PREAMBLE

Pattern recognition is the most basic description of what is done in the process of data mining. Usually, these patterns are stored in structured databases and organized into records, which are composed of rows and columns of data. The columns are attributes (numbers or text strings) associated with a table (entity), accessed by links between attributes among the tables (relations). This entity-relational structure of data is called a relational database. Large relational databases store huge quantities of data in data warehouses in large companies. Despite the rather large amount of business information that exists in data warehouses, the vast majority of business data is stored in documents that are virtually unstructured. According to a study by Merrill Lynch and Gartner, 85–90% of all corporate data are stored in some sort of unstructured form (i.e., as text) (McKnight, 2005). This is where text mining fits into the picture: it is the process of discovering new, previously unknown, potentially useful information from a variety of unstructured data sources including business documents, customer comments, web pages, and XML files.