Data Mining in Human Resources

Marvin D. Troutt
Kent State University, USA

Lori K. Long
Kent State University, USA

INTRODUCTION

In this paper, we briefly review and update our earlier work (Long & Troutt, 2003) on the topic of data mining in the human resources area. To gain efficiency, many organizations have turned to technology to automate many HR processes (Hendrickson, 2003). As a result of this automation, HR professionals are able to make more informed strategic HR decisions (Bussler & Davis, 2002). While HR professionals may no longer need to manage the manual processing of data, they should not abandon their ties to data collected on and about the organization's employees. Using HR data in decision-making provides a firm with the opportunity to make more informed strategic decisions. If a firm can extract useful or unique information on the behavior and potential of their people from HR data, they can contribute to the firm's strategic planning process. The challenge is identifying useful information in vast human resources databases that are the result of the automation of HR related transaction processing.

Data mining is essentially the extracting of knowledge based on patterns of data in very large databases and is an analytical technique that may become a valuable tool for HR professionals. Organizations that employ thousands of employees and track employment related information might find valuable information patterns contained within their databases to provide insights in such areas as employee retention and compensation planning. To develop an understanding of the potential of data mining HR information in a firm, we will identify opportunities as well as concerns in applying data mining techniques to HR Information Systems.

BACKGROUND

In this section, we review existing work on the topic. The human resource information systems (HRIS) of most organizations today feature relational database systems that allow data to be stored in separate files that can be linked by common elements such as name or identification number. The relational database provides organizations with the ability to keep a virtually limitless amount of data on employees. It also allows organizations to access the data in a variety of ways. For example, a firm can retrieve data on a particular employee or they can retrieve data on a certain group of employees through conducting a search based on a specific parameter such as job classification. The development of relational databases in organizations along with advances in storage technology has resulted in organizations collecting a large amount of data on employees.

While these calculations are helpful to quantify the value of some HR practices, the bottom-line impact of HR practices is not always so clear. One can evaluate the cost per hire, but does that information provide any reference to the value of that hire? Should a greater value be assessed to an employee who stays with the organization for an extended period of time? Most data analysis retrieved from HRIS does not provide an opportunity to seek out additional relationships beyond those that the system was originally designed to identify.

Traditional data analysis methods often involve manual work and interpretation of data that is slow, expensive and highly subjective (Fayyad, Piatsky-Shapiro, & Smyth, 1996a). For example, if an HR professional is interested in analyzing the cost of turnover, they might have to extract data from several different sources such as accounting records, termination reports and personnel hiring records. That data is then combined, reconciled and evaluated. This process creates many opportunities for errors. As business databases have grown in size, the traditional approach has grown more impractical.

Data mining has been used successfully in many functional areas such as finance and marketing. HRIS applications in many organizations provide an as yet unexplored opportunity to apply data mining techniques (Patterson & Lindsay, 2003). While most applications provide opportunities to generate ad-hoc or standardized reports from specific sets of data, the relationships between the data sets are rarely explored. It is this type of relationship that data mining seeks to discover.

The blind application of data mining techniques can easily lead to the discovery of meaningless and invalid patterns. If one searches long enough in any data set, it is likely possible to find patterns that appear to hold but

are not necessarily statistically significant or useful (Fayyad *et al.,* 1996a). There has not been any specific exploration of applying these techniques to human resource applications; however, there are some guidelines in the process that are transferable to an HRIS. Feelders, Daniels and Holsheimer (2000) outline six important steps in the data mining process: 1) problem definition, 2) acquisition of background knowledge, 3) selection of data, 4) pre-processing of data, 5) analysis and interpretation, and 6) reporting and use. At each of these steps, we will look at important considerations as they relate to data mining human resources databases. Further, we will examine some specific legal and ethical considerations of data mining in the HR context.

The formulation of the questions to be explored is an important aspect of the data mining process. As mentioned earlier, with enough searching or application of sufficiently many techniques, one might be able to find useless or ungeneralizable patterns in almost any set of data. Therefore, the effectiveness of a data mining project is improved through establishing some general outlines of inquiry prior to start the project. To this extent, data mining and the more traditional statistical studies are similar. Thus, careful attention to the scientific method and sound research methods are to be followed. A widely respected source of guidelines on research methods is the book by Kerlinger and Lee (2000).

A certain level of expertise is necessary to carefully evaluate questions posed in a data mining project. Obviously, a requirement is data mining and statistical expertise, but one must also have some intimate understanding of the data that is available, along with its business context. Furthermore, some subject matter expertise is needed to determine useful questions, select relevant data and interpret results (Feelders *et al.,* 2000). For example, a firm with interest in evaluating the success of an affirmative action program needs to understand the Equal Employment Opportunity (EEO) classification system to know what data is relevant.

Another important consideration in the process of developing a question to look at is the role of causality (Feelders *et al.,* 2000). A subject matter expert's involvement is important in interpreting the results of the data analysis. For example, a firm might find a pattern indicating a relationship between high compensation levels and extended length of service. The question then becomes, do employees stay with the company longer because they receive high compensation? Or do employees receive higher compensation if they stay longer with the company? An expert in the area can take the relationship discovered and build upon it with additional information available in the organization to help understand the cause and effect of the specific relationship identified.

Selecting and preparing the data is the next step in the data mining process. Some organizations have independent Human Resource Information Systems that feature multiple databases that are not connected to each other. This type of system is sometimes selected to offer greater flexibility to remote organizational locations or sub-groups with unique information needs (Anthony *et al.,* 1996). The possible inconsistency of the design of the databases could make data mining difficult when multiple databases exist. Data warehousing can prevent this problem and an organization may need to create a data warehouse before they begin a data-mining project. The advantage gained in first developing the data warehouse or mart is that most of the data editing is effectively done in advance.

Another challenge in mining data is dealing with the issues of missing or noisy data. Data quality may be insufficient if data is collected without any specific analysis in mind (Feelders *et al.,* 2000). This is especially true for human resource information. Typically when HR data is collected, the purpose is some kind of administrative need such as payroll processing. The need of data for the required transaction is the only consideration in the type of data to collect. Future analysis needs and the value in the data collected is not usually considered. Missing data may also be a problem, especially if the system administrator does not have control over data input. Many organizations have taken advantage of web-based technology to allow for employee input and updating of their own data (Hendrickson, 2003). Employees may choose not to enter certain types of data resulting in missing data. However, a data warehouse or datamart may help to prevent or systemized the handling of many of these problems.

There are many types of algorithms in use in data mining. The choice of the algorithm depends on the intended use of the extracted knowledge (Brodley, Lane, & Stough, 1999). The goals of data mining can be broken down into two main categories. Some applications seek to verify the hypothesis formulated by the user. The other main goal is the discovery or uncovering new patterns systematically (Fayyad *et al*., 1996a). Within discovery, the data can be used to either predict future behavior or describe patterns in an understandable form. A complete discussion of data mining techniques is beyond the scope of this paper. However, the following techniques have the potential to be applicable for data mining of human resources information.

Clustering and classification is an example of a set of data mining techniques borrowed from classical statistical methods that can help describe patterns in information. Clustering seeks to identify a small set of exhaustive and mutual exclusive categories to describe the data that is present (Fayyad *et al*., 1996a). This might be a useful application to human resource data if you were trying to