

- **Term-by-document matrix (occurrence matrix).** A common representation schema of the frequency-based relationship between the terms and documents in tabular format where terms are listed in rows, documents are listed in columns, and the frequency between the terms and documents is listed in cells as integer values.
- **Singular-value decomposition (latent semantic indexing).** A dimensionality reduction method used to transform the term-by-document matrix to a manageable size by generating an intermediate representation of the frequencies using a matrix manipulation method similar to principal component analysis.

Application Case 5.1

Text Mining for Patent Analysis

A patent is a set of exclusive rights granted by a country to an inventor for a limited period of time in exchange for a disclosure of an invention (note that the procedure for granting patents, the requirements placed on the patentee, and the extent of the exclusive rights vary widely from country to country). The disclosure of these inventions is critical to future advancements in science and technology. If carefully analyzed, patent documents can help identify emerging technologies, inspire novel solutions, foster symbiotic partnerships, and enhance overall awareness of business' capabilities and limitations.

Patent analysis is the use of analytical techniques to extract valuable knowledge from patent databases. Countries or groups of countries that maintain patent databases (e.g., the United States, the European Union, Japan) add tens of millions of new patents each year. It is nearly impossible to efficiently process such enormous amounts of semi-structured data (patent documents usually contain partially structured and partially textual data). Patent analysis with semiautomated software tools is one way to ease the processing of these very large databases.

A Representative Example of Patent Analysis

Eastman Kodak employs more than 5,000 scientists, engineers, and technicians around the world. During the twentieth century, these knowledge workers and their predecessors claimed nearly 20,000 patents, putting the company among the top 10 patent holders in the world. Being in the business of constant change, the company knows that success (or mere

survival) depends on its ability to apply more than a century's worth of knowledge about imaging science and technology to new uses and to secure those new uses with patents.

Appreciating the value of patents, Kodak not only generates new patents but also analyzes those created by others. Using dedicated analysts and state-of-the-art software tools (including specialized text mining tools from ClearForest Corp.), Kodak continuously digs deep into various data sources (patent databases, new release archives, and product announcements) in order to develop a holistic view of the competitive landscape. Proper analysis of patents can bring companies like Kodak a wide range of benefits:

- It enables competitive intelligence. Knowing what competitors are doing can help a company to develop countermeasures.
- It can help the company make critical business decisions, such as what new products, product lines, and/or technologies to get into or what mergers and acquisitions to pursue.
- It can aid in identifying and recruiting the best and brightest new talent, those whose names appear on the patents that are critical to the company's success.
- It can help the company to identify the unauthorized use of its patents, enabling it to take action to protect its assets.
- It can identify complementary inventions to build symbiotic partnerships or to facilitate mergers and/or acquisitions.
- It prevents competitors from creating similar products and it can help protect the company from patent infringement lawsuits.

Using patent analysis as a rich source of knowledge and a strategic weapon (both defensive as well as offensive), Kodak not only survives but excels in its market segment defined by innovation and constant change.

QUESTIONS FOR DISCUSSION

1. Why is it important for companies to keep up with patent filings?

2. How did Kodak use text analytics to better analyze patents?
3. What were the challenges, the proposed solution, and the obtained results?

Sources: P. X. Chiem, "Kodak Turns Knowledge Gained About Patents into Competitive Intelligence," *Knowledge Management*, 2001, pp. 11–12; Y-H. Tsenga, C-J. Linb, and Y-I. Linc, "Text Mining Techniques for Patent Analysis," *Information Processing & Management*, Vol. 43, No. 5, 2007, pp. 1216–1245.

SECTION 5.2 REVIEW QUESTIONS

1. What is text analytics? How does it differ from text mining?
2. What is text mining? How does it differ from data mining?
3. Why is the popularity of text mining as an analytics tool increasing?
4. What are some of the most popular application areas of text mining?

5.3 NATURAL LANGUAGE PROCESSING

Some of the early text mining applications used a simplified representation called *bag-of-words* when introducing structure to a collection of text-based documents in order to classify them into two or more predetermined classes or to cluster them into natural groupings. In the bag-of-words model, text, such as a sentence, paragraph, or complete document, is represented as a collection of words, disregarding the grammar or the order in which the words appear. The bag-of-words model is still used in some simple document classification tools. For instance, in spam filtering an e-mail message can be modeled as an unordered collection of words (a bag-of-words) that is compared against two different predetermined bags. One bag is filled with words found in spam messages and the other is filled with words found in legitimate e-mails. Although some of the words are likely to be found in both bags, the "spam" bag will contain spam-related words such as *stock*, *Viagra*, and *buy* much more frequently than the legitimate bag, which will contain more words related to the user's friends or workplace. The level of match between a specific e-mail's bag-of-words and the two bags containing the descriptors determines the membership of the e-mail as either spam or legitimate.

Naturally, we (humans) do not use words without some order or structure. We use words in sentences, which have semantic as well as syntactic structure. Thus, automated techniques (such as text mining) need to look for ways to go beyond the bag-of-words interpretation and incorporate more and more semantic structure into their operations. The current trend in text mining is toward including many of the advanced features that can be obtained using natural language processing.

It has been shown that the bag-of-words method may not produce good enough information content for text mining tasks (e.g., classification, clustering, association). A good example of this can be found in evidence-based medicine. A critical component of evidence-based medicine is incorporating the best available research findings into the clinical decision-making process, which involves appraisal of the information collected from the printed media for validity and relevance. Several researchers from the University of Maryland developed evidence assessment models using a bag-of-words method (Lin and Demner, 2005). They employed popular machine-learning methods along with more than half a million research articles collected from MEDLINE (Medical