INTRODUCTION

As databases are pervading every parcel of reality, they record what happens in and around organizations all over the world. Databases store the detailed history of organizations, institutions, governments, and individuals. An efficient and agile analysis of the data recorded in a database can no longer be done manually.

Knowledge discovery from databases (KDD) is a collection of technologies that aim at extracting non-trivial, implicit, previously unknown, and potentially useful information (Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996) from raw data stored in databases. The extracted patterns, models, or trends can be used to better understand the data, and hence, the context of an organization, and to predict future behaviors in this context that could improve decision making. KDD can be used to answer questions such as, Is there a group of customers buying a special kind of product? Which sequence of financial products improves the chance of contracting a mortgage? Which telephone call patterns suggest a future churn? Are there relevant associations between risk factors in coronary diseases? How can I assess if my e-mail messages are more or less likely to be spam (junk mail)?

The previous questions cannot be answered by other tools usually associated with database technology, such as OLAP tools, decision support systems, executive information systems, and so forth. The key difference is that KDD does not convert information into (more aggregated or interwoven) information, but generates inductive models (under the form of rules, equations, or other kinds of knowledge) that could be sufficiently consistent with the data. In other words, KDD is not a deductive process but an inductive one.

BACKGROUND

The more and more rapid evolution of the context of organizations forces the revision of their knowledge relentlessly (what used to work before no longer works). This provisional character of knowledge helps explain why knowledge discovery from databases, still in its inception, is so successful. The models and patterns that can be obtained by data mining methods are useful for virtually any area dealing with information: finance, insurance, banking, commerce, marketing, industry, private and public healthcare, medicine, bioengineering, telecommunications, and many other areas (M. J. A. Berry & Linoff, 2004).

The name KDD dates back to the early 1990s and is not a new “technology” itself; KDD is a heterogeneous area that integrates many techniques from several different fields without prejudices, incorporating tools from statistics, machine learning, databases, decision support systems, data visualization, World Wide Web research, among others, in order to obtain novel, valid, and intelligible patterns from data (Berthold & Hand 2002; Dunham 2003; Han & Kamber, 2001; Hand, Mannila, & Smyth, 2000).

PROCESS OF KNOWLEDGE DISCOVERY FROM DATABASES

The KDD process is a complex, elaborate process that comprises several stages (Dunham, 2003; Han & Kamber, 2001): data preparation (including data integration, selection, cleansing, and transformation), data mining, model evaluation and deployment (including model interpretation, use, dissemination and monitoring).

Cross-industry standard process for data mining (CRISP-DM; see http://www.crisp-dm.org for more information) is a standard reference that serves as a guide to carry out a knowledge discovery project and comprises all the stages mentioned (see Figure 1).

Figure 1. Stages of the knowledge discovery process
According to this process, data mining is just a stage of the knowledge discovery process. The data mining stage, also called the **modeling stage**, converts the prepared data (usually in the form of a “minable view” with an assigned task) into one or more models. This stage is thereby the most characteristic of the whole process, and data mining is frequently used as a synonym for all the process. The process is cyclic because, after a complete cycle, the goals can be revised or extended to start the whole process again.

Many references to KDD choose the data preparation stage as the start of the knowledge discovery process (Dunham, 2003; Fayyad, 1996; Han & Kamber, 2001). The CRISP-DM standard, however, emphasizes that business understanding and data understanding (which includes data integration) must be handled before data preparation. This is reasonable because the first thing to know before starting with a KDD project is what the business needs are, to establish the business goals according to the business context, to see whether there is (or if we can get) enough data to solve them, and, in this case, to specify the data mining objectives.

For example, consider a distribution company that has a problem of inadequate stocks, which generate higher costs and frequently make some perishable products expire. From this problem, one business objective could be to **reduce the stock level of perishables**. If there is an internal database with enough information about the orders performed by each customer in the last 3 years, we can complement this with external information about the reference market sale prices of these product categories, it could be sufficient to start looking for models to address the business objective. This business objective could be translated into one or more data mining objectives (e.g., predict how many perishable products by category the customer is to buy week by week, from the orders performed during the last three years and the current market sale price).

Before starting the discussion about each specific KDD stage, it is important to highlight that many key issues on the success of a KDD project have to do with a good understanding of the goals and the feasibility of these goals, as well as a precise assessment of the resources needed (e.g., data, human, software, organizational). Table 1 shows the most important issues for success in a KDD project.

Additionally, a major reason for a possible failure in KDD projects may stem from an excessive focus on technology, that is, implementing data mining because others do, without recognizing what the needs of the organization are and without understanding the resources and data necessary to cover them.

### Data Integration and Preparation

Once the data mining goals are clear and the data that will be required to achieve them is located, it is necessary to get all the data and integrate them. Usually, the data can come from many sources, either internal or external. The typical internal data, and generally the main source of

---

**Table 1. Keys to success in a knowledge discovery project**

- **Business needs must drive the knowledge discovery project.** The business problems and, hence, the **business objectives must be clearly stated**. These business objectives will help one understand the data that will be needed and the scope of the project.
- **Business objectives must be translated into specific data mining objectives,** which must be relevant for the organization. The data mining objectives must be accompanied with a specification on the required quality of the models to be extracted in terms of a set of metrics and features: expected error, reliability, costs, relevance, comprehensibility, and so forth.
- **The knowledge discovery project must be integrated with other plans in the organization and must have the unconditional support from the organization executives.**
- **Data quality** is crucial. The integration of data from several sources (internal or external), its neatness and an adequate organization and availability (using a data warehouse, if necessary) is a sine qua non in knowledge discovery.
- **The use of integrated and friendly tools** is also decisive, helping the process on many issues, not only for the specific stages of the process but also on other issues, such as documentation, communication tools, workflow tools, and so on.
- **The need of a heterogeneous team,** comprising not only professionals with a specific data mining training but also professionals from statistics, databases and business. A strong leadership among the group is also essential.

To translate the positive results of knowledge discovery into positive results for the organization, it is necessary to use a **holistic evaluation and an ambitious deployment of the models** to the know-how and daily operation of the organization.
Related Content

FOOM - Functional and Object-Oriented Analysis and Design of Information Systems: An Integrated Methodology
www.igi-global.com/article/foom-functional-object-oriented-analysis/3258?camid=4v1a

A Scalable Algorithm for One-to-One, Onto, and Partial Schema Matching with Uninterpreted Column Names and Column Values
Boris Rabinovich and Mark Last (2014). Journal of Database Management (pp. 1-16).
www.igi-global.com/article/a-scalable-algorithm-for-one-to-one-onto-and-partial-schema-matching-with-uninterpreted-column-names-and-column-values/138623?camid=4v1a

Extended Action Rule Discovery Based on Single Classification Rules and Reducts
www.igi-global.com/chapter/extended-action-rule-discovery-based/8039?camid=4v1a

Legal Protection of the Web Page as a Database
www.igi-global.com/chapter/legal-protection-web-page-database/8054?camid=4v1a