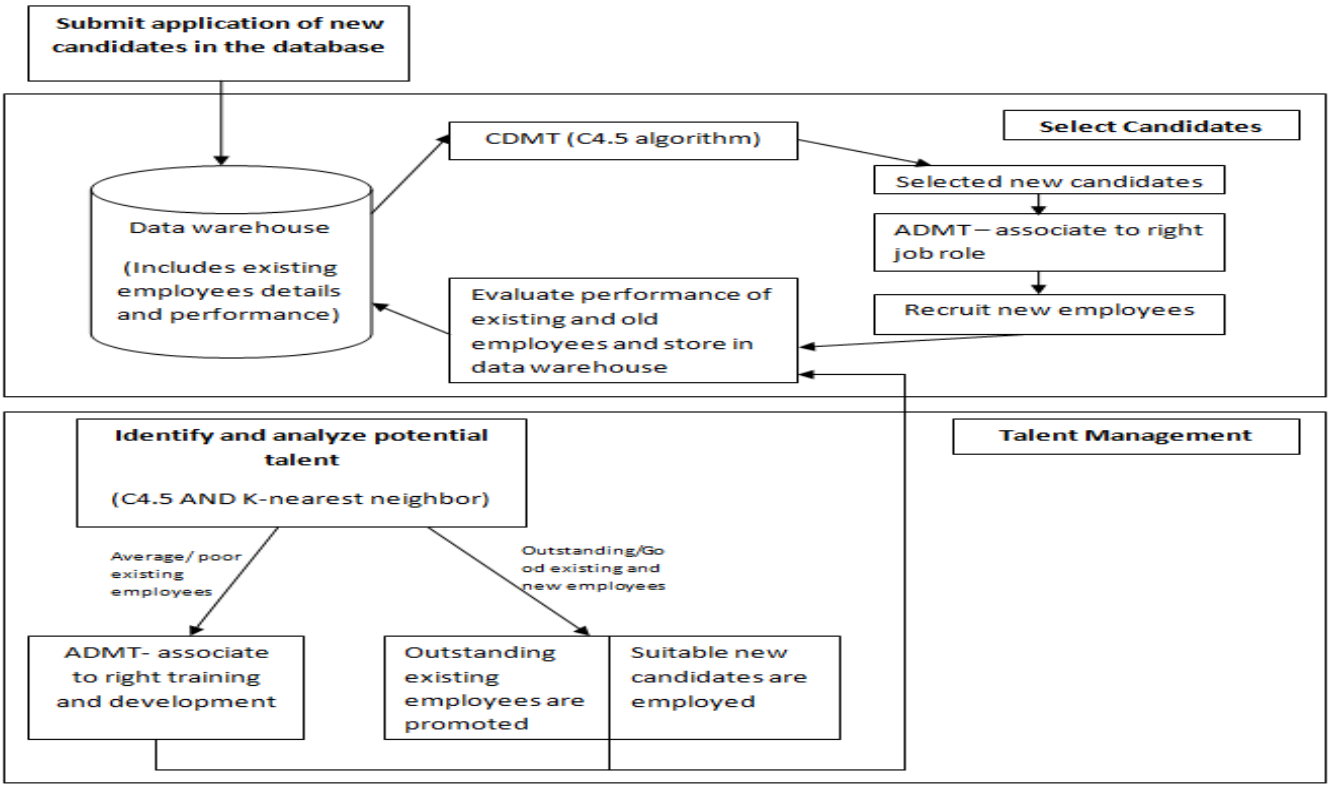

Figure 2: Proposed Framework- Integrating with data mining

VI. Conclusion

This paper brings out the application of data mining in HR in an educational environment reflecting particular to human talent prediction, manpower planning and recruitment wherein application of C4.5 algorithm, K-nearest neighbor and Apriori algorithm were found suitable in this aspect. The paper has also analyzed and proposed a model which will enhance and regulate the system.

Data mining techniques can aid an organization to make intelligent and quick managerial decisions and improve the organization's performance. Therefore more data mining techniques should be applied to the different problem domains in HR field of research to broaden the horizon of academic and practice work on data mining in HR. Other data mining techniques like Fuzzy logic can be considered for future work on classification techniques. The proposed framework can be implemented and its results to be analyzed in terms of efficiency and effectiveness to an organization and its employees

Acknowledgment

I would like to express my gratitude to many people who have helped me in different ways with the development of this paper. Without their continuous support and guidance, the completion of my research leading to this paper would be impossible.

I would like to thank Dr. Sophia Rahaman for her continuous support and guidance in completing this research study. Her teachings and guidance are treasures which I have gained during this study. This shall definitely help me throughout my life and I am very grateful to her for always having faith on me.

References

[1] Adhikari, D. R. (2010). Knowledge management in academic institutions. International Journal of Educational Management .

[2] Azar, A. S. (2013). A model for personnel selection with a data mining approach: a case study on commercial bank. Journal of Human Resource Management .

[3] Ramageri, M. B. (2010). DATA MINING TECHNIQUES AND APPLICATIONS. Indian Journal of Computer Science and Engineering .

[4] Khasawneh, S. (2011). Human capital planning in higher education institutions. International Journal of Educational Management .

[5] Jantan, H., Hamdan, A. R., & Othman, Z. A. (2011). Data Mining Classification Techniques for Human Talent Forecasting. Knowledge-Oriented Applications in Data Mining .

[6] Jantan, H., Hamdan, A. R., & Othman, Z. A. (2010). Human Talent Prediction in HRM using C4.5 Classification Algorithm. International Journal on Computer Science and Engineering , 2526-2534.

[7] Danping, Z., & Jin, D. (2011). The Data Mining of the Human Resources Data Warehouse in University Based on Association Rule. JOURNAL OF COMPUTERS.

[8] KOVÁCS, L., LIZÁK, M., & KOLCZA, G. (2004). Selection with the help of data mining. Production Systems and Information Engineering , 91-105.

[9] FAN, J., & WEN, P. (2007). APPLICATION OF C4.5 ALGORITHM IN WEB-BASED LEARNING ASSESSMENT SYSTEM. Sixth International

Conference on Machine Learning and Cybernetics. Hong Kong.

[10] Yılmaz, N., & Alptekin, G. I. (2013). The Effect of Clustering in the Apriori Data Mining Algorithm : A case study. Proceedings of the World Congress on Engineering 2013 Vol III. London.

[11] Liu, Y. (2010). Study on Application of Apriori Algorithm in Data Mining. Second International Conference on Computer Modeling and Simulation .