T

# Text Mining in the Context of Business Intelligence

**Hércules Antonio do Prado**
*Brazilian Enterprise for Agricultural Research and Catholic University of Brasília, Brazil*

**José Palazzo Moreira de Oliveira**
*Federal University of Rio Grande do Sul, Brazil*

**Edilson Ferneda**
*Catholic University of Brasília, Brazil*

**Leandro Krug Wives**
*Centro Universitário Feevale and Federal University of Rio Grande do Sul, Brazil*

**Edilberto Magalhães Silva**
*Brazilian Public News Agency, Brazil*

**Stanley Loh**
*Catholic University of Pelotas and Lutheran University of Brazil, Brazil*

## INTRODUCTION

Information about the external environment and organizational processes are among the most worthwhile input for business intelligence (BI). Nowadays, companies have plenty of information in structured or textual forms, either from external monitoring or from the corporative systems. In the last years, the structured part of this information stock has been massively explored by means of data-mining (DM) techniques (Wang, 2003), generating models that enable the analysts to gain insights on the solutions for organizational problems. On the text-mining (TM) side, the rhythm of new applications development did not go so fast. In an informal poll carried out in 2002 (Kdnuggets), just 4% of the knowledge-discovery-from-databases (KDD) practitioners were applying TM techniques. This fact is as intriguing as surprising if one considers that 80% of all information available in an organization comes in textual form (Tan, 1999).

In their popular model to explain the phases of technology adoption (Figure 1), Moore and McKenna (1999) discuss the existence of a chasm between the "early adopters, visionaries," and the "early majority pragmatists" phases that a technology has to cross in order to become extensively adopted. From our point of view, TM is crossing this chasm yet. Although there is the existence of mature tools in the market, and an increasing number of successful case studies have been presented (Ferneda, Prado, & Silva, 2003; Fliedl & Weber, 2002; Dini & Mazzini, 2002; Prado, Oliveira, Ferneda, Wives, Silva, & Loh, 2004), it seems that the community is still leaving the second phase. However, the results presented in the case studies point out that the broad adoption of TM will happen in the near future.

## BACKGROUND

When studying the relations between TM and BI, it is necessary to take into account an important intermediate layer between them: the knowledge-management (KM) process. KM refers to the set of activities responsible for carrying the information along the organization and making knowledge available where it is necessary.

To clarify the relations between TM and BI, under the point of view of a KM model, we adopted the generic KM model (Figure 2) proposed by Stollenwerk (2001). The model is made up of seven processes: (a) identification and development of the critical abilities, (b) capture of knowledge, skills, and experiences to create and maintain skills, (c) selection and validation that filter, evaluate, and summarize the acquired knowledge for future use, (d) organization and storage to assure the quick and correct recovery of the stored knowledge, (e) sharing that makes easy the access to information and knowledge, (f) application in which the knowledge is applied in real situations, and (g) creation that comprises the activities of sharing tacit knowledge, creating concepts, building archetypes, and cross-leveling knowledge. Involving the mentioned processes, there exist the aspects of leadership, organiza-

*Figure 1. Moore's and McKenna's (1999) life cycle of technology adoption*

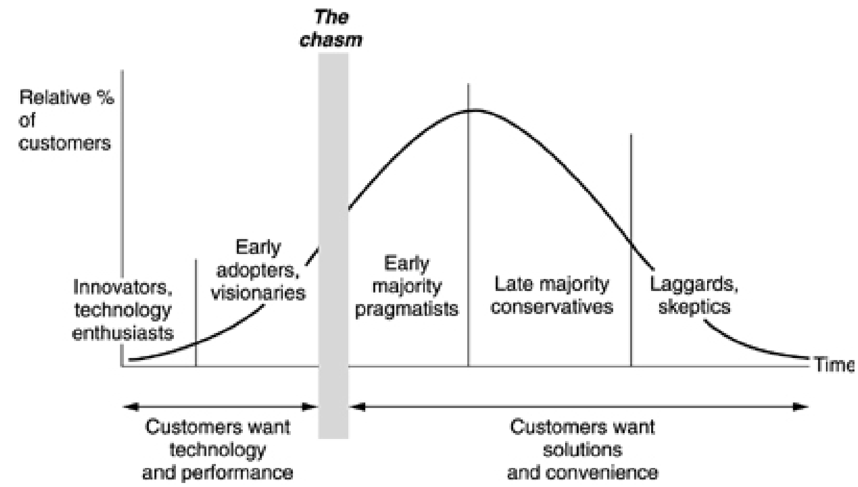


*Figure 2. Generic KM model of Stollenwerk (2001)*


tional culture, measuring and compensation, and technology. The main relation between TM and KM is located in the creation process. By applying the TM techniques discussed in the next section, it is possible to find patterns that, adequately interpreted, can leverage the concept-creation activity.

## METHODS AND TECHNIQUES FOR TEXT MINING

Text mining can be defined as the application of computational methods and techniques over textual data in order to find relevant and intrinsic information and previously unknown knowledge.

Text-mining techniques can be organized into four categories: classification, association analysis, information extraction, and clustering. Classification techniques consist of the allocation of objects into predefined classes or categories. They are used to identify the class or category of texts in tasks such as *topic spotting* (the identification of a known topic or subject in a document) and *document routing* or *filtering* (the selection of relevant documents to a process or to someone).

Association analysis is used to identify correlation or dependencies among elements or attributes (words or concepts present in documents). It helps the identification of words or concepts that co-occur together and, consequently, to understand the contents of a document or set of documents and their relationships.