Discovering Implicit Knowledge from Data Warehouses

M. Mehdi Owrang O.
American University, USA

INTRODUCTION

Today, every corporation faces the problem of how to acquire, store, and share information. Knowledge management (KM) has been introduced to accomplish these tasks (Adams, 2004; Barquin, 2000; Frappaolo & Wilson, 2004). Fundamental to KM is the realization that knowledge exists in two basic forms: explicit and tacit (Adams, 2004; Barquin, 2000; Frappaolo & Wilson, 2004; Orr, 2004). Organizations have data, in the form of operational databases and/or data warehouses, which contain implicit knowledge. Some knowledge believed to be tacit (experiential and intuitive) can be transformed into explicit knowledge. Getting to implicit knowledge requires taking a look at tacit knowledge resources (i.e., domain experts or data warehouses) to determine whether that knowledge could be codified if it were subjected to some type of mining and translation process. Then, it requires implementing that mining/translation process. The majority of an organization's knowledge is presumed to be tacit. Yet, the majority of the KM applications seem to focus on the explicit knowledge base: working on existing corporate knowledge or making individuals more effective at sharing explicit knowledge (Frappaolo & Wilson, 2004). Efforts have been put in creating an organized explicit knowledge repository, called data warehousing (Bischoff & Alexander, 1997) that is continuously fed and leveraged. Knowledge management is not truly possible without data warehousing (Barquin, 2000). It is the real-time access to an enterprise's integrated data stores through data warehousing that complements an individual's tacit knowledge of how something is done.

Knowledge discovery is defined as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data (Adriaans & Zantinge, 1996; Agrawal, Imielinski & Swami, 1993; Brachman et al., 1996; Fayyad, 1996; Inmon, 1996). The automatic knowledge acquisition in a nondata warehouse environment has been on the operational databases which contain the most recent data about the organizations. Summary and historical data, which are essential for accurate and complete knowledge discovery, are generally absent in the operational databases. A data warehouse is an ideal environment for rule discovery since it contains the cleaned, integrated, detailed, summarized, historical, and metadata (Bischoff & Alexander, 1997; Inmon, 1996; Meredith & Khader, 1996; Parsaye, 1996).

In this article, we are looking at the discovery of implicit knowledge from the data warehouses. Most of the success of knowledge discovery resides in the ability of the system to elicit the right level of detail as well as accuracy from the data warehouse which has the implicit data. We look at the knowledge discovery process on detailed, summary, and historical data. Also, we show how the discovered knowledge from these data sources can complement and validate each other.

KNOWLEDGE DISCOVERY IN DATA WAREHOUSES

Knowledge discovery on operational relational databases could lead to inaccurate and incomplete discovered knowledge. Without first warehousing its data, an organization has lots of information that is not integrated and has little summary or history information. The effectiveness of knowledge discovery on such data is limited. A data warehouse environment integrates data from a variety of source databases into a target database that is optimally designed for decision support. A data warehouse includes integrated data, detailed and summary data,

historical data, and metadata. Each of these elements enhances the knowledge discovery process (Adriaans & Zantinge, 1996; Barquin & Edelstein, 1997; Bischoff & Alexander, 1997; Meredith & Khader, 1996).

There are several benefits in rule discovery in a data warehouse environment. First, in a data warehouse environment, the validation of the data is done in a more rigorous and systematic manner. Using metadata, many data redundancies from different application areas are identified and removed. The cleansing process will remove duplication and reconcile differences between various styles of data collection. Second, data warehouses are not concerned with the update anomalies since update of data is not done. This means that at the physical level of design, we can take liberties to optimize the access of data, particularly in dealing with the issues of normalization and physical denormalization. Universal relations can be built in the data warehouse environment for the purposes of rule discovery, which could minimize the chance of undetecting hidden patterns.

Figure 1 shows a general framework for knowledge discovery in a data warehouse environment. External data, domain knowledge (data that is not explicitly stored in the database, that is, male patient cannot be pregnant), and domain expert are other essential components to be added in order to provide an effective knowledge discovery process in a data warehouse environment.

KNOWLEDGE DISCOVERY FROM DETAILED DATA

Most of the knowledge discovery has been done on the operational relational databases. An operational database stores the most recent and detailed data. In addition, the goal of the relational databases are to provide a platform for querying data about uniquely identified objects. However, such uniqueness constraints are not desirable in a knowledge discovery environment. In fact, they are harmful since from the data mining point of view, we are interested in the frequency with which objects occur (Adriaans & Zantinge, 1996). In the following, we discuss two main problems associated with the knowledge discovery in the operational relational databases, namely, the possibility of discovering incorrect and incomplete knowledge.

INCORRECT KNOWLEDGE DISCOVERY FROM RELATIONAL DATABASES

In general, summary data (aggregation) is never found in the operational environment. Without discovery process on summary data, we may discover incorrect knowledge from detailed operational data. Discovering rules based just on current detail data may not depict the actual trends on data. The problem is that statistical significance is usually used

Figure 1. A framework for knowledge discovery in a data warehouse environment