CHAPTER

7

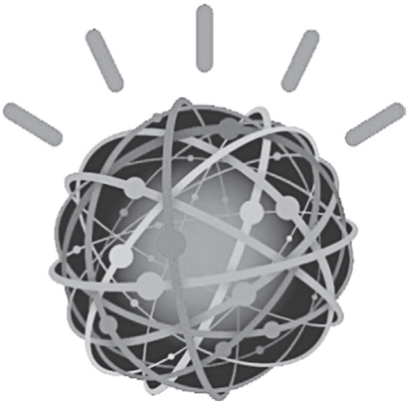
Text Analytics, Text Mining, and Sentiment Analysis

LEARNING OBJECTIVES

- Describe text mining and understand the need for text mining
- Differentiate among text analytics, text mining, and data mining
- Understand the different application areas for text mining
- Know the process for carrying out a text mining project
- Appreciate the different methods to introduce structure to text-based data
- Describe sentiment analysis
- Develop familiarity with popular applications of sentiment analysis
- Learn the common methods for sentiment analysis
- Become familiar with speech analytics as it relates to sentiment analysis

This chapter provides a rather comprehensive overview of text mining and one of its most popular applications, sentiment analysis, as they both relate to business analytics and decision support systems. Generally speaking, sentiment analysis is a derivative of text mining, and text mining is essentially a derivative of data mining. Because textual data is increasing in volume more than the data in structured databases, it is important to know some of the techniques used to extract actionable information from this large quantity of unstructured data.

7.1 OPENING VIGNETTE: Machine Versus Men on *Jeopardy!*: The Story of Watson



Can machine beat the best of man in what man is supposed to be the best at? Evidently, yes, and the machine's name is Watson. Watson is an extraordinary computer system (a novel combination of advanced hardware and software) designed to answer questions posed in natural human language. It was developed in 2010 by an IBM Research team as part of a DeepQA project and was named after IBM's first president, Thomas J. Watson.

BACKGROUND

Roughly 3 years ago, IBM Research was looking for a major research challenge to rival the scientific and popular interest of Deep Blue, the computer chess-playing champion, which would also have clear relevance to IBM business interests. The goal was to advance computer science by exploring new ways for computer technology to affect science, business, and society. Accordingly, IBM Research undertook a challenge to build a computer system that could compete at the human champion level in real time on the American TV quiz show, *Jeopardy!* The extent of the challenge included fielding a real-time automatic contestant on the show, capable of listening, understanding, and responding—not merely a laboratory exercise.

COMPETING AGAINST THE BEST

In 2011, as a test of its abilities, Watson competed on the quiz show *Jeopardy!*, which was the first ever human-versus-machine matchup for the show. In a two-game, combined-point match (broadcast in three *Jeopardy!* episodes during February 14–16), Watson beat Brad Rutter, the biggest all-time money winner on *Jeopardy!*, and Ken Jennings, the record holder for the longest championship streak (75 days). In these episodes, Watson consistently outperformed its human opponents on the game's signaling device, but had trouble responding to a few categories, notably those having short clues containing only a few words. Watson had access to 200 million pages of structured and unstructured content consuming four terabytes of disk storage. During the game Watson was not connected to the Internet.

Meeting the *Jeopardy!* Challenge required advancing and incorporating a variety of QA technologies (text mining and natural language processing) including parsing, question classification, question decomposition, automatic source acquisition and evaluation, entity and relation detection, logical form generation, and knowledge representation and reasoning. Winning at *Jeopardy!* required accurately computing confidence in your answers. The questions and content are ambiguous and noisy and none of the individual algorithms are perfect. Therefore, each component must produce a confidence in its output, and individual component confidences must be combined to compute the overall confidence of the final answer. The final confidence is used to determine whether the computer system should risk choosing to answer at all. In *Jeopardy!* parlance, this confidence is used to determine whether the computer will "ring in" or "buzz in" for a question. The confidence must be computed during the time the question is read and before the opportunity to buzz in. This is roughly between 1 and 6 seconds with an average around 3 seconds.

HOW DOES WATSON DO IT?

The system behind Watson, which is called DeepQA, is a massively parallel, text mining–focused, probabilistic evidence-based computational architecture. For the *Jeopardy!* challenge, Watson used more than 100 different techniques for analyzing natural language, identifying sources, finding and generating hypotheses, finding and scoring evidence, and merging and ranking hypotheses. What is far more important than any particular technique that they used was how they combine them in DeepQA such that overlapping approaches can bring their strengths to bear and contribute to improvements in accuracy, confidence, and speed.

DeepQA is an architecture with an accompanying methodology, which is not specific to the *Jeopardy!* challenge. The overarching principles in DeepQA are massive parallelism, many experts, pervasive confidence estimation, and integration of the-latest-and-greatest in text analytics.

- ***Massive parallelism:*** Exploit massive parallelism in the consideration of multiple interpretations and hypotheses.
- ***Many experts:*** Facilitate the integration, application, and contextual evaluation of a wide range of loosely coupled probabilistic question and content analytics.
- ***Pervasive confidence estimation:*** No component commits to an answer; all components produce features and associated confidences, scoring different question and content interpretations. An underlying confidence-processing substrate learns how to stack and combine the scores.
- ***Integrate shallow and deep knowledge:*** Balance the use of strict semantics and shallow semantics, leveraging many loosely formed ontologies.

Figure 7.1 illustrates the DeepQA architecture at a very high level. More technical details about the various architectural components and their specific roles and capabilities can be found in Ferrucci et al. (2010).

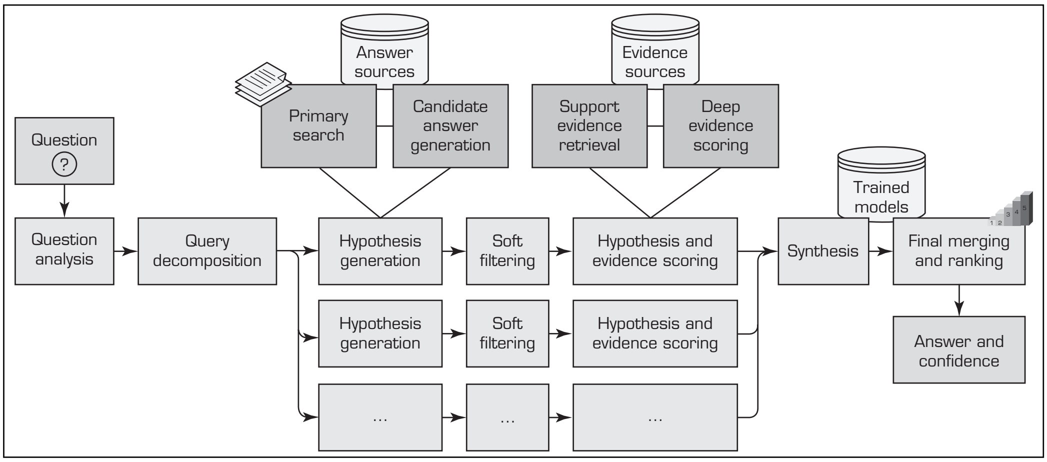


FIGURE 7.1 A High-Level Depiction of DeepQA Architecture.

CONCLUSION

The *Jeopardy!* challenge helped IBM address requirements that led to the design of the DeepQA architecture and the implementation of Watson. After 3 years of intense research and development by a core team of about 20 researchers, Watson is performing at human expert levels in terms of precision, confidence, and speed at the *Jeopardy!* quiz show.

IBM claims to have developed many computational and linguistic algorithms to address different kinds of issues and requirements in QA. Even though the internals of these algorithms are not known, it is imperative that they made the most out of text analytics and text mining. Now IBM is working on a version of Watson to take on surmountable problems in healthcare and medicine (Feldman et al., 2012).

QUESTIONS FOR THE OPENING VIGNETTE

1. What is Watson? What is special about it?
2. What technologies were used in building Watson (both hardware and software)?
3. What are the innovative characteristics of DeepQA architecture that made Watson superior?
4. Why did IBM spend all that time and money to build Watson? Where is the ROI?
5. Conduct an Internet search to identify other previously developed "smart machines" (by IBM or others) that compete against the best of man. What technologies did they use?

WHAT WE CAN LEARN FROM THIS VIGNETTE

It is safe to say that computer technology, on both the hardware and software fronts, is advancing faster than anything else in the last 50-plus years. Things that were too big, too complex, impossible to solve are now well within the reach of information technology. One of those enabling technologies is perhaps text analytics/text mining. We created databases to structure the data so that it can be processed by computers. Text, on the other hand, has always been meant for humans to process. Can machines do the things that require human creativity and intelligence, and which were not originally designed for machines? Evidently, yes! Watson is a great example of the distance that we have traveled in addressing the impossible. Computers are now intelligent enough to take on men at what we think men are the best at. Understanding the question that was posed in spoken human language, processing and digesting it, searching for an answer, and replying within a few seconds was something that we could not have imagined possible before Watson actually did it. In this chapter, you will learn the tools and techniques embedded in Watson and many other smart machines to create miracles in tackling problems that were once believed impossible to solve.

Sources: D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, N. Schlaefer, and C. Welty, "Building Watson: An Overview of the DeepQA Project," AI Magazine, Vol. 31, No. 3, 2010; DeepQA, DeepQA Project: FAQ, IBM Corporation, 2011, research.ibm.com/deepqa/faq.shtml (accessed January 2013); and S. Feldman, J. Hanover, C. Burghard, and D. Schubmehl, "Unlocking the Power of Unstructured Data," IBM white paper, 2012, www-01.ibm.com/software/ebusiness/jstart/downloads/unlockingUnstructuredData.pdf (accessed February 2013).

7.2 TEXT ANALYTICS AND TEXT MINING CONCEPTS AND DEFINITIONS

The information age that we are living in is characterized by the rapid growth in the amount of data and information collected, stored, and made available in electronic format. The vast majority of business data is stored in text documents that are virtually unstructured. According to a study by Merrill Lynch and Gartner, 85 percent of all corporate data is captured and stored in some sort of unstructured form (McKnight, 2005). The same study also stated that this unstructured data is doubling in size every 18 months. Because knowledge is power in today's business world, and knowledge is derived from data and information, businesses that effectively and efficiently tap into their text data sources will have the necessary knowledge to make better decisions, leading to a competitive advantage over those businesses that lag behind. This is where the need for text analytics and text mining fits into the big picture of today's businesses.

Even though the overarching goal for both text analytics and text mining is to turn unstructured textual data into actionable information through the application of natural language processing (NLP) and analytics, their definitions are somewhat different, at least to some experts in the field. According to them, text analytics is a broader concept that includes information retrieval (e.g., searching and identifying relevant documents for a given set of key terms) as well as information extraction, data mining, and Web mining, whereas text mining is primarily focused on discovering new and useful knowledge from the textual data sources. Figure 7.2 illustrates the relationships between text analytics and text mining along with other related application areas. The bottom of Figure 7.2 lists the main disciplines (the foundation of the house) that play a critical role in the development of these increasingly more popular application areas. Based on this definition of text analytics and text mining, one could simply formulate the difference between the two as follows:

Text Analytics = Information Retrieval + Information Extraction + Data Mining + Web Mining,

or simply

Text Analytics = Information Retrieval + Text Mining

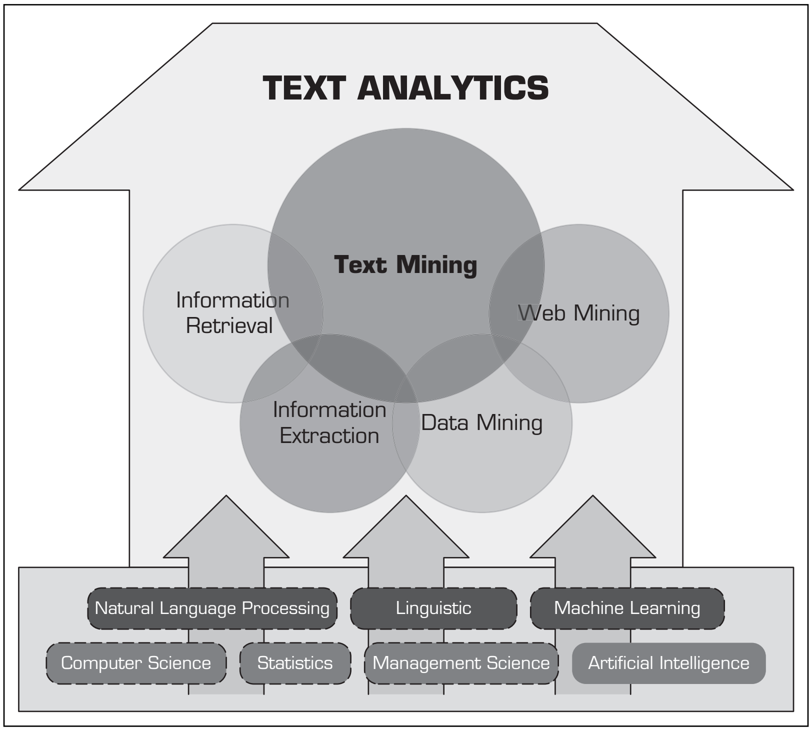


FIGURE 7.2 Text Analytics, Related Application Areas, and Enabling Disciplines.

Compared to text mining, text analytics is a relatively new term. With the recent emphasis on *analytics*, as has been the case in many other related technical application areas (e.g., consumer analytics, completive analytics, visual analytics, social analytics, and so forth), the text field has also wanted to get on the analytics bandwagon. While the term *text analytics* is more commonly used in a business application context, text mining is frequently used in academic research circles. Even though they may be defined somewhat differently at times, text analytics and text mining are usually used synonymously, and we (the authors of this book) concur with this.

Text mining (also known as *text data mining* or *knowledge discovery in textual databases*) is the semi-automated process of extracting patterns (useful information and knowledge) from large amounts of unstructured data sources. Remember that data mining is the process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases, where the data are organized in records structured by categorical, ordinal, or continuous variables. Text mining is the same as data mining in that it has the same purpose and uses the same processes, but with text mining the input to the process is a collection of unstructured (or less structured) data files such as Word documents, PDF files, text excerpts, XML files, and so on. In essence, text mining can be thought of as a process (with two main steps) that starts with imposing structure on the text-based data sources, followed by extracting relevant information and knowledge from this structured text-based data using data mining techniques and tools.

The benefits of text mining are obvious in the areas where very large amounts of textual data are being generated, such as law (court orders), academic research (research articles), finance (quarterly reports), medicine (discharge summaries), biology (molecular interactions), technology (patent files), and marketing (customer comments). For example, the free-form text-based interactions with customers in the form of complaints (or praises) and warranty claims can be used to objectively identify product and service characteristics that are deemed to be less than perfect and can be used as input to better product development and service allocations. Likewise, market outreach programs and focus groups generate large amounts of data. By not restricting product or service feedback to a codified form, customers can present, in their own words, what they think about a company's products and services. Another area where the automated processing of unstructured text has had a lot of impact is in electronic communications and e-mail. Text mining not only can be used to classify and filter junk e-mail, but it can also be used to automatically prioritize e-mail based on importance level as well as generate automatic responses (Weng and Liu, 2004). The following are among the most popular application areas of text mining:

- ***Information extraction.*** Identification of key phrases and relationships within text by looking for predefined objects and sequences in text by way of pattern matching. Perhaps the most commonly used form of information extraction is *named entity extraction*. Named entity extraction includes *named entity recognition* (recognition of known entity names—for people and organizations, place names, temporal expressions, and certain types of numerical expressions, using existing knowledge of the domain), *co-reference resolution* (detection of co-reference and anaphoric links between text entities), and *relationship extraction* (identification of relations between entities).
- ***Topic tracking.*** Based on a user profile and documents that a user views, text mining can predict other documents of interest to the user.
- ***Summarization.*** Summarizing a document to save time on the part of the reader.
- ***Categorization.*** Identifying the main themes of a document and then placing the document into a predefined set of categories based on those themes.
- ***Clustering.*** Grouping similar documents without having a predefined set of categories.

- ***Concept linking.*** Connects related documents by identifying their shared concepts and, by doing so, helps users find information that they perhaps would not have found using traditional search methods.
- ***Question answering.*** Finding the best answer to a given question through knowledge-driven pattern matching.

See Technology Insights 7.1 for explanations of some of the terms and concepts used in text mining. Application Case 7.1 describes the use of text mining in patent analysis.

TECHNOLOGY INSIGHTS 7.1 Text Mining Lingo

The following list describes some commonly used text mining terms:

- ***Unstructured data (versus structured data).*** Structured data has a predetermined format. It is usually organized into records with simple data values (categorical, ordinal, and continuous variables) and stored in databases. In contrast, **unstructured data** does not have a predetermined format and is stored in the form of textual documents. In essence, the structured data is for the computers to process while the unstructured data is for humans to process and understand.
- ***Corpus.*** In linguistics, a **corpus** (plural *corpora*) is a large and structured set of texts (now usually stored and processed electronically) prepared for the purpose of conducting knowledge discovery.
- ***Terms.*** A *term* is a single word or multiword phrase extracted directly from the corpus of a specific domain by means of natural language processing (NLP) methods.
- ***Concepts.*** *Concepts* are features generated from a collection of documents by means of manual, statistical, rule-based, or hybrid categorization methodology. Compared to terms, concepts are the result of higher level abstraction.
- ***Stemming.*** **Stemming** is the process of reducing inflected words to their stem (or base or root) form. For instance, *stemmer, stemming*, and *stemmed* are all based on the root *stem*.
- ***Stop words.*** **Stop words** (or *noise words*) are words that are filtered out prior to or after processing of natural language data (i.e., text). Even though there is no universally accepted list of stop words, most natural language processing tools use a list that includes articles (*a, am, the, of,* etc.), auxiliary verbs (*is, are, was, were,* etc.), and context-specific words that are deemed not to have differentiating value.
- ***Synonyms and polysemes.*** Synonyms are syntactically different words (i.e., spelled differently) with identical or at least similar meanings (e.g., *movie, film*, and *motion picture*). In contrast, **polysemes**, which are also called *homonyms*, are syntactically identical words (i.e., spelled exactly the same) with different meanings (e.g., *bow* can mean "to bend forward," "the front of the ship," "the weapon that shoots arrows," or "a kind of tied ribbon").
- ***Tokenizing.*** A *token* is a categorized block of text in a sentence. The block of text corresponding to the token is categorized according to the function it performs. This assignment of meaning to blocks of text is known as **tokenizing**. A token can look like anything; it just needs to be a useful part of the structured text.
- ***Term dictionary.*** A collection of terms specific to a narrow field that can be used to restrict the extracted terms within a corpus.
- ***Word frequency.*** The number of times a word is found in a specific document.
- ***Part-of-speech tagging.*** The process of marking up the words in a text as corresponding to a particular part of speech (such as nouns, verbs, adjectives, adverbs, etc.) based on a word's definition and the context in which it is used.
- ***Morphology.*** A branch of the field of linguistics and a part of natural language processing that studies the internal structure of words (patterns of word-formation within a language or across languages).
- ***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 7.1

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 semistructured 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.

(*Continued*)

Application Case 7.1 (Continued)

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. 1216–1247.

SECTION 7.2 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?

7.3 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 Literature Analysis and Retrieval System Online). In their models, they represented each abstract as a bag-of-words, where each stemmed term represented a feature. Despite using popular classification methods with proven experimental design methodologies, their prediction results were not much better than simple guessing, which may indicate that the bag-of-words is not generating a good enough representation of the research articles in this domain; hence, more advanced techniques such as natural language processing are needed.

Natural language processing (NLP) is an important component of text mining and is a subfield of artificial intelligence and computational linguistics. It studies the problem of "understanding" the natural human language, with the view of converting depictions of human language (such as textual documents) into more formal representations (in the form of numeric and symbolic data) that are easier for computer programs to manipulate. The goal of NLP is to move beyond syntax-driven text manipulation (which is often called "word counting") to a true understanding and processing of natural language that considers grammatical and semantic constraints as well as the context.

The definition and scope of the word "understanding" is one of the major discussion topics in NLP. Considering that the natural human language is vague and that a true understanding of meaning requires extensive knowledge of a topic (beyond what is in the words, sentences, and paragraphs), will computers ever be able to understand natural language the same way and with the same accuracy that humans do? Probably not! NLP has come a long way from the days of simple word counting, but it has an even longer way to go to really understanding natural human language. The following are just a few of the challenges commonly associated with the implementation of NLP:

- ***Part-of-speech tagging.*** It is difficult to mark up terms in a text as corresponding to a particular part of speech (such as nouns, verbs, adjectives, adverbs, etc.) because the part of speech depends not only on the definition of the term but also on the context within which it is used.
- ***Text segmentation.*** Some written languages, such as Chinese, Japanese, and Thai, do not have single-word boundaries. In these instances, the text-parsing task requires the identification of word boundaries, which is often a difficult task. Similar challenges in speech segmentation emerge when analyzing spoken language, because sounds representing successive letters and words blend into each other.
- ***Word sense disambiguation.*** Many words have more than one meaning. Selecting the meaning that makes the most sense can only be accomplished by taking into account the context within which the word is used.
- ***Syntactic ambiguity.*** The grammar for natural languages is ambiguous; that is, multiple possible sentence structures often need to be considered. Choosing the most appropriate structure usually requires a fusion of semantic and contextual information.
- ***Imperfect or irregular input.*** Foreign or regional accents and vocal impediments in speech and typographical or grammatical errors in texts make the processing of the language an even more difficult task.
- ***Speech acts.*** A sentence can often be considered an action by the speaker. The sentence structure alone may not contain enough information to define this action. For example, "Can you pass the class?" requests a simple yes/no answer, whereas "Can you pass the salt?" is a request for a physical action to be performed.

It is a longstanding dream of the artificial intelligence community to have algorithms that are capable of automatically reading and obtaining knowledge from text. By applying a learning algorithm to parsed text, researchers from Stanford University's NLP lab have developed methods that can automatically identify the concepts and relationships between those concepts in the text. By applying a unique procedure to large amounts of text, their algorithms automatically acquire hundreds of thousands of items of world knowledge and use them to produce significantly enhanced repositories for WordNet. **WordNet** is a laboriously hand-coded database of English words, their definitions, sets of synonyms, and various semantic relations between synonym sets. It is a major resource for NLP applications, but it has proven to be very expensive to build and maintain manually. By automatically inducing knowledge into WordNet, the potential exists to make WordNet an even greater and more comprehensive resource for NLP at a fraction of the cost. One prominent area where the benefits of NLP and WordNet are already being harvested is in customer relationship management (CRM). Broadly speaking, the goal of CRM is to maximize customer value by better understanding and effectively responding to their actual and perceived needs. An important area of CRM, where NLP is making a significant impact, is sentiment analysis. **Sentiment analysis** is a technique used to detect favorable and unfavorable opinions toward specific products and services using large numbers of textual data sources (customer feedback in the form of Web postings). A detailed coverage of sentiment analysis and WordNet is given in Section 7.7.

Text mining is also used in assessing public complaints. Application Case 7.2 provides an example where text mining is used to anticipate and address public complaints in Hong Kong.

Application Case 7.2

Text Mining Improves Hong Kong Government's Ability to Anticipate and Address Public Complaints

The 1823 Call Centre of the Hong Kong government's Efficiency Unit acts as a single point of contact for handling public inquiries and complaints on behalf of many government departments. 1823 operates round-the-clock, including during Sundays and public holidays. Each year, it answers about 2.65 million calls and 98,000 e-mails, including inquiries, suggestions, and complaints. "Having received so many calls and e-mails, we gather substantial volumes of data. The next step is to make sense of the data," says the Efficiency Unit's assistant director, W. F. Yuk. "Now, with SAS text mining technologies, we can obtain deep insights through uncovering the hidden relationship between words and sentences of complaints information, spot emerging trends and public concerns, and produce high-quality complaints intelligence for the departments we serve."

Building a "Complaints Intelligence System"

The Efficiency Unit aims to be the preferred consulting partner for all government bureaus and departments and to advance the delivery of world-class public services to the people of Hong Kong. The Unit launched the 1823 Call Centre in 2001. One of 1823's main functions is handling complaints—10 percent of the calls received last year were complaints. The Efficiency Unit recognized that there are social messages hidden in the complaints data, which provides important feedback on public service and highlights opportunities for service improvement. Rather than simply handling calls and e-mails, the Unit seeks to use the complaints information collected to gain a better understanding of daily issues for the public.

"We previously compiled some reports on complaint statistics for reference by government departments," says Yuk. "However, through 'eyeball' observations, it was absolutely impossible to effectively reveal new or more complex potential public issues and identify their root causes, as most of the complaints were recorded in unstructured textual format," says Yuk. Aiming to build a platform, called the Complaints Intelligence System, the Unit required a robust and powerful suite of text processing and mining solutions that could uncover the trends, patterns, and relationships inherent in the complaints.

Uncovering Root Causes of Issues from Unstructured Data

The Efficiency Unit chose to deploy SAS Text Miner, which can access and analyze various text formats, including e-mails received by the 1823 Call Centre. "The solution consolidates all information and uncovers hidden relationships through statistical modeling analyses," says Yuk. "It helps us understand hidden social issues so that government departments can discover them before they become serious, and thus seize the opportunities for service improvement."

Equipped with text analytics, the departments can better understand underlying issues and quickly respond even as situations evolve. Senior management can access accurate, up-to-date information from the Complaints Intelligence System.

Performance Reports at Fingertips

With the platform for SAS Business Analytics in place, the Efficiency Unit gets a boost from the system's ability to instantly generate reports. For instance, it previously took a week to compile reports on key performance indicators such as abandoned call rate, customer satisfaction rate, and first-time resolution rate. Now, these reports can be created at the click of a mouse through performance dashboards, as all complaints information is consolidated into the Complaints Intelligence System. This enables effective monitoring of the 1823 Call Centre's operations and service quality.

Strong Language Capabilities, Customized Services

Of particular importance in Hong Kong, SAS Text Miner has strong language capabilities—supporting English and traditional and simplified Chinese—and can perform automated spelling correction. The solution is also aided by the SAS capability of developing customized lists of synonyms such as the full and short forms of different government departments and to parse Chinese text for similar or identical terms whose meanings and connotations change, often dramatically, depending on the context in which they are used. "Also, throughout this 4-month project, SAS has proved to be our trusted partner," said Yuk. "We are satisfied with the comprehensive support provided by the SAS Hong Kong team."

Informed Decisions Develop Smart Strategies

"Using SAS Text Miner, 1823 can quickly discover the correlations among some key words in the complaints," says Yuk. "For instance, we can spot districts with frequent complaints received concerning public health issues such as dead birds found in residential areas. We can then inform relevant government departments and property management companies, so that they can allocate adequate resources to step up cleaning work to avoid spread of potential pandemics.

"The public's views are of course extremely important to the government. By decoding the 'messages' through statistical and root-cause analyses of complaints data, the government can better understand the voice of the people, and help government departments improve service delivery, make informed decisions, and develop smart strategies. This in turn helps boost public satisfaction with the government, and build a quality city," said W. F. Yuk, Assistant Director, Hong Kong Efficiency Unit.

Questions for Discussion

1. How did the Hong Kong government use text mining to better serve its constituents?
2. What were the challenges, the proposed solution, and the obtained results?

Sources: SAS Institute, Customer Success Story, **sas.com/success/pdf/hongkongeu.pdf** (accessed February 2013); and **enterpriseinnovation.net/whitepaper/text-mining-improves-hong-kong-governments-ability-anticipate-and-address-public**.

NLP has successfully been applied to a variety of domains for a variety of tasks via computer programs to automatically process natural human language that previously could only be done by humans. Following are among the most popular of these tasks:

- ***Question answering.*** The task of automatically answering a question posed in natural language; that is, producing a human-language answer when given a human-language question. To find the answer to a question, the computer program may use either a prestructured database or a collection of natural language documents (a text corpus such as the World Wide Web).
- ***Automatic summarization.*** The creation of a shortened version of a textual document by a computer program that contains the most important points of the original document.
- ***Natural language generation.*** Systems convert information from computer databases into readable human language.
- ***Natural language understanding.*** Systems convert samples of human language into more formal representations that are easier for computer programs to manipulate.
- ***Machine translation.*** The automatic translation of one human language to another.
- ***Foreign language reading.*** A computer program that assists a nonnative language speaker to read a foreign language with correct pronunciation and accents on different parts of the words.
- ***Foreign language writing.*** A computer program that assists a nonnative language user in writing in a foreign language.
- ***Speech recognition.*** Converts spoken words to machine-readable input. Given a sound clip of a person speaking, the system produces a text dictation.
- ***Text-to-speech.*** Also called *speech synthesis*, a computer program automatically converts normal language text into human speech.
- ***Text proofing.*** A computer program reads a proof copy of a text in order to detect and correct any errors.
- ***Optical character recognition.*** The automatic translation of images of handwritten, typewritten, or printed text (usually captured by a scanner) into machine-editable textual documents.

The success and popularity of text mining depend greatly on advancements in NLP in both generation as well as understanding of human languages. NLP enables the extraction of features from unstructured text so that a wide variety of data mining techniques can be used to extract knowledge (novel and useful patterns and relationships) from it. In that sense, simply put, text mining is a combination of NLP and data mining.

SECTION 7.3 REVIEW QUESTIONS

1. What is natural language processing?
2. How does NLP relate to text mining?
3. What are some of the benefits and challenges of NLP?
4. What are the most common tasks addressed by NLP?

7.4 TEXT MINING APPLICATIONS

As the amount of unstructured data collected by organizations increases, so does the value proposition and popularity of text mining tools. Many organizations are now realizing the importance of extracting knowledge from their document-based data repositories through the use of text mining tools. Following are only a small subset of the exemplary application categories of text mining.

Marketing Applications

Text mining can be used to increase cross-selling and up-selling by analyzing the unstructured data generated by call centers. Text generated by call center notes as well as transcriptions of voice conversations with customers can be analyzed by text mining algorithms to extract novel, actionable information about customers' perceptions toward a company's products and services. Additionally, blogs, user reviews of products at independent Web sites, and discussion board postings are a gold mine of customer sentiments. This rich collection of information, once properly analyzed, can be used to increase satisfaction and the overall lifetime value of the customer (Coussement and Van den Poel, 2008).

Text mining has become invaluable for customer relationship management. Companies can use text mining to analyze rich sets of unstructured text data, combined with the relevant structured data extracted from organizational databases, to predict customer perceptions and subsequent purchasing behavior. Coussement and Van den Poel (2009) successfully applied text mining to significantly improve the ability of a model to predict customer churn (i.e., customer attrition) so that those customers identified as most likely to leave a company are accurately identified for retention tactics.

Ghani et al. (2006) used text mining to develop a system capable of inferring implicit and explicit attributes of products to enhance retailers' ability to analyze product databases. Treating products as sets of attribute–value pairs rather than as atomic entities can potentially boost the effectiveness of many business applications, including demand forecasting, assortment optimization, product recommendations, assortment comparison across retailers and manufacturers, and product supplier selection. The proposed system allows a business to represent its products in terms of attributes and attribute values without much manual effort. The system learns these attributes by applying supervised and semi-supervised learning techniques to product descriptions found on retailers' Web sites.

Security Applications

One of the largest and most prominent text mining applications in the security domain is probably the highly classified ECHELON surveillance system. As rumor has it, ECHELON is assumed to be capable of identifying the content of telephone calls, faxes, e-mails, and other types of data and intercepting information sent via satellites, public switched telephone networks, and microwave links.

In 2007, EUROPOL developed an integrated system capable of accessing, storing, and analyzing vast amounts of structured and unstructured data sources in order to track transnational organized crime. Called the Overall Analysis System for Intelligence Support (OASIS), this system aims to integrate the most advanced data and text mining technologies available in today's market. The system has enabled EUROPOL to make significant progress in supporting its law enforcement objectives at the international level (EUROPOL, 2007).

The U.S. Federal Bureau of Investigation (FBI) and the Central Intelligence Agency (CIA), under the direction of the Department for Homeland Security, are jointly developing a supercomputer data and text mining system. The system is expected to create a gigantic data warehouse along with a variety of data and text mining modules to meet the knowledge-discovery needs of federal, state, and local law enforcement agencies. Prior to this project, the FBI and CIA each had its own separate databases, with little or no interconnection.

Another security-related application of text mining is in the area of **deception detection**. Applying text mining to a large set of real-world criminal (person-of-interest) statements, Fuller et al. (2008) developed prediction models to differentiate deceptive statements from truthful ones. Using a rich set of cues extracted from the textual statements, the model predicted the holdout samples with 70 percent accuracy, which is believed to be a significant success considering that the cues are extracted only from textual statements (no verbal or visual cues are present). Furthermore, compared to other deception-detection techniques, such as polygraph, this method is nonintrusive and widely applicable to not only textual data, but also (potentially) to transcriptions of voice recordings. A more detailed description of text-based deception detection is provided in Application Case 7.3.

Application Case 7.3

Mining for Lies

Driven by advancements in Web-based information technologies and increasing globalization, computer-mediated communication continues to filter into everyday life, bringing with it new venues for deception. The volume of text-based chat, instant messaging, text messaging, and text generated by online communities of practice is increasing rapidly. Even e-mail continues to grow in use. With the massive growth of text-based communication, the potential for people to deceive others through computer-mediated communication has also grown, and such deception can have disastrous results.

Unfortunately, in general, humans tend to perform poorly at deception-detection tasks. This phenomenon is exacerbated in text-based communications. A large part of the research on deception detection (also known as *credibility assessment*) has involved face-to-face meetings and interviews. Yet, with the growth of text-based communication, text-based deception-detection techniques are essential.

Techniques for successfully detecting deception—that is, lies—have wide applicability. Law enforcement can use decision support tools and techniques to investigate crimes, conduct security screening in airports, and monitor communications of suspected terrorists. Human resources professionals might use deception detection tools to screen applicants. These tools and techniques also have the potential to screen e-mails to uncover fraud or other wrongdoings committed by corporate officers. Although some people believe that they can readily identify those who are not being truthful, a summary of deception research showed that, on average, people are only 54 percent accurate in making veracity determinations (Bond and DePaulo, 2006). This figure may actually be worse when humans try to detect deception in text.

Using a combination of text mining and data mining techniques, Fuller et al. (2008) analyzed person-of-interest statements completed by people involved in crimes on military bases. In these statements, suspects and witnesses are required to write their recollection of the event in their own words. Military law enforcement personnel searched archival data for statements that they could conclusively identify as being truthful or deceptive. These decisions were made on the basis of corroborating evidence and case resolution. Once labeled as truthful or deceptive, the law enforcement personnel removed identifying information and gave the statements to the research team. In total, 371 usable statements were received for analysis. The text-based deception-detection method used by Fuller et al. (2008) was based on a process known as *message feature mining*, which relies on elements of data and text mining techniques. A simplified depiction of the process is provided in Figure 7.3.

First, the researchers prepared the data for processing. The original handwritten statements had to be transcribed into a word processing file. Second, features (i.e., cues) were identified. The researchers identified 31 features representing categories or types of language that are relatively independent of the text content and that can be readily analyzed by automated means. For example, first-person pronouns such as *I* or *me* can be identified without analysis of the surrounding text. Table 7.1 lists the categories and an example list of features used in this study.

The features were extracted from the textual statements and input into a flat file for further processing. Using several feature-selection methods along with *10*-fold cross-validation, the researchers compared the prediction accuracy of three popular data mining methods. Their results indicated that neural network models performed the best, with 73.46 percent prediction accuracy on test data samples; decision trees performed second best, with 71.60 percent accuracy; and logistic regression was last, with 67.28 percent accuracy.

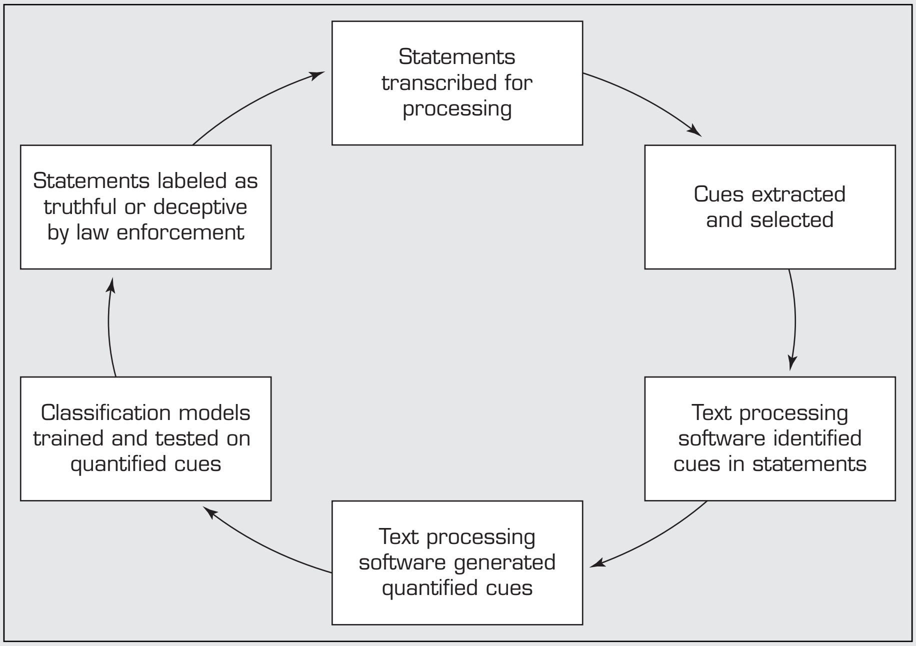

Statements transcribed for processing
Cues extracted and selected
Text processing software identified cues in statements
Text processing software generated quantified cues
Classification models trained and tested on quantified cues
Statements labeled as truthful or deceptive by law enforcement

FIGURE 7.3 Text-Based Deception-Detection Process. *Source:* C. M. Fuller, D. Biros, and D. Delen, "Exploration of Feature Selection and Advanced Classification Models for High-Stakes Deception Detection," in *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS),* January 2008, Big Island, HI, IEEE Press, pp. 80–99.

TABLE 7.1 Categories and Examples of Linguistic Features Used in Deception Detection

Number	Construct (Category)	Example Cues
1	Quantity	Verb count, noun-phrase count, etc.
2	Complexity	Average number of clauses, average sentence length, etc.
3	Uncertainty	Modifiers, modal verbs, etc.
4	Nonimmediacy	Passive voice, objectification, etc.
5	Expressivity	Emotiveness
6	Diversity	Lexical diversity, redundancy, etc.
7	Informality	Typographical error ratio
8	Specificity	Spatiotemporal information, perceptual information, etc.
9	Affect	Positive affect, negative affect, etc.

The results indicate that automated text-based deception detection has the potential to aid those who must try to detect lies in text and can be successfully applied to real-world data. The accuracy of these techniques exceeded the accuracy of most other deception-detection techniques even though it was limited to textual cues.

(*Continued*)

Application Case 7.3 (Continued)

Questions for Discussion

1. Why is it difficult to detect deception?
2. How can text/data mining be used to detect deception in text?
3. What do you think are the main challenges for such an automated system?

Sources: C. M. Fuller, D. Biros, and D. Delen, "Exploration of Feature Selection and Advanced Classification Models for High-Stakes Deception Detection," in *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS),* 2008, Big Island, HI, IEEE Press, pp. 80–99; C. F. Bond and B. M. DePaulo, "Accuracy of Deception Judgments," *Personality and Social Psychology Reports,* Vol. 10, No. 3, 2006, pp. 214–234.

Biomedical Applications

Text mining holds great potential for the medical field in general and biomedicine in particular for several reasons. First, the published literature and publication outlets (especially with the advent of the open source journals) in the field are expanding at an exponential rate. Second, compared to most other fields, the medical literature is more standardized and orderly, making it a more "minable" information source. Finally, the terminology used in this literature is relatively constant, having a fairly standardized ontology. What follows are a few exemplary studies where text mining techniques were successfully used in extracting novel patterns from biomedical literature.

Experimental techniques such as DNA microarray analysis, serial analysis of gene expression (SAGE), and mass spectrometry proteomics, among others, are generating large amounts of data related to genes and proteins. As in any other experimental approach, it is necessary to analyze this vast amount of data in the context of previously known information about the biological entities under study. The literature is a particularly valuable source of information for experiment validation and interpretation. Therefore, the development of automated text mining tools to assist in such interpretation is one of the main challenges in current bioinformatics research.

Knowing the location of a protein within a cell can help to elucidate its role in biological processes and to determine its potential as a drug target. Numerous location-prediction systems are described in the literature; some focus on specific organisms, whereas others attempt to analyze a wide range of organisms. Shatkay et al. (2007) proposed a comprehensive system that uses several types of sequence- and text-based features to predict the location of proteins. The main novelty of their system lies in the way in which it selects its text sources and features and integrates them with sequence-based features. They tested the system on previously used data sets and on new data sets devised specifically to test its predictive power. The results showed that their system consistently beat previously reported results.

Chun et al. (2006) described a system that extracts disease–gene relationships from literature accessed via MedLine. They constructed a dictionary for disease and gene names from six public databases and extracted relation candidates by dictionary matching. Because dictionary matching produces a large number of false positives, they developed a method of machine learning–based named entity recognition (NER) to filter out false recognitions of disease/gene names. They found that the success of relation extraction is heavily dependent on the performance of NER filtering and that the filtering improved the precision of relation extraction by 26.7 percent, at the cost of a small reduction in recall.

Figure 7.4 shows a simplified depiction of a multilevel text analysis process for discovering gene–protein relationships (or protein–protein interactions) in the biomedical literature (Nakov et al., 2005). As can be seen in this simplified example that uses a simple sentence from biomedical text, first (at the bottom three levels) the text is tokenized

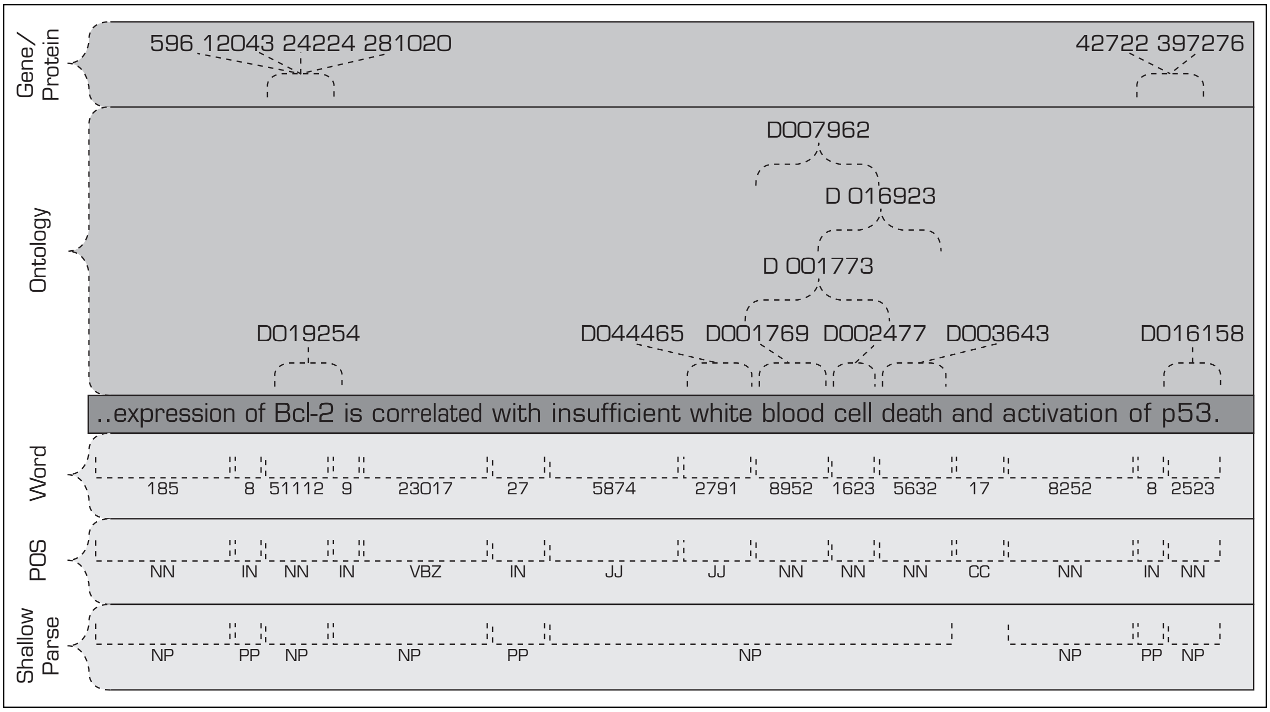

Gene/Protein
596 12043 24224 281020
42722 397276
Ontology
D007962
D 016923
D 001773
D019254
D044465 D001769 D002477 D003643
D016158
..expression of Bcl-2 is correlated with insufficient white blood cell death and activation of p53.
Word
185 8 51112 9 23017 27 5874 2791 8952 1623 5632 17 8252 8 2523
POS
NN IN NN IN VBZ IN JJ JJ NN NN NN CC NN IN NN
Shallow Parse
NP PP NP NP PP NP NP PP NP

FIGURE 7.4 Multilevel Analysis of Text for Gene/Protein Interaction Identification. *Source:* P. Nakov, A. Schwartz, B. Wolf, and M. A. Hearst, "Supporting Annotation Layers for Natural Language Processing," *Proceedings of the Association for Computational Linguistics (ACL),* interactive poster and demonstration sessions, 2005, Ann Arbor, MI, Association for Computational Linguistics, pp. 65–68.

using **part-of-speech tagging** and shallow-parsing. The tokenized terms (words) are then matched (and interpreted) against the hierarchical representation of the domain ontology to derive the gene–protein relationship. Application of this method (and/or some variation of it) to the biomedical literature offers great potential to decode the complexities in the Human Genome Project.

Academic Applications

The issue of text mining is of great importance to publishers who hold large databases of information requiring indexing for better retrieval. This is particularly true in scientific disciplines, in which highly specific information is often contained within written text. Initiatives have been launched, such as *Nature's* proposal for an Open Text Mining Interface (OTMI) and the National Institutes of Health's common Journal Publishing Document Type Definition (DTD), which would provide semantic cues to machines to answer specific queries contained within text without removing publisher barriers to public access.

Academic institutions have also launched text mining initiatives. For example, the National Centre for Text Mining, a collaborative effort between the Universities of Manchester and Liverpool, provides customized tools, research facilities, and advice on text mining to the academic community. With an initial focus on text mining in the biological and biomedical sciences, research has since expanded into the social sciences. In the United States, the School of Information at the University of California, Berkeley, is developing a program called BioText to assist bioscience researchers in text mining and analysis.

As described in this section, text mining has a wide variety of applications in a number of different disciplines. See Application Case 7.4 for an example of how a financial services firm is using text mining to improve its customer service performance.

Application Case 7.4

Text Mining and Sentiment Analysis Help Improve Customer Service Performance

The company is a financial services firm that provides a broad range of solutions and services to a global customer base. The company has a comprehensive network of facilities around the world, with over 5000 associates assisting their customers. Customers lodge service requests by telephone, email, or through an online chat interface.

As a B2C service provider, the company strives to maintain high standards for effective communication between their associates and customers, and tries to monitor customer interactions at every opportunity. The broad objective of this service performance monitoring is to maintain satisfactory quality of service over time and across the organization. To this end, the company has devised a set of standards for service excellence, to which all customer interactions are expected to adhere. These standards comprise different qualitative measures of service levels (e.g., associates should use clear and understandable language, associates should always maintain a professional and friendly demeanor, etc.) Associates' performances are measured based on compliance with these quality standards. Organizational units at different levels, like teams, departments, and the company as a whole, also receive scores based on associate performances. The evaluations and remunerations of not only the associates but also of management are influenced by these service performance scores.

Challenge

Continually monitoring service levels is essential for service quality control. Customer surveys are an excellent way of gathering feedback about service levels. An even richer source of information is the corpus of associate-customer interactions. Historically the company manually evaluated a sample of associate-customer interactions and survey responses for compliance with excellence standards. This approach, in addition to being subjective and error-prone, was time- and labor-intensive. Advances in machine learning and computational linguistics offer an opportunity to objectively evaluate all customer interactions in a timely manner.

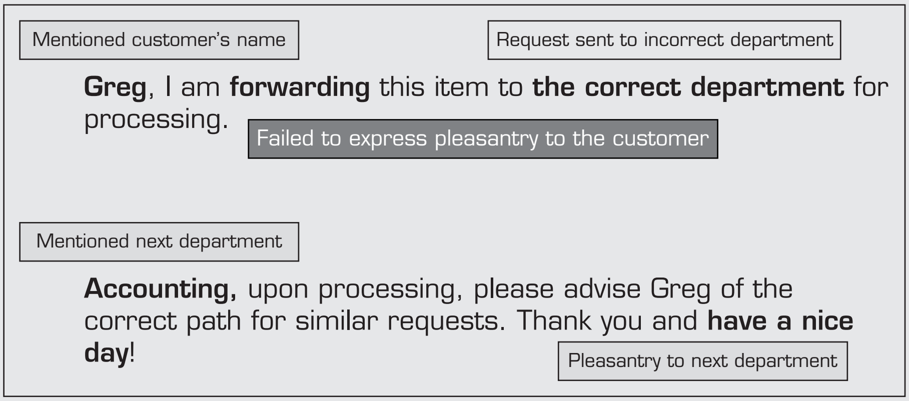
The company needs a system for (1) automatically evaluating associate-customer interactions for compliance with quality standards and (2) analyzing survey responses to extract positive and negative feedback. The analysis must be able to account for the wide diversity of expression in natural language (e.g., pleasant and reassuring tone, acceptable language, appropriate abbreviations, addressing all of the customers' issues, etc.).

Solution

PolyAnalyst 6.5™ by Megaputer Intelligence is a data mining and analysis platform that provides a comprehensive set of tools for analyzing structured and unstructured data. PolyAnalyst's text analysis tools are used for extracting complex word patterns, grammatical and semantic relationships, and expressions of sentiment. The results of these text analyses are then classified into context-specific themes to identify actionable issues, which can be assigned to relevant individuals responsible for their resolution. The system can be programmed to provide feedback in case of insufficient classification so that analyses can be modified or amended. The relationships between structured fields and text analysis results are also established in order to identify patterns and interactions. The system publishes the results of analyses through graphical, interactive, web-based reports. Users create analysis scenarios using a drag-and-drop graphical user interface (GUI). These scenarios are reusable solutions that can be programmed to automate the analysis and report generation process.

A set of specific criteria were designed to capture and automatically detect compliance with the company's Quality Standards. The figure below displays an example of an associate's response, as well as the quality criteria that it succeeds or fails to match.

As illustrated above, this comment matches several criteria while failing to match one, and contributes accordingly to the associate's performance score. These scores are then automatically calculated and aggregated across various organizational units. It is relatively easy to modify the system in case of changes in quality standards, and the changes can be quickly applied to historical data. The system also has an integrated case management system, which generates email alerts in case of



Greg, I am **forwarding** this item to **the correct department** for processing.

Accounting, upon processing, please advise Greg of the correct path for similar requests. Thank you and **have a nice day**!

Mentioned customer's name
Request sent to incorrect department
Failed to express pleasantry to the customer
Mentioned next department
Pleasantry to next department

drops in service quality and allows users to track the progress of issue resolution.

Tangible Results

1. Completely automated analysis; saves time.
2. Analysis of entire dataset (> 1 million records per year); no need for sampling.
3. 45% cost savings over traditional analysis.
4. Weekly processing. In the case of traditional analysis, data could only be processed monthly due to time and resource constraints.
5. Analysis not subjective to the analyst.
   a. Increased accuracy.
   b. Increased uniformity.
6. Greater accountability. Associates can review the analysis and raise concerns in case of discrepancies.

Future Directions

Currently the corpus of associate-customer interactions does not include transcripts of phone conversations. By incorporating speech recognition capability, the system can become a one-stop destination for analyzing all customer interactions. The system could also potentially be used in real-time, instead of periodic analyses.

Questions for Discussion

1. How did the financial services firm use text mining and text analytics to improve its customer service performance?
2. What were the challenges, the proposed solution, and the obtained results?

Source: Megaputer, Customer Success Story, megaputer.com (accessed September 2013).

SECTION 7.4 REVIEW QUESTIONS

1. List and briefly discuss some of the text mining applications in marketing.

2. How can text mining be used in security and counterterrorism?

3. What are some promising text mining applications in biomedicine?

7.5 TEXT MINING PROCESS

In order to be successful, text mining studies should follow a sound methodology based on best practices. A standardized process model is needed similar to CRISP-DM, which is the industry standard for data mining projects (see Chapter 5). Even though most parts of CRISP-DM are also applicable to text mining projects, a specific process model for text mining would include much more elaborate data preprocessing activities. Figure 7.5 depicts a high-level context diagram of a typical text mining process (Delen and Crossland, 2008). This context diagram presents the scope of the process, emphasizing its interfaces with the larger environment. In essence, it draws boundaries around the specific process to explicitly identify what is included in (and excluded from) the text mining process.

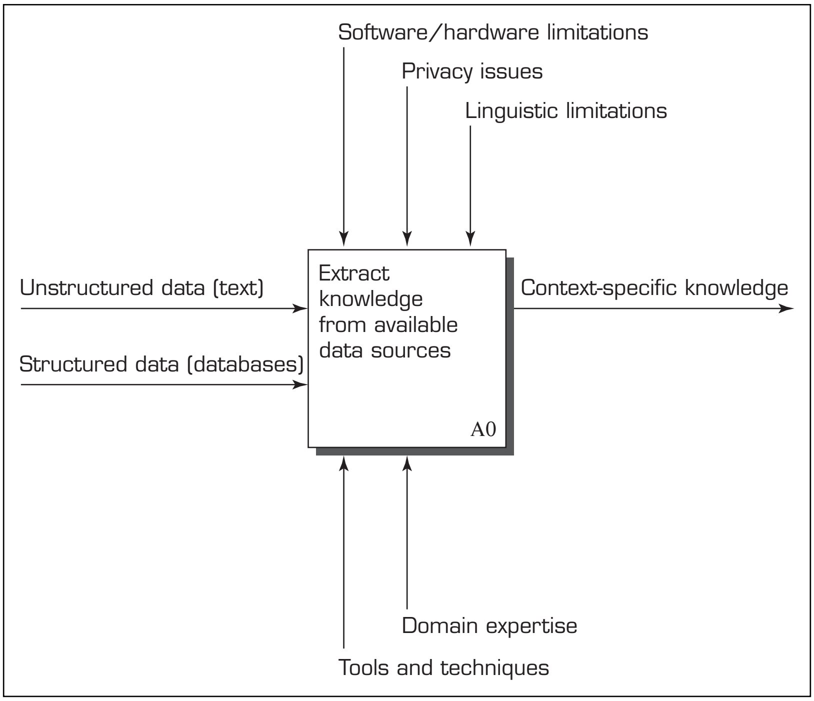

Software/hardware limitations
Privacy issues
Linguistic limitations
Unstructured data (text)
Structured data (databases)
Extract knowledge from available data sources
A0
Context-specific knowledge
Domain expertise
Tools and techniques

FIGURE 7.5 Context Diagram for the Text Mining Process.

As the context diagram indicates, the input (inward connection to the left edge of the box) into the text-based knowledge-discovery process is the unstructured as well as structured data collected, stored, and made available to the process. The output (outward extension from the right edge of the box) of the process is the context-specific knowledge that can be used for decision making. The controls, also called the *constraints* (inward connection to the top edge of the box), of the process include software and hardware limitations, privacy issues, and the difficulties related to processing the text that is presented in the form of natural language. The mechanisms (inward connection to the bottom edge of the box) of the process include proper techniques, software tools, and domain expertise. The primary purpose of text mining (within the context of knowledge discovery) is to process unstructured (textual) data (along with structured data, if relevant to the problem being addressed and available) to extract meaningful and actionable patterns for better decision making.

At a very high level, the text mining process can be broken down into three consecutive tasks, each of which has specific inputs to generate certain outputs (see Figure 7.6). If, for some reason, the output of a task is not what is expected, a backward redirection to the previous task execution is necessary.

Task 1: Establish the Corpus

The main purpose of the first task activity is to collect all of the documents related to the context (domain of interest) being studied. This collection may include textual documents, XML files, e-mails, Web pages, and short notes. In addition to the readily available textual data, voice recordings may also be transcribed using speech-recognition algorithms and made a part of the text collection.

Once collected, the text documents are transformed and organized in a manner such that they are all in the same representational form (e.g., ASCII text files) for computer processing. The organization of the documents can be as simple as a collection of digitized text excerpts stored in a file folder or it can be a list of links to a collection of Web pages in a specific domain. Many commercially available text mining software tools

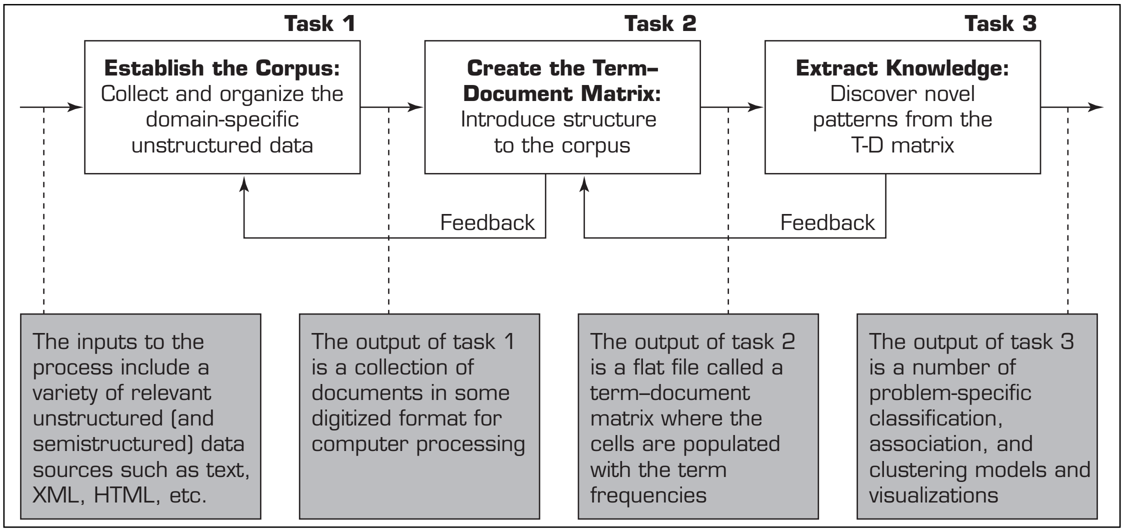

Task 1
Task 2
Task 3
Establish the Corpus: Collect and organize the domain-specific unstructured data
Create the Term–Document Matrix: Introduce structure to the corpus
Extract Knowledge: Discover novel patterns from the T-D matrix
Feedback
Feedback
The inputs to the process include a variety of relevant unstructured (and semistructured) data sources such as text, XML, HTML, etc.
The output of task 1 is a collection of documents in some digitized format for computer processing
The output of task 2 is a flat file called a term–document matrix where the cells are populated with the term frequencies
The output of task 3 is a number of problem-specific classification, association, and clustering models and visualizations

FIGURE 7.6 The Three-Step Text Mining Process.

could accept these as input and convert them into a flat file for processing. Alternatively, the flat file can be prepared outside the text mining software and then presented as the input to the text mining application.

Task 2: Create the Term–Document Matrix

In this task, the digitized and organized documents (the corpus) are used to create the **term–document matrix (TDM)**. In the TDM, rows represent the documents and columns represent the terms. The relationships between the terms and documents are characterized by indices (i.e., a relational measure that can be as simple as the number of occurrences of the term in respective documents). Figure 7.7 is a typical example of a TDM.

Terms / Documents	Investment Risk	Project Management	Software Engineering	Development	SAP	. . .
Document 1	1			1		
Document 2		1				
Document 3			3		1	
Document 4		1				
Document 5			2	1		
Document 6	1			1		
. . .						

FIGURE 7.7 A Simple Term–Document Matrix.

The goal is to convert the list of organized documents (the corpus) into a TDM where the cells are filled with the most appropriate indices. The assumption is that the essence of a document can be represented with a list and frequency of the terms used in that document. However, are all terms important when characterizing documents? Obviously, the answer is "no." Some terms, such as articles, auxiliary verbs, and terms used in almost all of the documents in the corpus, have no differentiating power and therefore should be excluded from the indexing process. This list of terms, commonly called *stop terms* or *stop words,* is specific to the domain of study and should be identified by the domain experts. On the other hand, one might choose a set of predetermined terms under which the documents are to be indexed (this list of terms is conveniently called *include terms* or *dictionary*). Additionally, synonyms (pairs of terms that are to be treated the same) and specific phrases (e.g., "Eiffel Tower") can also be provided so that the index entries are more accurate.

Another filtration that should take place to accurately create the indices is *stemming*, which refers to the reduction of words to their roots so that, for example, different grammatical forms or declinations of a verb are identified and indexed as the same word. For example, stemming will ensure that *modeling* and *modeled* will be recognized as the word *model*.

The first generation of the TDM includes all of the unique terms identified in the corpus (as its columns), excluding the ones in the stop term list; all of the documents (as its rows); and the occurrence count of each term for each document (as its cell values). If, as is commonly the case, the corpus includes a rather large number of documents, then there is a very good chance that the TDM will have a very large number of terms. Processing such a large matrix might be time-consuming and, more importantly, might lead to extraction of inaccurate patterns. At this point, one has to decide the following: (1) What is the best representation of the indices? and (2) How can we reduce the dimensionality of this matrix to a manageable size?

REPRESENTING THE INDICES Once the input documents are indexed and the initial word frequencies (by document) computed, a number of additional transformations can be performed to summarize and aggregate the extracted information. The raw term frequencies generally reflect on how salient or important a word is in each document. Specifically, words that occur with greater frequency in a document are better descriptors of the contents of that document. However, it is not reasonable to assume that the word counts themselves are proportional to their importance as descriptors of the documents. For example, if a word occurs one time in document *A*, but three times in document *B*, then it is not necessarily reasonable to conclude that this word is three times as important a descriptor of document *B* as compared to document *A*. In order to have a more consistent TDM for further analysis, these raw indices need to be normalized. As opposed to showing the actual frequency counts, the numerical representation between terms and documents can be normalized using a number of alternative methods. The following are a few of the most commonly used normalization methods (StatSoft, 2009):

- ***Log frequencies.*** The raw frequencies can be transformed using the log function. This transformation would "dampen" the raw frequencies and how they affect the results of subsequent analysis.

  $$f(wf) = 1 + \log(wf) \quad \text{for} \quad wf > 0$$

  In the formula, wf is the raw word (or term) frequency and $f(wf)$ is the result of the log transformation. This transformation is applied to all of the raw frequencies in the TDM where the frequency is greater than zero.

- ***Binary frequencies.*** Likewise, an even simpler transformation can be used to enumerate whether a term is used in a document.

  $$f(wf) = 1 \quad \text{for} \quad wf > 0$$

The resulting TDM matrix will contain only 1s and 0s to indicate the presence or absence of the respective words. Again, this transformation will dampen the effect of the raw frequency counts on subsequent computations and analyses.

- ***Inverse document frequencies.*** Another issue that one may want to consider more carefully and reflect in the indices used in further analyses is the relative document frequencies (*df*) of different terms. For example, a term such as *guess* may occur frequently in all documents, whereas another term, such as *software,* may appear only a few times. The reason is that one might make *guesses* in various contexts, regardless of the specific topic, whereas *software* is a more semantically focused term that is only likely to occur in documents that deal with computer software. A common and very useful transformation that reflects both the specificity of words (document frequencies) as well as the overall frequencies of their occurrences (term frequencies) is the so-called **inverse document frequency** (Manning and Schutze, 2009). This transformation for the *i*th word and *j*th document can be written as:

$$idf(i,j) = \begin{cases} 0 & if\ wf_{ij} = 0 \\ (1 + \log(wf_{ij}))\log\dfrac{N}{df_i} & if\ wf_{ij} \geq 1 \end{cases}$$

In this formula, N is the total number of documents, and df_{i} is the document frequency for the *i*th word (the number of documents that include this word). Hence, it can be seen that this formula includes both the dampening of the simple-word frequencies via the log function (described here) and a weighting factor that evaluates to 0 if the word occurs in all documents [i.e., $\log(N/N = 1) = 0$], and to the maximum value when a word only occurs in a single document [i.e., $\log(N/1) = \log(N)$]. It can easily be seen how this transformation will create indices that reflect both the relative frequencies of occurrences of words as well as their semantic specificities over the documents included in the analysis. This is the most commonly used transformation in the field.

REDUCING THE DIMENSIONALITY OF THE MATRIX Because the TDM is often very large and rather sparse (most of the cells filled with zeros), another important question is "How do we reduce the dimensionality of this matrix to a manageable size?" Several options are available for managing the matrix size:

- A domain expert goes through the list of terms and eliminates those that do not make much sense for the context of the study (this is a manual, labor-intensive process).
- Eliminate terms with very few occurrences in very few documents.
- Transform the matrix using singular value decomposition.

Singular value decomposition (SVD), which is closely related to principal components analysis, reduces the overall dimensionality of the input matrix (number of input documents by number of extracted terms) to a lower dimensional space, where each consecutive dimension represents the largest degree of variability (between words and documents) possible (Manning and Schutze, 1999). Ideally, the analyst might identify the two or three most salient dimensions that account for most of the variability (differences) between the words and documents, thus identifying the latent semantic space that organizes the words and documents in the analysis. Once such dimensions are identified, the underlying "meaning" of what is contained (discussed or described) in the documents has been extracted. Specifically, assume that matrix A represents an $m \times n$ term occurrence matrix where m is the number of input documents and n is the number of terms selected for analysis. The SVD computes the $m \times r$ orthogonal matrix U, $n \times r$ orthogonal matrix V, and $r \times r$ matrix D, so that $A = UDV'$ and r is the number of eigen values of $A'A$.

Task 3: Extract the Knowledge

Using the well-structured TDM, and potentially augmented with other structured data elements, novel patterns are extracted in the context of the specific problem being addressed. The main categories of knowledge extraction methods are classification, clustering, association, and trend analysis. A short description of these methods follows.

CLASSIFICATION Arguably the most common knowledge-discovery topic in analyzing complex data sources is the **classification** (or categorization) of certain objects. The task is to classify a given data instance into a predetermined set of categories (or classes). As it applies to the domain of text mining, the task is known as *text categorization,* where for a given set of categories (subjects, topics, or concepts) and a collection of text documents the goal is to find the correct topic (subject or concept) for each document using models developed with a training data set that includes both the documents and actual document categories. Today, automated text classification is applied in a variety of contexts, including automatic or semiautomatic (interactive) indexing of text, spam filtering, Web page categorization under hierarchical catalogs, automatic generation of metadata, detection of genre, and many others.

The two main approaches to text classification are knowledge engineering and machine learning (Feldman and Sanger, 2007). With the knowledge-engineering approach, an expert's knowledge about the categories is encoded into the system either declaratively or in the form of procedural classification rules. With the machine-learning approach, a general inductive process builds a classifier by learning from a set of reclassified examples. As the number of documents increases at an exponential rate and as knowledge experts become harder to come by, the popularity trend between the two is shifting toward the machine-learning approach.

CLUSTERING **Clustering** is an unsupervised process whereby objects are classified into "natural" groups called *clusters*. Compared to categorization, where a collection of preclassified training examples is used to develop a model based on the descriptive features of the classes in order to classify a new unlabeled example, in clustering the problem is to group an unlabelled collection of objects (e.g., documents, customer comments, Web pages) into meaningful clusters without any prior knowledge.

Clustering is useful in a wide range of applications, from document retrieval to enabling better Web content searches. In fact, one of the prominent applications of clustering is the analysis and navigation of very large text collections, such as Web pages. The basic underlying assumption is that relevant documents tend to be more similar to each other than to irrelevant ones. If this assumption holds, the clustering of documents based on the similarity of their content improves search effectiveness (Feldman and Sanger, 2007):

- ***Improved search recall.*** Clustering, because it is based on overall similarity as opposed to the presence of a single term, can improve the recall of a query-based search in such a way that when a query matches a document its whole cluster is returned.
- ***Improved search precision.*** Clustering can also improve search precision. As the number of documents in a collection grows, it becomes difficult to browse through the list of matched documents. Clustering can help by grouping the documents into a number of much smaller groups of related documents, ordering them by relevance, and returning only the documents from the most relevant group (or groups).

The two most popular clustering methods are scatter/gather clustering and query-specific clustering:

- ***Scatter/gather.*** This document browsing method uses clustering to enhance the efficiency of human browsing of documents when a specific search query cannot be formulated. In a sense, the method dynamically generates a table of contents for the collection and adapts and modifies it in response to the user selection.
- ***Query-specific clustering.*** This method employs a hierarchical clustering approach where the most relevant documents to the posed query appear in small tight clusters that are nested in larger clusters containing less similar documents, creating a spectrum of relevance levels among the documents. This method performs consistently well for document collections of realistically large sizes.

ASSOCIATION A formal definition and detailed description of **association** was provided in the chapter on data mining (Chapter 5). Associations, or *association rule learning in data mining,* is a popular and well-researched technique for discovering interesting relationships among variables in large databases. The main idea in generating association rules (or solving market-basket problems) is to identify the frequent sets that go together.

In text mining, associations specifically refer to the direct relationships between concepts (terms) or sets of concepts. The concept set association rule $A \Rightarrow B$, relating two frequent concept sets A and C, can be quantified by the two basic measures of support and confidence. In this case, confidence is the percentage of documents that include all the concepts in C within the same subset of those documents that include all the concepts in A. Support is the percentage (or number) of documents that include all the concepts in A and C. For instance, in a document collection the concept "Software Implementation Failure" may appear most often in association with "Enterprise Resource Planning" and "Customer Relationship Management" with significant support (4%) and confidence (55%), meaning that 4 percent of the documents had all three concepts represented together in the same document and of the documents that included "Software Implementation Failure," 55 percent of them also included "Enterprise Resource Planning" and "Customer Relationship Management."

Text mining with association rules was used to analyze published literature (news and academic articles posted on the Web) to chart the outbreak and progress of bird flu (Mahgoub et al., 2008). The idea was to automatically identify the association among the geographic areas, spreading across species, and countermeasures (treatments).

TREND ANALYSIS Recent methods of trend analysis in text mining have been based on the notion that the various types of concept distributions are functions of document collections; that is, different collections lead to different concept distributions for the same set of concepts. It is therefore possible to compare two distributions that are otherwise identical except that they are from different subcollections. One notable direction of this type of analyses is having two collections from the same source (such as from the same set of academic journals) but from different points in time. Delen and Crossland (2008) applied **trend analysis** to a large number of academic articles (published in the three highest-rated academic journals) to identify the evolution of key concepts in the field of information systems.

As described in this section, a number of methods are available for text mining. Application Case 7.5 describes the use of a number of different techniques in analyzing a large set of literature.

Application Case 7.5

Research Literature Survey with Text Mining

Researchers conducting searches and reviews of relevant literature face an increasingly complex and voluminous task. In extending the body of relevant knowledge, it has always been important to work hard to gather, organize, analyze, and assimilate existing information from the literature, particularly from one's home discipline. With the increasing abundance of potentially significant research being reported in related fields, and even in what are traditionally deemed to be nonrelated fields of study, the researcher's task is ever more daunting, if a thorough job is desired.

In new streams of research, the researcher's task may be even more tedious and complex. Trying to ferret out relevant work that others have reported may be difficult, at best, and perhaps even near impossible if traditional, largely manual reviews of published literature are required. Even with a legion of dedicated graduate students or helpful colleagues, trying to cover all potentially relevant published work is problematic.

Many scholarly conferences take place every year. In addition to extending the body of knowledge of the current focus of a conference, organizers often desire to offer additional mini-tracks and workshops. In many cases, these additional events are intended to introduce the attendees to significant streams of research in related fields of study and to try to identify the "next big thing" in terms of research interests and focus. Identifying reasonable candidate topics for such mini-tracks and workshops is often subjective rather than derived objectively from the existing and emerging research.

In a recent study, Delen and Crossland (2008) proposed a method to greatly assist and enhance the efforts of the researchers by enabling a semi-automated analysis of large volumes of published literature through the application of text mining. Using standard digital libraries and online publication search engines, the authors downloaded and collected all of the available articles for the three major journals in the field of management information systems: *MIS Quarterly* (MISQ), *Information Systems Research* (ISR), and the *Journal of Management Information Systems* (JMIS). In order to maintain the same time interval for all three journals (for potential comparative longitudinal studies), the journal with the most recent starting date for its digital publication availability was used as the start time for this study (i.e., JMIS articles have been digitally available since 1994). For each article, they extracted the title, abstract, author list, published keywords, volume, issue number, and year of publication. They then loaded all of the article data into a simple database file. Also included in the combined data set was a field that designated the journal type of each article for likely discriminatory analysis. Editorial notes, research notes, and executive overviews were omitted from the collection. Table 7.2 shows how the data was presented in a tabular format.

In the analysis phase, they chose to use only the abstract of an article as the source of information extraction. They chose not to include the keywords listed with the publications for two main reasons: (1) under normal circumstances, the abstract would already include the listed keywords, and therefore inclusion of the listed keywords for the analysis would mean repeating the same information and potentially giving them unmerited weight; and (2) the listed keywords may be terms that authors would like their article to be associated with (as opposed to what is really contained in the article), therefore potentially introducing unquantifiable bias to the analysis of the content.

The first exploratory study was to look at the longitudinal perspective of the three journals (i.e., evolution of research topics over time). In order to conduct a longitudinal study, they divided the 12-year period (from 1994 to 2005) into four 3-year periods for each of the three journals. This framework led to 12 text mining experiments with 12 mutually exclusive data sets. At this point, for each of the 12 data sets they used text mining to extract the most descriptive terms from these collections of articles represented by their abstracts. The results were tabulated and examined for time-varying changes in the terms published in these three journals.

As a second exploration, using the complete data set (including all three journals and all four

TABLE 7.2 Tabular Representation of the Fields Included in the Combined Data Set

Journal	Year	Author(s)	Title	Vol/No	Pages	Keywords	Abstract
MISQ	2005	A. Malhotra, S. Gossain, and O. A. El Sawy	Absorptive capacity configurations in supply chains: Gearing for partner-enabled market knowledge creation	29/1	145–187	knowledge management supply chain absorptive capacity interorganizational information systems configuration approaches	The need for continual value innovation is driving supply chains to evolve from a pure transactional focus to leveraging interorganization partnerships for sharing
ISR	1999	D. Robey and M. C. Boudtreau	Accounting for the contradictory organizational consequences of information technology: Theoretical directions and methodological implications		165–185	organizational transformation impacts of technology organization theory research methodology intraorganizational power electronic communication misimplementation culture systems	Although much contemporary thought considers advanced information technologies as either determinants or enablers of radical organizational change, empirical studies have revealed inconsistent findings to support the deterministic logic implicit in such arguments. This paper reviews the contradictory…
JMIS	2001	R. Aron and E. K. Clemons	Achieving the optimal balance between investment in quality and investment in self-promotion for information products		65–88	information products Internet advertising product positioning signaling signaling games	When producers of goods (or services) are confronted by a situation in which their offerings no longer perfectly match consumer preferences, they must determine the extent to which the advertised features of…

(Continued)

Application Case 7.5 (Continued)

periods), they conducted a clustering analysis. Clustering is arguably the most commonly used text mining technique. Clustering was used in this study to identify the natural groupings of the articles (by putting them into separate clusters) and then to list the most descriptive terms that characterized those clusters. They used singular value decomposition to reduce the dimensionality of the term-by-document matrix and then an expectation-maximization algorithm to create the clusters. They conducted several experiments to identify the *optimal* number of clusters, which turned out to be nine. After the construction of the nine clusters, they analyzed the content of those clusters from two perspectives: (1) representation of the journal type (see Figure 7.8) and (2) representation of time. The idea was to explore the potential differences and/or commonalities among the three journals and potential changes in the emphasis on those clusters; that is, to answer questions such as "Are there clusters that represent different research themes specific to a single journal?" and "Is there a time-varying characterization of those clusters?" They discovered and discussed several interesting patterns using tabular and graphical representation of their findings (for further information see Delen and Crossland, 2008).

Questions for Discussion

1. How can text mining be used to ease the task of literature review?
2. What are the common outcomes of a text mining project on a specific collection of journal articles? Can you think of other potential outcomes not mentioned in this case?

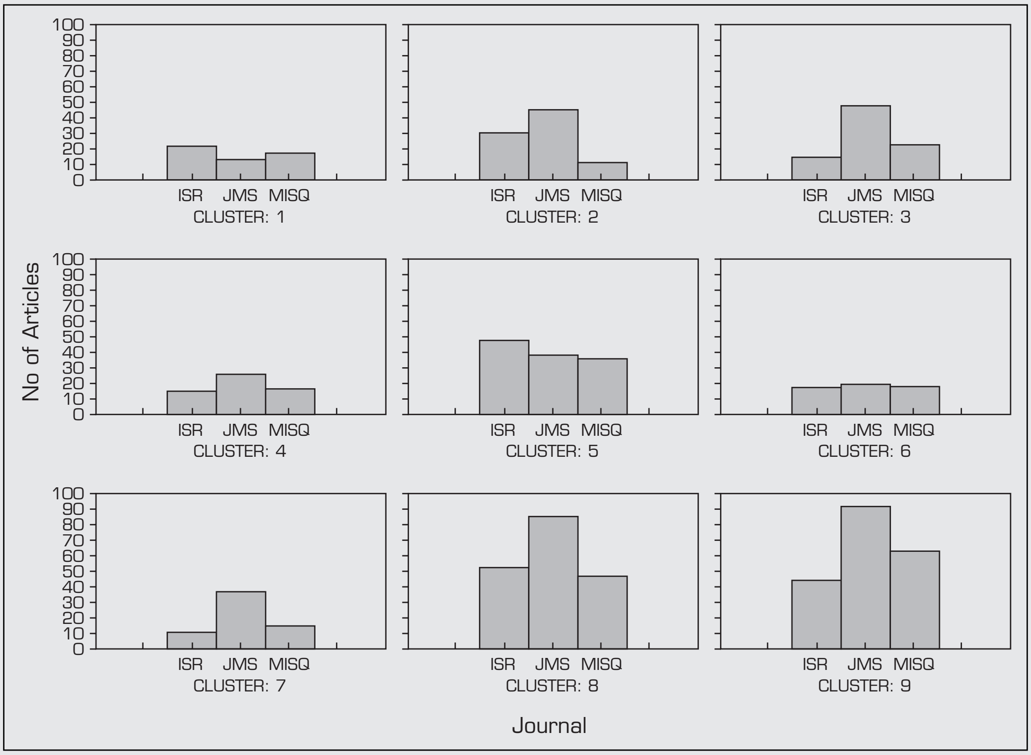


FIGURE 7.8 Distribution of the Number of Articles for the Three Journals over the Nine Clusters. *Source:* D. Delen and M. Crossland, "Seeding the Survey and Analysis of Research Literature with Text Mining," *Expert Systems with Applications,* Vol. 34, No. 3, 2008, pp. 1707–1720.

SECTION 7.5 REVIEW QUESTIONS

1. What are the main steps in the text mining process?
2. What is the reason for normalizing word frequencies? What are the common methods for normalizing word frequencies?
3. What is singular value decomposition? How is it used in text mining?
4. What are the main knowledge extraction methods from corpus?

7.6 TEXT MINING TOOLS

As the value of text mining is being realized by more and more organizations, the number of software tools offered by software companies and nonprofits is also increasing. Following are some of the popular text mining tools, which we classify as commercial software tools and free (and/or open source) software tools.

Commercial Software Tools

The following are some of the most popular software tools used for text mining. Note that many companies offer demonstration versions of their products on their Web sites.

1. ClearForest offers text analysis and visualization tools.
2. IBM offers SPSS Modeler and data and text analytics toolkits.
3. Megaputer Text Analyst offers semantic analysis of free-form text, summarization, clustering, navigation, and natural language retrieval with search dynamic refocusing.
4. SAS Text Miner provides a rich suite of text processing and analysis tools.
5. KXEN Text Coder (KTC) offers a text analytics solution for automatically preparing and transforming unstructured text attributes into a structured representation for use in KXEN Analytic Framework.
6. The Statistica Text Mining engine provides easy-to-use text mining functionality with exceptional visualization capabilities.
7. VantagePoint provides a variety of interactive graphical views and analysis tools with powerful capabilities to discover knowledge from text databases.
8. The WordStat analysis module from Provalis Research analyzes textual information such as responses to open-ended questions, interviews, etc.
9. Clarabridge text mining software provides end-to-end solutions for customer experience professionals wishing to transform customer feedback for marketing, service, and product improvements.

Free Software Tools

Free software tools, some of which are open source, are available from a number of nonprofit organizations:

1. RapidMiner, one of the most popular free, open source software tools for data mining and text mining, is tailored with a graphically appealing, drag-and-drop user interface.
2. Open Calais is an open source toolkit for including semantic functionality within your blog, content management system, Web site, or application.
3. GATE is a leading open source toolkit for text mining. It has a free open source framework (or SDK) and graphical development environment.
4. LingPipe is a suite of Java libraries for the linguistic analysis of human language.
5. S-EM (Spy-EM) is a text classification system that learns from positive and unlabeled examples.
6. Vivisimo/Clusty is a Web search and text-clustering engine.

Often, innovative application of text mining comes from the collective use of several software tools. Application Case 7.6 illustrates a few customer case study synopses where text mining and advanced analytics are used to address a variety of business challenges.

Application Case 7.6

A Potpourri of Text Mining Case Synopses

1. Alberta's Parks Division gains insight from unstructured data

Business Issue:

Alberta's Parks Division was relying on manual processes to respond to stakeholders, which was time-consuming and made it difficult to glean insight from unstructured data sources.

Solution:

Using SAS Text Miner, the Parks Division is able to reduce a three-week process down to a couple of days, and discover new insights in a matter of minutes.

Benefits:

The solution has not only automated manual tasks, but also provides insight into both structured and unstructured data sources that was previously not possible.

"We now have opportunities to channel customer communications into products and services that meet their needs. Having the analytics will enable us to better support changes in program delivery," said Roy Finzel, Manager of Business Integration and Analysis, Alberta Tourism, Parks and Recreation.

For more details, please go to http://www.sas.com/success/alberta-parks2012.html

2. American Honda Saves Millions by Using Text and Data Mining

Business Issue:

One of the most admired and recognized automobile brands in the United States, American Honda wanted to detect and contain warranty and call center issues before they become widespread.

Solution:

SAS Text Miner helps American Honda spot patterns in a wide range of data and text to pinpoint problems early, ensuring safety, quality, and customer satisfaction.

Benefits:

"SAS is helping us make discoveries so that we can address the core issues before they ever become problems—and we can make sure that we are addressing the right causes. We're talking about hundreds of millions of dollars in savings," said Tracy Cermack, Project Manager in the Service Engineering Information Department, American Honda Motor Co. *For more details, please go to http://www.sas.com/success/honda.html*

3. MaspexWadowice Group Analyzes Online Brand Image with Text Mining

Business Issue:

MaspexWadowice Group, a dominant player among food and beverage manufacturers in Central and Eastern Europe, wanted to analyze social media channels to monitor a product's brand image and see how it compares with its general perception in the market.

Solution:

MaspexWadowice Group choose to use SAS Text Miner, which is a part of the SAS Business Analytics capabilities, to tap into social media data sorces.

Benefits:

Maspex gained a competitive advantage through better consumer insights, resulting in more effective and efficient marketing efforts.

"This will allow us to plan and implement our marketing and communications activities more effectively, in particular those using a Web-based channel," said Marcin Lesniak, Research Manager, MaspexWadowice Group.

For more details, please go to http://www.sas.com/success/maspex-wadowice.html

4. Viseca Card Services Reduces Fraud Loss with Text Analytics

Business Issue:

Switzerland's largest credit card company aimed to prevent losses by detecting and preventing fraud on Viseca Card Services' 1 million credit cards and more than 100,000 daily transactions.

Solution:

They choose to use a suite of analytics tools from SAS including SAS® Enterprise Miner™, SAS® Enterprise Guide®, SAS Text Miner, and SAS BI Server.

Benefits:

Eighty-one percent of all fraud cases are found within a day, and total fraud loss has been reduced by 15 percent. Even as the number of fraud cases across the industry has doubled, Viseca Card Services has reduced loss per fraud case by 40 percent.

"Thanks to SAS Analytics our total fraud loss has been reduced by 15 percent. We have one of the best fraud prevention ratings in Switzerland and our business case for fraud prevention is straightforward: Our returns are simply more than our investment," said Marcel Bieler, Business Analyst, Viseca Card Services.

For more details, please go to http://www.sas.com/success/Visecacardsvcs.html

5. Improving Quality with Text Mining and Advanced Analytics

Business Issue:

Whirlpool Corp., the world's leading manufacturer and marketer of major home appliances, wanted to reduce service calls by finding defects through warranty analysis and correcting them quickly.

Solution:

SAS Warranty Analysis and early-warning tools on the SAS Enterprise BI Server distill and analyze warranty claims data to quickly detect product issues. The tools used in this project included SAS Enterprise BI Server, SAS Warranty Analysis, SAS Enterprise Guide, and SAS Text Miner.

Benefits:

Whirlpool Corp. aims to cut overall cost of quality, and SAS is playing a significant part in that objective. Expectations of the SAS Warranty Analysis solution include a significant reduction in Whirlpool's issue detection-to-correction cycle, a three-month decrease in initial issue detection, and a potential to cut overall warranty expenditures with significant quality, productivity and efficiency gains.

"SAS brings a level of analytics to business intelligence that no one else matches," said John Kerr, General Manager of Quality and Operational Excellence, Whirlpool Corp.

For more details, please go to http://www.sas.com/success/whirlpool.html

Questions for Discussion

1. What do you think are the common characteristics of the kind of challenges these five companies were facing?
2. What are the types of solution methods and tools proposed in these case synopses?
3. What do you think are the key benefits of using text mining and advanced analytics (compared to the traditional way to do the same)?

Sources: SAS, **www.sas.com/success/** (accessed September 2013).

SECTION 7.6 REVIEW QUESTIONS

1. What are some of the most popular text mining software tools?

2. Why do you think most of the text mining tools are offered by statistics companies?

3. What do you think are the pros and cons of choosing a free text mining tool over a commercial tool?

7.7 SENTIMENT ANALYSIS OVERVIEW

We, humans, are social beings. We are adept at utilizing a variety of means to communicate. We often consult financial discussion forums before making an investment decision; ask our friends for their opinions on a newly opened restaurant or a newly released movie; and conduct Internet searches and read consumer reviews and expert reports before making a big purchase like a house, a car, or an appliance. We rely on others' opinions to make better decisions, especially in an area where we don't have a lot of knowledge or experience. Thanks to the growing availability and popularity of opinion-rich Internet resources such as social media outlets (e.g., Twitter, Facebook, etc.), online review sites, and personal blogs, it is now easier than ever to find opinions of others (thousands of them, as a matter of fact) on everything from the latest gadgets to political and public figures. Even though not everybody expresses opinions over the Internet, due mostly to the fast-growing numbers and capabilities of social communication channels, the numbers are increasing exponentially.

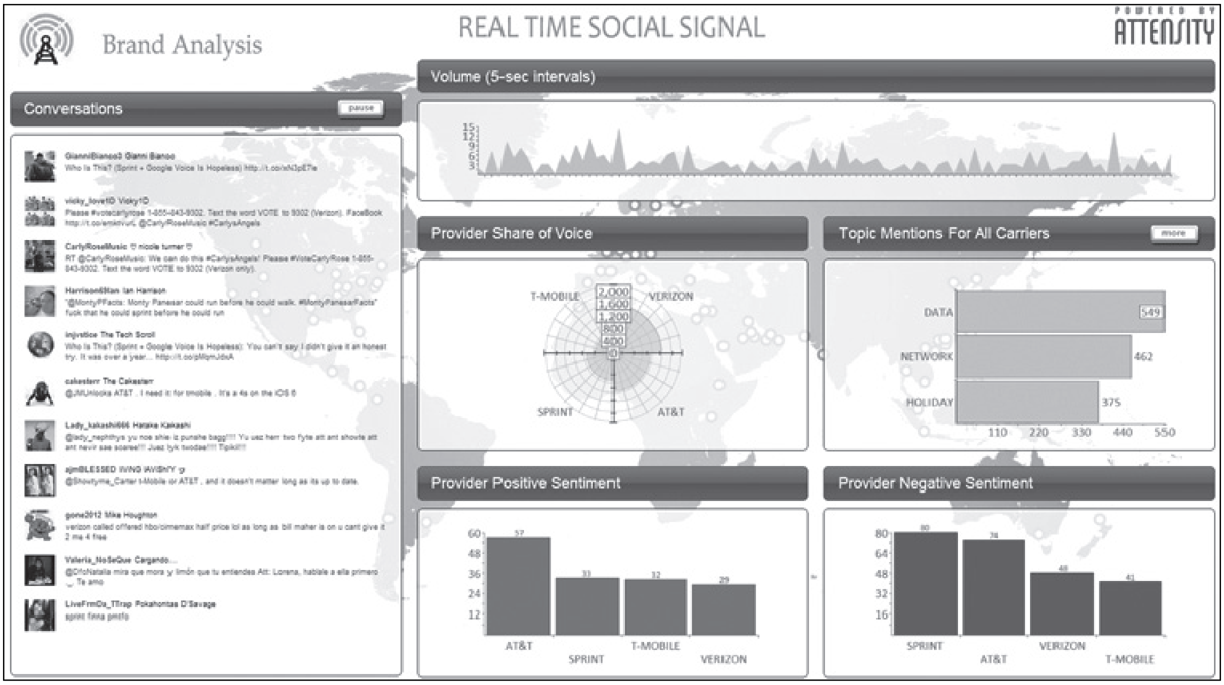
Sentiment is a difficult word to define. It is often linked to or confused with other terms like *belief, view, opinion*, and *conviction*. Sentiment suggests a settled opinion reflective of one's feelings (Mejova, 2009). Sentiment has some unique properties that set it apart from other concepts that we may want to identify in text. Often we want to categorize text by topic, which may involve dealing with whole taxonomies of topics. Sentiment classification, on the other hand, usually deals with two classes (positive versus negative), a range of polarity (e.g., star ratings for movies), or even a range in strength of opinion (Pang and Lee, 2008). These classes span many topics, users, and documents. Although dealing with only a few classes may seem like an easier task than standard text analysis, it is far from the truth.

As a field of research, sentiment analysis is closely related to computational linguistics, natural language processing, and text mining. Sentiment analysis has many names. It's often referred to as *opinion mining, subjectivity analysis,* and *appraisal extraction,* with some connections to affective computing (computer recognition and expression of emotion). The sudden upsurge of interest and activity in the area of sentiment analysis (i.e., opinion mining), which deals with the automatic extraction of opinions, feelings, and subjectivity in text, is creating opportunities and threats for businesses and individuals alike. The ones who embrace and take advantage of it will greatly benefit from it. Every opinion put on the Internet by an individual or a company will be accredited to the originator (good or bad) and will be retrieved and mined by others (often automatically by computer programs).

Sentiment analysis is trying to answer the question "What do people feel about a certain topic?" by digging into opinions of many using a variety of automated tools. Bringing together researchers and practitioners in business, computer science, computational linguistics, data mining, text mining, psychology, and even sociology, sentiment analysis aims to expand traditional fact-based text analysis to new frontiers, to realize opinion-oriented information systems. In a business setting, especially in marketing and customer relationship management, sentiment analysis seeks to detect favorable and unfavorable opinions toward specific products and/or services using large numbers of textual data sources (customer feedback in the form of Web postings, tweets, blogs, etc.).

Sentiment that appears in text comes in two flavors: explicit, where the subjective sentence directly expresses an opinion ("It's a wonderful day"), and implicit, where the text implies an opinion ("The handle breaks too easily"). Most of the earlier work done in sentiment analysis focused on the first kind of sentiment, since it was easier to analyze. Current trends are to implement analytical methods to consider both implicit and explicit sentiments. Sentiment polarity is a particular feature of text that sentiment analysis primarily focuses on. It is usually dichotomized into two—positive and negative—but polarity can also be thought of as a range. A document containing several opinionated statements would have a mixed polarity overall, which is different from not having a polarity at all (being objective) (Mejova, 2009).

Timely collection and analysis of textual data, which may be coming from a variety of sources—ranging from customer call center transcripts to social media postings—is a crucial part of the capabilities of proactive and customer-focused companies, nowadays.



GianniBianco3 Gianni Banco
Who Is This? (Sprint + Google Voice Is Hopeless) http://t.co/xN3pE7e

vicky_love1D Vicky1D
Please #votecarlyrose 1-855-843-9302. Text the word VOTE to 9302 (Verizon). FaceBook http://t.co/emktvuL @CarlyRoseMusic #CarlysAngels

CarlyRoseMusic ♡ nicole turner ♡
RT @CarlyRoseMusic: We can do this #CarlysAngels! Please #VoteCarlyRose 1-855-843-9302. Text the word VOTE to 9302 (Verizon only).

Harrison69Ian Ian Harrison
"@MontyPFacts: Monty Panesar could run before he could walk. #MontyPanesarFacts" fuck that he could sprint before he could run

injustice The Tech Scroll
Who Is This? (Sprint + Google Voice Is Hopeless): You can't say I didn't give it an honest try. It was over a year... http://t.co/pMomJdsA

cakesterr The Cakesterr
@JMUnlocks AT&T . I need it for tmobile . It's a 4s on the iOS 6

Lady_kakashi666 Hatake Kakashi
@lady_nephthys yu noe shie iz punshe bagg!!!! Yu usz herr two fyte att ant showte att ant nevir see soaree!!! Juss lyk twodae!!!! Tgikil!!!

ajmBLESSED IWNG IAYSH'Y g
@Showtyme_Carter t-Mobile or AT&T , and it doesn't matter long as its up to date.

gone2012 Mike Houghton
verizon called offered hbo/cinemax half price lol as long as bill maher is on u cant give it 2 me 4 free

Valeria_NoSeQue Cargando...
@DiioNatalia mira que mora y limón que tu entiendes Att: Lorena, hablale a ella primero ._. Te amo

LiveFrmDa_TTrap Pokahontas D'Savage
sprint fixxa pretty

FIGURE 7.9 A Sample Social Media Dashboard for Continuous Brand Analysis *Source:* Attensity.

These real-time analyses of textual data are often visualized in easy-to-understand dashboards. Attensity is one of those companies that provide such end-to-end solutions to companies' text analytics needs (Figure 7.9 shows an example social media analytics dashboard created by Attensity). Application Case 7.7 provides an Attensity's customer success story, where a large consumer product manufacturer used text analytics and sentiment analysis to better connect with their customers.

Application Case 7.7

Whirlpool Achieves Customer Loyalty and Product Success with Text Analytics

Background

Every day, a substantial amount of new customer feedback data—rich in sentiment, customer issues, and product insights—becomes available to organizations through e-mails, repair notes, CRM notes, and online in social media. Within that data exists a wealth of insight into how customers feel about products, services, brands, and much more. That data also holds information about potential issues that could easily impact a product's long-term success and a company's bottom line. This data is invaluable to marketing, product, and service managers across every industry.

Attensity, a premier text analytics solution provider, combines the company's rich text analytics applications within customer-specific BI platforms. The result is an intuitive solution that enables customers to fully leverage critical data assets to discover invaluable business insight and to foster better and faster decision making.

(*Continued*)

Application Case 7.7 (Continued)

Whirlpool is the world's leading manufacturer and marketer of major home appliances, with annual sales of approximately $19 billion, 67,000 employees, and nearly 70 manufacturing and technology research centers around the world. Whirlpool recognizes that consumers lead busy, active lives, and continues to create solutions that help consumers optimize productivity and efficiency in the home. In addition to designing appliance solutions based on consumer insight, Whirlpool's brand is dedicated to creating ENERGY STAR–qualified appliances like the Resource Saver side-by-side refrigerator, which recently was rated the #1 brand for side-by-side refrigerators.

Business Challenge

Customer satisfaction and feedback are at the center of how Whirlpool drives its overarching business strategy. As such, gaining insight into customer satisfaction and product feedback is paramount. One of Whirlpool's goals is to more effectively understand and react to customer and product feedback data, originating from blogs, e-mails, reviews, forums, repair notes, and other data sources. Whirlpool also strives to enable its managers to report on longitudinal data, and be able to compare issues by brand over time. Whirlpool has entrusted Attensity's text analytics solutions; and with that, Whirlpool listens and acts on customer data in their service department, their innovation and product developments groups, and in market every day.

Methods and the Benefits

To face its business requirements head-on, Whirlpool uses Attensity products for deep text analytics of their multi-channel customer data, which includes e-mails, CRM notes, repair notes, warranty data, and social media. More than 300 business users at Whirlpool use text analytics solutions every day to get to the root cause of product issues and receive alerts on emerging issues. Users of Attensity's analytics products at Whirlpool include product/service managers, corporate/product safety staff, consumer advocates, service quality staff, innovation managers, the Category Insights team, and all of Whirlpool's manufacturing divisions (across five countries).

Attensity's Text Analytics application has played a particularly critical role for Whirlpool. Whirlpool relies on the application to conduct deep analysis of the voice of the customer, with the goal of identifying product quality issues and innovation opportunities, and drive those insights more broadly across the organization. Users conduct in-depth analysis of customer data and then extend access to that analysis to business users all over the world.

Whirlpool has been able to more proactively identify and mitigate quality issues before issues escalate and claims are filed. Whirlpool has also been able to avoid recalls, which has the dual benefit of increased customer loyalty and reduced costs (realizing 80% savings on their costs of recalls due to early detection). Having insight into customer feedback and product issues has also resulted in more efficient customer support and ultimately in better products. Whirlpool's customer support agents now receive fewer product service support calls, and when agents do receive a call, it's easier for them to leverage the interaction to improve products and services.

The process of launching new products has also been enhanced by having the ability to analyze its customers' needs and fit new products and services to those needs appropriately. When a product is launched, Whirlpool can use external customer feedback data to stay on top of potential product issues and address them in a timely fashion.

Michael Page, development and testing manager for Quality Analytics at Whirpool Corporation affirms these types of benefits: "Attensity's products have provided immense value to our business. We've been able to proactively address customer feedback and work toward high levels of customer service and product success."

Questions for Discussion

1. How did Whirlpool use capabilities of text analytics to better understand their customers and improve product offerings?
2. What were the challenges, the proposed solution, and the obtained results?

Source: Source: Attensity, Customer Success Story, **www.attensity.com/2010/08/21/whirlpool-2/** (accessed August 2013).

SECTION 7.7 REVIEW QUESTIONS

1. What is sentiment analysis? How does it relate to text mining?

2. What are the sources of data for sentiment analysis?

3. What are the common challenges that sentiment analysis has to deal with?

7.8 SENTIMENT ANALYSIS APPLICATIONS

Compared to traditional sentiment analysis methods, which were survey based or focus group centered, costly, and time-consuming (and therefore driven from small samples of participants), the new face of text analytics–based sentiment analysis is a limit breaker. Current solutions automate very large-scale data collection, filtering, classification, and clustering methods via natural language processing and data mining technologies that handle both factual and subjective information. Sentiment analysis is perhaps the most popular application of text analytics, tapping into data sources like tweets, Facebook posts, online communities, discussion boards, Web logs, product reviews, call center logs and recording, product rating sites, chat rooms, price comparison portals, search engine logs, and newsgroups. The following applications of sentiment analysis are meant to illustrate the power and the widespread coverage of this technology.

VOICE OF THE CUSTOMER (VOC) **Voice of the customer (VOC)** is an integral part of an analytic CRM and customer experience management systems. As the enabler of VOC, sentiment analysis can access a company's product and service reviews (either continuously or periodically) to better understand and better manage the customer complaints and praises. For instance, a motion picture advertising/marketing company may detect the negative sentiments toward a movie that is about to open in theatres (based on its trailers), and quickly change the composition of trailers and advertising strategy (on all media outlets) to mitigate the negative impact. Similarly, a software company may detect the negative buzz regarding the bugs found in their newly released product early enough to release patches and quick fixes to alleviate the situation.

Often, the focus of VOC is individual customers, their service- and support-related needs, wants, and issues. VOC draw data from the full set of customer touch points, including e-mails, surveys, call center notes/recordings, and social media postings, and match customer voices to transactions (inquiries, purchases, returns) and individual customer profiles captured in enterprise operational systems. VOC, mostly driven by sentiment analysis, is a key element of **customer experience management** initiatives, where the goal is to create an intimate relationship with the customer.

VOICE OF THE MARKET (VOM) **Voice of the market** is about understanding aggregate opinions and trends. It's about knowing what stakeholders—customers, potential customers, influencers, whoever—are saying about your (and your competitors') products and services. A well-done VOM analysis helps companies with competitive intelligence and product development and positioning.

VOICE OF THE EMPLOYEE (VOE) Traditionally VOE has been limited to employee satisfaction surveys. Text analytics in general (and sentiment analysis in particular) is a huge enabler of assessing the VOE. Using rich, opinionated textual data is an effective and efficient way to listen to what employees are saying. As we all know, happy employees empower customer experience efforts and improve customer satisfaction.

BRAND MANAGEMENT Brand management focuses on listening to social media where anyone (past/current/prospective customers, industry experts, other authorities) can post opinions that can damage or boost your reputation. There are a number of relatively newly launched start-up companies that offer analytics-driven brand management services for others. Brand management is product and company (rather than customer) focused. It attempts to shape perceptions rather than to manage experiences using sentiment analysis techniques.

FINANCIAL MARKETS Predicting the future values of individual (or a group of) stocks has been an interesting and seemingly unsolvable problem. What makes a stock (or a group of stocks) move up or down is anything but an exact science. Many believe that the stock market is mostly sentiment driven, making it anything but rational (especially for short-term stock movements). Therefore, use of sentiment analysis in financial markets has gained significant popularity. Automated analysis of market sentiments using social media, news, blogs, and discussion groups seems to be a proper way to compute the market movements. If done correctly, sentiment analysis can identify short-term stock movements based on the buzz in the market, potentially impacting liquidity and trading.

POLITICS As we all know, opinions matter a great deal in politics. Because political discussions are dominated by quotes, sarcasm, and complex references to persons, organizations, and ideas, politics is one of the most difficult, and potentially fruitful, areas for sentiment analysis. By analyzing the sentiment on election forums, one may predict who is more likely to win or lose. Sentiment analysis can help understand what voters are thinking and can clarify a candidate's position on issues. Sentiment analysis can help political organizations, campaigns, and news analysts to better understand which issues and positions matter the most to voters. The technology was successfully applied by both parties to the 2008 and 2012 American presidential election campaigns.

GOVERNMENT INTELLIGENCE Government intelligence is another application that has been used by intelligence agencies. For example, it has been suggested that one could monitor sources for increases in hostile or negative communications. Sentiment analysis can allow the automatic analysis of the opinions that people submit about pending policy or government-regulation proposals. Furthermore, monitoring communications for spikes in negative sentiment may be of use to agencies like Homeland Security.

OTHER INTERESTING AREAS Sentiments of customers can be used to better design e-commerce sites (product suggestions, upsell/cross-sell advertising), better place advertisements (e.g., placing dynamic advertisement of products and services that consider the sentiment on the page the user is browsing), and manage opinion- or review-oriented search engines (i.e., an opinion-aggregation Web site, an alternative to sites like Epinions, summarizing user reviews). Sentiment analysis can help with e-mail filtration by categorizing and prioritizing incoming e-mails (e.g., it can detect strongly negative or flaming e-mails and forward them to the proper folder), as well as citation analysis, where it can determine whether an author is citing a piece of work as supporting evidence or as research that he or she dismisses.

SECTION 7.8 REVIEW QUESTIONS

1. What are the most popular application areas for sentiment analysis? Why?
2. How can sentiment analysis be used for brand management?
3. What would be the expected benefits and beneficiaries of sentiment analysis in politics?
4. How can sentiment analysis be used in predicting financial markets?

7.9 SENTIMENT ANALYSIS PROCESS

Because of the complexity of the problem (underlying concepts, expressions in text, context in which the text is expressed, etc.), there is no readily available standardized process to conduct sentiment analysis. However, based on the published work in the field of sensitivity analysis so far (both on research methods and range of applications), a multi-step, simple logical process, as given in Figure 7.10, seems to be an appropriate methodology for sentiment analysis. These logical steps are iterative (i.e., feedback, corrections, and iterations are part of the discovery process) and experimental in nature, and once completed and combined, capable of producing desired insight about the opinions in the text collection.

STEP 1: SENTIMENT DETECTION After the retrieval and preparation of the text documents, the first main task in sensitivity analysis is the detection of objectivity. Here the goal is to differentiate between a fact and an opinion, which may be viewed as classification of text as objective or subjective. This may also be characterized as calculation of O-S Polarity (Objectivity-Subjectivity Polarity, which may be represented with a numerical value ranging from 0 to 1). If the objectivity value is close to 1, then there is no opinion to mine (i.e., it is a fact); therefore, the process goes back and grabs the next text data to analyze. Usually opinion

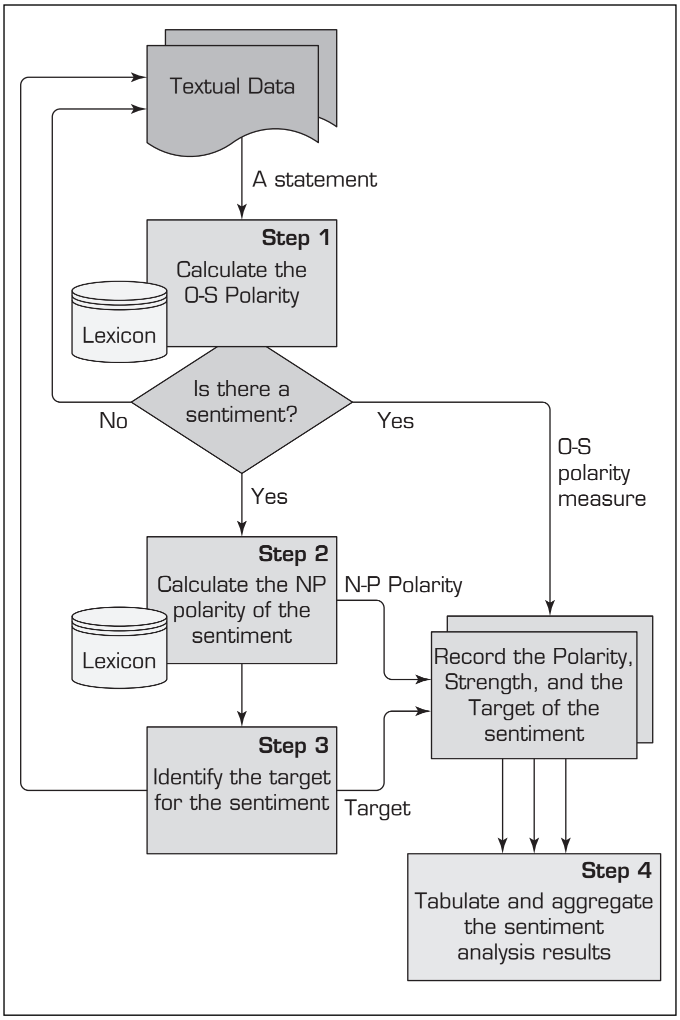

Textual Data
A statement
Step 1
Calculate the O-S Polarity
Lexicon
Is there a sentiment?
No
Yes
Yes
O-S polarity measure
Step 2
Calculate the NP polarity of the sentiment
Lexicon
N-P Polarity
Record the Polarity, Strength, and the Target of the sentiment
Step 3
Identify the target for the sentiment
Target
Step 4
Tabulate and aggregate the sentiment analysis results

FIGURE 7.10 A Multi-Step Process to Sentiment Analysis.

detection is based on the examination of adjectives in text. For example, the polarity of "what a wonderful work" can be determined relatively easily by looking at the adjective.

STEP 2: N-P POLARITY CLASSIFICATION The second main task is that of polarity classification. Given an opinionated piece of text, the goal is to classify the opinion as falling under one of two opposing sentiment polarities, or locate its position on the continuum between these two polarities (Pang and Lee, 2008). When viewed as a binary feature, polarity classification is the binary classification task of labeling an opinionated document as expressing either an overall positive or an overall negative opinion (e.g., thumbs up or thumbs down). In addition to the identification of N-P polarity, one should also be interested in identifying the strength of the sentiment (as opposed to just positive, it may be expressed as mildly, moderately, strongly, or very strongly positive). Most of this research was done on product or movie reviews where the definitions of "positive" and "negative" are quite clear. Other tasks, such as classifying news as "good" or "bad," present some difficulty. For instance an article may contain negative news without explicitly using any subjective words or terms. Furthermore, these classes usually appear intermixed when a document expresses both positive and negative sentiments. Then the task can be to identify the main (or dominating) sentiment of the document. Still, for lengthy texts, the tasks of classification may need to be done at several levels: term, phrase, sentence, and perhaps document level. For those, it is common to use the outputs of one level as the inputs for the next higher layer. Several methods used to identify the polarity and strengths of the polarity are explained in the next section.

STEP 3: TARGET IDENTIFICATION The goal of this step is to accurately identify the target of the expressed sentiment (e.g., a person, a product, an event, etc.). The difficulty of this task depends largely on the domain of the analysis. Even though it is usually easy to accurately identify the target for product or movie reviews, because the review is directly connected to the target, it may be quite challenging in other domains. For instance, lengthy, general-purpose text such as Web pages, news articles, and blogs do not always have a predefined topic that they are assigned to, and often mention many objects, any of which may be deduced as the target. Sometimes there is more than one target in a sentiment sentence, which is the case in comparative texts. A subjective comparative sentence orders objects in order of preferences—for example, "This laptop computer is better than my desktop PC." These sentences can be identified using comparative adjectives and adverbs (more, less, better, longer), superlative adjectives (most, least, best), and other words (such as same, differ, win, prefer, etc.). Once the sentences have been retrieved, the objects can be put in an order that is most representative of their merits, as described in text.

STEP 4: COLLECTION AND AGGREGATION Once the sentiments of all text data points in the document are identified and calculated, in this step they are aggregated and converted to a single sentiment measure for the whole document. This aggregation may be as simple as summing up the polarities and strengths of all texts, or as complex as using semantic aggregation techniques from natural language processing to come up with the ultimate sentiment.

Methods for Polarity Identification

As mentioned in the previous section, **polarity identification**—identifying the polarity of a text—can be made at the word, term, sentence, or document level. The most granular level for polarity identification is at the word level. Once the polarity identification is made at the word level, then it can be aggregated to the next higher level, and then the next until the level of aggregation desired from the sentiment analysis is reached. There seem to be two dominant techniques used for identification of polarity at the word/term level, each having its advantages and disadvantages:

1. Using a lexicon as a reference library (either developed manually or automatically, by an individual for a specific task or developed by an institution for general use)
2. Using a collection of training documents as the source of knowledge about the polarity of terms within a specific domain (i.e., inducing predictive models from opinionated textual documents)

Using a Lexicon

A lexicon is essentially the catalog of words, their synonyms, and their meanings for a given language. In addition to lexicons for many other languages, there are several general-purpose lexicons created for English. Often general-purpose lexicons are used to create a variety of special-purpose lexicons for use in sentiment analysis projects. Perhaps the most popular general-purpose lexicon is WordNet, created at Princeton University, which has been extended and used by many researchers and practitioners for sentiment analysis purposes. As described on the WordNet Web site (**wordnet.princeton.edu**), it is a large lexical database of English, including nouns, verbs, adjectives, and adverbs grouped into sets of cognitive synonyms (i.e., synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.

An interesting extension of WordNet was created by Esuli and Sebastiani (2006) where they added polarity (Positive-Negative) and objectivity (Subjective-Objective) labels for each term in the lexicon. To label each term, they classify the synset (a group of synonyms) to which this term belongs using a set of ternary classifiers (a measure that attaches to each object exactly one out of three labels), each of them capable of deciding whether a synset is Positive, or Negative, or Objective. The resulting scores range from 0.0 to 1.0, giving a graded evaluation of opinion-related properties of the terms. These can be summed up visually as in Figure 7.11. The edges of the triangle represent one of the three classifications (positive, negative, and objective). A term can be located in this space as a point, representing the extent to which it belongs to each of the classifications.

A similar extension methodology is used to create SentiWordNet, a publicly available lexicon specifically developed for opinion mining (sentiment analysis) purposes.

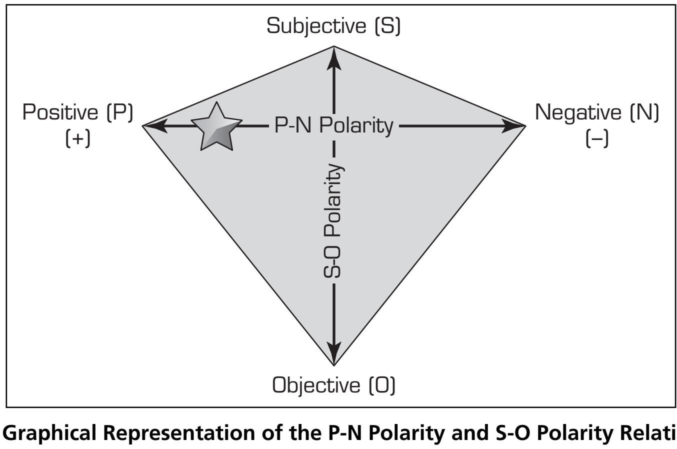


FIGURE 7.11 A Graphical Representation of the P-N Polarity and S-O Polarity Relationship.

SentiWordNet assigns to each synset of WordNet three sentiment scores: positivity, negativity, objectivity. More about SentiWordNet can be found at **sentiwordnet.isti.cnr.it**.

Another extension to WordNet is WordNet-Affect, developed by Strapparava and Valitutti (Strapparava and Valitutti, 2004). They label WordNet synsets using affective labels representing different affective categories like emotion, cognitive state, attitude, feeling, and so on. WordNet has also been directly used in sentiment analysis. For example, Kim and Hovy (Kim and Hovy, 2004) and Hu and Liu (Hu and Liu, 2005) generate lexicons of positive and negative terms by starting with a small list of "seed" terms of known polarities (e.g., love, like, nice, etc.) and then using the antonymy and synonymy properties of terms to group them into either of the polarity categories.

Using a Collection of Training Documents

It is possible to perform sentiment classification using statistical analysis and machine-learning tools that take advantage of the vast resources of labeled (manually by annotators or using a star/point system) documents available. Product review Web sites like Amazon, C-NET, ebay, RottenTomatoes, and the Internet Movie Database (IMDB) have all been extensively used as sources of annotated data. The star (or tomato, as it were) system provides an explicit label of the overall polarity of the review, and it is often taken as a gold standard in algorithm evaluation.

A variety of manually labeled textual data is available through evaluation efforts such as the Text REtrieval Conference (TREC), NII Test Collection for IR Systems (NTCIR), and Cross Language Evaluation Forum (CLEF). The data sets these efforts produce often serve as a standard in the text mining community, including for sentiment analysis researchers. Individual researchers and research groups have also produced many interesting data sets. Technology Insights 7.2 lists some of the most popular ones. Once an already labeled textual data set is obtained, a variety of predictive modeling and other machine-learning algorithms can be used to train sentiment classifiers. Some of the most popular algorithms used for this task include artificial neural networks, support vector machines, *k*-nearest neighbor, Naive Bayes, decision trees, and expectation maximization-based clustering.

Identifying Semantic Orientation of Sentences and Phrases

Once the semantic orientation of individual words has been determined, it is often desirable to extend this to the phrase or sentence the word appears in. The simplest way to accomplish such aggregation is to use some type of averaging for the polarities of words in the phrases or sentences. Though rarely applied, such aggregation can be as complex as using one or more machine-learning techniques to create a predictive relationship between the words (and their polarity values) and phrases or sentences.

Identifying Semantic Orientation of Document

Even though the vast majority of the work in this area is done in determining semantic orientation of words and phrases/sentences, some tasks like summarization and information retrieval may require semantic labeling of the whole document (REF). Similar to the case in aggregating sentiment polarity from word level to phrase or sentence level, aggregation to document level is also accomplished by some type of averaging. Sentiment orientation of the document may not make sense for very large documents; therefore, it is often used on small to medium-sized documents posted on the Internet.

TECHNOLOGY INSIGHTS 7.2 Large Textual Data Sets for Predictive Text Mining and Sentiment Analysis

Congressional Floor-Debate Transcripts: Published by Thomas et al. (Thomas and B. Pang, 2006); contains political speeches that are labeled to indicate whether the speaker supported or opposed the legislation discussed.

Economining: Published by Stern School at New York University; consists of feedback postings for merchants at Amazon.com.

Cornell Movie-Review Data Sets: Introduced by Pang and Lee (Pang and Lee, 2008); contains 1,000 positive and 1,000 negative automatically derived document-level labels, and 5,331 positive and 5,331 negative sentences/snippets.

Stanford—Large Movie Review Data Set: A set of 25,000 highly polar movie reviews for training, and 25,000 for testing. There is additional unlabeled data for use as well. Raw text and already processed bag-of-words formats are provided. (See: **http://ai.stanford.edu/~amaas/data/sentiment.**)

MPQA Corpus: Corpus and Opinion Recognition System corpus; contains 535 manually annotated news articles from a variety of news sources containing labels for opinions and private states (beliefs, emotions, speculations, etc.).

Multiple-Aspect Restaurant Reviews: Introduced by Snyder and Barzilay (Snyder and Barzilay, 2007); contains 4,488 reviews with an explicit 1-to-5 rating for five different aspects: food, ambiance, service, value, and overall experience.

SECTION 7.9 REVIEW QUESTIONS

1. What are the main steps in carrying out sentiment analysis projects?
2. What are the two common methods for polarity identification? What is the main difference between the two?
3. Describe how special lexicons are used in identification of sentiment polarity.

7.10 SENTIMENT ANALYSIS AND SPEECH ANALYTICS

Speech analytics is a growing field of science that allows users to analyze and extract information from both live and recorded conversations. It is being used effectively to gather intelligence for security purposes, to enhance the presentation and utility of rich media applications, and perhaps most significantly, to deliver meaningful and quantitative business intelligence through the analysis of the millions of recorded calls that occur in customer contact centers around the world.

Sentiment analysis, as it applies to speech analytics, focuses specifically on assessing the emotional states expressed in a conversation and on measuring the presence and strength of positive and negative feelings that are exhibited by the participants. One common use of sentiment analysis within contact centers is to provide insight into a customer's feelings about an organization, its products, services, and customer service processes, as well as an individual agent's behavior. Sentiment analysis data can be used across an organization to aid in customer relationship management, agent training, and in identifying and resolving troubling issues as they emerge.

How Is It Done?

The core of automated sentiment analysis centers around creating a model to describe how certain features and content in the audio relate to the sentiments being felt and expressed by the participants in the conversation. Two primary methods have been deployed to predict sentiment within audio: acoustic/phonetic and linguistic modeling.

THE ACOUSTIC APPROACH The acoustic approach to sentiment analysis relies on extracting and measuring a specific set of features (e.g., tone of voice, pitch or volume, intensity and rate of speech) of the audio. These features can in some circumstances provide basic indicators of sentiment. For example, the speech of a surprised speaker tends to become somewhat faster, louder, and higher in pitch. Sadness and depression are presented as slower, softer, and lower in pitch (see Moore et al., 2008). An angry caller may speak much faster, much louder, and will increase the pitch of stressed vowels. There is a wide variety of audio features that can be measured. The most common ones are as follows:

- Intensity: energy, sound pressure level
- Pitch: variation of fundamental frequency
- Jitter: variation in amplitude of vocal fold movements
- Shimmer: variation in frequency of vocal fold movements
- Glottal pulse: glottal-source spectral characteristics
- HNR: harmonics-to-noise ratio
- Speaking rate: number of phonemes, vowels, syllables, or words per unit of time

When developing an acoustic analysis tool, the system must be built on a model that defines the sentiments being measured. The model is based on a database of the audio features (some of which are listed here) and how their presence may indicate each of the sentiments (as simple as positive, negative, neutral, or refined, such as fear, anger, sadness, hurt, surprise, relief, etc.) that are being measured. To create this database, each single-emotion example is preselected from an original set of recordings, manually reviewed, and annotated to identify which sentiment it represents. The final acoustic analysis tools are then trained (using data mining techniques) and a predictive model is tested and validated using a different set of the same annotated recordings.

As sophisticated as it sounds, the acoustic approach has its deficiencies. First, because acoustic analysis relies on identifying the audio characteristics of a call, the quality of the audio can significantly impact the ability to identify these features. Second, speakers often express blended emotions, such as both empathy and annoyance (as in "I do understand, madam, but I have no miracle solution"), which are extremely difficult to classify based solely on their acoustic features. Third, acoustic analysis is often incapable of recognizing and adjusting for the variety of ways that different callers may express the same sentiment. Finally, its time-demanding and laborious process make it impractical for use with live audio streams.

THE LINGUISTIC APPROACH Conversely, the linguistic approach focuses on the explicit indications of sentiment and context of the spoken content within the audio; linguistic models acknowledge that, when in a charged state, the speaker has a higher probability of using specific words, exclamations, or phrases in a particular order. The features that are most often analyzed in a linguistic model include:

- Lexical: words, phrases, and other linguistic patterns
- Disfluencies: filled pauses, hesitation, restarts, and nonverbals such as laughter or breathing
- Higher semantics: taxonomy/ontology, dialogue history, and pragmatics

The simplest method, in the linguistic approach, is to catch within the audio a limited number of specific keywords (a specific lexicon) that has domain-specific sentiment significance. This approach is perhaps the least popular due to its limited applicability and less-than-desired prediction accuracy. Alternatively, as with the acoustic approach, a model is built based on understanding which linguistic elements are predictors of particular sentiments, and this model is then run against a series of recordings to determine the sentiments that are contained therein. The challenge with this approach is in collecting the linguistic information contained in any corpus of audio. This has traditionally been done using a large vocabulary continuous speech recognition (LVCSR) system, often referred to as speech-to-text. However, LVCSR systems are prone to creating significant error in the textual indexes they create. In addition, the level of computational effort they require—that is, the amount of computer processing power needed to analyze large amounts of audio content—has made them very expensive to deploy for mass audio analysis.

Yet, another approach to linguistic analysis is that of phonetic indexing and search. Among the significant advantages associated with this approach to linguistic modeling is the method's ability to maintain a high degree of accuracy no matter what the quality of the audio source, and its incorporation of conversational context through the use of structured queries during analysis (Nexidia, 2009).

Application Case 7.8 is a great example to how analytically savvy companies find ways to better "listen" and improve their customers' experience.

Application Case 7.8

Cutting Through the Confusion: Blue Cross Blue Shield of North Carolina Uses Nexidia's Speech Analytics to Ease Member Experience in Healthcare

Introduction

With the passage of the healthcare law, many health plan members were perplexed by new rules and regulations and concerned about the effects mandates would have on their benefits, copays, and providers. In an attempt to ease concerns, health plans such as Blue Cross Blue Shield of North Carolina (BCBSNC) published literature, updated Web sites, and sent various forms of communication to members to further educate them on the changes. However, members continued to reach out via the contact center, seeking answers regarding current claims and benefits and how their health insurance coverage might be affected in the future. As the law moves forward, members will be more engaged in making their own decisions about healthcare plans and about where to seek care, thus becoming better consumers. The transformation to healthcare consumerism has made it crucial for health plan contact centers to diligently work to optimize the customer experience.

BCBSNC became concerned that despite its best efforts to communicate changes, confusion remained among its nearly 4 million members, which was driving unnecessary calls into its contact center, which could lead to a decrease in member satisfaction. Also, like all plans, BCBSNC was looking to trim costs associated with its contact center, as the health reform law mandates health plans spend a minimum of 80 percent of all premium payments on healthcare. This rule leaves less money for administrative expenses, like the contact center.

However, BCBSNC saw an opportunity to leverage its partnership with Nexidia, a leading provider of customer interaction analytics, and use speech analytics to better understand the cause and depth of member confusion. The use of speech analytics was a more attractive option for BCBSNC than asking their customer service professionals to more thoroughly document the nature of the calls within the contact center desktop application, which would have decreased efficiency and increased contact center administrative expenses. By identifying the specific root cause of the interactions when members called the contact center, BCBSNC would be able to take corrective actions to reduce call volumes and costs and improve the members' experience.

Alleviating the Confusion

BCBSNC has been ahead of the curve on engaging and educating its members and providing exemplary customer service. The health plan knew it needed to work vigorously to maintain its customer

(*Continued*)

Application Case 7.8 (Continued)

service track record as the healthcare mandates began. The first step was to better understand how members perceived the value they received from BCBSNC and their overall opinion of the company. To accomplish this, BCBSNC elected to conduct sentiment analysis to get richer insights into members' opinions and interactions.

When conducting sentiment analysis, two strategies can be used to garner results. The acoustic model relies on measuring specific characteristics of the audio, such as sound, tone of voice, pitch, volume, intensity, and rate of speech. The other strategy, used by Nexidia, is linguistic modeling, which focuses directly on spoken sentiment. Acoustic modeling results in inaccurate data because of poor recording quality, background noise, and a person's inability to change tone or cadence to reflect his or her emotion. The linguistic approach, which focuses directly on words or phrases used to convey a feeling, has proven to be most effective.

Since BCBSNC suspected its members may perceive their health coverage as confusing, BCBSNC utilized Nexidia to put together structured searches for words or phrases used by callers to express confusion: "I'm a little confused," "I don't understand," "I don't get it," and "Doesn't make sense." The results were the exact percentage of calls containing this sentiment and helped BCBSNC specifically isolate those circumstances and coverage instances where callers were more likely to be confused with a benefit or claim. BCBSNC filtered their "confusion calls" from their overall call volume so these calls were available for further analysis.

The next step was to use speech analytics to get to the root cause of what was driving the disconnection and develop strategies to alleviate the confusion. BCBSNC used Nexidia's dictionary independent phonetic indexing and search solution, allowing for all processed audio to be searched for any word or phrase, to create additional structured searches. These searches further classified the call drivers, and when combined with targeted listening, BCBSNC pinpointed the problems.

The findings revealed that literature created by BCBSNC used industry terms that members were unfamiliar with and didn't clearly explain their benefits, claims processes, and deductibles. Additionally, information on the Web site was neither easily located nor understood, and members were unable to "self-serve," resulting in unnecessary contact center interaction. Further, adding to BCBSNC's troubles, when Nexidia's speech analytics combined the unstructured call data with the structured data associated with the call, it showed "confusion calls" had a significantly higher average talk time (ATT), resulting in a higher cost to serve for BCBSNC.

The Results

By listening to, and more specifically understanding, the confusion of its members regarding benefits, BCBSNC began implementing strategies to improve member communication and customer experience. The health plan has developed more reader-friendly literature and simplified the layout to highlight pertinent information. BCBSNC also has implemented Web site redesigns to support easier navigation and education. As a result of the modifications, BCBSNC projects a 10 to 25 percent drop in "confusion calls," resulting in a better customer service experience and a lower cost to serve. Utilizing Nexidia's analytic solution to continuously monitor and track changes will be paramount to BCBSNC's continued success as a leading health plan.

"Because there is so much to do in healthcare today and because of the changes under way in the industry, you really want to invest in the consumer experience so that customers can get the most out of their health care coverage," says Gretchen Gray, director of Customer and Consumer Experience at BCBSNC. "I believe that unless you use [Nexidia's] approach, I don't know how you pick your priorities and focus. Speech analytics is one of the main tools we have where we can say, 'here is where we can have the most impact and here's what I need to do better or differently to assist my customers.'"

Questions for Discussion

1. For a large company like BCBSNC with a lot of customers, what does "listening to customer" mean?
2. What were the challenges, the proposed solution, and the obtained results for BCBSNC?

Source: Used with permission from **Nexidia.com**.

SECTION 7.10 REVIEW QUESTIONS

1. What is speech analytics? How does it relate to sentiment analysis?
2. Describe the acoustic approach to speech analytics.
3. Describe the linguistic approach to speech analytics.

Chapter Highlights

- Text mining is the discovery of knowledge from unstructured (mostly text-based) data sources. Given that a great deal of information is in text form, text mining is one of the fastest growing branches of the business intelligence field.
- Companies use text mining and Web mining to better understand their customers by analyzing their feedback left on Web forms, blogs, and wikis.
- Text mining applications are in virtually every area of business and government, including marketing, finance, healthcare, medicine, and homeland security.
- Text mining uses natural language processing to induce structure into the text collection and then uses data mining algorithms such as classification, clustering, association, and **sequence discovery** to extract knowledge from it.
- Successful application of text mining requires a structured methodology similar to the CRISP-DM methodology in data mining.
- Text mining is closely related to information extraction, natural language processing, and document summarization.
- Text mining entails creating numeric indices from unstructured text and then applying data mining algorithms to these indices.
- Sentiment can be defined as a settled opinion reflective of one's feelings.
- Sentiment classification usually deals with differentiating between two classes, positive and negative.
- As a field of research, sentiment analysis is closely related to computational linguistics, natural language processing, and text mining. It may be used to enhance search results produced by search engines.
- Sentiment analysis is trying to answer the question of "What do people feel about a certain topic?" by digging into opinions of many using a variety of automated tools.
- Voice of the customer is an integral part of an analytic CRM and customer experience management systems, and is often powered by sentiment analysis.
- Voice of the market is about understanding aggregate opinions and trends at the market level.
- Brand management focuses on listening to social media where anyone can post opinions that can damage or boost your reputation.
- Polarity identification in sentiment analysis is accomplished either by using a lexicon as a reference library or by using a collection of training documents.
- WordNet is a popular general-purpose lexicon created at Princeton University.
- SentiWordNet is an extension of WordNet to be used for sentiment identification.
- Speech analytics is a growing field of science that allows users to analyze and extract information from both live and recorded conversations.
- The acoustic approach to sentiment analysis relies on extracting and measuring a specific set of features (e.g., tone of voice, pitch or volume, intensity and rate of speech) of the audio.

Key Terms

association
classification
clustering
corpus
customer experience management (CEM)
deception detection
inverse document frequency
natural language processing (NLP)
part-of-speech tagging
polarity identification
polyseme
sentiment

sentiment analysis
SentiWordNet
sequence discovery
singular value decomposition (SVD)
speech analytics
stemming
stop words
term–document matrix (TDM)
text mining
tokenizing
trend analysis
unstructured data
voice of customer (VOC)
voice of the market
WordNet

Questions for Discussion

1. Explain the relationships among data mining, text mining, and sentiment analysis.
2. What should an organization consider before making a decision to purchase text mining software?
3. Discuss the differences and commonalities between text mining and sentiment analysis.
4. In your own words, define *text mining* and discuss its most popular applications.
5. Discuss the similarities and differences between the data mining process (e.g., CRISP-DM) and the three-step, high-level text mining process explained in this chapter.
6. What does it mean to introduce structure into the text-based data? Discuss the alternative ways of introducing structure into text-based data.
7. What is the role of natural language processing in text mining? Discuss the capabilities and limitations of NLP in the context of text mining.
8. List and discuss three prominent application areas for text mining. What is the common theme among the three application areas you chose?
9. What is sentiment analysis? How does it relate to text mining?
10. What are the sources of data for sentiment analysis?
11. What are the common challenges that sentiment analysis has to deal with?
12. What are the most popular application areas for sentiment analysis? Why?
13. How can sentiment analysis be used for brand management?
14. What would be the expected benefits and beneficiaries of sentiment analysis in politics?
15. How can sentiment analysis be used in predicting financial markets?
16. What are the main steps in carrying out sentiment analysis projects?
17. What are the two common methods for polarity identification? What is the main difference between the two?
18. Describe how special lexicons are used in identification of sentiment polarity.
19. What is speech analytics? How does it relate to sentiment analysis?
20. Describe the acoustic approach to speech analytics.
21. Describe the linguistic approach to speech analytics.

Exercises

Teradata University Network (TUN) and Other Hands-On Exercises

1. Visit **teradatauniversitynetwork.com**. Identify cases about text mining. Describe recent developments in the field. If you cannot find enough cases at the Teradata University network Web site, broaden your search to other Web-based resources.
2. Go to **teradatauniversitynetwork.com** or locate white papers, Web seminars, and other materials related to text mining. Synthesize your findings into a short written report.
3. Browse the Web and your library's digital databases to identify articles that make the natural linkage between text/Web mining and contemporary business intelligence systems.
4. Go to **teradatauniversitynetwork.com** and find a case study named "eBay Analytics." Read the case carefully, extend your understanding of the case by searching the Internet for additional information, and answer the case questions.
5. Go to **teradatauniversitynetwork.com** and find a sentiment analysis case named "How Do We Fix and App Like That!" Read the description and follow the directions to download the data and the tool to carry out the exercise.

Team Assignments and Role-Playing Projects

1. Examine how textual data can be captured automatically using Web-based technologies. Once captured, what are the potential patterns that you can extract from these unstructured data sources?
2. Interview administrators in your college or executives in your organization to determine how text mining and Web mining could assist them in their work. Write a proposal describing your findings. Include a preliminary cost–benefits analysis in your report.
3. Go to your library's online resources. Learn how to download attributes of a collection of literature (journal articles) in a specific topic. Download and process the data using a methodology similar to the one explained in Application Case 7.5.
4. Find a readily available sentiment text data set (see Technology Insights 7.2 for a list of popular data sets) and

download it into your computer. If you have an analytics tool that is capable of text mining, use that; if not, download RapidMiner (**rapid-i.com**) and install it. Also install the text analytics add-on for RapidMiner. Process the downloaded data using your text mining tool (i.e., convert the data into a structured form). Build models and assess the sentiment detection accuracy of several classification models (e.g., support vector machines, decision trees, neural networks, logistic regression, etc.). Write a detailed report where you explain your finings and your experiences.

Internet Exercises

1. Survey some text mining tools and vendors. Start with **clearforest.com** and **megaputer.com**. Also consult with **dmreview.com** and identify some text mining products and service providers that are not mentioned in this chapter.
2. Find recent cases of successful text mining and Web mining applications. Try text and Web mining software vendors and consultancy firms and look for cases or success stories. Prepare a report summarizing five new case studies.
3. Go to **statsoft.com**. Select Downloads and download at least three white papers on applications. Which of these applications may have used the data/text/Web mining techniques discussed in this chapter?
4. Go to **sas.com**. Download at least three white papers on applications. Which of these applications may have used the data/text/Web mining techniques discussed in this chapter?
5. Go to **ibm.com**. Download at least three white papers on applications. Which of these applications may have used the data/text/Web mining techniques discussed in this chapter?
6. Go to **teradata.com**. Download at least three white papers on applications. Which of these applications may have used the data/text/Web mining techniques discussed in this chapter?
7. Go to **fairisaac.com**. Download at least three white papers on applications. Which of these applications may have used the data/text/Web mining techniques discussed in this chapter?
8. Go to **salfordsystems.com**. Download at least three white papers on applications. Which of these applications may have used the data/text/Web mining techniques discussed in this chapter?
9. Go to **clarabridge.com**. Download at least three white papers on applications. Which of these applications may have used text mining in a creative way?
10. Go to **kdnuggets.com**. Explore the sections on applications as well as software. Find names of at least three additional packages for data mining and text mining.

End-of-Chapter Application Case

BBVA Seamlessly Monitors and Improves its Online Reputation

BBVA is a global group that offers individual and corporate customers a comprehensive range of financial and nonfinancial products and services. It enjoys a solid leadership position in the Spanish market, where it first began its activities over 150 years ago. It also has a leading franchise in South America; it is the largest financial institution in Mexico; one of the 15 largest U.S. commercial banks and one of the few large international groups operating in China and Turkey. BBVA employs approximately 104,000 people in over 30 countries around the world, and has more than 47 million customers and 900,000 shareholders.

Looking for tools to reduce reputational risks

BBVA is interested in knowing what existing clients—and possible new ones—think about it through social media. Therefore, the bank has implemented an automated consumer insight solution to monitor and measure the impact of brand perception online—whether this be customer comments on social media sites (Twitter, Facebook, forums, blogs, etc.), the voices of experts in online articles about BBVA and its competitors, or references to BBVA on news sites—to detect possible risks to its reputation or to possible business opportunities.

Insights derived from this analytical tool give BBVA the opportunity to address reputational challenges and continue to build on positive opinions. For example, the bank can now respond to negative (or positive) brand perception by focusing its communication strategies on particular Internet sites, countering—or backing up—the most outspoken authors on Twitter, boards and blogs.

Finding a way forward

In 2009, BBVA began monitoring the web with an IBM social media research asset called Corporate Brand Reputation Analysis (COBRA), as a pilot between IBM and the bank's Innovation department. This pilot proved highly successful for different areas of the bank, including the Communications, Brand & Reputation, Corporate Social Responsibility, Consumer Insight, and Online Banking departments.

The BBVA Communication department then decided to tackle a new project, deploying a single tool that would enable the entire group to analyze online mentions of BBVA and monitor the bank's brand perception in various online communities.

The bank decided to implement IBM Cognos Consumer Insight to unify all its branches worldwide and allow them to use the same samples, models, and taxonomies. IBM Global Business Services is currently helping the bank to implement the solution, as well as design the focus of the analysis adapted to each country's requirements.

IBM Cognos Consumer Insight will allow BBVA to monitor the voices of current and potential clients on social media websites such as Twitter, Facebook and message boards, identify expert opinions about BBVA and its competitors on blogs, and control the presence of the bank in news channels to gain insights and detect possible reputational risks. All this new information will be distributed among the business departments of BBVA, enabling the bank to take a holistic view across all areas of its business.

Seamless focus on online reputation

The solution has now been rolled out in Spain, and BBVA's Online Communications team is already seeing its benefits.

"Huge amounts of data are being posted on Twitter every day, which makes it a great source of information for us," states the Online Communications Department of this bank. "To make effective use of this resource, we needed to find a way to capture, store and analyze the data in a better, faster and more detailed fashion. We believe that IBM Cognos Consumer Insight will help us to differentiate and categorize all the data we collect according to pre-established criteria, such as author, date, country and subject. This enables us to focus only on comments and news items that are actually relevant, whether in a positive, negative or neutral sense."

The content of the comments is subsequently analyzed using custom Spanish and English dictionaries, in order to identify whether the sentiments expressed are positive or negative. "What is great about this solution is that it helps us to focus our actions on the most important topics of online discussions and immediately plan the correct and most suitable reaction," adds the Department, "By building on what we accomplished in the initial COBRA project, the new solution enables BBVA to seamlessly monitor comments and postings, improve its decision-making processes, and thereby strengthen its online reputation."

"When BBVA detects a negative comment, a reputational risk arises," explains Miguel Iza Moreno, Business Analytics and Optimization Consultant at IBM Global Business Services. "Cognos Consumer Insight provides a reporting system which identifies the origin of a negative statement and BBVA sets up an internal protocol to decide how to react. This can happen through press releases, direct communication with users or, in some cases, no action is deemed to be required; the solution also highlights those cases in which the negative comment is considered 'irrelevant' or 'harmless'. The same procedure applies to positive comments—the solution allows the bank to follow a standard and structured process, which, based on positive insights, enables it to strengthen its reputation.

"Following the successful deployment in Spain, BBVA will be able to easily replicate the Cognos Consumer Insight solution in other countries, providing a single solution that will help to consolidate and reaffirm the bank's reputation management strategy," says the Department.

Tangible Results

Starting with the COBRA pilot project, the solution delivered visible benefits during the first half of 2011. Positive feedback about the company increased by more than one percent while negative feedback was reduced by 1.5 percent—suggesting that hundreds of customers and stakeholders across Spain are already enjoying a more satisfying experience from BBVA. Moreover, global monitoring improved, providing greater reliability when comparing results between branches and countries. Similar benefits are expected from the Cognos Consumer Insight project, and the initial results are expected shortly.

"BBVA is already seeing a remarkable improvement in the way that information is gathered and analyzed, which we are sure will translate into the same kind of tangible benefits we saw from the COBRA pilot project," states the bank, "For the time being, we have already achieved what we needed the most: a single tool which unifies the online measuring of our business strategies, enabling more detailed, structured and controlled online data analysis."

Questions for the End-of-Chapter Application Case

1. How did BBVA use text mining?
2. What were BBVA's challenges? How did BBVA overcome them with text mining and social media analysis?
3. In what other areas, in your opinion, can BBVA use text mining?

Source: IBM Customer Success Story, "BBVA seamlessly monitors and improves its online reputation" at **http://www-01.ibm.com/software/success/cssdb.nsf/CS/STRD-8NUD29?OpenDocument&Site=corp&cty=en_us** (accessed August 2013).

References

Chun, H. W., Y. Tsuruoka, J. D. Kim, R. Shiba, N. Nagata, and T. Hishiki. (2006). "Extraction of Gene-Disease Relations from Medline Using Domain Dictionaries and Machine Learning." *Proceedings of the 11th Pacific Symposium on Biocomputing,* pp. 4–15.

Cohen, K. B., and L. Hunter. (2008). "Getting Started in Text Mining." *PLoS Compututional Biology,* Vol. 4, No. 1, pp. 1–10.

Coussement, K., and D. Van Den Poel. (2008). "Improving Customer Complaint Management by Automatic Email

Classification Using Linguistic Style Features as Predictors." *Decision Support Systems*, Vol. 44, No. 4, pp. 870–882.

Coussement, K., and D. Van Den Poel. (2009). "Improving Customer Attrition Prediction by Integrating Emotions from Client/Company Interaction Emails and Evaluating Multiple Classifiers." *Expert Systems with Applications,* Vol. 36, No. 3, pp. 6127–6134.

Delen, D., and M. Crossland. (2008). "Seeding the Survey and Analysis of Research Literature with Text Mining." *Expert Systems with Applications,* Vol. 34, No. 3, pp. 1707–1720.

Etzioni, O. (1996). "The World Wide Web: Quagmire or Gold Mine?" *Communications of the ACM,* Vol. 39, No. 11, pp. 65–68.

EUROPOL. (2007). "EUROPOL Work Program 2007." **statewatch.org/news/2006/apr/europol-work-programme-2007.pdf** (accessed October 2008).

Feldman, R., and J. Sanger. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Boston: ABS Ventures.

Fuller, C. M., D. Biros, and D. Delen. (2008). "Exploration of Feature Selection and Advanced Classification Models for High-Stakes Deception Detection." *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS)*, Big Island, HI: IEEE Press, pp. 80–99.

Ghani, R., K. Probst, Y. Liu, M. Krema, and A. Fano. (2006). "Text Mining for Product Attribute Extraction." *SIGKDD Explorations,* Vol. 8, No. 1, pp. 41–48.

Grimes, S. (2011, February 17). "Seven Breakthrough Sentiment Analysis Scenarios." *InformationWeek*.

Han, J., and M. Kamber. (2006). *Data Mining: Concepts and Techniques,* 2nd ed. San Francisco: Morgan Kaufmann.

Kanayama, H., and T. Nasukawa. (2006). "Fully Automatic Lexicon Expanding for Domain-oriented Sentiment Analysis, EMNLP: Empirical Methods in Natural Language Processing." **trl.ibm.com/projects/textmining/takmi/sentiment_analysis_e.htm.**

Kleinberg, J. (1999). "Authoritative Sources in a Hyperlinked Environment." *Journal of the ACM,* Vol. 46, No. 5, pp. 604–632.

Lin, J., and D. Demner-Fushman. (2005). "'Bag of Words' Is Not Enough for Strength of Evidence Classification." *AMIA Annual Symposium Proceedings,* pp. 1031–1032. **pubmedcentral.nih.gov/articlerender.fcgi?artid=1560897.**

Mahgoub, H., D. Rösner, N. Ismail, and F. Torkey. (2008). "A Text Mining Technique Using Association Rules Extraction." *International Journal of Computational Intelligence,* Vol. 4, No. 1, pp. 21–28.

Manning, C. D., and H. Schutze. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Masand, B. M., M. Spiliopoulou, J. Srivastava, and O. R. Zaïane. (2002). "Web Mining for Usage Patterns and Profiles." *SIGKDD Explorations,* Vol. 4, No. 2, pp. 125–132.

McKnight, W. (2005, January 1). "Text Data Mining in Business Intelligence." *Information Management Magazine.* **information-management.com/issues/20050101/1016487-1.html** (accessed May 22, 2009).

Mejova, Y. (2009). "Sentiment Analysis: An Overview." Comprehensive exam paper. **www.cs.uiowa.edu/~ymejova/publications/CompsYelenaMejova.pdf** (accessed February 2013).

Miller, T. W. (2005). *Data and Text Mining: A Business Applications Approach*. Upper Saddle River, NJ: Prentice Hall.

Nakov, P., A. Schwartz, B. Wolf, and M. A. Hearst. (2005). "Supporting Annotation Layers for Natural Language Processing." *Proceedings of the ACL,* interactive poster and demonstration sessions, Ann Arbor, MI. Association for Computational Linguistics, pp. 65–68.

Nasraoui, O., M. Spiliopoulou, J. Srivastava, B. Mobasher, and B. Masand. (2006). "WebKDD 2006: Web Mining and Web Usage Analysis Post-Workshop Report." *ACM SIGKDD Explorations Newsletter*, Vol. 8, No. 2, pp. 84–89.

Nexidia (2009). "State of the art: Sentiment analysis" Nexidia White Paper, http://nexidia.com/files/resource_files/nexidia_sentiment_analysis_wp_8269.pdf (accessed February 2013).

Pang, B., and L. Lee. (2008). "Opinion Mining and Sentiment Analysis." Now Pub. **http://books.google.com**.

Peterson, E. T. (2008). "The Voice of Customer: Qualitative Data as a Critical Input to Web Site Optimization." **foreseeresults.com/Form_Epeterson_WebAnalytics.html** (accessed May 22, 2009).

Shatkay, H., A. Höglund, S. Brady, T. Blum, P. Dönnes, and O. Kohlbacher. (2007). "SherLoc: High-Accuracy Prediction of Protein Subcellular Localization by Integrating Text and Protein Sequence Data." *Bioinformatics,* Vol. 23, No. 11, pp. 1410–1417.

SPSS. "Merck Sharp & Dohme." **spss.com/success/template_view.cfm?Story_ID=185** (accessed May 15, 2009).

StatSoft. (2009). *Statistica Data and Text Miner User Manual.* Tulsa, OK: StatSoft, Inc.

Turetken, O., and R. Sharda. (2004). "Development of a Fisheye-Based Information Search Processing Aid (FISPA) for Managing Information Overload in the Web Environment." *Decision Support Systems,* Vol. 37, No. 3, pp. 415–434.

Weng, S. S., and C. K. Liu. (2004) "Using Text Classification and Multiple Concepts to Answer E-Mails." *Expert Systems with Applications,* Vol. 26, No. 4, pp. 529–543.

Zhou, Y., E. Reid, J. Qin, H. Chen, and G. Lai. (2005). "U.S. Domestic Extremist Groups on the Web: Link and Content Analysis." *IEEE Intelligent Systems,* Vol. 20, No. 5, pp. 44–51.

CHAPTER

8

Web Analytics, Web Mining, and Social Analytics

LEARNING OBJECTIVES

- Define *Web mining* and understand its taxonomy and its application areas
- Differentiate between Web content mining and Web structure mining
- Understand the internals of Web search engines
- Learn the details about search engine optimization
- Define *Web usage mining* and learn its business application
- Describe the Web analytics maturity model and its use cases
- Understand social networks and social analytics and their practical applications
- Define *social network analysis* and become familiar with its application areas
- Understand social media analytics and its use for better customer engagement

This chapter is all about Web mining and its application areas. As you will see, Web mining is one of the fastest growing technologies in business intelligence and business analytics. Under the umbrella of Web mining, in this chapter, we will cover Web analytics, search engines, social analytics and their enabling methods, algorithms, and technologies.

8.1 OPENING VIGNETTE: Security First Insurance Deepens Connection with Policyholders

Security First Insurance is one of the largest homeowners' insurance companies in Florida. Headquartered in Ormond Beach, it employs more than 80 insurance professionals to serve its nearly 190,000 customers.

CHALLENGE

Being There for Customers Storm After Storm, Year After Year

Florida has more property and people exposed to hurricanes than any state in the country. Each year, the Atlantic Ocean averages 12 named storms and nine named hurricanes. Security First is one of a few Florida homeowners' insurance companies that has the financial strength to withstand multiple natural disasters. "One of our promises is to be there for our customers, storm after storm, year after year," says Werner Kruck, chief operating officer for Security First.

During a typical month, Security First processes 700 claims. However, in the aftermath of a hurricane, that number can swell to tens of thousands within days. It can be a challenge for the company to quickly scale up to handle the influx of customers trying to file post-storm insurance claims for damaged property and possessions. In the past, customers submitted claims primarily by phone and sometimes email. Today, policyholders use any means available to connect with an agent or claims representative, including posting a question or comment on the company's Facebook page or Twitter account.

Although Security First provides ongoing monitoring of its Facebook and Twitter accounts, as well as its multiple email addresses and call centers, the company knew that the communication volume after a major storm required a more aggressive approach. "We were concerned that if a massive number of customers contacted us through email or social media after a hurricane, we would be unable to respond quickly and appropriately," Kruck says. "We need to be available to our customers in whatever way they want to contact us." In addition, Security First recognized the need to integrate its social media responses into the claims process and document those responses to comply with industry regulations.

SOLUTION

Providing Responsive Service No Matter How Customers Get in Touch

Security First contacted IBM Business Partner Integritie for help with harnessing social media to improve the customer experience. Integritie configured a solution built on key IBM Enterprise Content Management software components, featuring IBM Content Analytics with Enterprise Search, IBM Content Collector for Email and IBM® FileNet® Content Manager software. Called Social Media Capture (SMC4), the Integritie solution offers four critical capabilities for managing social media platforms: capture, control, compliance and communication. For example, the SMC4 solution logs all social networking interaction for Security First, captures content, monitors incoming and outgoing messages and archives all communication for compliance review.

Because the solution uses open IBM Enterprise Content Management software, Security First can easily link it to critical company applications, databases and processes.

For example, Content Collector for Email software automatically captures email content and attachments and sends an email back to the policyholder acknowledging receipt. In addition, Content Analytics with Enterprise Search software sifts through and analyzes the content of customers' posts and emails. The software then captures information gleaned from this analysis directly into claims documents to begin the claims process. Virtually all incoming communication from the company's web, the Internet and emails is pulled into a central FileNet Content Manager software repository to maintain, control and link to the appropriate workflow. "We can bring the customer conversation and any pictures and attachments into our policy and claims management system and use it to trigger our claims process and add to our documentation," says Kruck.

Prioritizing Communications with Access to Smarter Content

People whose homes have been damaged or destroyed by a hurricane are often displaced quickly, with little more than the clothes on their backs. Grabbing an insurance policy on the way out the door is often an afterthought. They're relying on their insurance companies to have the information they need to help them get their lives back in order as quickly as possible. When tens of thousands of policyholders require assistance within a short period of time, Security First must triage requests quickly. The Content Analytics with Enterprise Search software that anchors the SMC4 solution provides the information necessary to help the company identify and address the most urgent cases first. The software automatically sifts through data in email and social media posts, tweets and comments using text mining, text analytics, natural language processing and sentiment analytics to detect words and tones that identify significant property damage or that convey distress. Security First can then prioritize the messages and route them to the proper personnel to provide reassurance, handle complaints or process a claim. "With access to smarter content, we can respond to our customers in a more rapid, efficient and personalized way," says Kruck. "When customers are having a bad experience, it's really important to get to them quickly with the level of assistance appropriate to their particular situations."

RESULTS

Successfully Addressing Potential Compliance Issues

Companies in all industries must stay compliant with new and emerging regulatory requirements regarding social media. The text analysis capabilities provided in the IBM software help Security First filter inappropriate incoming communications and audit outbound communications, avoiding potential issues with message content. The company can be confident that the responses its employees provide are compliant and controlled based on both Security First policies and industry regulations.

Security First can designate people or roles in the organization that are authorized to create and submit responses. The system automatically verifies these designations and analyzes outgoing message content, stopping any ineffective or questionable communications for further review. "Everything is recorded for compliance, so we can effectively track and maintain the process. We have the ability to control which employees respond, their level of authority and the content of their responses," says Kruck.

These capabilities give Security First the confidence to expand its use of social media. Because compliance is covered, the company can focus on additional opportunities for direct dialog with customers. Before this solution, Security First filtered customer communications through agents. Now it can reach out to customers directly and proactively as a company.

"We're one of the first insurance companies in Florida to make ourselves available to customers whenever, wherever and however they choose to communicate. We're also managing internal processes more effectively and proactively, reaching out to customers in a controlled and compliant manner," says Kruck.

Some of the prevailing business benefits of creative use of Web and social analytics include:

- Turns social media into an actionable communications channel during a major disaster
- Speeds claims processes by initiating claims with information from email and social media posts
- Facilitates prioritizing urgent cases by analyzing social media content for sentiments
- Helps ensure compliance by automatically documenting social media communications

QUESTIONS FOR THE OPENING VIGNETTE

1. What does Security First do?
2. What were the main challenges Security First was facing?
3. What was the proposed solution approach? What types of analytics were integrated in the solution?
4. Based on what you learn from the vignette, what do you think are the relationships between Web analytics, text mining, and sentiment analysis?
5. What were the results Security First obtained? Were any surprising benefits realized?

WHAT WE CAN LEARN FROM THIS VIGNETTE

Web analytics is becoming a way of life for many businesses, especially the ones that are directly facing the consumers. Companies are expected to find new and innovative ways to connect with their customers, understand their needs, wants, and opinions, and proactively develop products and services that fit well with them. In this day and age, asking customers to tell you exactly what they like and dislike is not a viable option. Instead, businesses are expected to deduce that information by applying advanced analytics tools to invaluable data generated on the Internet and social media sites (along with corporate databases). Security First realized the need to revolutionize their business processes to be more effective and efficient in the way that they deal with their customers and customer claims. They not only used what the Internet and social media have to offer, but also tapped into the customer call records/recordings and other relevant transaction databases. This vignette illustrates the fact that analytics technologies are advanced enough to bring together many different data sources to create a holistic view of the customer. And that is perhaps the greatest success criterion for today's businesses. In the following sections, you will learn about many of the Web-based analytical techniques that make it all happen.

Source: IBM Customer Success Story, "Security First Insurance deepens connection with policyholders" accessed at **http://www-01.ibm.com/software/success/cssdb.nsf/CS/SAKG-975H4N?OpenDocument&Site=default&cty=en_us** (accessed August 2013)..

8.2 WEB MINING OVERVIEW

The Internet has forever changed the landscape of business as we know it. Because of the highly connected, flattened world and broadened competitive field, today's companies are increasingly facing greater opportunities (being able to reach customers and markets that they may have never thought possible) and bigger challenge (a globalized and ever-changing competitive marketplace). Ones with the vision and capabilities to deal with such a volatile environment are greatly benefiting from it, while others who resist are having a hard time surviving. Having an engaged presence on the Internet is not a choice anymore: It is a business requirement. Customers are expecting companies to offer their products and/or services over the Internet. They are not only buying products and services but also talking about companies and sharing their transactional and usage experiences with others over the Internet.

The growth of the Internet and its enabling technologies has made data creation, data collection, and data/information/opinion exchange easier. Delays in service, manufacturing, shipping, delivery, and customer inquiries are no longer private incidents and are accepted as necessary evils. Now, thanks to social media tools and technologies on the Internet, everybody knows everything. Successful companies are the ones who embrace these Internet technologies and use them for the betterment of their business processes so that they can better communicate with their customers, understanding their needs and wants and serving them thoroughly and expeditiously. Being customer focused and keeping customers happy have never been as important a concept for businesses as they are now, in this age of the Internet and social media.

The World Wide Web (or, for short, the Web) serves as an enormous repository of data and information on virtually everything one can conceive—business, personal, you name it; an abundant amount of it is there. The Web is perhaps the world's largest data and text repository, and the amount of information on the Web is growing rapidly. A lot of interesting information can be found online: whose homepage is linked to which other pages, how many people have links to a specific Web page, and how a particular site is organized. In addition, each visitor to a Web site, each search on a search engine, each click on a link, and each transaction on an e-commerce site create additional data. Although unstructured textual data in the form of Web pages coded in HTML or XML is the dominant content of the Web, the Web infrastructure also contains hyperlink information (connections to other Web pages) and usage information (logs of visitors' interactions with Web sites), all of which provide rich data for knowledge discovery. Analysis of this information can help us make better use of Web sites and also aid us in enhancing relationships and value for the visitors to our own Web sites.

Because of its sheer size and complexity, mining the Web is not an easy undertaking by any means. The Web also poses great challenges for effective and efficient knowledge discovery (Han and Kamber, 2006):

- ***The Web is too big for effective data mining.*** The Web is so large and growing so rapidly that it is difficult to even quantify its size. Because of the sheer size of the Web, it is not feasible to set up a data warehouse to replicate, store, and integrate all of the data on the Web, making data collection and integration a challenge.
- ***The Web is too complex.*** The complexity of a Web page is far greater than a page in a traditional text document collection. Web pages lack a unified structure. They contain far more authoring style and content variation than any set of books, articles, or other traditional text-based document.
- ***The Web is too dynamic.*** The Web is a highly dynamic information source. Not only does the Web grow rapidly, but its content is constantly being updated. Blogs, news stories, stock market results, weather reports, sports scores, prices, company advertisements, and numerous other types of information are updated regularly on the Web.
- ***The Web is not specific to a domain.*** The Web serves a broad diversity of communities and connects billions of workstations. Web users have very different backgrounds, interests, and usage purposes. Most users may not have good knowledge of the structure of the information network and may not be aware of the heavy cost of a particular search that they perform.
- ***The Web has everything.*** Only a small portion of the information on the Web is truly relevant or useful to someone (or some task). It is said that 99 percent of the information on the Web is useless to 99 percent of Web users. Although this may not seem obvious,

it is true that a particular person is generally interested in only a tiny portion of the Web, whereas the rest of the Web contains information that is uninteresting to the user and may swamp desired results. Finding the portion of the Web that is truly relevant to a person and the task being performed is a prominent issue in Web-related research.

These challenges have prompted many research efforts to enhance the effectiveness and efficiency of discovering and using data assets on the Web. A number of index-based Web search engines constantly search the Web and index Web pages under certain keywords. Using these search engines, an experienced user may be able to locate documents by providing a set of tightly constrained keywords or phrases. However, a simple keyword-based search engine suffers from several deficiencies. First, a topic of any breadth can easily contain hundreds or thousands of documents. This can lead to a large number of document entries returned by the search engine, many of which are marginally relevant to the topic. Second, many documents that are highly relevant to a topic may not contain the exact keywords defining them. As we will cover in more detail later in this chapter, compared to keyword-based Web search, Web mining is a prominent (and more challenging) approach that can be used to substantially enhance the power of Web search engines because Web mining can identify authoritative Web pages, classify Web documents, and resolve many ambiguities and subtleties raised in keyword-based Web search engines.

Web mining (or Web data mining) is the process of discovering intrinsic relationships (i.e., interesting and useful information) from Web data, which are expressed in the form of textual, linkage, or usage information. The term *Web mining* was first used by Etzioni (1996); today, many conferences, journals, and books focus on Web data mining. It is a continually evolving area of technology and business practice. Web mining is essentially the same as data mining that uses data generated over the Web. The goal is to turn vast repositories of business transactions, customer interactions, and Web site usage data into actionable information (i.e., knowledge) to promote better decision making throughout the enterprise. Because of the increased popularity of the term *analytics,* nowadays many have started to call Web mining *Web analytics*. However, these two terms are not the same. Although Web analytics is primarily Web site usage data focused, Web mining is inclusive of all data generated via the Internet, including transaction, social, and usage data. While Web analytics aims to describe what has happened on the Web site (employing a predefined, metrics-driven descriptive analytics methodology), Web mining aims to discover previously unknown patterns and relationships (employing a novel predictive or prescriptive analytics methodology). From a big-picture perspective, Web analytics can be considered a part of Web mining. Figure 8.1 presents a simple taxonomy of Web mining, where it is divided into three main areas: Web content mining, Web structure mining, and Web usage mining. In the figure, the data sources used in these three main areas are also specified. Although these three areas are shown separately, as you will see in the following section, they are often used collectively and synergistically to address business problems and opportunities.

As Figure 8.1 indicates, Web mining relies heavily on data mining and text mining and their enabling tools and techniques, which we have covered in detail in the previous two chapters (Chapters 6 and 7). The figure also indicates that these three generic areas are further extended into several very well-known application areas. Some of these areas were explained in the previous chapters, and some of the others will be covered in detail in this chapter.

SECTION 8.2 REVIEW QUESTIONS

1. What are some of the main challenges the Web poses for knowledge discovery?
2. What is Web mining? How does it differ from regular data mining or text mining?
3. What are the three main areas of Web mining?
4. Identify three application areas for Web mining (at the bottom of Figure 8.1). Based on your own experiences, comment on their use cases in business settings.

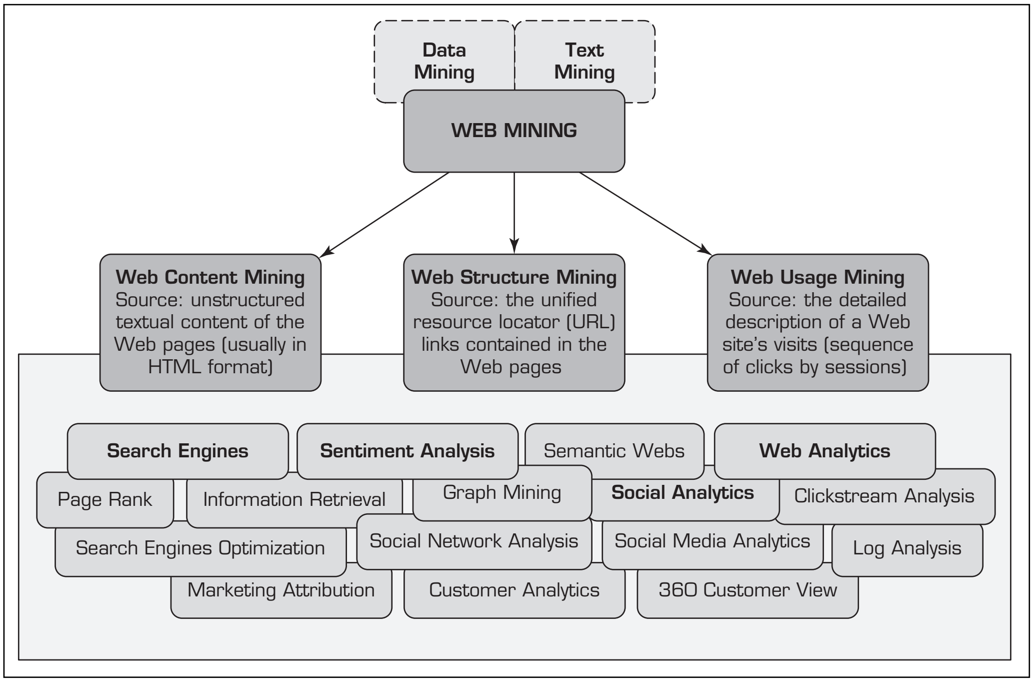


FIGURE 8.1 A Simple Taxonomy of Web Mining.

8.3 WEB CONTENT AND WEB STRUCTURE MINING

Web content mining refers to the extraction of useful information from Web pages. The documents may be extracted in some machine-readable format so that automated techniques can extract some information from these Web pages. **Web crawlers** (also called **spiders**) are used to read through the content of a Web site automatically. The information gathered may include document characteristics similar to what are used in text mining, but it may also include additional concepts, such as the document hierarchy. Such an automated (or semiautomated) process of collecting and mining Web content can be used for competitive intelligence (collecting intelligence about competitors' products, services, and customers). It can also be used for information/news/opinion collection and summarization, sentiment analysis, automated data collection, and structuring for predictive modeling. As an illustrative example to using Web content mining as an automated data collection tool, consider the following. For more than 10 years, two of the three authors of this book (Drs. Sharda and Delen) have been developing models to predict the financial success of Hollywood movies before their theatrical release. The data that they use for training the models come from several Web sites, each of which has a different hierarchical page structure. Collecting a large set of variables on thousands of movies (from the past several years) from these Web sites is a time-demanding, error-prone process. Therefore, they use Web content mining and spiders as an enabling technology to automatically collect, verify, validate (if the specific data item is available on more than one Web site, then the values are validated against each other and anomalies are captured and recorded), and store these values in a relational database. That way, they ensure the quality of the data while saving valuable time (days or weeks) in the process.

In addition to text, Web pages also contain hyperlinks pointing one page to another. Hyperlinks contain a significant amount of hidden human annotation that can potentially help to automatically infer the notion of centrality or *authority*. When a Web page developer includes a link pointing to another Web page, this may be regarded as the developer's endorsement of the other page. The collective endorsement of a given page by different developers on the Web may indicate the importance of the page and may naturally lead to the discovery of authoritative Web pages (Miller, 2005). Therefore, the vast amount of Web linkage information provides a rich collection of information about the relevance, quality, and structure of the Web's contents, and thus is a rich source for Web mining.

Web content mining can also be used to enhance the results produced by search engines. In fact, search is perhaps the most prevailing application of Web content mining and Web structure mining. A search on the Web to obtain information on a specific topic (presented as a collection of keywords or a sentence) usually returns a few relevant, high-quality Web pages and a larger number of unusable Web pages. Use of a relevance index based on keywords and authoritative pages (or some measure of it) will improve the search results and ranking of relevant pages. The idea of authority (or **authoritative pages**) stems from earlier information retrieval work using citations among journal articles to evaluate the impact of research papers (Miller, 2005). Though that was the origin of the idea, there are significant differences between the citations in research articles and hyperlinks on Web pages. First, not every hyperlink represents an endorsement (some links are created for navigation purposes and some are for paid advertisement). While this is true, if the majority of the hyperlinks are of the endorsement type, then the collective opinion will still prevail. Second, for commercial and competitive interests, one authority will rarely have its Web page point to rival authorities in the same domain. For example, Microsoft may prefer not to include links on its Web pages to Apple's Web sites, because this may be regarded as endorsement of its competitor's authority. Third, authoritative pages are seldom particularly descriptive. For example, the main Web page of Yahoo! may not contain the explicit self-description that it is in fact a Web search engine.

The structure of Web hyperlinks has led to another important category of Web pages called a **hub**. A hub is one or more Web pages that provide a collection of links to authoritative pages. Hub pages may not be prominent and only a few links may point to them; however, they provide links to a collection of prominent sites on a specific topic of interest. A hub could be a list of recommended links on an individual's homepage, recommended reference sites on a course Web page, or a professionally assembled resource list on a specific topic. Hub pages play the role of implicitly conferring the authorities on a narrow field. In essence, a close symbiotic relationship exists between good hubs and authoritative pages; a good hub is good because it points to many good authorities, and a good authority is good because it is being pointed to by many good hubs. Such relationships between hubs and authorities make it possible to automatically retrieve high-quality content from the Web.

The most popular publicly known and referenced algorithm used to calculate hubs and authorities is **hyperlink-induced topic search (HITS)**. It was originally developed by Kleinberg (1999) and has since been improved on by many researchers. HITS is a link-analysis algorithm that rates Web pages using the hyperlink information contained within them. In the context of Web search, the HITS algorithm collects a base document set for a specific query. It then recursively calculates the hub and authority values for each document. To gather the base document set, a root set that matches the query is fetched from a search engine. For each document retrieved, a set of documents that points to the original document and another set of documents that is pointed to by the original document are added to the set as the original document's neighborhood. A recursive process of document identification and link analysis continues until the hub and authority values converge. These values are then used to index and prioritize the document collection generated for a specific query.

Web structure mining is the process of extracting useful information from the links embedded in Web documents. It is used to identify authoritative pages and hubs, which are the cornerstones of the contemporary page-rank algorithms that are central to popular search engines such as Google and Yahoo!. Just as links going to a Web page may indicate a site's popularity (or authority), links within the Web page (or the compete Web site) may indicate the depth of coverage of a specific topic. Analysis of links is very important in understanding the interrelationships among large numbers of Web pages, leading to a better understanding of a specific Web community, clan, or clique. Application Case 8.1 describes a project that used both Web content mining and Web structure mining to better understand how U.S. extremist groups are connected.

SECTION 8.3 REVIEW QUESTIONS

1. What is Web content mining? How can it be used for competitive advantage?
2. What is an "authoritative page"? What is a "hub"? What is the difference between the two?
3. What is Web structure mining? How does it differ from Web content mining?

Application Case 8.1

Identifying Extremist Groups with Web Link and Content Analysis

We normally search for answers to our problems outside of our immediate environment. Often, however, the trouble stems from within. In taking action against global terrorism, domestic extremist groups often go unnoticed. However, domestic extremists pose a significant threat to U.S. security because of the information they possess, as well as their increasing ability, through the use of the Internet, to reach out to extremist groups around the world.

Keeping tabs on the content available on the Internet is difficult. Researchers and authorities need superior tools to analyze and monitor the activities of extremist groups. Researchers at the University of Arizona, with support from the Department of Homeland Security and other agencies, have developed a Web mining methodology to find and analyze Web sites operated by domestic extremists in order to learn about these groups through their use of the Internet. Extremist groups use the Internet to communicate, to access private messages, and to raise money online.

The research methodology begins by gathering a superior-quality collection of relevant extremist and terrorist Web sites. Hyperlink analysis is performed, which leads to other extremist and terrorist Web sites. The interconnectedness with other Web sites is crucial in estimating the similarity of the objectives of various groups. The next step is content analysis, which further codifies these Web sites based on various attributes, such as communications, fund raising, and ideology sharing, to name a few.

Based on link analysis and content analysis, researchers have identified 97 Web sites of U.S. extremist and hate groups. Often, the links between these communities do not necessarily represent any cooperation between them. However, finding numerous links between common interest groups helps in clustering the communities under a common banner. Further research using data mining to automate the process has a global aim, with the goal of identifying links between international hate and extremist groups and their U.S. counterparts.

Questions for Discussion

1. How can Web link/content analysis be used to identify extremist groups?
2. What do you think are the challenges and the potential solution to such intelligence gathering activities?

Source: Y. Zhou, E. Reid, J. Qin, H. Chen, and G. Lai, "U.S. Domestic Extremist Groups on the Web: Link and Content Analysis," *IEEE Intelligent Systems*, Vol. 20, No. 5, September/October 2005, pp. 44–51.

8.4 SEARCH ENGINES

In this day and age, there is no denying the importance of Internet search engines. As the size and complexity of the World Wide Web increases, finding what you want is becoming a complex and laborious process. People use search engines for a variety of reasons. We use them to learn about a product or a service before committing to buy (including who else is selling it, what the prices are at different locations/sellers, the common issues people are discussing about it, how satisfied previous buyers are, what other products or services might be better, etc.) and search for places to go, people to meet, and things to do. In a sense, search engines have become the centerpiece of most Internet-based transactions and other activities. The incredible success and popularity of Google, the most popular search engine company, is a good testament to this claim. What is somewhat a mystery to many is how a search engine actually does what it is meant to do. In simplest terms, a **search engine** is a software program that searches for documents (Internet sites or files) based on the keywords (individual words, multi-word terms, or a complete sentence) that users have provided that have to do with the subject of their inquiry. Search engines are the workhorses of the Internet, responding to billions of queries in hundreds of different languages every day.

Technically speaking, *search engine* is the popular term for information retrieval system. Although Web search engines are the most popular, search engines are often used in a context other than the Web, such as desktop search engines or document search engines. As you will see in this section, many of the concepts and techniques that we covered in the text analytics and text mining chapter (Chapter 7) also apply here. The overall goal of a search engine is to return one or more documents/pages (if more than one documents/pages applies, a rank-order list is often provided) that best match the user's query. The two metrics that are often used to evaluate search engines are *effectiveness* (or quality—finding the right documents/pages) and *efficiency* (or speed—returning a response quickly). These two metrics tend to work in reverse direction; improving one tends to worsen the other. Often, based on user expectation, search engines focus on one at the expense of the other. Better search engines are the ones that excel in both at the same time. Because search engines not only search but, in fact, find and return the documents/pages, perhaps a more appropriate name for them would be "finding engines."

Anatomy of a Search Engine

Now let us dissect a search engine and look inside it. At the highest level, a search engine system is composed of two main cycles: a development cycle and a responding cycle (see the structure of a typical Internet search engine in Figure 8.2). While one is interfacing

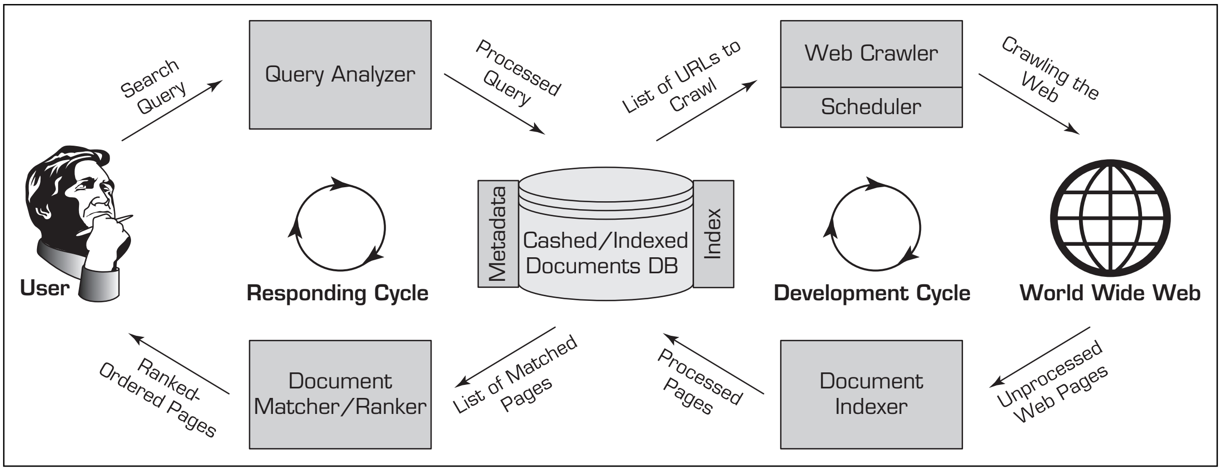


FIGURE 8.2 Structure of a Typical Internet Search Engine.

with the World Wide Web, the other is interfacing with the user. One can think of the development cycle as a production process (manufacturing and inventorying documents/pages) and the responding cycle as a retailing process (providing customers/users with what they want). In the following section these two cycles are explained in more detail.

1. Development Cycle

The two main components of the development cycle are the Web crawler and document indexer. The purpose of this cycle is to create a huge database of documents/pages organized and indexed based on their content and information value. The reason for developing such a repository of documents/pages is quite obvious: Due to its sheer size and complexity, searching the Web to find pages in response to a user query is not practical (or feasible within a reasonable time frame); therefore, search engines "cashes the Web" into their database, and uses the cashed version of the Web for searching and finding. Once created, this database allows search engines to rapidly and accurately respond to user queries.

Web Crawler

A Web crawler (also called a spider or a Web spider) is a piece of software that systematically browses (crawls through) the World Wide Web for the purpose of finding and fetching Web pages. Often Web crawlers copy all the pages they visit for later processing by other functions of a search engine.

A Web crawler starts with a list of URLs to visit, which are listed in the scheduler and often are called the seeds. These URLs may come from submissions made by Webmasters or, more often, they come from the internal hyperlinks of previously crawled documents/pages. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit (i.e., the scheduler). URLs in the scheduler are recursively visited according to a set of policies determined by the specific search engine. Because there are large volumes of Web pages, the crawler can only download a limited number of them within a given time; therefore, it may need to prioritize its downloads.

Document Indexer

As the documents are found and fetched by the crawler, they are stored in a temporary staging area for the document indexer to grab and process. The document indexer is responsible for processing the documents (Web pages or document files) and placing them into the document database. In order to convert the documents/pages into the desired, easily searchable format, the document indexer performs the following tasks.

STEP 1: PREPROCESSING THE DOCUMENTS Because the documents fetched by the crawler may all be in different formats, for the ease of processing them further, in this step they all are converted to some type of standard representation. For instance, different content types (text, hyperlink, image, etc.) may be separated from each other, formatted (if necessary), and stored in a place for further processing.

STEP 2: PARSING THE DOCUMENTS This step is essentially the application of text mining (i.e., computational linguistic, natural language processing) tools and techniques to a collection of documents/pages. In this step, first the standardized documents are parsed into its components to identify index-worthy words/terms. Then, using a set of rules, the words/terms are indexed. More specifically, using tokenization rules, the words/terms/entities are extracted from the sentences in these documents. Using proper lexicons, the spelling errors and other anomalies in these words/terms are corrected. Not all the terms are discriminators. The nondiscriminating words/terms (also known as stop words) are eliminated from the list of index-worthy words/terms. Because the same word/term can be in many different forms, stemming is applied to reduce the words/terms to their root forms. Again, using lexicons and other language-specific resources (e.g., WordNet), synonyms and homonyms are identified and the word/term collection is processed before moving into the indexing phase.

STEP 3: CREATING THE TERM-BY-DOCUMENT MATRIX In this step, the relationships between the words/terms and documents/pages are identified. The weight can be as simple as assigning 1 for presence or 0 for absence of the word/term in the document/page. Usually more sophisticated weight schemas are used. For instance, as opposed to binary, one may choose to assign frequency of occurrence (number of times the same word/term is found in a document) as a weight. As we have seen in Chapter 7, text mining research and practice have clearly indicated that the best weighting may come from the use of *term-frequency* divided by *inverse-document-frequency* (TF/IDF). This algorithm measures the frequency of occurrence of each word/term within a document, and then compares that frequency against the frequency of occurrence in the document collection. As we all know, not all high-frequency words/term are good document discriminators; and a good document discriminator in a domain may not be one in another domain. Once the weighing schema is determined, the weights are calculated and the term-by-document index file is created.

2. Response Cycle

The two main components of the responding cycle are the query analyzer and the document matcher/ranker.

Query Analyzer

The query analyzer is responsible for receiving a search request from the user (via the search engine's Web server interface) and converting it into a standardized data structure, so that it can be easily queried/matched against the entries in the document database. How the query analyzer does what it is supposed to do is quite similar to what the document indexer does (as we have just explained). The query analyzer parses the search string into individual words/terms using a series of tasks that include tokenization, removal of stop words, stemming, and word/term disambiguation (identification of spelling errors, synonyms, and homonyms). The close similarity between the query analyzer and document indexer is not coincidental. In fact, it is quite logical, because both are working off of the document database; one is putting in documents/pages using a specific index structures, and the other is converting a query string into the same structure so that it can be used to quickly locate most relevant documents/pages.

Document Matcher/Ranker

This is where the structured query data is matched against the document database to find the most relevant documents/pages and also rank them in the order of relevance/importance. The proficiency of this step is perhaps the most important component when different search engines are compared to one another. Every search engine has its own (often proprietary) algorithm that it uses to carry out this important step.

The early search engines used a simple keyword match against the document database and returned a list of ordered documents/pages, where the determinant of the order was a function that used the number of words/terms matched between the query and the document along with the weights of those words/terms. The quality and the usefulness of the search results were not all that good. Then, in 1997, the creators of Google came up with a new algorithm, called PageRank. As the name implies, PageRank is an algorithmic way to rank-order documents/pages based on their relevance and value/importance. Technology Insights 8.1 provides a high-level description of this patented algorithm. Even

TECHNOLOGY INSIGHTS 8.1 PageRank Algorithm

PageRank is a link analysis algorithm—named after Larry Page, one of the two inventors of Google, which started as a research project at Stanford University in 1996—used by the Google Web search engine. PageRank assigns a numerical weight to each element of a hyperlinked set of documents, such as the ones found on the World Wide Web, with the purpose of measuring its relative importance within a given collection.

It is believed that PageRank has been influenced by citation analysis, where citations in scholarly works are examined to discover relationships among researchers and their research topics. The applications of citation analysis ranges from identification of prominent experts in a given field of study to providing invaluable information for a transparent review of academic achievements, which can be used for merit review, tenure, and promotion decisions. The PageRank algorithm aims to do the same thing: identifying reputable/important/valuable documents/pages that are highly regarded by other documents/pages. A graphical illustration of PageRank is shown in Figure 8.3.

How Does PageRank Work?

Computationally speaking, PageRank extends the citation analysis idea by not counting links from all pages equally and by normalizing by the number of links on a page. PageRank is defined as follows:

Assume page A has pages P_1 through P_n pointing to it (with *hyperlinks*, which is similar to *citations* in citation analysis). The parameter d is a damping/smoothing factor that can assume values between 0 and 1. Also *C(A)* is defined as the number of links going out of page *A*. The simple formula for the PageRank for page *A* can be written as follows:

$$PageRank(A) = (1 - d) + d\sum_{i=1}^{n} \frac{PageRank(P_i)}{C(P_i)}$$

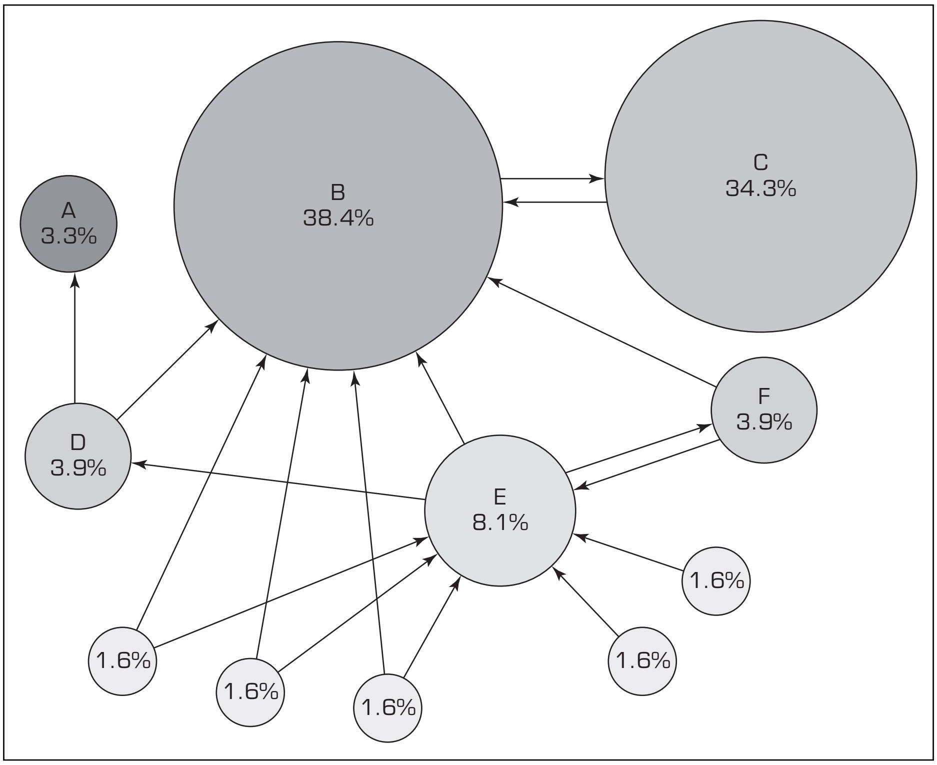


FIGURE 8.3 A Graphical Example for the PageRank Algorithm.

Note that the PageRanks form a probability distribution over Web pages, so the sum of all Web pages' PageRanks will be 1. *PageRank(A)* can be calculated using a simple iterative algorithm and corresponds to the principal eigenvector of the normalized link matrix of the Web. The algorithm is so computationally efficient that a PageRank for 26 million Web pages can be computed in a few hours on a medium-size workstation (Brin and Page, 2012). Of course, there are more details to the actual calculation of PageRank in Google. Most of those details are either not publicly available or are beyond the scope of this simple explanation.

Justification of the Formulation

PageRank can be thought of as a model of user behavior. It assumes there is a *random surfer* who is given a Web page at random and keeps clicking on hyperlinks, never hitting *back* but eventually getting bored and starting on another random page. The probability that the random surfer visits a page is its PageRank. And, the *d* damping factor is the probability at each page the *random surfer* will get bored and request another random page. One important variation is to only add the damping factor *d* to a single page, or a group of pages. This allows for personalization and can make it nearly impossible to deliberately mislead the system in order to get a higher ranking.

Another intuitive justification is that a page can have a high PageRank if there are many pages that point to it, or if there are some pages that point to it and have a high PageRank. Intuitively, pages that are well cited from many places around the Web are worth looking at. Also, pages that have perhaps only one citation from something like the Yahoo! homepage are also generally worth looking at. If a page was not high quality, or was a broken link, it is quite likely that Yahoo!'s homepage would not link to it. The formulation of PageRank handles both of these cases and everything in between by recursively propagating weights through the link structure of the Web.

though PageRank is an innovative way to rank documents/pages, it is an augmentation to the process of retrieving relevant documents from the database and ranking them based on the weights of the words/terms. Google does all of these collectively and more to come up with the most relevant list of documents/pages for a given search request. Once an ordered list of documents/pages is created, it is pushed back to the user in an easily digestible format. At this point, users may choose to click on any of the documents in the list, and it may not be the one at the top. If they click on a document/page link that is not at the top of the list, then can we assume that the search engine did not do a good job ranking them? Perhaps, yes. Leading search engines like Google monitor the performance of their search results by capturing, recording, and analyzing postdelivery user actions and experiences. These analyses often lead to more and more rules to further refine the ranking of the documents/pages so that the links at the top are more preferable to the end users.

How Does Google Do It?

Even though complex low-level computational details are trade secrets and are not known to the public, the high-level structure of the Google search system is well-known and quite simple. From the infrastructure standpoint, the Google search system runs on a distributed network of tens of thousands of computers/servers and can, therefore, carry out its heavy workload effectively and efficiently using sophisticated parallel processing algorithms (a method of computation in which many calculations can be distributed to many servers and performed simultaneously, significantly speeding up data processing). At the highest level, the Google search system has three distinct parts (**googleguide.com**):

1. Googlebot, a Web crawler that roams the Internet to find and fetch Web pages
2. The indexer, which sorts every word on every page and stores the resulting index of words in a huge database

3. The query processor, which compares your search query to the index and recommends the documents that it considers most relevant

1. ***Googlebot*** Googlebot is Google's Web crawling robot, which finds and retrieves pages on the Web and hands them off to the Google indexer. It's easy to imagine Googlebot as a little spider scurrying across the strands of cyberspace, but in reality Googlebot doesn't traverse the Web at all. It functions, much like your Web browser, by sending a request to a Web server for a Web page, downloading the entire page, and then handing it off to Google's indexer. Googlebot consists of many computers requesting and fetching pages much more quickly than you can with your Web browser. In fact, Googlebot can request thousands of different pages simultaneously. To avoid overwhelming Web servers, or crowding out requests from human users, Googlebot deliberately makes requests of each individual Web server more slowly than it's capable of doing.

   When Googlebot fetches a page, it removes all the links appearing on the page and adds them to a queue for subsequent crawling. Googlebot tends to encounter little spam because most Web authors link only to what they believe are high-quality pages. By harvesting links from every page it encounters, Googlebot can quickly build a list of links that can cover broad reaches of the Web. This technique, known as *deep crawling,* also allows Googlebot to probe deep within individual sites. Because of their massive scale, deep crawls can reach almost every page in the Web. To keep the index current, Google continuously recrawls popular frequently changing Web pages at a rate roughly proportional to how often the pages change. Such crawls keep an index current and are known as *fresh crawls*. Newspaper pages are downloaded daily; pages with stock quotes are downloaded much more frequently. Of course, fresh crawls return fewer pages than the deep crawl. The combination of the two types of crawls allows Google to both make efficient use of its resources and keep its index reasonably current.
2. ***Google Indexer*** Googlebot gives the indexer the full text of the pages it finds. These pages are stored in Google's index database. This index is sorted alphabetically by search term, with each index entry storing a list of documents in which the term appears and the location within the text where it occurs. This data structure allows rapid access to documents that contain user query terms. To improve search performance, Google ignores common words, called *stop words* (such as *the*, *is*, *on*, *or*, *of*, *a*, *an*, as well as certain single digits and single letters). Stop words are so common that they do little to narrow a search, and therefore they can safely be discarded. The indexer also ignores some punctuation and multiple spaces, as well as converting all letters to lowercase, to improve Google's performance.
3. ***Google Query Processor*** The query processor has several parts, including the user interface (search box), the "engine" that evaluates queries and matches them to relevant documents, and the results formatter.

Google uses a proprietary algorithm, called PageRank, to calculate the relative rank order of a given collection of Web pages. PageRank is Google's system for ranking Web pages. A page with a higher PageRank is deemed more important and is more likely to be listed above a page with a lower PageRank. Google considers over a hundred factors in computing a PageRank and determining which documents are most relevant to a query, including the popularity of the page, the position and size of the search terms within the page, and the proximity of the search terms to one another on the page.

Google also applies machine-learning techniques to improve its performance automatically by learning relationships and associations within the stored data. For example, the *spelling-correcting system* uses such techniques to figure out likely alternative spellings. Google closely guards the formulas it uses to calculate relevance; they're tweaked to improve quality and performance, and to outwit the latest devious techniques used by spammers.

Indexing the full text of the Web allows Google to go beyond simply matching single search terms. Google gives more priority to pages that have search terms near each other and in the same order as the query. Google can also match multi-word phrases and sentences. Because Google indexes HTML code in addition to the text on the page, users can restrict searches on the basis of where query words appear (e.g., in the title, in the URL, in the body, and in links to the page, options offered by Google's Advanced Search Form and Using Search Operators).

Understanding the internals of popular search engines helps companies, who rely on search engine traffic, better design their e-commerce sites to improve their chances of getting indexed and highly ranked by search providers. Application Case 8.2 gives an illustrative example of such a phenomenon, where an entertainment company increased its search-originated customer traffic by 1500 percent.

Application Case 8.2

IGN Increases Search Traffic by 1500 Percent

IGN Entertainment operates the Internet's largest network of destinations for video gaming, entertainment, and community geared toward teens and 18- to 34-year-old males. The company's properties include IGN.com, GameSpy, AskMen.com, RottenTomatoes, FilePlanet, TeamXbox, 3D Gamers, VE3D, and Direct2Drive—more than 70 community sites and a vast array of online forums. IGN Entertainment is also a leading provider of technology for online game play in video games.

The Challenge

When this company contacted SEO Inc. in summer 2003, the site was an established and well-known site in the gaming community. The site also had some good search engine rankings and was getting approximately 2.5 million unique visitors per month. At the time IGN used proprietary in-house content management and a team of content writers. The pages that were generated when new game reviews and information were added to the site were not very well optimized. In addition, there were serious architectural issues with the site, which prevented search engine spiders from thoroughly and consistently crawling the site.

IGN's goals were to "dominate the search rankings for keywords related to any video games and gaming systems reviewed on the site." IGN wanted to rank high in the search engines, and most specifically, Google, for any and all game titles and variants on those game titles' phrases. IGN's revenue is generated from advertising sales, so more traffic leads to more inventory for ad sales, more ads being sold, and therefore more revenue. In order to generate more traffic, IGN knew that it needed to be much more visible when people used the search engines.

The Strategy

After several conversations with the IGN team, SOE Inc. created a customized optimization package that was designed to achieve their ranking goals and also fit the client's budget. Because IGN.com had architectural problems and a proprietary CMS (content management system), it was decided that SEO Inc. would work with their IT and Web development team at their location. This allowed SEO to send their team to the IGN location for several days to learn how their system worked and partner with their in-house programmers to improve the system and, hence, improve search engine optimization. In addition, SEO created customized SEO best practices and architected these into their proprietary CMS. SEO also trained their content writers and page developers on SEO best practices. When new games and pages are added to the site, they are typically getting ranked within weeks, if not days.

(*Continued*)

Application Case 8.2 (Continued)

The Results

This was a true and quick success story. Organic search engine rankings skyrocketed and thousands of previously not-indexed pages were now being crawled regularly by search engine spiders. Some of the specific results were as follows:

- Unique visitors to the site doubled within the first 2 months after the optimization was completed.
- There was a 1500 percent increase in organic search engine traffic.
- Massive growth in traffic and revenues enabled acquisition of additional Web properties including Rottentomatoes.com and Askmen.com

IGN was acquired by News Corp in September 2005 for $650 million.

Questions for Discussion

1. How did IGN dramatically increase search traffic to its Web portals?
2. What were the challenges, the proposed solution, and the obtained results?

Source: SOE Inc., Customer Case Study, **seoinc.com/seo/case-studies/ign** (accessed March 2013).

SECTION 8.4 REVIEW QUESTIONS

1. What is a search engine? Why are they important for today's businesses?

2. What is the relationship between search engines and text mining?

3. What are the two main cycles in search engines? Describe the steps in each cycle.

4. What is a Web crawler? What is it used for? How does it work?

5. How does a query analyzer work? What is PageRank algorithm and how does it work?

8.5 SEARCH ENGINE OPTIMIZATION

Search engine optimization (SEO) is the intentional activity of affecting the visibility of an e-commerce site or a Web site in a search engine's natural (unpaid or organic) search results. In general, the higher ranked on the search results page, and more frequently a site appears in the search results list, the more visitors it will receive from the search engine's users. As an Internet marketing strategy, SEO considers how search engines work, what people search for, the actual search terms or keywords typed into search engines, and which search engines are preferred by their targeted audience. Optimizing a Web site may involve editing its content, HTML, and associated coding to both increase its relevance to specific keywords and to remove barriers to the indexing activities of search engines. Promoting a site to increase the number of backlinks, or inbound links, is another SEO tactic.

In the early days, in order to be indexed, all Webmasters needed to do was to submit the address of a page, or URL, to the various engines, which would then send a "spider" to "crawl" that page, extract links to other pages from it, and return information found on the page to the server for indexing. The process, as explained before, involves a search engine spider downloading a page and storing it on the search engine's own server, where a second program, known as an indexer, extracts various information about the page, such as the words it contains and where these are located, as well as any weight for specific words, and all links the page contains, which are then placed into a scheduler for crawling at a later date. Nowadays search engines are no longer relying on Webmasters submitting URLs (even though they still can); instead, they are proactively and continuously crawling the Web, and finding, fetching, and indexing everything about it.

Being indexed by search engines like Google, Bing, and Yahoo! is not good enough for businesses. Getting ranked on the most widely used search engines (see Technology Insights 8.2 for a list of most widely used search engines) and getting ranked higher than your competitors are what make the difference. A variety of methods can increase the ranking of a Web page within the search results. Cross-linking between pages of the same Web site to provide more links to the most important pages may improve its visibility. Writing content that includes frequently searched keyword phrases, so as to be relevant to a wide variety of search queries, will tend to increase traffic. Updating content so as to keep search engines crawling back frequently can give additional weight to a site. Adding relevant keywords to a Web page's metadata, including the title tag and metadescription, will tend to improve the relevancy of a site's search listings, thus increasing traffic. URL normalization of Web pages so that they are accessible via multiple URLs and using canonical link element and redirects can help make sure links to different versions of the URL all count toward the page's link popularity score.

Methods for Search Engine Optimization

In general, SEO techniques can be classified into two broad categories: techniques that search engines recommend as part of good site design, and those techniques of which search engines do not approve. The search engines attempt to minimize the effect of the latter, which is often called *spamdexing* (also known as *search spam, search engine spam,* or *search engine poisoning*). Industry commentators have classified these methods, and the practitioners who employ them, as either white-hat SEO or black-hat SEO

TECHNOLOGY INSIGHTS 8.2 Top 15 Most Popular Search Engines (March 2013)

Here are the 15 most popular search engines as derived from eBizMBA Rank **(ebizmba.com/articles/search-engines)**, which is a constantly updated average of each Web site's Alexa Global Traffic Rank, and U.S. Traffic Rank from both Compete and Quantcast.

Rank	Name	Estimated Unique Monthly Visitors
1	Google	900,000,000
2	Bing	165,000,000
3	Yahoo! Search	160,000,000
4	Ask	125,000,000
5	AOL Search	33,000,000
6	MyWebSearch	19,000,000
7	blekko	9,000,000
8	Lycos	4,300,000
9	Dogpile	2,900,000
10	WebCrawler	2,700,000
11	Info	2,600,000
12	Infospace	2,000,000
13	Search	1,450,000
14	Excite	1,150,000
15	GoodSearch	1,000,000

(Goodman, 2005). White hats tend to produce results that last a long time, whereas black hats anticipate that their sites may eventually be banned either temporarily or permanently once the search engines discover what they are doing.

An SEO technique is considered white hat if it conforms to the search engines' guidelines and involves no deception. Because search engine guidelines are not written as a series of rules or commandments, this is an important distinction to note. White-hat SEO is not just about following guidelines, but about ensuring that the content a search engine indexes and subsequently ranks is the same content a user will see. White-hat advice is generally summed up as creating content for users, not for search engines, and then making that content easily accessible to the spiders, rather than attempting to trick the algorithm from its intended purpose. White-hat SEO is in many ways similar to Web development that promotes accessibility, although the two are not identical.

Black-hat SEO attempts to improve rankings in ways that are disapproved by the search engines, or involve deception. One black-hat technique uses text that is hidden, either as text colored similar to the background, in an invisible div, or positioned off-screen. Another method gives a different page depending on whether the page is being requested by a human visitor or a search engine, a technique known as *cloaking*. Search engines may penalize sites they discover using black-hat methods, either by reducing their rankings or eliminating their listings from their databases altogether. Such penalties can be applied either automatically by the search engines' algorithms, or by a manual site review. One example was the February 2006 Google removal of both BMW Germany and Ricoh Germany for use of unapproved practices (Cutts, 2006). Both companies, however, quickly apologized, fixed their practices, and were restored to Google's list.

For some businesses SEO may generate significant return on investment. However, one should keep in mind that search engines are not paid for organic search traffic, their algorithms change constantly, and there are no guarantees of continued referrals. Due to this lack of certainty and stability, a business that relies heavily on search engine traffic can suffer major losses if the search engine decides to change its algorithms and stop sending visitors. According to Google's CEO, Eric Schmidt, in 2010, Google made over 500 algorithm changes—almost 1.5 per day. Because of the difficulty in keeping up with changing search engine rules, companies that rely on search traffic practice one or more of the following: (1) Hire a company that specializes in search engine optimization (there seem to be an abundant number of those nowadays) to continuously improve your site's appeal to changing practices of the search engines; (2) pay the search engine providers to be listed on the paid sponsors sections; and (3) consider liberating yourself from dependence on search engine traffic.

Either originating from a search engine (organically or otherwise) or coming from other sites and places, what is most important for an e-commerce site is to maximize the likelihood of customer transactions. Having a lot of visitors without sales is not what a typical e-commerce site is built for. Application Case 8.3 is about a large Internet-based shopping mall where detailed analysis of customer behavior (using clickstreams and other data sources) is used to significantly improve the conversion rate.

SECTION 8.5 REVIEW QUESTIONS

1. What is "search engine optimization"? Who benefits from it?

2. Describe the old and new ways of indexing performed by search engines.

3. What are the things that help Web pages rank higher in the search engine results?

4. What are the most commonly used methods for search engine optimization?

Application Case 8.3

Understanding Why Customers Abandon Shopping Carts Results in $10 Million Sales Increase

Lotte.com, the leading Internet shopping mall in Korea with 13 million customers, has developed an integrated Web traffic analysis system using SAS for Customer Experience Analytics. As a result, Lotte.com has been able to improve the online experience for its customers, as well as generate better returns from its marketing campaigns. Now, Lotte.com executives can confirm results anywhere, anytime, as well as make immediate changes.

With almost 1 million Web site visitors each day, Lotte.com needed to know how many visitors were making purchases and which channels were bringing the most valuable traffic. After reviewing many diverse solutions and approaches, Lotte.com introduced its integrated Web traffic analysis system using the SAS for Customer Experience Analytics solution. This is the first online behavioral analysis system applied in Korea.

With this system, Lotte.com can accurately measure and analyze Web site visitor numbers (UV), page view (PV) status of site visitors and purchasers, the popularity of each product category and product, clicking preferences for each page, the effectiveness of campaigns, and much more. This information enables Lotte.com to better understand customers and their behavior online, and conduct sophisticated, cost-effective targeted marketing.

Commenting on the system, Assistant General Manager Jung Hyo-hoon of the Marketing Planning Team for Lotte.com said, "As a result of introducing the SAS system of analysis, many 'new truths' were uncovered around customer behavior, and some of them were 'inconvenient truths.'" He added, "Some site-planning activities that had been undertaken with the expectation of certain results actually had a low reaction from customers, and the site planners had a difficult time recognizing these results."

Benefits

Introducing the SAS for Customer Experience Analytics solution fully transformed the Lotte.com Web site. As a result, Lotte.com has been able to improve the online experience for its customers as well as generate better returns from its marketing campaigns. Now, Lotte.com executives can confirm results anywhere, anytime, as well as make immediate changes.

Since implementing SAS for Customer Experience Analytics, Lotte.com has seen many benefits:

A Jump in Customer Loyalty

A large amount of sophisticated activity information can be collected under a visitor environment, including quality of traffic. Deputy Assistant General Manager Jung said that "by analyzing actual valid traffic and looking only at one to two pages, we can carry out campaigns to heighten the level of loyalty, and determine a certain range of effect, accordingly." He added, "In addition, it is possible to classify and confirm the order rate for each channel and see which channels have the most visitors."

Optimized Marketing Efficiency Analysis

Rather than just analyzing visitor numbers only, the system is capable of analyzing the conversion rate (shopping cart, immediate purchase, wish list, purchase completion) compared to actual visitors for each campaign type (affiliation or e-mail, banner, keywords, and others), so detailed analysis of channel effectiveness is possible. Additionally, it can confirm the most popular search words used by visitors for each campaign type, location, and purchased products. The page overlay function can measure the number of clicks and number of visitors for each item in a page to measure the value for each location in a page. This capability enables Lotte.com to promptly replace or renew low traffic items.

Enhanced Customer Satisfaction and Customer Experience Lead to Higher Sales

Lotte.com built a customer behavior analysis database that measures each visitor, what pages are visited, how visitors navigate the site, and what activities are undertaken to enable diverse analysis and improve site efficiency. In addition, the database captures customer demographic information, shopping cart size and conversion rate, number of orders, and number of attempts.

By analyzing which stage of the ordering process deters the most customers and fixing those stages, conversion rates can be increased. Previously, analysis was done only on placed orders. By analyzing the movement pattern of visitors before ordering and at the point where breakaway occurs, customer

(Continued)

Application Case 8.3 (Continued)

behavior can be forecast, and sophisticated marketing activities can be undertaken. Through a pattern analysis of visitors, purchases can be more effectively influenced and customer demand can be reflected in real time to ensure quicker responses. Customer satisfaction has also improved as Lotte .com has better insight into each customer's behaviors, needs, and interests.

Evaluating the system, Jung commented, "By finding out how each customer group moves on the basis of the data, it is possible to determine customer service improvements and target marketing subjects, and this has aided the success of a number of campaigns." However, the most significant benefit of the system is gaining insight into individual customers and various customer groups. By understanding when customers will make purchases and the manner in which they navigate throughout the Web page, targeted channel marketing and better customer experience can now be achieved.

Plus, when SAS for Customer Experience Analytics was implemented by **Lotte.com**'s largest overseas distributor, it resulted in a first-year sales increase of 8 million euros (US$10 million) by identifying the causes of shopping-cart abandonment.

Source: SAS, Customer Success Stories, **sas.com/success/lotte .html** (accessed March 2013).

8.6 WEB USAGE MINING (WEB ANALYTICS)

Web usage mining (also called **Web analytics**) is the extraction of useful information from data generated through Web page visits and transactions. Masand et al. (2002) state that at least three types of data are generated through Web page visits:

1. Automatically generated data stored in server access logs, referrer logs, agent logs, and client-side cookies
2. User profiles
3. Metadata, such as page attributes, content attributes, and usage data.

Analysis of the information collected by Web servers can help us better understand user behavior. Analysis of this data is often called **clickstream analysis**. By using the data and text mining techniques, a company might be able to discern interesting patterns from the clickstreams. For example, it might learn that 60 percent of visitors who searched for "hotels in Maui" had searched earlier for "airfares to Maui." Such information could be useful in determining where to place online advertisements. Clickstream analysis might also be useful for knowing *when* visitors access a site. For example, if a company knew that 70 percent of software downloads from its Web site occurred between 7 and 11 P.M., it could plan for better customer support and network bandwidth during those hours. Figure 8.4 shows the process of extracting knowledge from clickstream data and how the generated knowledge is used to improve the process, improve the Web site, and, most important, increase the customer value.

Web mining has wide a range of business applications. For instance, Nasraoui (2006) listed the following six most common applications:

1. Determine the lifetime value of clients.
2. Design cross-marketing strategies across products.
3. Evaluate promotional campaigns.
4. Target electronic ads and coupons at user groups based on user access patterns.
5. Predict user behavior based on previously learned rules and users' profiles.
6. Present dynamic information to users based on their interests and profiles.

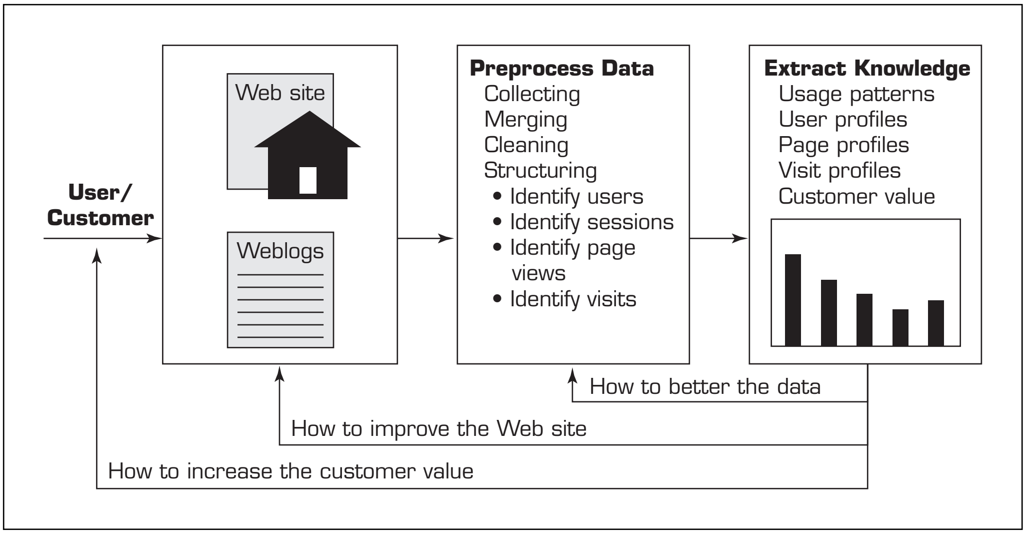


FIGURE 8.4 Extraction of Knowledge from Web Usage Data.

Amazon.com provides an excellent example of how Web usage history can be leveraged dynamically. A registered user who revisits Amazon.com is greeted by name. This is a simple task that involves recognizing the user by reading a cookie (i.e., a small text file written by a Web site on the visitor's computer). Amazon.com also presents the user with a choice of products in a personalized store, based on previous purchases and an association analysis of similar users. It also makes special "Gold Box" offers that are good for a short amount of time. All these recommendations involve a detailed analysis of the visitor as well as the user's peer group developed through the use of clustering, sequence pattern discovery, association, and other data and text mining techniques.

Web Analytics Technologies

There are numerous tools and technologies for Web analytics in the marketplace. Because of their power to measure, collect, and analyze Internet data to better understand and optimize Web usage, the popularity of Web analytics tools is increasing. Web analytics holds the promise to revolutionize how business is done on the Web. Web analytics is not just a tool for measuring Web traffic; it can also be used as a tool for e-business and market research, and to assess and improve the effectiveness of an e-commerce Web site. Web analytics applications can also help companies measure the results of traditional print or broadcast advertising campaigns. It can help estimate how traffic to a Web site changes after the launch of a new advertising campaign. Web analytics provides information about the number of visitors to a Web site and the number of page views. It helps gauge traffic and popularity trends, which can be used for market research.

There are two main categories of web analytics; off-site and on-site. Off-site Web analytics refers to Web measurement and analysis about you and your products that takes place outside your Web site. It includes the measurement of a Web site's potential audience (prospect or opportunity), share of voice (visibility or word-of-mouth), and buzz (comments or opinions) that is happening on the Internet.

What is more mainstream is on-site Web analytics. Historically, Web analytics has referred to on-site visitor measurement. However, in recent years this has blurred, mainly because vendors are producing tools that span both categories. On-site Web analytics measure a visitors' behavior once they are on your Web site. This includes its drivers and conversions—for example, the degree to which different landing pages are associated with online purchases. On-site Web analytics measure the performance of your Web site in a commercial context. This data collected on the Web site is then compared against key performance indicators for performance, and used to improve a Web site's or marketing campaign's audience response. Even though Google Analytics is the most widely-used on-site Web analytics service, there are others provided by Yahoo! and Microsoft, and newer and better tools are emerging constantly that provide additional layers of information.

For on-site Web analytics, there are two technical ways of collecting the data. The first and more traditional method is the server log file analysis, where the Web server records file requests made by browsers. The second method is page tagging, which uses JavaScript embedded in the site page code to make image requests to a third-party analytics-dedicated server whenever a page is rendered by a Web browser (or when a mouse click occurs). Both collect data that can be processed to produce Web traffic reports. In addition to these two main streams, other data sources may also be added to augment Web site behavior data. These other sources may include e-mail, direct-mail campaign data, sales and lead history, or social media–originated data. Application Case 8.4 shows how Allegro improved Web site performance by 500 percent with analysis of Web traffic data.

Application Case 8.4

Allegro Boosts Online Click-Through Rates by 500 Percent with Web Analysis

The Allegro Group is headquartered in Posnan, Poland, and is considered the largest non-eBay online marketplace in the world. Allegro, which currently offers over 75 proprietary Web sites in 11 European countries around the world, hosts over 15 million products and generates over 500 million page views per day. The challenge it faced was how to match the right offer to the right customer while still being able to support the extraordinary amount of data it held.

Problem

In today's marketplace, buyers have a wide variety of retail, catalog, and online options for buying their goods and services. Allegro is an e-marketplace with over 20 million customers who themselves buy from a network of over 30 thousand professional retail sellers using the Allegro network of e-commerce and auction sites. Allegro had been supporting its internal recommendation engine solely by applying rules provided by its re-sellers.

The challenge was for Allegro to increase its income and gross merchandise volume from its current network, as measured by two key performance indicators.

- ***Click-Thru Rates (CTR):*** The number of clicks on a product ad divided by the number of times the product is displayed.
- ***Conversion Rates:*** The number of completed sales transactions of a product divided by the number of customers receiving the product ad.

Solution

The online retail industry has evolved into the premier channel for personalized product recommendations. To succeed in this increasingly competitive e-commerce environment, Allegro realized that it needed to create a new, highly personalized solution integrating predictive analytics and campaign management into a real-time recommendation system.

Allegro decided to apply Social Network Analysis (SNA) as the analytic methodology underlying its product recommendation system. SNA focuses on the relationships or links between nodes (individuals or products) in a network, rather than the nodes' attributes as in traditional statistical methods. SNA was used to group similar products into communities based on their commonalities; then, communities were weighted based on visitor click paths, items placed in shopping carts, and purchases to create predictive attributes. The graph in Figure 8.5 displays a few of the product communities generated by Allegro using the KXEN's InfiniteInsight Social product for social network analysis (SNA).

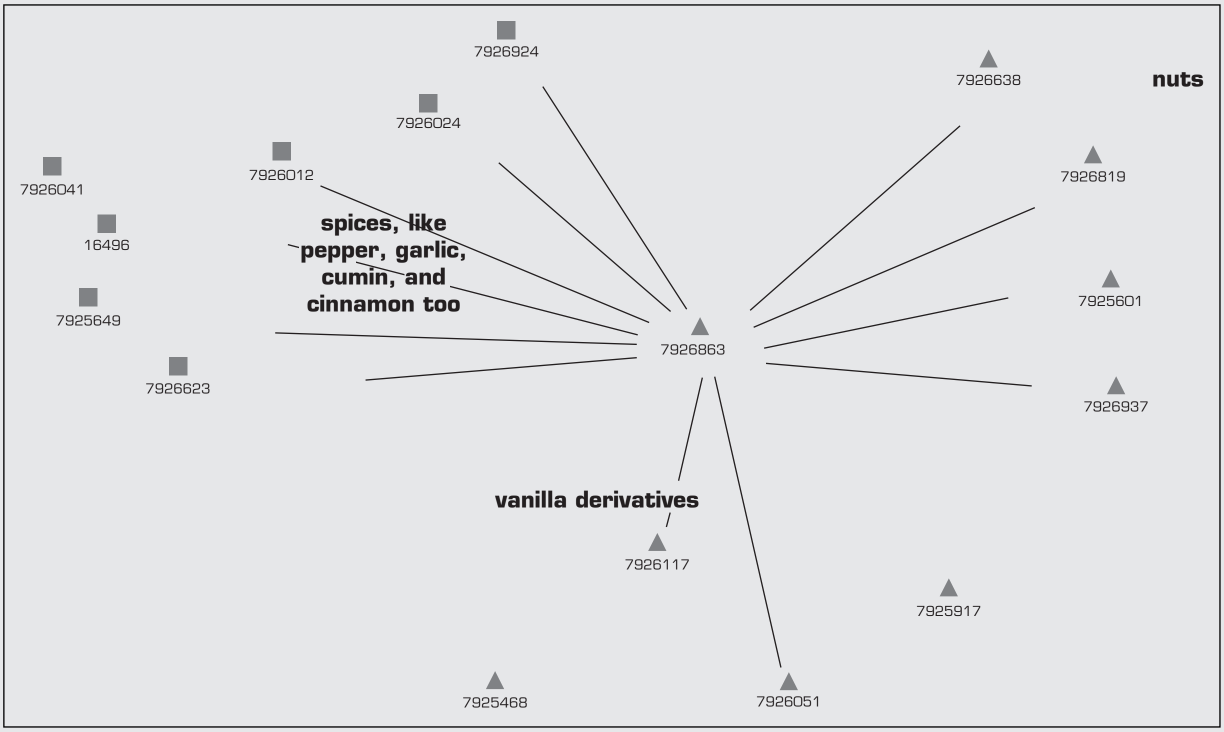


FIGURE 8.5 The Product Communities Generated by Allegro Using KXEN's InfiniteInsight . *Source:* KXEN.

Statistical classification models were then built using KXEN InfiniteInsight Modeler to predict conversion propensity for each product based on these SNA product communities and individual customer attributes. These conversion propensity scores are then used by Allegro to define personalized offers presented to millions of Web site visitors in real time.

Some of the challenges Allegro faced applying social network analysis included:

- Need to build multiple networks, depending on the product group categories
  – Very large differences in the frequency distribution of particular products and their popularity (clicks, transactions)
- Automatic setting of optimal parameters, such as the minimum number of occurrences of items (support)
- Automation through scripting
- Overconnected products (best-sellers, megahub communities).

Implementing this solution also presented its own challenges including:

- Different rule sets are produced per Web page placement
- Business owners decide appropriate weightings of rule sets for each type of placement / business strategy
- Building 160k rules every week
- Automatic conversion of social network analyses into rules and table-ization of rules

Results

As a result of implementing social network analysis in its automated real-time recommendation process, Allegro has seen a marked improvement in all areas.

Today Allegro offers 80 million personalized product recommendations daily, and its page views have increased by over 30 percent. But it's in the

(*Continued*)

Application Case 4.4 (Continued)

Rule ID	Antecedent product ID	Consequent product ID	Rule support	Rule confidence	Rule KI	Belong to the same product community?
1	DIGITAL CAMERA	LENS	21213	20%	0.76	YES
2	DIGITAL CAMERA	MEMORY CARD	3145	18%	0.64	NO
3	PINK SHOES	PINK DRESS	4343	38%	0.55	NO
…	…	…	…	…	…	…

numbers delivered by Allegro's two most critical KPIs that the results are most obvious:

- Click-through rate (CTR) has increased by more than 500 percent as compared to 'best seller' rules.
- Conversion rates are up by a factor of over 40X.

Questions for Discussion

1. How did Allegro significantly improve click-through rates with Web analytics?
2. What were the challenges, the proposed solution, and the obtained results?

Source: **kxen.com/customers/allegro** (accessed July 2013).

Web Analytics Metrics

Using a variety of data sources, Web analytics programs provide access to a lot of valuable marketing data, which can be leveraged for better insights to grow your business and better document your ROI. The insight and intelligence gained from Web analytics can be used to effectively manage the marketing efforts of an organization and its various products or services. Web analytics programs provide nearly real-time data, which can document your marketing campaign successes or empower you to make timely adjustments to your current marketing strategies.

While Web analytics provides a broad range of metrics, there are four categories of metrics that are generally actionable and can directly impact your business objectives (TWG, 2013). These categories include:

- Web site usability: How were they using my Web site?
- Traffic sources: Where did they come from?
- Visitor profiles: What do my visitors look like?
- Conversion statistics: What does all this mean for the business?

Web Site Usability

Beginning with your Web site, let's take a look at how well it works for your visitors. This is where you can learn how "user-friendly" it really is or whether or not you are providing the right content.

1. ***Page views.*** The most basic of measurements, this metric is usually presented as the "average page views per visitor." If people come to your Web site and don't view many pages, then your Web site may have issues with its design or structure. Another explanation for low page views is a disconnect in the marketing messages that brought them to the site and the content that is actually available.

2. ***Time on site.*** Similar to page views, it's a fundamental measurement of a visitor's interaction with your Web site. Generally, the longer a person spends on your Web site, the better it is. That could mean they're carefully reviewing your content, utilizing interactive components you have available, and building toward an informed decision to buy, respond, or take the next step you've provided. On the contrary, the time on site also needs to be examined against the number of pages viewed to make sure the visitor isn't spending his or her time trying to locate content that should be more readily accessible.

3. *Downloads.* This includes PDFs, videos, and other resources you make available to your visitors. Consider how accessible these items are as well as how well they're promoted. If your Web statistics, for example, reveal that 60 percent of the individuals who watch a demo video also make a purchase, then you'll want to strategize to increase viewership of that video.

4. *Click map.* Most analytics programs can show you the percentage of clicks each item on your Web page received. This includes clickable photos, text links in your copy, downloads, and, of course, any navigation you may have on the page. Are they clicking the most important items?

5. *Click paths.* Although an assessment of click paths is more involved, it can quickly reveal where you might be losing visitors in a specific process. A well-designed Web site uses a combination of graphics and information architecture to encourage visitors to follow "predefined" paths through your Web site. These are not rigid pathways but rather intuitive steps that align with the various processes you've built into the Web site. One process might be that of "educating" a visitor who has minimum understanding of your product or service. Another might be a process of "motivating" a returning visitor to consider an upgrade or repurchase. A third process might be structured around items you market online. You'll have as many process pathways in your Web site as you have target audiences, products, and services. Each can be measured through Web analytics to determine how effective they are.

Traffic Sources

Your Web analytics program is an incredible tool for identifying where your Web traffic originates. Basic categories such as search engines, referral Web sites, and visits from bookmarked pages (i.e., direct) are compiled with little involvement by the marketer. With a little effort, however, you can also identify Web traffic that was generated by your various offline or online advertising campaigns.

1. *Referral Web sites.* Other Web sites that contain links that send visitors directly to your Web site are considered referral Web sites. Your analytics program will identify each referral site your traffic comes from, and a deeper analysis will help you determine which referrals produce the greatest volume, the highest conversions, the most new visitors, etc.

2. *Search engines.* Data in the search engine category is divided between paid search and organic (or natural) search. You can review the top keywords that generated Web traffic to your site and see if they are representative of your products and services. Depending upon your business, you might want to have hundreds (or thousands) of keywords that draw potential customers. Even the simplest product search can have multiple variations based on how the individual phrases the search query.

3. *Direct.* Direct searches are attributed to two sources. An individual who bookmarks one of your Web pages in their favorites and clicks that link will be recorded as a direct search. Another source occurs when someone types your URL directly into their browser. This happens when someone retrieves your URL from a business card, brochure, print ad, radio commercial, etc. That's why it's good strategy to use coded URLs.

4. *Offline campaigns.* If you utilize advertising options other than Web-based campaigns, your Web analytics program can capture performance data if you'll include a mechanism for sending them to your Web site. Typically, this is a dedicated URL that you include in your advertisement (i.e., "**www.mycompany.com/offer50**") that delivers those visitors to a specific landing page. You now have data on how many responded to that ad by visiting your Web site.

5. *Online campaigns.* If you are running a banner ad campaign, search engine advertising campaign, or even e-mail campaigns, you can measure individual campaign effectiveness by simply using a dedicated URL similar to the offline campaign strategy.

Visitor Profiles

One of the ways you can leverage your Web analytics into a really powerful marketing tool is through segmentation. By blending data from different analytics reports, you'll begin to see a variety of user profiles emerge.

1. *Keywords.* Within your analytics report, you can see what keywords visitors used in search engines to locate your Web site. If you aggregate your keywords by similar attributes, you'll begin to see distinct visitor groups that are using your Web site. For example, the particular search phrase that was used can indicate how well they understand your product or its benefits. If they use words that mirror your own product or service descriptions, then they probably are already aware of your offerings from effective advertisements, brochures, etc. If the terms are more general in nature, then your visitor is seeking a solution for a problem and has happened upon your Web site. If this second group of searchers is sizable, then you'll want to ensure that your site has a strong education component to convince them they've found their answer and then move them into your sales channel.

2. *Content groupings.* Depending upon how you group your content, you may be able to analyze sections of your Web site that correspond with specific products, services, campaigns, and other marketing tactics. If you conduct a lot of trade shows and drive traffic to your Web site for specific product literature, then your Web analytics will highlight the activity in that section.

3. *Geography.* Analytics permits you to see where your traffic geographically originates, including country, state, and city locations. This can be especially useful if you use geo-targeted campaigns or want to measure your visibility across a region.

4. *Time of day.* Web traffic generally has peaks at the beginning of the workday, during lunch, and toward the end of the workday. It's not unusual, however, to find strong Web traffic entering your Web site up until the late evening. You can analyze this data to determine when people browse versus buy and also make decisions on what hours you should offer customer service.

5. *Landing page profiles.* If you structure your various advertising campaigns properly, you can drive each of your targeted groups to a different landing page, which your Web analytics will capture and measure. By combining these numbers with the demographics of your campaign media, you can know what percentage of your visitors fit each demographic.

Conversion Statistics

Each organization will define a "conversion" according to its specific marketing objectives. Some Web analytics programs use the term "goal" to benchmark certain Web site objectives, whether that be a certain number of visitors to a page, a completed registration form, or an online purchase.

1. *New visitors.* If you're working to increase visibility, you'll want to study the trends in your new visitors data. Analytics identifies all visitors as either new or returning.

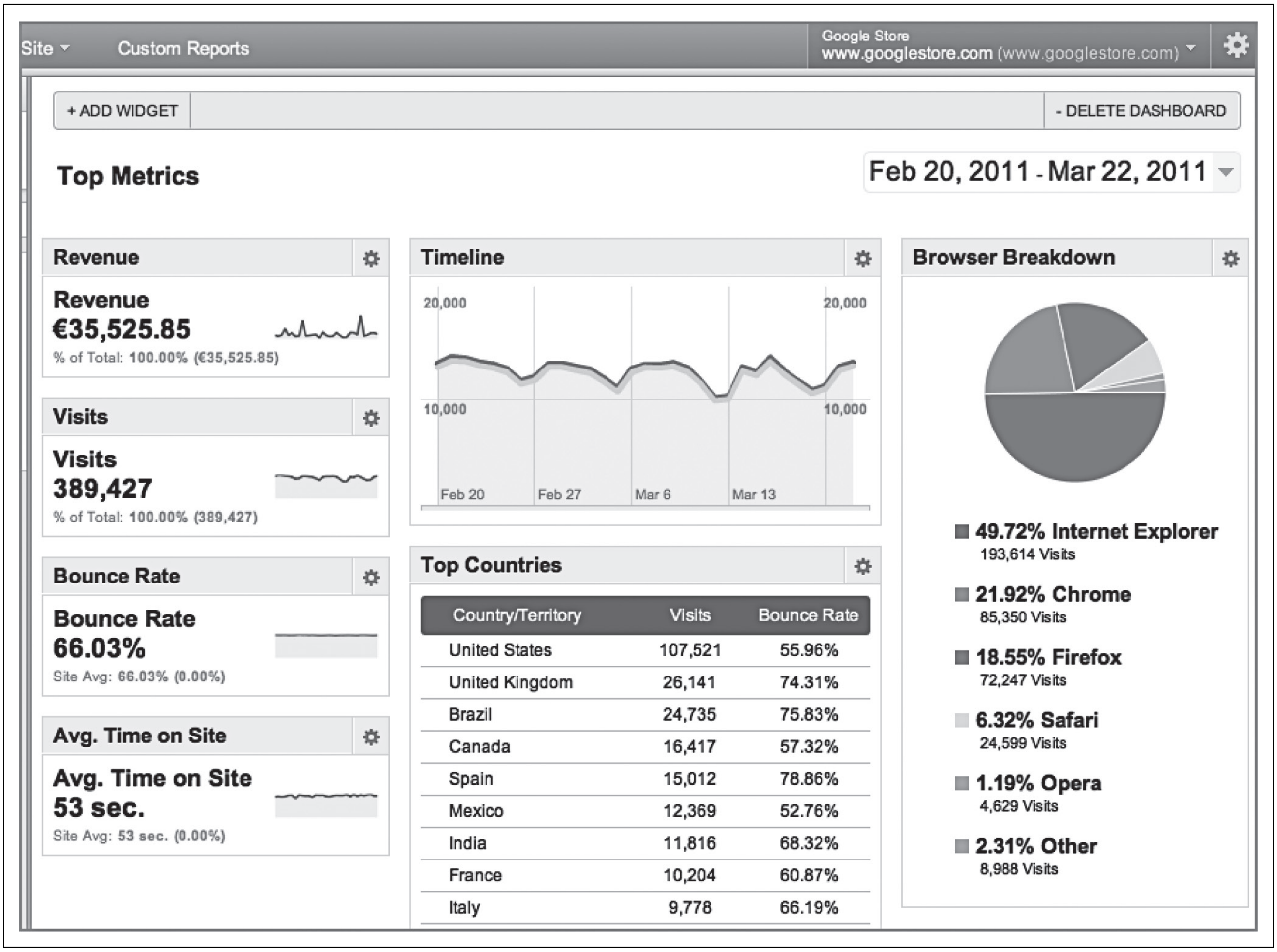
2. *Returning visitors.* If you're involved in loyalty programs or offer a product that has a long purchase cycle, then your returning visitors data will help you measure progress in this area.

3. *Leads.* Once a form is submitted and a thank-you page is generated, you have created a lead. Web analytics will permit you to calculate a completion rate (or abandonment rate) by dividing the number of completed forms by the number of Web visitors that came to your page. A low completion percentage would indicate a page that needs attention.

4. *Sales/conversions.* Depending upon the intent of your Web site, you can define a "sale" by an online purchase, a completed registration, an online submission, or any number of other Web activities. Monitoring these figures will alert you to any changes (or successes!) that occur further upstream.

5. *Abandonment/exit rates.* Just as important as those moving through your Web site are those who began a process and quit or came to your Web site and left after a page or two. In the first case, you'll want to analyze where the visitor terminated the process and whether there are a number of visitors quitting at the same place. Then investigate the situation for resolution. In the latter case, a high exit rate on a Web site or a specific page generally indicates an issue with expectations. Visitors click to your Web site based on some message contained in an advertisement, a presentation, etc., and expect some continuity in that message. Make sure you're advertising a message that your Web site can reinforce and deliver.

Within each of these items are metrics that can be established for your specific organization. You can create a weekly dashboard that includes specific numbers or percentages that will indicate where you're succeeding—or highlight a marketing challenge that should be addressed. When these metrics are evaluated consistently and used in conjunction with other available marketing data, they can lead you to a highly quantified marketing program. Figure 8.6 shows a Web analytics dashboard created with freely available Google Analytics tools.



Google Store
www.googlestore.com (www.googlestore.com)

Top Metrics

Feb 20, 2011 - Mar 22, 2011

Revenue

Revenue
€35,525.85
% of Total: 100.00% (€35,525.85)

Visits

Visits
389,427
% of Total: 100.00% (389,427)

Bounce Rate

Bounce Rate
66.03%
Site Avg: 66.03% (0.00%)

Avg. Time on Site

Avg. Time on Site
53 sec.
Site Avg: 53 sec. (0.00%)

Timeline

Top Countries

Country/Territory	Visits	Bounce Rate
United States	107,521	55.96%
United Kingdom	26,141	74.31%
Brazil	24,735	75.83%
Canada	16,417	57.32%
Spain	15,012	78.86%
Mexico	12,369	52.76%
India	11,816	68.32%
France	10,204	60.87%
Italy	9,778	66.19%

Browser Breakdown

- ■ 49.72% Internet Explorer
  193,614 Visits
- ■ 21.92% Chrome
  85,350 Visits
- ■ 18.55% Firefox
  72,247 Visits
- □ 6.32% Safari
  24,599 Visits
- ■ 1.19% Opera
  4,629 Visits
- ■ 2.31% Other
  8,988 Visits

FIGURE 8.6 A Sample Web Analytics Dashboard.

SECTION 8.6 REVIEW QUESTIONS

1. What are the three types of data generated through Web page visits?
2. What is clickstream analysis? What is it used for?
3. What are the main applications of Web mining?
4. What are commonly used Web analytics metrics? What is the importance of metrics?

8.7 WEB ANALYTICS MATURITY MODEL AND WEB ANALYTICS TOOLS

The term "maturity" relates to the degree of proficiency, formality, and optimization of business models, moving "ad hoc" practices to formally defined steps and optimal business processes. A maturity model is a formal depiction of critical dimensions and their competency levels of a business practice. Collectively, these dimensions and levels define the maturity level of an organization in that area of practice. It often describes an evolutionary improvement path from ad hoc, immature practices to disciplined, mature processes with improved quality and efficiency.

A good example of maturity models is the BI Maturity Model developed by The Data Warehouse Institute (TDWI). In the TDWI BI Maturity Model the main purpose was to gauge where organization data warehousing initiatives are at a point in time and where it should go next. It was represented in a six-stage framework (Management Reporting ➔ Spreadmarts ➔ Data Marts ➔ Data Warehouse ➔ Enterprise Data Warehouse ➔ BI Services). Another related example is the simple business analytics maturity model, moving from simple descriptive measures to predicting future outcomes, to obtaining sophisticated decision systems (i.e., Descriptive Analytics ➔ Predictive Analytics ➔ Prescriptive Analytics).

For Web analytics perhaps the most comprehensive model was proposed by Stéphane Hamel (2009). In this model, Hamel used six dimensions—(1) Management, Governance and Adoption, (2) Objectives Definition, (3) Scoping, (4) The Analytics Team and Expertise, (5) The Continuous Improvement Process and Analysis Methodology, (6) Tools, Technology and Data Integration—and for each dimension he used six levels of proficiency/competence. Figure 8.7 shows Hamel's six dimensions and the respective proficiency levels.

The proficiency/competence levels have different terms/labels for each of the six dimensions, describing specifically what each level means. Essentially, the six levels are indications of analytical maturity ranging from "0–Analytically Impaired" to "5–Analytical Competitor." A short description of each of the six levels of competencies is given here (Hamel, 2009):

1. *Impaired:* Characterized by the use of out-of-the-box tools and reports; limited resources lacking formal training (hands-on skills) and education (knowledge). Web analytics is used on an ad hoc basis and is of limited value and scope. Some tactical objectives are defined, but results are not well communicated and there are multiple versions of the truth.

2. *Initiated:* Works with metrics to optimize specific areas of the business (such as marketing or the e-commerce catalogue). Resources are still limited, but the process is getting streamlined. Results are communicated to various business stakeholders (often director level). However, Web analytics might be supporting obsolete business processes and, thus, be limited in the ability to push for optimization beyond the online channel. Success is mostly anecdotal.

3. *Operational:* Key performance indicators and dashboards are defined and aligned with strategic business objectives. A multidisciplinary team is in place and uses various sources of information such as competitive data, voice of customer, and social media or mobile analysis. Metrics are exploited and explored through segmentation and multivariate testing. The Internet channel is being optimized; personas are being defined.

Results start to appear and be considered at the executive level. Results are centrally driven, but broadly distributed.

4. ***Integrated:*** Analysts can now correlate online and offline data from various sources to provide a near 360-degree view of the whole value chain. Optimization encompasses complete processes, including back-end and front-end. Online activities are defined from the user perspective and persuasion scenarios are defined. A continuous improvement process and problem-solving methodologies are prevalent. Insight and recommendations reach the CXO level.

5. ***Competitor:*** This level is characterized by several attributes of companies with a strong analytical culture (Davenport and Harris, 2007):

a. One or more senior executives strongly advocate fact-based decision making and analytics
b. Widespread use of not just descriptive statistics, but predictive modeling and complex optimization techniques
c. Substantial use of analytics across multiple business functions or processes
d. Movement toward an enterprise-level approach to managing analytical tools, data, and organizational skills and capabilities.

6. ***Addicted:*** This level matches Davenport's "Analytical Competitor" characteristics: deep strategic insight, continuous improvement, integrated, skilled resources, top management commitment, fact-based culture, continuous testing, learning, and most important: far beyond the boundaries of the online channel.

In Figure 8.7, one can mark the level of proficiency in each of the six dimensions to create their organization's maturity model (which would look like a spider diagram).

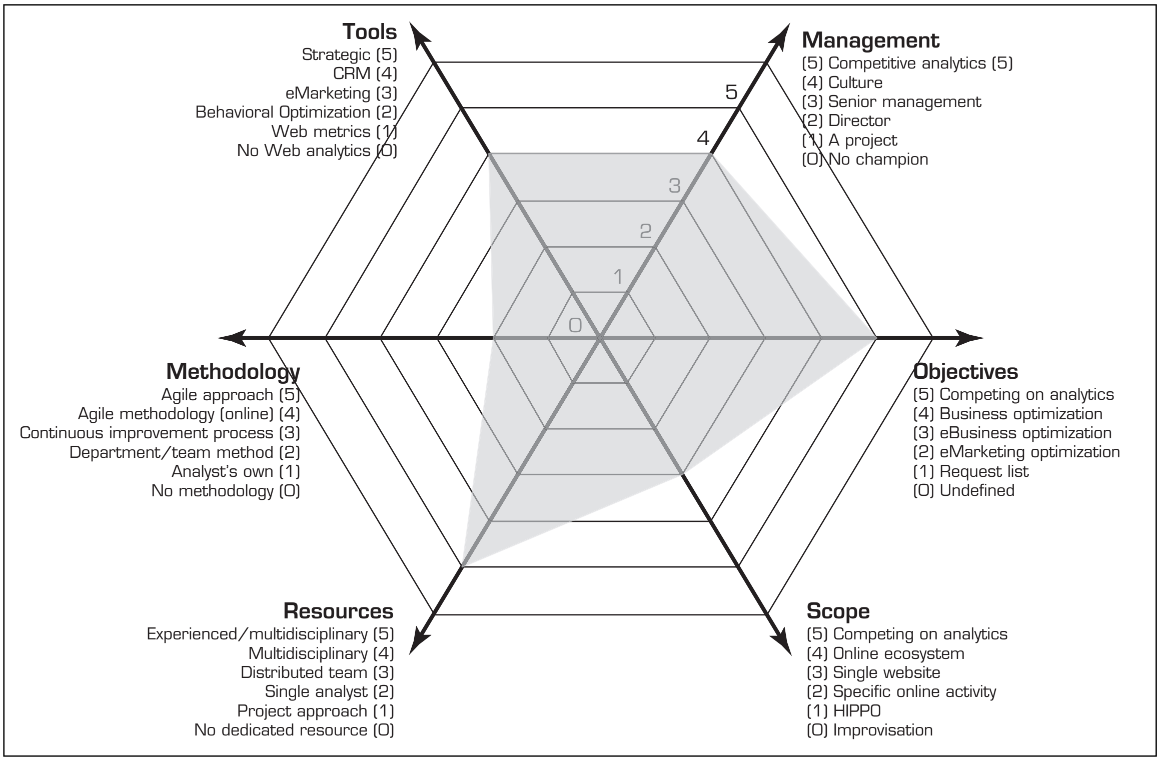


FIGURE 8.7 A Framework for Web Analytics Maturity Model.

Such an assessment can help organizations better understand at what dimensions they are lagging behind, and take corrective actions to mitigate it.

Web Analytics Tools

There are plenty of Web analytics applications (downloadable software tools and Web-based/on-demand service platforms) in the market. Companies (large, medium, or small) are creating products and services to grab their fair share from the emerging Web analytics marketplace. What is the most interesting is that many of the most popular Web analytics tools are free—yes, free to download and use for whatever reasons, commercial or nonprofit. The following are among the most popular free (or almost free) Web analytics tools:

GOOGLE WEB ANALYTICS (GOOGLE.COM/ANALYTICS) This is a service offered by Google that generates detailed statistics about a Web site's traffic and traffic sources and measures conversions and sales. The product is aimed at marketers as opposed to the Webmasters and technologists from which the industry of Web analytics originally grew. It is the most widely used Web analytics service. Even though the basic service is free of charge, the premium version is available for a fee.

YAHOO! WEB ANALYTICS (WEB.ANALYTICS.YAHOO.COM) Yahoo! Web analytics is Yahoo!'s alternative to the dominant Google Analytics. It's an enterprise-level, robust Web-based third-party solution that makes accessing data easy, especially for multiple-user groups. It's got all the things you'd expect from a comprehensive Web analytics tool, such as pretty graphs, custom-designed (and printable) reports, and real-time data tracking.

OPEN WEB ANALYTICS (OPENWEBANALYTICS.COM) Open Web Analytics (OWA) is a popular open source Web analytics software that anyone can use to track and analyze how people use Web sites and applications. OWA is licensed under GPL and provides Web site owners and developers with easy ways to add Web analytics to their sites using simple Javascript, PHP, or REST-based APIs. OWA also comes with built-in support for tracking Web sites made with popular content management frameworks such as WordPress and MediaWiki.

PIWIK (PIWIK.ORG) Piwik is the one of the leading self-hosted, decentralized, open source Web analytics platforms, used by 460,000 Web sites in 150 countries. Piwik was founded by Matthieu Aubry in 2007. Over the last 6 years, more talented and passionate members of the community have joined the team. As is the case in many open source initiatives, they are actively looking for new developers, designers, datavis architects, and sponsors to join them.

FIRESTAT (FIRESTATS.CC) FireStats is a simple and straightforward Web analytics application written in PHP/MySQL. It supports numerous platforms and set-ups including C# sites, Django sites, Drupal, Joomla!, WordPress, and several others. FireStats has an intuitive API that assists developers in creating their own custom apps or publishing platform components.

SITE METER (SITEMETER.COM) Site Meter is a service that provides counter and tracking information for Web sites. By logging IP addresses and using JavaScript or HTML to track visitor information, Site Meter provides Web site owners with information about their visitors, including how they reached the site, the date and time of their visit, and more.

WOOPRA (WOOPRA.COM) Woopra is a real-time customer analytics service that provides solutions for sales, service, marketing, and product teams. The platform is designed to help organizations optimize the customer life cycle by delivering live, granular behavioral data for individual Web site visitors and customers. It ties this individual-level data to aggregate analytics reports for a full life-cycle view that bridges departmental gaps.

AWSTATS (AWSTATS.ORG) AWStats is an open source Web analytics reporting tool, suitable for analyzing data from Internet services such as Web, streaming media, mail, and FTP servers. AWStats parses and analyzes server log files, producing HTML reports. Data is visually presented within reports by tables and bar graphs. Static reports can be created through a command line interface, and on-demand reporting is supported through a Web browser CGI program.

SNOOP (REINVIGORATE.NET) Snoop is a desktop-based application that runs on the Mac OS X and Windows XP/Vista platforms. It sits nicely on your system status bar/system tray, notifying you with audible sounds whenever something happens. Another outstanding Snoop feature is the Name Tags option, which allows you to "tag" visitors for easier identification. So when Joe over at the accounting department visits your site, you'll instantly know.

MOCHIBOT (MOCHIBOT.COM) MochiBot is a free Web analytics/tracking tool especially designed for Flash assets. With MochiBot, you can see who's sharing your Flash content, how many times people view your content, as well as help you track where your Flash content is to prevent piracy and content theft. Installing MochiBot is a breeze; you simply copy a few lines of ActionScript code in the .FLA files you want to monitor.

In addition to these free Web analytics tools, Table 8.1 provides a list of commercially available Web analytics tools.

TABLE 8.1 Commercial Web Analytics Software Tools

Product Name	Description	URL
Angoss Knowledge WebMiner	Combines ANGOSS Knowledge STUDIO and clickstream analysis	**angoss.com**
ClickTracks	Visitor patterns can be shown on Web site	**clicktracks.com, now at Lyris.com**
LiveStats from DeepMetrix	Real-time log analysis, live demo on site	**deepmetrix.com**
Megaputer WebAnalyst	Data and text mining capabilities	**megaputer.com/site/textanalyst.php**
MicroStrategy Web Traffic Analysis Module	Traffic highlights, content analysis, and Web visitor analysis reports	**microstrategy.com/Solutions/Applications/WTAM**
SAS Web Analytics	Analyzes Web site traffic	**sas.com/solutions/webanalytics**
SPSS Web Mining for Clementine	Extraction of Web events	**www-01.ibm.com/software/analytics/spss/**
WebTrends	Data mining of Web traffic information.	**webtrends.com**
XML Miner	A system and class library for mining data and text expressed in XML, using fuzzy logic expert system rules	**scientio.com**

Putting It All Together—A Web Site Optimization Ecosystem

It seems that just about everything on the Web can be measured—every click can be recorded, every view can be captured, and every visit can be analyzed—all in an effort to continually and automatically optimize the online experience. Unfortunately, the notions of "infinite measurability" and "automatic optimization" in the online channel are far more complex than most realize. The assumption that any single application of Web mining techniques will provide the necessary range of insights required to understand Web site visitor behavior is deceptive and potentially risky. Ideally, a holistic view to customer experience is needed that can only be captured using both quantitative and qualitative data. Forward-thinking companies have already taken steps toward capturing and analyzing a holistic view of the customer experience, which has led to significant gains, both in terms of incremental financial growth and increasing customer loyalty and satisfaction.

According to Peterson (2008), the inputs for Web site optimization efforts can be classified along two axes describing the nature of the data and how that data can be used. On one axis are data and information—data being primarily quantitative and information being primarily qualitative. On the other axis are measures and actions—measures being reports, analysis, and recommendations all designed to drive actions, the actual changes being made in the ongoing process of site and marketing optimization. Each quadrant created by these dimensions leverages different technologies and creates different outputs, but much like a biological ecosystem, each technological niche interacts with the others to support the entire online environment (see Figure 8.8).

Most believe that the Web site optimization ecosystem is defined by the ability to log, parse, and report on the clickstream behavior of site visitors. The underlying technology of this ability is generally referred to as *Web analytics*. Although Web analytics

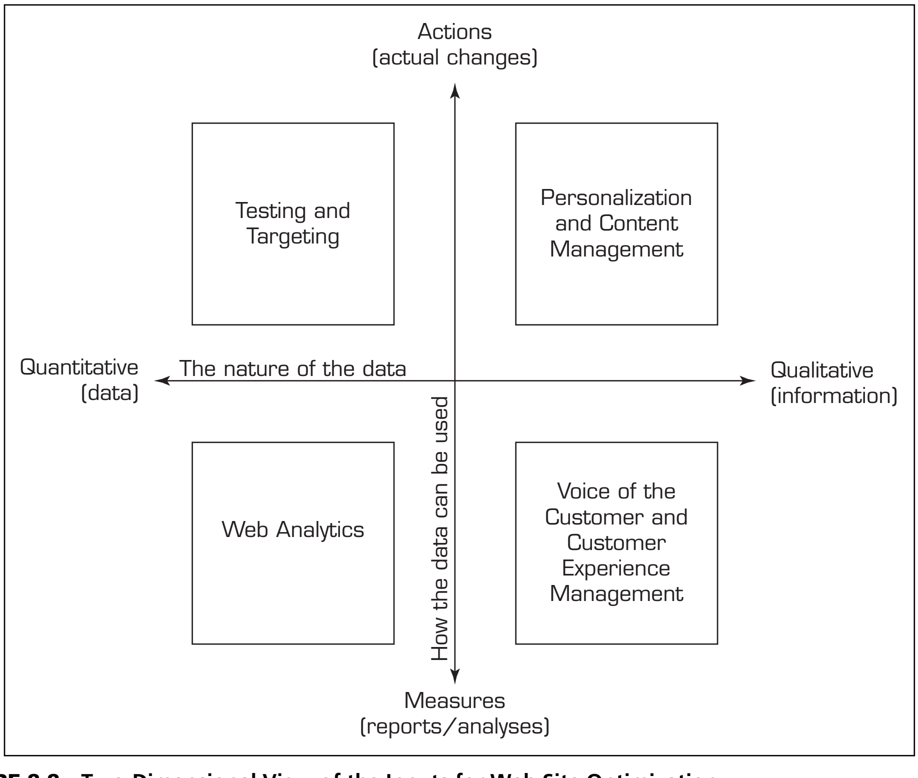


FIGURE 8.8 Two-Dimensional View of the Inputs for Web Site Optimization.

tools provide invaluable insights, understanding visitor behavior is as much a function of qualitatively determining interests and intent as it is quantifying clicks from page to page. Fortunately there are two other classes of applications designed to provide a more qualitative view of online visitor behavior designed to report on the overall user experience and report direct feedback given by visitors and customers: **customer experience management (CEM)** and **voice of customer (VOC)**:

- Web analytics applications focus on "where and when" questions by aggregating, mining, and visualizing large volumes of data, by reporting on online marketing and visitor acquisition efforts, by summarizing page-level visitor interaction data, and by summarizing visitor flow through defined multistep processes.
- Voice of customer applications focus on "who and how" questions by gathering and reporting direct feedback from site visitors, by benchmarking against other sites and offline channels, and by supporting predictive modeling of future visitor behavior.
- Customer experience management applications focus on "what and why" questions by detecting Web application issues and problems, by tracking and resolving business process and usability obstacles, by reporting on site performance and availability, by enabling real-time alerting and monitoring, and by supporting deep diagnosis of observed visitor behavior.

All three applications are needed to have a complete view of the visitor behavior where each application plays a distinct and valuable role. Web analytics, CEM, and VOC applications form the foundation of the Web site optimization ecosystem that supports the online business's ability to positively influence desired outcomes (a pictorial representation of this process view of the Web site optimization ecosystem is given in Figure 8.9). These similar-yet-distinct applications each contribute to a site operator's ability to recognize, react, and respond to the ongoing challenges faced by every Web site owner. Fundamental to the optimization process is measurement, gathering data and information that can then be transformed into tangible analysis, and recommendations for improvement using Web mining tools and techniques. When used properly, these applications allow for convergent validation—combining different sets of data collected for the same audience to provide a richer and deeper understanding of audience behavior. The convergent validation model—one

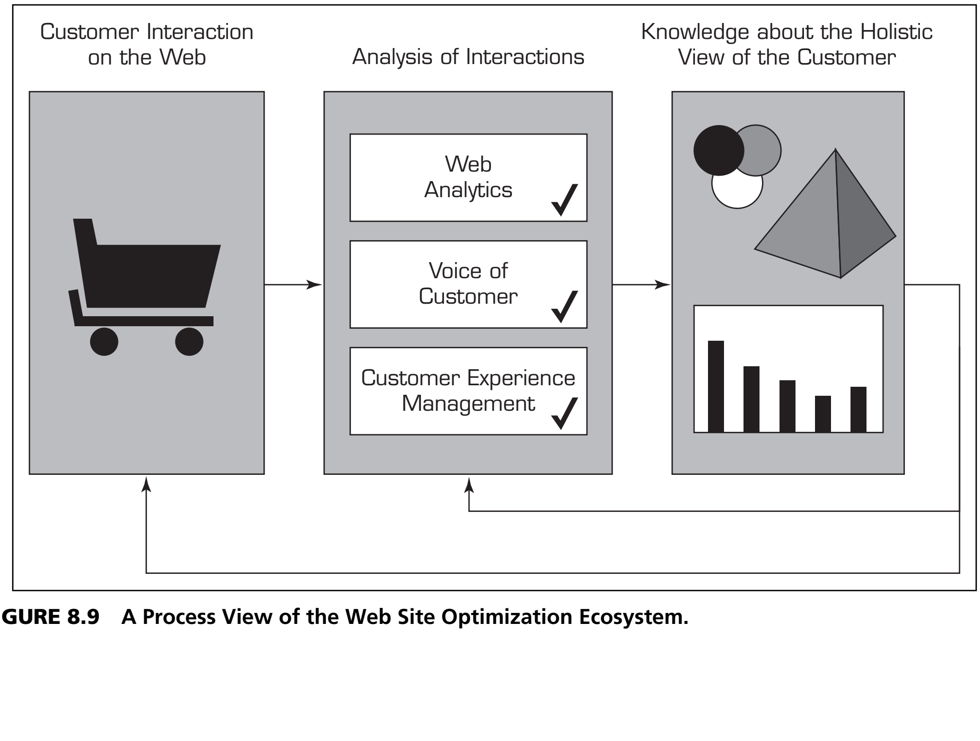

Customer Interaction on the Web
Analysis of Interactions
Knowledge about the Holistic View of the Customer
Web Analytics ✓
Voice of Customer ✓
Customer Experience Management ✓

FIGURE 8.9 A Process View of the Web Site Optimization Ecosystem.

where multiple sources of data describing the same population are integrated to increase the depth and richness of the resulting analysis—forms the framework of the Web site optimization ecosystem. On one side of the spectrum are the primarily qualitative inputs from VOC applications; on the other side are the primarily quantitative inputs from CEM bridging the gap by supporting key elements of data discovery. When properly implemented, all three systems sample data from the same audience. The combination of these data—either through data integration projects or simply via the process of conducting good analysis—supports far more actionable insights than any of the ecosystem members individually.

A Framework for Voice of the Customer Strategy

Voice of the customer (VOC) is a term usually used to describe the analytic process of capturing a customer's expectations, preferences, and aversions. It essentially is a market research technique that produces a detailed set of customer wants and needs, organized into a hierarchical structure, and then prioritized in terms of relative importance and satisfaction with current alternatives. Attensity, one of the innovative service providers in the analytics marketplace, developed an intuitive framework for VOC strategy that they called LARA, which stands for Listen, Analyze, Relate, and Act. It is a methodology that outlines a process by which organizations can take user-generated content (UGC), whether generated by consumers talking in Web forums, on micro-blogging sites like Twitter and social networks like Facebook, or in feedback surveys, e-mails, documents, research, etc., and using it as a business asset in a business process. Figure 8.10 shows a pictorial depiction of this framework.

LISTEN To "listen" is actually a process in itself that encompasses both the capability to listen to the open Web (forums, blogs, tweets, you name it) and the capability to seamlessly access enterprise information (CRM notes, documents, e-mails, etc.). It takes a listening post, deep federated search capabilities, scraping and enterprise class data integration, and a strategy to determine who and what you want to listen to.

ANALYZE This is the hard part. How can you take all of this mass of unstructured data and make sense of it? This is where the "secret sauce" of text analytics comes into play. Look for solutions that include keyword, statistical, and natural language approaches

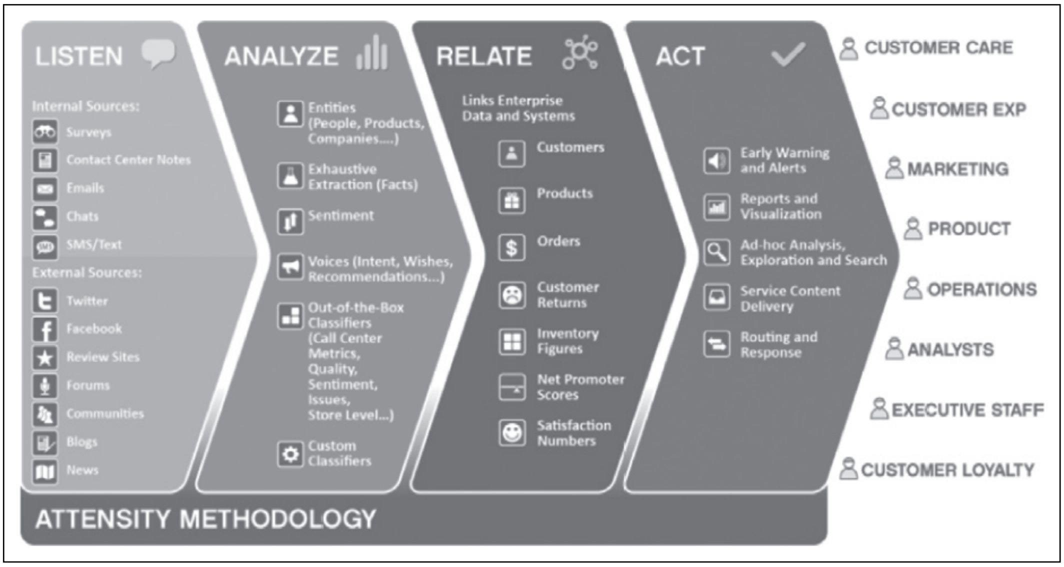


FIGURE 8.10 **Voice of the Customer Strategy Framework.** *Source:* **Attensity.com.** Used with permission.

that will allow you to essentially tag or barcode every word and the relationships among words, making it data that can be accessed, searched, routed, counted, analyzed, charted, reported on, and even reused. Keep in mind that, in addition to technical capabilities, it has to be easy to use, so that your business users can focus on the insights, not the technology. It should have an engine that doesn't require the user to define keywords or terms that they want their system to look for or include in a rule base. Rather, it should *automatically* identify terms ("facts," people, places, things, etc.) *and their relationships with other terms or combinations of terms*—making it easy to use, maintain, and also be more accurate, so you can rely on the insights as actionable.

RELATE Now that you have found the insights and can analyze the unstructured data, the real value comes when you can connect those insights to your "structured" data: your customers (which customer segment is complaining about your product most?); your products (which product is having the issue?); your parts (is there a problem with a specific part manufactured by a specific partner?); your locations (is the customer who is tweeting about wanting a sandwich near your nearest restaurant?); and so on. Now you can ask questions of your data and get deep, actionable insights.

ACT Here is where it gets exciting, and your business strategy and rules are critical. What do you do with the new customer insight you've obtained? How do you leverage the problem resolution content created by a customer that you just identified? How do you connect with a customer who is uncovering issues that are important to your business or who is asking for help? How do you route the insights to the right people? And, how do you engage with customers, partners, and influencers once you understand what they are saying? You understand it; now you've got to act.

SECTION 8.7 REVIEW QUESTIONS

1. What is a maturity model?
2. List and comment on the six stages of TDWI's BI maturity framework.
3. What are the six dimensions used in Hamel's Web analytics maturity model?
4. Describe Attensity's framework for VOC strategy. List and describe the four stages.

8.8 SOCIAL ANALYTICS AND SOCIAL NETWORK ANALYSIS

Social analytics may mean different things to different people, based on their worldview and field of study. For instance, the dictionary definition of social analytics refers to a philosophical perspective developed by the Danish historian and philosopher Lars-Henrik Schmidt in the 1980s. The theoretical object of the perspective is *socius*, a kind of "commonness" that is neither a universal account nor a communality shared by every member of a body (Schmidt, 1996). Thus, social analytics differs from traditional philosophy as well as sociology. It might be viewed as a perspective that attempts to articulate the contentions between philosophy and sociology.

Our definition of social analytics is somewhat different; as opposed to focusing on the "social" part (as is the csae in its philosophical definition), we are more interested in the "analytics" part of the term. Gartner defined social analytics as "monitoring, analyzing, measuring and interpreting digital interactions and relationships of people, topics, ideas and content." Social analytics include mining the textual content created in social media (e.g., sentiment analysis, natural language processing) and analyzing socially established networks (e.g., influencer identification, profiling, prediction) for the purpose of gaining insight about existing and potential customers' current and future behaviors, and about the likes and dislikes toward a firm's products and services. Based on this definition and the current practices, social analytics can be classified into two different, but not necessarily mutually exclusive, branches: social network analysis and social media analytics.

Social Network Analysis

A **social network** is a social structure composed of individuals/people (or groups of individuals or organizations) linked to one another with some type of connections/relationships. The social network perspective provides a holistic approach to analyzing structure and dynamics of social entities. The study of these structures uses social network analysis to identify local and global patterns, locate influential entities, and examine network dynamics. Social networks and the analysis of them is essentially an interdisciplinary field that emerged from social psychology, sociology, statistics, and graph theory. Development and formalization of the mathematical extent of social network analysis dates back to the 1950s; the development of foundational theories and methods of social networks dates back to the 1980s (Scott and Davis, 2003). Social network analysis is now one of the major paradigms in business analytics, consumer intelligence, and contemporary sociology, and is also employed in a number of other social and formal sciences.

A social network is a theoretical construct useful in the social sciences to study relationships between individuals, groups, organizations, or even entire societies (social units). The term is used to describe a social structure determined by such interactions. The ties through which any given social unit connects represent the convergence of the various social contacts of that unit. In general, social networks are self-organizing, emergent, and complex, such that a globally coherent pattern appears from the local interaction of the elements (individuals and groups of individuals) that make up the system.

Following are a few typical social network types that are relevant to business activities.

COMMUNICATION NETWORKS Communication studies are often considered a part of both the social sciences and the humanities, drawing heavily on fields such as sociology, psychology, anthropology, information science, biology, political science, and economics. Many communications concepts describe the transfer of information from one source to another, and thus can be represented as a social network. Telecommunication companies are tapping into this rich information source to optimize their business practices and to improve customer relationships.

COMMUNITY NETWORKS Traditionally, community referred to a specific geographic location, and studies of community ties had to do with who talked, associated, traded, and attended social activities with whom. Today, however, there are extended "online" communities developed through social networking tools and telecommunications devices. Such tools and devices continuously generate large amounts of data, which can be used by companies to discover invaluable, actionable information.

CRIMINAL NETWORKS In criminology and urban sociology, much attention has been paid to the social networks among criminal actors. For example, studying gang murders and other illegal activities as a series of exchanges between gangs can lead to better understanding and prevention of such criminal activities. Now that we live in a highly connected world (thanks to the Internet), many of the criminal networks' formations and their activities are being watched/pursued by security agencies using state-of-the-art Internet tools and tactics. Even though the Internet has changed the landscape for criminal networks and law enforcement agencies, the traditional social and philosophical theories still apply to a large extent.

INNOVATION NETWORKS Business studies on diffusion of ideas and innovations in a network environment focus on the spread and use of ideas among the members of the social network. The idea is to understand why some networks are more innovative, and why some communities are early adopters of ideas and innovations (i.e., examining the impact of social network structure on influencing the spread of an innovation and innovative behavior).

Social Network Analysis Metrics

Social network analysis (SNA) is the systematic examination of social networks. Social network analysis views social relationships in terms of network theory, consisting of nodes (representing individuals or organizations within the network) and ties/connections (which represent relationships between the individuals or organizations, such as friendship, kinship, organizational position, etc.). These networks are often represented using social network diagrams, where nodes are represented as points and ties are represented as lines. Application Case 8.5 gets into the details of how SNA can be used to help telecommunication companies.

Over the years, various metrics (or measurements) have been developed to analyze social network structures from different perspectives. These metrics are often grouped into three categories: connections, distributions, and segmentation.

Application Case 8.5

Social Network Analysis Helps Telecommunication Firms

Because of the widespread use of free Internet tools and techniques (VoIP, video conferencing tools such as Skype, free phone calls within the United States by Google Voice, etc.), the telecommunication industry is going through a tough time. In order to stay viable and competitive, they need to make the right decisions and utilize their limited resources optimally. One of the key success factors for telecom companies is to maximize their profitability by listening and understanding the needs and wants of the customers, offering communication plans, prices, and features that they want at the prices that they are willing to pay.

These market pressures force telecommunication companies to be more innovative. As we all know, "necessity is the mother of invention." Therefore, many of the most promising use cases for social network analysis (SNA) are coming from the telecommunication companies. Using detailed call records that are already in their databases, they are trying to identify social networks and influencers. In order to identify the social networks, they are asking questions like "Who contacts whom?" "How often?" "How long?" "Both directions?" "On Net, off Net?" They are also trying to answer questions that lead to identification of influencers, such as "Who influenced whom how much on purchases?" "Who influences whom how much on churn?" and "Who will acquire others?" SNA metrics like degree (how many people are directly in a person's social network), density (how dense is the calling pattern within the calling circle), betweenness (how essential you are to facilitate communication within your calling circle), and centrality (how "important" you are in the social network) are often used answer these questions.

Here are some of the benefits that can be obtained from SNA:

- Manage customer churn
  - Reactive (reduce collateral churn)—Identify subscribers whose loyalty is threatened by churn around them.
  - Preventive (reduce influential churn)—Identify subscribers who, should they churn, would take a few friends with them.
- Improve cross-sell and technology transfer
  - – Reactive (leverage collateral adoption)—Identify subscribers whose affinity for products is increased due to adoption around them and stimulate them.
  - – Proactive (identify influencers for this adoption)—Identify subscribers who, should they adopt, would push a few friends to do the same.

(*Continued*)

Application Case 8.5 (Continued)

- Manage viral campaigns—Understand what leads to high-scale spread of messages about products and services, and use this information to your benefit.
- Improve acquisition—Identify who are most likely to recommend a (off-Net) friend to become a new subscriber of the operator. The recommendation itself, as well as the subscription, is incentivized for both the subscriber and the recommending person.
- Identify households, communities, and close-groups to better manage your relationships with them.
- Identify customer life-stages—Identifying social network changes and from there identifying life-stage changes such as moving, changing a job, going to a university, starting a relationships, getting married, etc.
- Identifying pre-churners—Detecting potential churners during the process of leaving and motivating them to stay with you.
- Gain competitor insights—Track dynamic changes in social networks based on competitor's marketing activities
- Others inducing identifying rotational churners (switching between operators)—Facilitating re- to postmigration, and tracking customer's networks dynamics over his/her life cycle.

Actual cases indicate that proper implementation of SNA can significantly lower churn, improve cross-sell, boost new customer acquisition, optimize pricing and, hence, maximize profit, and improve overall competitiveness.

Questions for Discussion

1. How can social network analysis be used in the telecommunications industry?
2. What do you think are the key challenges, potential solution, and probable results in applying SNA in telecommunications firms?

Source: Compiled from "More Things We Love About SNA: Return of the Magnificent 10," February 2013, presentation by Judy Bayer and Fawad Qureshi, Teradata.

Connections

Homophily: The extent to which actors form ties with similar versus dissimilar others. Similarity can be defined by gender, race, age, occupation, educational achievement, status, values, or any other salient characteristic.

Multiplexity: The number of content-forms contained in a tie. For example, two people who are friends and also work together would have a multiplexity of 2. Multiplexity has been associated with relationship strength.

Mutuality/reciprocity: The extent to which two actors reciprocate each other's friendship or other interaction.

Network closure: A measure of the completeness of relational triads. An individual's assumption of network closure (i.e., that their friends are also friends) is called *transitivity*. Transitivity is an outcome of the individual or situational trait of need for cognitive closure.

Propinquity: The tendency for actors to have more ties with geographically close others.

Distributions

Bridge: An individual whose weak ties fill a structural hole, providing the only link between two individuals or clusters. It also includes the shortest route when a longer one is unfeasible due to a high risk of message distortion or delivery failure.

Centrality: Refers to a group of metrics that aim to quantify the importance or influence (in a variety of senses) of a particular node (or group) within a network. Examples of common methods of measuring centrality include betweenness centrality, closeness centrality, eigenvector centrality, alpha centrality, and degree centrality.

Density: The proportion of direct ties in a network relative to the total number possible.

Distance: The minimum number of ties required to connect two particular actors.

Structural holes: The absence of ties between two parts of a network. Finding and exploiting a structural hole can give an entrepreneur a competitive advantage. This concept was developed by sociologist Ronald Burt and is sometimes referred to as an alternate conception of social capital.

Tie strength: Defined by the linear combination of time, emotional intensity, intimacy, and reciprocity (i.e., mutuality). Strong ties are associated with homophily, propinquity, and transitivity, while weak ties are associated with bridges.

Segmentation

Cliques and social circles: Groups are identified as *cliques* if every individual is directly tied to every other individual or *social circles* if there is less stringency of direct contact, which is imprecise, or as structurally cohesive blocks if precision is wanted.

Clustering coefficient: A measure of the likelihood that two members of a node are associates. A higher clustering coefficient indicates a greater *cliquishness*.

Cohesion: The degree to which actors are connected directly to each other by cohesive bonds. Structural cohesion refers to the minimum number of members who, if removed from a group, would disconnect the group.

SECTION 8.8 REVIEW QUESTIONS

1. What is meant by social analytics? Why is it an important business topic?
2. What is a social network? What is social network analysis?
3. List and briefly describe the most common social network types.
4. List and briefly describe the social network analysis metrics.

8.9 SOCIAL MEDIA DEFINITIONS AND CONCEPTS

Social media refers to the enabling technologies of social interactions among people in which they create, share, and exchange information, ideas, and opinions in virtual communities and networks. It is a group of Internet-based software applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content (Kaplan and Haenlein, 2010). Social media depends on mobile and other Web-based technologies to create highly interactive platforms for individuals and communities to share, co-create, discuss, and modify user-generated content. It introduces substantial changes to communication between organizations, communities, and individuals.

Since their emergence in the early 1990s, Web-based social media technologies have seen a significant improvement in both quality and quantity. These technologies take on many different forms, including online magazines, Internet forums, Web logs, social blogs, microblogging, wikis, social networks, podcasts, pictures, video, and product/service evaluations/ratings. By applying a set of theories in the field of media research (social presence, media richness) and social processes (self-presentation, self-disclosure), Kaplan and Haenlein (2010) created a classification scheme with six different types of social media: collaborative projects (e.g., Wikipedia), blogs and microblogs (e.g., Twitter), content communities (e.g., YouTube), social networking sites (e.g., Facebook), virtual game worlds (e.g., World of Warcraft), and virtual social worlds (e.g., Second Life).

Web-based social media are different from traditional/industrial media, such as newspapers, television, and film, as they are comparatively inexpensive and accessible to enable anyone (even private individuals) to publish or access/consume information. Industrial media generally require significant resources to publish information, as in most cases the articles (or books) go through many revisions before being published (as was the case in the publication of this very book). Here are some of the most prevailing characteristics that help differentiate between social and industrial media (Morgan et al., 2010):

Quality: In industrial publishing—mediated by a publisher—the typical range of quality is substantially narrower than in niche, unmediated markets. The main challenge posed by content in social media sites is the fact that the distribution of quality has high variance: from very high-quality items to low-quality, sometimes abusive, content.

Reach: Both industrial and social media technologies provide scale and are capable of reaching a global audience. Industrial media, however, typically use a centralized framework for organization, production, and dissemination, whereas social media are by their very nature more decentralized, less hierarchical, and distinguished by multiple points of production and utility.

Frequency: Compared to industrial media, updating and reposting on social media platforms is easier, faster, and cheaper, and therefore practiced more frequently, resulting in fresher content.

Accessibility: The means of production for industrial media are typically government and/or corporate (privately owned), and are costly, whereas social media tools are generally available to the public at little or no cost.

Usability: Industrial media production typically requires specialized skills and training. Conversely, most social media production requires only modest reinterpretation of existing skills; in theory, anyone with access can operate the means of social media production.

Immediacy: The time lag between communications produced by industrial media can be long (weeks, months, or even years) compared to social media (which can be capable of virtually instantaneous responses).

Updatability: Industrial media, once created, cannot be altered (once a magazine article is printed and distributed, changes cannot be made to that same article), whereas social media can be altered almost instantaneously by comments or editing.

How Do People Use Social Media?

Not only are the numbers on social networking sites growing, but so is the degree to which they are engaged with the channel. Brogan and Bastone (2011) presented research results that stratify users according to how actively they use social media and tracked evolution of these user segments over time. They listed six different engagement levels (Figure 8.11).

According to the research results, the online user community has been steadily migrating upwards on this engagement hierarchy. The most notable change is among Inactives. Forty-four percent of the online population fell into this category. Two years

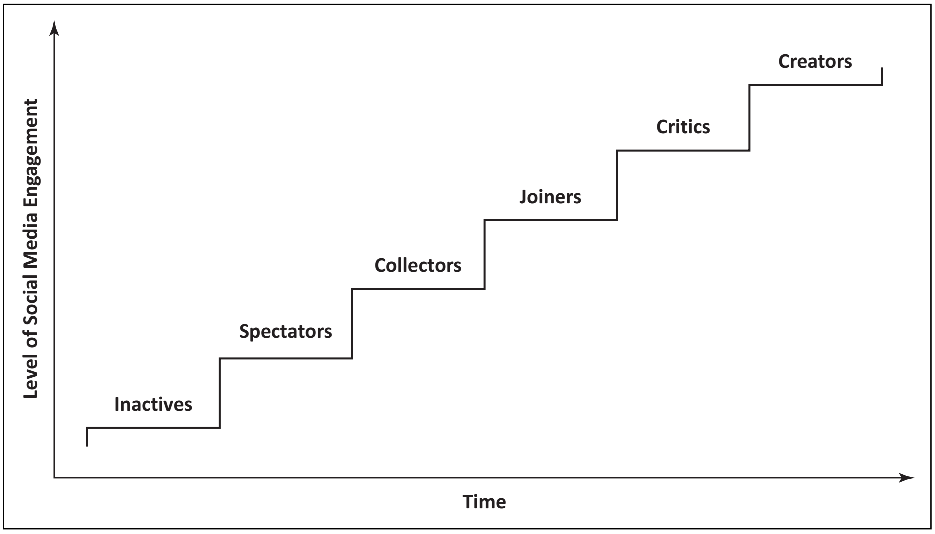


FIGURE 8.11 **Evolution of Social Media User Engagement.**

later, more than half of those Inactives had jumped into social media in some form or another. "Now roughly 82 percent of the adult population online is in one of the upper categories," said Bastone. "Social media has truly reached a state of mass adoption."

Application Case 8.6 shows the positive impact of social media at Lollapalooza.

Application Case 8.6

Measuring the Impact of Social Media at Lollapalooza

C3 Presents creates, books, markets, and produces live experiences, concerts, events, and just about anything that makes people stand up and cheer. Among others, they produce the Austin City Limits Music Festival, Lollapalooza, as well as more than 800 shows nationwide. They hope to see you up in front sometime.

An early adopter of social media as a way to drive event attendance, Lollapalooza organizer C3 Presents needed to know the impact of its social media efforts. They came to Cardinal Path for a social media measurement strategy and ended up with some startling insights.

The Challenge

When the Lollapalooza music festival decided to incorporate social media into their online marketing strategy, they did it with a bang. Using Facebook, MySpace, Twitter, and more, the Lollapalooza Web site was a first mover in allowing its users to engage and share through social channels that were integrated into the site itself.

After investing the time and resources in building out these integrations and their functionality, C3 wanted to know one simple thing: "Did it work?" To answer this, C3 Presents needed a measurement strategy that would provide a wealth of information about their social media implementation, such as:

- Which fans are using social media and sharing content?
- What social media is being used the most, and how?
- Are visitors that interact with social media more likely to buy a ticket?
- Is social media driving more traffic to the site? Is that traffic buying tickets?

The Solution

Cardinal Path was asked to architect and implement a solution based on an existing Google Analytics implementation that would to answer these questions.

A combination of customized event tracking, campaign tagging, custom variables, and a complex implementation and configuration was deployed to include the tracking of each social media outlet on the site.

(*Continued*)

Application Case 8.6 (Continued)

The Results

As a result of this measurement solution, it was easy to surface some impressive insights that helped C3 quantify the return on their social media investment:

- Users of the social media applications on Lollapalooza.com spent twice as much as non-users.
- Over 66 percent of the traffic referred from Facebook, MySpace, and Twitter was a result of sharing applications and Lollapalooza's messaging to its fans on those platforms.
- Fan engagement metrics such as time on site, bounce rate, page views per visit, and interaction goals improved significantly across the board as a result of social media applications.

Questions for Discussion

1. How did C3 Presents use social media analytics to improve its business?
2. What were the challenges, the proposed solution, and the obtained results?

Source: **www.cardinalpath.com/case-study/social-media-measurement** (accessed March 2013).

SECTION 8.9 REVIEW QUESTIONS

1. What is social media? How does it relate to Web 2.0?
2. What are the differences and commonalities between Web-based social media and traditional/industrial media?
3. How do people use social media? What are the evolutionary levels of engagement?

8.10 SOCIAL MEDIA ANALYTICS

Social media analytics refers to the systematic and scientific ways to consume the vast amount of content created by Web-based social media outlets, tools, and techniques for the betterment of an organization's competitiveness. Social media analytics is rapidly becoming a new force in organizations around the world, allowing them to reach out to and understand consumers as never before. In many companies, it is becoming the tool for integrated marketing and communications strategies.

The exponential growth of social media outlets, from blogs, Facebook, and Twitter to LinkedIn and YouTube, and analytics tools that tap into these rich data sources offer organizations the chance to join a conversation with millions of customers around the globe every day. This aptitude is why nearly two-thirds of the 2,100 companies who participated in a recent survey by Harvard Business Review (HBR) Analytic Services said they are either currently using social media channels or have social media plans in the works (HBR, 2010). But many still say social media is an experiment, as they try to understand how to best use the different channels, gauge their effectiveness, and integrate social media into their strategy.

Despite the vast potential social media analytics brings, many companies seem focused on social media activity primarily as a one-way promotional channel and have yet to capitalize on the ability to not only listen to, but also analyze, consumer conversations and turn the information into insights that impact the bottom line. Here are some of the results from the HBR Analytic Services survey (HBR, 2010):

- Three-quarters (75%) of the companies in the survey said they did not know where their most valuable customers were talking about them.
- Nearly one-third (31%) do not measure effectiveness of social media.
- Less than one-quarter (23%) are using social media analytic tools.
- A fraction (7%) of participating companies are able to integrate social media into their marketing activities.

While still searching for best practice and measurements, two-thirds of the companies surveyed are convinced their use of social media will grow, and many anticipate investing more in it next year, even as spending in traditional media declines. So what is it specifically that the companies are interested in measuring in social media?

Measuring the Social Media Impact

For organizations, small or large, there is valuable insight hidden in all the user-generated content on social media sites. But how do you dig it out of dozens of review sites, thousands of blogs, millions of Facebook posts, and billions of tweets? Once you do that, how do you measure the impact of your efforts? These questions can be addressed by the analytics extension of the social media technologies. Once you decide on your goal for social media (what it is that you want to accomplish), there is a multitude of tools to help you get there. These analysis tools usually fall into three broad categories:

- ***Descriptive analytics:*** Uses simple statistics to identify activity characteristics and trends, such as how many followers you have, how many reviews were generated on Facebook, and which channels are being used most often.
- ***Social network analysis:*** Follows the links between friends, fans, and followers to identify connections of influence as well as the biggest sources of influence.
- ***Advanced analytics:*** Includes predictive analytics and text analytics that examine the *content* in online conversations to identify themes, sentiments, and connections that would not be revealed by casual surveillance.

Sophisticated tools and solutions to social media analytics use all three categories of analytics (i.e., descriptive, predictive, and prescriptive) in a somewhat progressive fashion.

Best Practices in Social Media Analytics

As an emerging tool, social media analytics is practiced by companies in a somewhat haphazard fashion. Because there are not well established methodologies, everybody is trying to create their own by trial and error. What follows are some of the field-tested best practices for social media analytics proposed by Paine and Chaves (2012).

THINK OF MEASUREMENT AS A GUIDANCE SYSTEM, NOT A RATING SYSTEM Measurements are often used for punishment or rewards; they should not be. They should be about figuring out what the most effective tools and practices are, what needs to be discontinued because it doesn't work, and what needs to be done more because it does work very well. A good analytics system should tell you where you need to focus. Maybe all that emphasis on Facebook doesn't really matter, because that is not where your audience is. Maybe they are all on Twitter, or vice versa. According to Paine and Chaves, channel preference won't necessarily be intuitive, "We just worked with a hotel that had virtually no activity on Twitter for one brand but lots of Twitter activity for one of their higher brands." Without an accurate measurement tool, you would not know.

TRACK THE ELUSIVE SENTIMENT Customers want to take what they are hearing and learning from online conversations and act on it. The key is to be precise in extracting and tagging their intentions by measuring their sentiments. As we have seen in Chapter 7, text analytic tools can categorize online content, uncover linked concepts, and reveal the sentiment in a conversation as "positive," "negative," or "neutral," based on the words people use. Ideally, you would like to be able to attribute sentiment to a specific product, service, and business unit. The more precise you can get in understanding the tone and perception that people express, the more actionable the information becomes, because you are mitigating concerns about mixed polarity. A mixed-polarity phrase, such as "hotel in great location but bathroom was smelly" should not be tagged as "neutral" because you have positives and negatives offsetting each other. To be actionable, these types of phrases are to be treated separately; "bathroom was smelly" is something someone can own and improve upon. One can classify and categorize these sentiments, look at trends over time, and see significant differences in the way people speak either positively or negatively about you. Furthermore, you can compare sentiment about your brand to your competitors.

CONTINUOUSLY IMPROVE THE ACCURACY OF TEXT ANALYSIS An industry-specific text analytics package will already know the vocabulary of your business. The system will have linguistic rules built into it, but it learns over time and gets better and better. Much as you would tune a statistical model as you get more data, better parameters, or new techniques to deliver better results, you would do the same thing with the natural language processing that goes into sentiment analysis. You set up rules, taxonomies, categorization, and meaning of words; watch what the results look like; and then go back and do it again.

LOOK AT THE RIPPLE EFFECT It is one thing to get a great hit on a high-profile site, but that's only the start. There's a difference between a great hit that just sits there and goes away versus a great hit that is tweeted, retweeted, and picked up by influential bloggers. Analysis should show you which social media activities go "viral" and which quickly go dormant—and why.

LOOK BEYOND THE BRAND One of the biggest mistakes people make is to be concerned only about their brand. To successfully analyze and act on social media, you need to understand not just what is being said about your brand, but the broader conversation about the spectrum of issues surrounding your product or service, as well. Customers don't usually care about a firm's message or its brand; they care about themselves. Therefore, you should pay attention to what they are talking about, where they are talking, and where their interests are.

IDENTIFY YOUR MOST POWERFUL INFLUENCERS Organizations struggle to identify who has the most power in shaping public opinion. It turns out, your most important influencers are not necessarily the ones who advocate specifically for your brand; they are the ones who influence the whole realm of conversation about your topic. You need to understand whether they are saying nice things, expressing support, or simply making observations or critiquing. What is the nature of their conversations? How is my brand being positioned relative to the competition in that space?

LOOK CLOSELY AT THE ACCURACY OF YOUR ANALYTIC TOOL Until recently, computer-based automated tools were not as accurate as humans for sifting through online content. Even now, accuracy varies depending on the media. For product review sites, hotel review sites, and Twitter, it can reach anywhere between 80 to 90 percent accuracy, because the context is more boxed in. When you start looking at blogs and discussion forums, where the conversation is more wide-ranging, the software can deliver 60 to 70 percent accuracy (Paine and Chaves, 2012). These figures will increase over time, because the analytics tools are continually upgraded with new rules and improved algorithms to reflect field experience, new products, changing market conditions, and emerging patterns of speech.

INCORPORATE SOCIAL MEDIA INTELLIGENCE INTO PLANNING Once you have big-picture perspective and detailed insight, you can begin to incorporate this information into your planning cycle. But that is easier said than done. A quick audience poll revealed that very few people currently incorporate learning from online conversations into their planning cycles (Paine and Chaves, 2012). One way to achieve this is to find time-linked associations between social media metrics and other business activities or market events. Social media is typically either organically invoked or invoked by something your organization does; therefore, if you see a spike in activity at some point in time, you want to know what was behind that.

Application Case 8.7 shows an interesting case where eHarmony, one of the most popular online relationship service providers, uses social media analytics to better listen, understand, and service its customers.

Application Case 8.7

eHarmony Uses Social Media to Help Take the Mystery Out of Online Dating

eHarmony launched in the United States in 2000 and is now the number-one trusted relationship services provider in the United States. Millions of people have used eHarmony's Compatibility Matching System to find compatible long-term relationships; an average of 542 eHarmony members marry every day in the United States, as a result of being matched on the site.

The Challenge

Online dating has continued to increase in popularity, and with the adoption of social media the social media team at eHarmony saw an even greater opportunity to connect with both current and future members. The team at eHarmony saw social media as a chance to dispel any myths and preconceived notions about online dating and, more importantly, have some fun with their social media presence. "For us it's about being human, and sharing great content that will help our members and our social media followers," says Grant Langston, director of social media at eHarmony. "We believe that if there are conversations happening around our brand, we need to be there and be a part of that dialogue."

The Approach

eHarmony started using Salesforce Marketing Cloud to listen to conversations around the brand and around keywords like "bad date" or "first date." They also took to Facebook and Twitter to connect with members, share success stories—including engagement and wedding videos—and answer questions from those looking for dating advice.

"We wanted to ensure our team felt comfortable using social media to connect with our community so we set up guidelines for how to respond and proceed," explains Grant Langston. "We try to use humor and have some fun when we reach out to people through Twitter or Facebook. We think it makes a huge difference and helps make people feel more comfortable."

The Results

By using social media to help educate and create awareness around the benefits of online dating, eHarmony has built a strong and loyal community. The social media team now has eight staff members working to respond to social interactions and posts, helping them reach out to clients and respond to hundreds of posts a week. They plan to start creating Facebook apps that celebrate their members' success, and they are looking to create some new videos around some common dating mistakes. The social team at eHarmony is making all the right moves and their hard work is paying off for their millions of happy members.

Questions for Discussion

1. How did eHarmony use social media to enhance online dating?
2. What were the challenges, the proposed solution, and the obtained results?

Source: SalesForce Marketing Cloud, Case Study, **salesforcemarketingcloud.com; eharmony.com.**

Social Media Analytics Tools and Vendors

Monitoring social media, identifying interesting conversations among potential customers, and inferring what they are saying about your company, your products, and services is an essential yet a challenging task for many organizations. Generally speaking, there are two main paths that an organization can take to attain social media analytics (SMA) capabilities: in-house development or outsourcing. Because the SMA-related field is still evolving/maturing and because building an effective SMA system requires extensive knowledge in several related fields (e.g., Web, text mining, predictive analytics, reporting, visualization, performance management, etc.), with the exception of very large enterprises, most organizations choose the easier path: outsourcing.

Due to the astounding emphasis given to SMA, in the last few years we have witnessed an incredible emergence of start-up companies claiming to provide practical, cost-effective SMA solutions to organizations of all sizes and types. Because what they offered was not much more than just monitoring a few keywords about brands/products/services in social media, many of them did not succeed. While there still is a lot of uncertainty and churn in the marketplace, a significant number of them have survived and evolved to provide services that go beyond basic monitoring of a few brand names and keywords; they provide an integrated approach that helps many parts of the business, including product development, customer support, public outreach, lead generation, market research, and campaign management.

In the following section, we list and briefly define 10 SMA tools/vendors. This list is not meant to be "the absolute top 10" or the complete top-tier leaders in the market. It is to provide only 10 of the many successful SMA vendors and their respective tools/services with which we have some familiarity.

ATTENSITY360 Attensity360 operates on four key principles: listen, analyze, relate, and act. Attensity360 helps monitor trending topics, influencers, and the reach of your brand while recommending ways to join the conversation. Attensity Analyze applies text analytics to unstructured text to extract meaning and uncover trends. Attensity Respond helps automate the routing of incoming social media mentions into user-defined queues. Clients include Whirlpool, Vodofone, Versatel, TMobile, Oracle, and Wiley.

RADIAN6/SALESFORCE CLOUD Radian 6, purchased by Salesforce in 2011, works with brands to help them listen more intelligently to your consumers, competitors, and influencers with the goal of growing your business via detailed, real-time insights. Beyond their monitoring dashboard, which tracks mentions on more than 100 million social media sites, they offer an engagement console that allows you to coordinate your internal responses to external activity by immediately updating your blog, Twitter, and Facebook accounts all in one spot. Their clients include Red Cross, Adobe, AAA, Cirque du Soleil, H&R Block, March of Dimes, Microsoft, Pepsi, and Southwest Airlines.

SYSOMOS Managing conversations in real time, Sysomos's Heartbeat is a real-time monitoring and measurement tool that provides constantly updated snapshots of social media conversations delivered using a variety of user-friendly graphics. Heartbeat organizes conversations, manages workflow, facilitates collaboration, and provides ways to engage with key influencers. Their clients include IBM, HSBC, Roche, Ketchum, Sony Ericsson, Philips, ConAgra, Edelman, Shell Oil, Nokia, Sapient, Citi, and Interbrand. Owner: Marketwire.

COLLECTIVE INTELLECT Boulder, Colorado–based Collective Intellect, which started out by providing monitoring to financial firms, has evolved into a top-tier player in the marketplace of social media intelligence gathering. Using a combination of self-serve client dashboards and human analysis, Collective Intellect offers a robust monitoring and measurement tool suited to mid-size to large companies with its Social CRM Insights platform. Their clients include General Mills, NBC Universal, Pepsi, Walmart, Unilever, MillerCoors, Paramount, and Siemens.

WEBTRENDS Webtrends offers services geared toward monitoring, measuring, analyzing, profiling, and targeting audiences for a brand. The partner-based platform allows for crowd-sourced improvements and problem solving, creating transparency for their products and services. Their clients include CBS, NBC Universal, 20th Century Fox, AOL, Electronic Arts, Lifetime, and Nestle.

CRIMSON HEXAGON Cambridge, Massachusetts–based Crimson Hexagon taps into billions of conversations taking place in online media and turns them into actionable data for better brand understanding and improvement. Based on a technology licensed from Harvard, its VoxTrot Opinion is able to analyze vast amounts of qualitative information and determine quantitative proportion of opinion. Their clients include CNN, Hanes, AT&T, HP, Johnson & Johnson, Mashable, Microsoft, Monster, Thomson Reuters, Rubbermaid, Sybase, and *The Wall Street Journal.*

CONVERSEON New York–based social media consulting firm Converseon, named a leader in the social media monitoring sector by Forrester Research, builds tailored dashboards for its enterprise installations and offers professional services around every step of the social business intelligence process. Converseon starts with the technology and adds human analysis, resulting in high-quality data and impressive functionality. Their clients include Dow, Amway, Graco, and other major brands.

SPIRAL16 Spiral16 takes an in-depth look at who is saying what about a brand and compares results with those of top competitors. The goal is to help you monitor the effectiveness of your social media strategy, understand the sentiment behind conversations online, and mine large amounts of data. It uses impressive 3D displays and a standard dashboard. Their clients include Toyota, Lee, and Cadbury.

BUZZLOGIC BuzzLogic uses its technology platform to identify and organize the conversation universe, combining both conversation topic and audience to help brands reach audiences who are passionate on everything from the latest tech craze and cloud computing to parenthood and politics. Their clients include Starbucks, American Express, HBO, and HP.

SPROUTSOCIAL Founded in 2010 in Chicago, Illinois, SproutSocial is an innovative social media analytics company that provides social analytics services to many well-known firms and organizations. Their clients include Yahoo!, Nokia, Pepsi, St. Jude Children's Research Center, Hyatt Regency, McDonalds, and AMD. A sample screen shot of their social media solution dashboard is shown in Figure 8.12.

SECTION 8.10 REVIEW QUESTIONS

1. What is social media analytics? What type of data is analyzed with it?
2. What are the reasons/motivations behind the exponential growth of social media analytics?
3. How can you measure the impact of social media analytics?
4. List and briefly describe the best practices in social media analytics.
5. Why do you think social media analytics tools are usually offered as a service and not a tool?

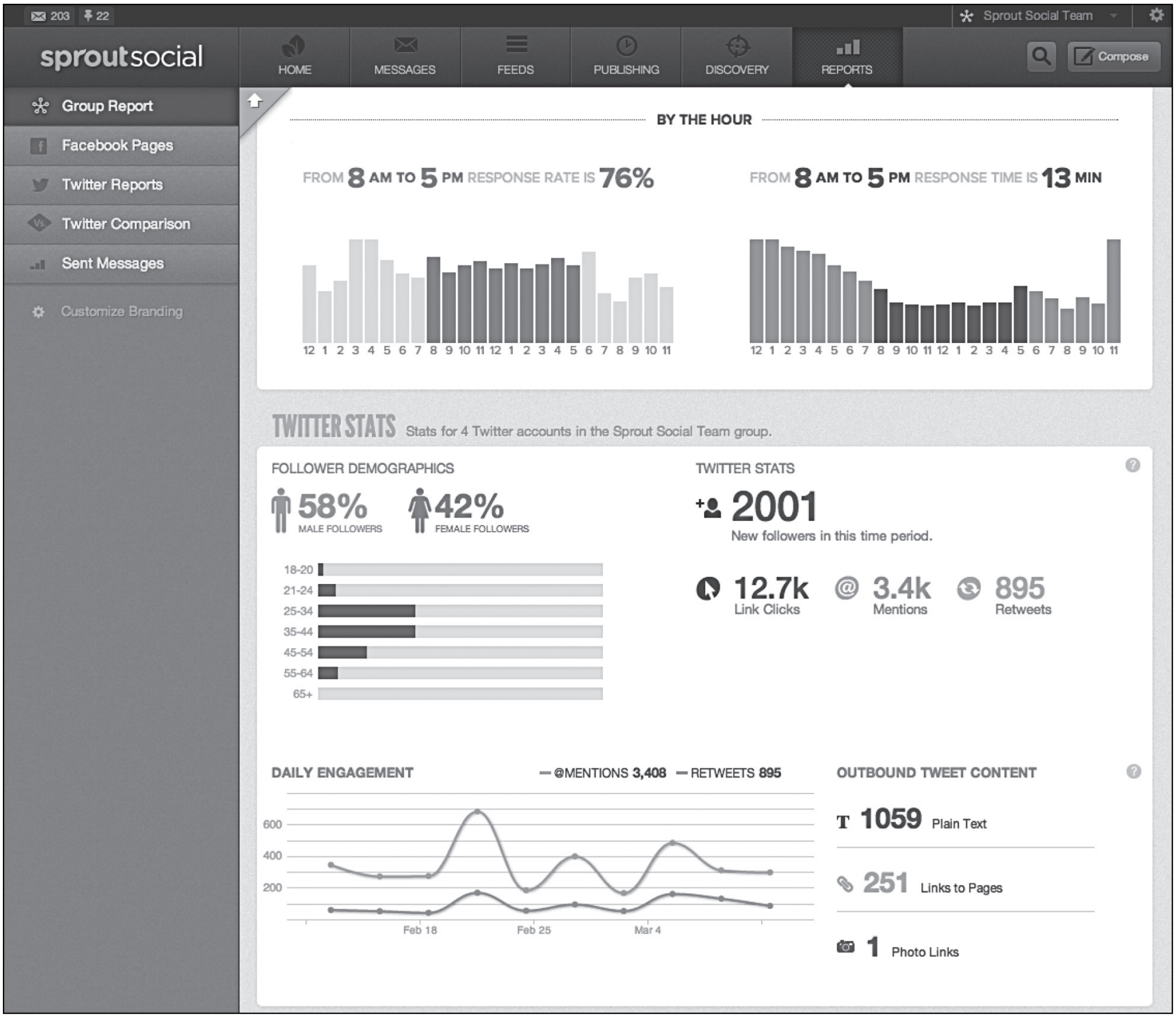


FIGURE 8.12 A Social Media Analytics Screenshot. *Source:* Courtesy of **sproutsocial.com.**

Chapter Highlights

- Web mining can be defined as the discovery and analysis of interesting and useful information from the Web, about the Web, and usually using Web-based tools.
- Web mining can be viewed as consisting of three areas: Web content mining, Web structure mining, and Web usage mining.
- Web content mining refers to the automatic extraction of useful information from Web pages. It may be used to enhance search results produced by search engines.
- Web structure mining refers to generating interesting information from the links included in Web pages. This is used in Google's page rank algorithm to order the display of pages, for example.
- Web structure mining can also be used to identify the members of a specific community and perhaps even the roles of the members in the community.
- Web usage mining refers to developing useful information through analysis of Web server logs, user profiles, and transaction information.

- Web usage mining can assist in better CRM, personalization, site navigation modifications, and improved business models.
- Text and Web mining are emerging as critical components of the next generation of business intelligence tools, enabling organizations to compete successfully.
- A search engine is a software program that searches for documents (Internet sites or files), based on keywords (individual words, multi-word terms, or a complete sentence) users have provided that have to do with the subject of their inquiry.
- PageRank is a link analysis algorithm named after Larry Page, who is one of the two inventors of Google, which began as a research project at Stanford University in 1996. PageRank is used by the Google Web search engine.
- Search engine optimization (SEO) is the intentional activity of affecting the visibility of an e-commerce site or a Web site in a search engine's natural (unpaid or organic) search results.
- A maturity model is a formal depiction of critical dimensions and their competency levels of a business practice.
- Voice of the customer (VOC) is a term usually used to describe the analytic process of capturing a customer's expectations, preferences, and aversions.
- Social analytics is the monitoring, analyzing, measuring, and interpreting of digital interactions and relationships among people, topics, ideas, and content.
- A social network is a social structure composed of individuals/people (or groups of individuals or organizations) linked to one another with some type of connections/relationships.
- Social media refers to the enabling technologies of social interactions among people in which they create, share, and exchange information, ideas, and opinions in virtual communities and networks.
- Social media analytics refers to the systematic and scientific ways to consume the vast amount of content created by Web-based social media outlets, tools, and techniques for the betterment of an organization's competitiveness.

Key Terms

authoritative pages
clickstream analysis
customer experience management (CEM)
hubs
hyperlink-induced topic search (HITS)
search engine
social network
spiders
voice of customer (VOC)
Web analytics
Web content mining
Web crawler
Web mining
Web structure mining
Web usage mining

Questions for Discussion

1. Explain how the size and complexity of the Web makes knowledge discovery challenging.
2. What are the limitations of a simple keyword-based search engine? How does Web mining address these deficiencies?
3. Define Web mining? How is it different from Web analytics?
4. How does Web content mining increase a company's efficiency and competitive advantage?
5. How can we collectively use Web content mining, Web structure mining, and Web usage mining for business gains?
6. What are authoritative pages, hubs, and hyperlink-induced topic search (HITS)? Discuss the differences between citations in research articles and hyperlinks on Web pages.
7. Define a search engine and discuss its anatomy.
8. What is PageRank algorithm? What is the relation between PageRank and citation analysis? How does Google use PageRank?
9. Define SEO? What does it involve? Discuss the most commonly used methods for SEO.
10. How would optimizing search engines help businesses? How does over-reliance on search engine traffic harm businesses and how can it be avoided?
11. What is Web Analytics? Discuss the application of Web analytics for businesses.
12. Discuss the impact of Web analytics metrics on market insight.

13. Discuss Web site optimization ecosystem?
14. Discuss social network analysis. Which types of social networks are relevant to businesses?
15. How can we measure the impact of social media analytics? How can social media intelligence be incorporated into planning?

Exercises

Teradata University Network (TUN) and Other Hands-on Exercises

1. Visit **teradatauniversitynetwork.com**. Identify cases about Web mining. Describe recent developments in the field. If you cannot find enough cases at the Teradata University network Web site, broaden your search to other Web-based resources.
2. Go to **teradatauniversitynetwork.com** or locate white papers, Web seminars, and other materials related to Web mining. Synthesize your findings into a short written report.
3. Browse the Web and your library's digital databases to identify articles that make the linkage between text/Web mining and contemporary business intelligence systems.

Team Assignments and Role-Playing Projects

1. Examine how Web-based data can be captured automatically using the latest technologies. Once captured, what are the potential patterns that you can extract from these content-rich, mostly unstructured data sources?
2. Interview administrators in your college or executives in your organization to determine how Web mining is assisting (or could assist) them in their work. Write a proposal describing your findings. Include a preliminary cost–benefit analysis in your report.
3. Go online, search for publicly available Web usage or social media data files, and download one of your choice. Then, download and use one of the free tools to analyze the data. Write your findings and experiences in a professionally organized report.

Internet Exercises

1. What types of online information can be collected using of Web crawlers? Discuss how organizations could use this information for decision making.
2. Write a brief report on the effectiveness of your educational institution's current SEO illustrating with your findings and recommendations. How can effective SEO strategies influence the number of potential students and their characteristics?
3. How would Web usage mining ensure content generation and user management on a site of your choice? Referring to the chapter, carry out a basic off-site analysis for the site and suggest steps to improve its user retention strategy.
4. Use social media Websites to gauge opinions about a clothing brand among a sample of your immediate friends. Identify examples of how social networks influence product selection.
5. Visit openwebanalytics.com and compare the functionality of this open source tool with its competitors. Distinguish between the two. Why would companies purchase web analytics solutions?
6. Go to **www.google.com/adwords** and download at least three success stories. Does the existence of services such as Google AdWords affect the need for SEO in any way? Justify your answer.
7. Go to **kdnuggets.com**. Explore the sections on applications as well as software. Find names of at least three additional packages for Web mining and social media analytics.

End-of-Chapter Application Case

Keeping Students on Track with Web and Predictive Analytics

What makes a college student stay in school? Over the years, educators have held a lot of theories—from student demographics to attendance in college prep courses—but they've lacked the hard data to prove conclusively what really drives retention. As a result, colleges and universities have struggled to understand how to lower dropout rates and keep students on track all the way to graduation.

That situation is changing at American Public University System (APUS), an online university serving 70,000 distance learners from the United States and more than 100 countries. APUS is breaking new ground by using analytics to zero in on those factors that most influence a student's decision to stay in school or drop out.

"By leveraging the power of predictive analytics, we can predict the probability that any given student will drop out," says Phil Ice, APUS' director of course design, research, and development. "This translates into actionable business intelligence that we can deploy across the enterprise to create and maintain the conditions for maximum student retention."

Rich Data Sources

As an online university, APUS has a rich store of student information available for analysis. "All of the activities are technology mediated here," says Ice, "so we have a very nice record of what goes on with the student. We can pull demographic data, registration data, course level data, and more." Ice and his colleagues then develop metrics that help APUS analyze and build predictive models of student retention.

One of the measures, for example, looks at the last time a student logged into the online system after starting a class. If too many days have passed, it may be a sign the student is about to drop out. Educators at APUS combine such online activity data with a sophisticated end-of-course survey to build a complete model of student satisfaction and retention.

The course survey yields particularly valuable data for APUS. Designed around a theoretical framework known as the Community of Inquiry, the survey seeks to understand the student's learning experience by analyzing three interdependent elements: social, cognitive, and teaching presence. Through the survey, Ice says, "We hone in on things such as a student's perception of being able to build effective community."

Increasing Accuracy

It turns out that the student's sense of being part of a larger community—his or her "social presence"—is one of the key variables affecting the student's likelihood of staying in school. Another one is the student's perception of the effectiveness of online learning. In fact, when fed into IBM SPSS Modeler and measured against disenrolment rates, these two factors together accounted for nearly 25 percent of the overall statistical variance, meaning they are strong predictors of student attrition. With the adoption of advanced analytics in its retention efforts, APUS predicts with approximately 80 percent certainty whether a given student is going to drop out.

Some of the findings generated by its predictive models actually came as a surprise to APUS. For example, it had long been assumed that gender and ethnicity were good predictors of attrition, but the models proved otherwise. Educators also assumed that a preparatory course called "College 100: Foundations of Online Learning" was a major driver of retention, but an in-depth analysis using Modeler came to a different conclusion. "When we ran the numbers, we found that students who took it were not retained the way we thought they were," says Dr. Frank McCluskey, provost and executive vice president at APUS. "IBM SPSS predictive analytics told us that our guess had been wrong."

Strategic Course Adjustments

The next step for APUS is to put its new-found predictive intelligence to work. Already the university is building online "dashboards" that are putting predictive analytics into the hands of deans and other administrators who can design and implement strategies for boosting retention. Specific action plans could include targeting individual at-risk students with special communications and counseling. Analysis of course surveys can also help APUS adjust course content to better engage students and provide feedback to instructors to help improve their teaching methods.

Survey and modeling results are reinforcing the university's commitment to enriching the student's sense of community—a key retention factor. Online courses, for example, are being refined to promote more interactions among students, and social media and online collaboration tools are being deployed to boost school spirit. "We have an online student lounge, online student clubs, online student advisors," says McCluskey. "We want to duplicate a campus fully and completely, where students can grow in all sorts of ways, learn things, exchange ideas—maybe even books—and get to know each other."

Smart Decisions

While predictive modeling gives APUS an accurate picture of the forces driving student attrition, tackling the problem means deciding among an array of possible intervention strategies. To help administrators sort out the options, APUS plans to implement a Decision Management System, a solution that turns IBM SPSS Modeler's predictive power into intelligent, data-driven decisions. The solution will draw from Modeler's analysis of at-risk students and suggest the best intervention strategies for any given budget.

APUS also plans to delve deeper into the surveys by mining the open-ended text responses that are part of each questionnaire (IBM SPSS Text Analytics for Surveys will help with that initiative). All of these data-driven initiatives aim to increase student learning, enhance the student's experience, and build an environment that encourages retention at APUS. It's no coincidence that achieving that goal also helps grow the university's bottom line. "Attracting and enrolling new students is expensive, so losing them is costly for us and for students as well," McCluskey says. "That is why we are excited about using predictive analytics to keep retention rates as high as possible."

Questions for the End-of-Chapter Application Case

1. Describe challenges that APUS was facing. Discuss the ramifications of such challenges.
2. What types of data did APUS tap into? What do you think are the main obstacles one would have to overcome when using data that comes from different domains and sources?
3. What solutions did they pursue? What tools and technologies did they use?
4. What were the results? Do you think these results are also applicable to other educational institutions? Why? Why not?
5. What additional analysis are they planning on conducting? Can you think of other data analyses that they could apply and benefit from?

Source: IBM, Customer Success Stories, **www-01.ibm.com/software/success/cssdb.nsf/CS/GREE-8F8M76** (accessed March 2013).

References

Brin, S., and L. Page. (2012). "Reprint of the Anatomy of a Large-Scale Hypertextual Web Search Engine." *Computer Networks,* Vol. 56, No. 18, pp. 3825–3833,

Brogan, C., and Bastone, J. (2011). "Acting on Customer Intelligence from Social Media: The New Edge for Building Customer Loyalty and Your Brand." SAS white paper. **sas.com/resources/whitepaper/wp_21122.pdf** (accessed March 2013).

Coussement, K., and D. Van Den Poel. (2009). "Improving Customer Attrition Prediction by Integrating Emotions from Client/Company Interaction Emails and Evaluating Multiple Classifiers." *Expert Systems with Applications,* Vol. 36, No. 3, pp. 6127–6134.

Cutts, M. (2006, February 4). "Ramping Up on International Webspam." **mattcutts.com/blog. mattcutts.com/blog/ramping-up-on-international-webspam** (accessed March 2013).

Etzioni, O. (1996). "The World Wide Web: Quagmire or Gold Mine?" *Communications of the ACM,* Vol. 39, No. 11, pp. 65–68.

Goodman, A. (2005). "Search Engine Showdown: Black Hats Versus White Hats at SES." SearchEngineWatch. **searchenginewatch.com/article/2066090/Search-Engine-Showdown-Black-Hats-vs.-White-Hats-at-SES** (accessed February 2013).

HBR. (2010). "The New Conversation: Taking Social Media from Talk to Action," A SAS-sponsored research report by Harvard Business Review Analytic Services. **sas.com/resources/whitepaper/wp_23348.pdf** (accessed March 2013).

Kaplan, A. M., and M. Haenlein. (2010). "Users of the World, Unite! The Challenges and Opportunities of Social Media." *Business Horizons,* Vol. 53, No. 1, pp. 59–68.

Kleinberg, J. (1999). "Authoritative Sources in a Hyperlinked Environment." *Journal of the ACM,* Vol. 46, No. 5, pp. 604–632.

Lars-Henrik, S. (1996). "Commonness Across Cultures," in Anindita Niyogi Balslev. *Cross-Cultural Conversation: Initiation*. Oxford University Press.

Liddy, E. (2012). "How a Search Engine Works." **www.infotoday.com/searcher/may01/liddy.htm** (accessed November 2012).

Masand, B. M., M. Spiliopoulou, J. Srivastava, and O. R. Zaïane. (2002). "Web Mining for Usage Patterns and Profiles." *SIGKDD Explorations,* Vol. 4, No. 2, pp. 125–132.

Morgan, N., G. Jones, and A. Hodges. (2010). "The Complete Guide to Social Media from the Social Media Guys." **thesocialmediaguys.co.uk/wp-content/uploads/downloads/2011/03/CompleteGuidetoSocialMedia.pdf** (accessed February 2013).

Nasraoui, O., M. Spiliopoulou, J. Srivastava, B. Mobasher, and B. Masand. (2006). "WebKDD 2006: Web Mining and Web Usage Analysis Post-Workshop Report." *ACM SIGKDD Explorations Newsletter*, Vol. 8, No. 2, pp. 84–89.

Paine, K. D., and M. Chaves. (2012). "Social Media Metrics." SAS white paper. **sas.com/resources/whitepaper/wp_19861.pdf** (accessed February 2013).

Peterson, E. T. (2008). "The Voice of Customer: Qualitative Data as a Critical Input to Web Site Optimization." **foreseeresults.com/Form_Epeterson_WebAnalytics.html** (accessed May 2009).

Scott, W. Richard, and Gerald F. Davis. (2003). "Networks In and Around Organizations." *Organizations and Organizing*. Upper Saddle River, NJ: Pearson Prentice Hall.

The Westover Group. (2013). "20 Key Web Analytics Metrics and How to Use Them." **www.thewestovergroup.com** (accessed February 2013).

Turetken, O., and R. Sharda. (2004). "Development of a Fisheye-Based Information Search Processing Aid (FISPA) for Managing Information Overload in the Web Environment." *Decision Support Systems,* Vol. 37, No. 3, pp. 415–434.

Zhou, Y., E. Reid, J. Qin, H. Chen, and G. Lai. (2005). "U.S. Domestic Extremist Groups on the Web: Link and Content Analysis." *IEEE Intelligent Systems,* Vol. 20, No. 5, pp. 44–51.

PART

IV

Prescriptive Analytics

LEARNING OBJECTIVES FOR PART IV

- Understand the applications of prescriptive analytics techniques in combination with reporting and predictive analytics
- Understand the concepts of analytical models for selected decision problems including linear programming and analytic hierarchy process
- Recognize the concepts of heuristic search methods and simulation models for decision support
- Understand the concepts and applications of automated rule systems and expert system technologies
- Gain familiarity with knowledge management and collaboration support systems

This part extends the decision support applications beyond reporting and data mining methods. It includes coverage of selected techniques that can be employed in combination with predictive models to help support decision making. We focus on techniques that can be implemented relatively easily using either spreadsheet tools or by using stand-alone software tools. Of course, there is much additional detail to be learned about management science models, but the objective of this part is to simply illustrate what is possible and how it has been implemented in some real settings. We also include coverage of automated decision systems that implement the results from various models described in this book, and expert system technologies. Finally, we conclude this part with a discussion of knowledge management issues and group support systems. Technologies and issues in knowledge management and group support systems directly impact the delivery and practice of prescriptive analytics.

CHAPTER

9

Model-Based Decision Making: Optimization and Multi-Criteria Systems

LEARNING OBJECTIVES

- Understand the basic concepts of analytical decision modeling
- Describe how prescriptive models interact with data and the user
- Understand some different, well-known model classes
- Understand how to structure decision making with a few alternatives
- Describe how spreadsheets can be used for analytical modeling and solution
- Explain the basic concepts of optimization and when to use them
- Describe how to structure a linear programming model
- Describe how to handle multiple goals
- Explain what is meant by sensitivity analysis, what-if analysis, and goal seeking
- Describe the key issues of multi-criteria decision making

In this chapter we describe selected techniques employed in prescriptive analytics. We present this material with a note of caution: Modeling can be a very difficult topic and is as much an art as a science. The purpose of this chapter is not necessarily for you to *master the topics* of modeling and analysis. Rather, the material is geared toward *gaining familiarity* with the important concepts as they relate to DSS and their use in decision making. It is important to recognize that the modeling we discuss here is only cursorily related to the concepts of data modeling. You should not confuse the two. We walk through some basic concepts and definitions of modeling before introducing the influence diagrams, which can aid a decision maker in sketching a model of a situation and even solving it. We next introduce the idea of modeling directly in spreadsheets. We then discuss the structure and application of some successful time-proven models and methodologies: optimization, decision analysis, decision trees, and analytic hierarchy process. This chapter includes the following sections: