Warning

concerning copyright restrictions

The Copyright law of the United states (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or reproduction.

One of three specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research.

If electronic transmission of reserve material is used for purposes in excess of what constitutes “fair use”, that user may be liable for copyright infringement.

This policy is in effect for the following document:

Miller, T. W.
Text Mining
Data and Text Mining: A Business Applications Approach, 2005. pp. 159–173.

No further transmission or distribution of this material is permitted.

CHAPTER

4 TEXT MINING

> Looking specifically at data mining, we see that the greatest need is not to enlarge our knowledge bases, however desirable that may be, but to be able to access in them the specific knowledge we need, whether or not we know it exists or even what it is.... We will design better data miners only if we remember that the scarce factor is not information but human attention, and that the task of the data miner is to conserve it for its most important uses. (Simon 2002, p. xvii)

Kim goes to the library looking for a book called *Management Consulting: A Guide to the Profession*. She searches an electronic catalog for the title. A Library of Congress call number identifies the book. A map points to the section of the library where the book is located. Shelves of books are neatly labeled with call number ranges. Each book spine has a unique call number. Kim's book search is easy because she knows what is relevant (the title), has a way to find it (an index or catalog), and has a well-organized place to find it (the library).

What if there were no obvious organization for the library's books? Imagine a library with hundreds of thousands of books and no electronic catalog, no titles or author names, no call numbers, no labels on shelves and book spines. How long would it take Kim to find a particular book? How long would it take to find relevant sources of information? This is the situation with much information in business.

The text mine of many firms is an unstructured collection of documents. Aside from files stored within functional areas—customer transactions under sales or order entry, new product ideas under research and development, complaints under customer support, litigation with the comptroller or legal counsel—there may be little to aid the employee in search of information. What does an employee need to do to get relevant information? Go to the functional area, find someone with experience in the area, ask relevant questions, and hope for the best.

Much of business information is text. There are internal memos, facsimiles, electronic mail messages, written contracts, and letters to and from suppliers and customers. What exists is raw text, paper documents in file cabinets and archive boxes, scanned text

with no obvious way to find it, electronic files without an index. Few firms maintain text databases. Fewer still would know what to do with a text database if they had one.

Data mining is today's job. Business databases beg for analysis. Many data fields are numeric or can be easily recoded as numeric. Data are organized into rows and columns. Databases have structure; there are keyed fields that point to relevant information.

Text mining is the job of the future. It begins with raw, unstructured text lacking the organization of a traditional database. Rather than being divided into discrete records and fields of numeric data, dates, or character strings, raw text is a string of characters separated by spaces and punctuation marks. There is no data dictionary or index to guide us to the relevant information—no dictionary or index until we create one. The tools of text mining help us to move from raw text to something that can be searched for answers to business questions.

This chapter introduces text mining and defines the major tasks of text mining. It reviews applications under the general headings of text categorization, information retrieval, and text measurement. For firms willing to invest time and money in the technology, text mining offers many benefits. Text mining is on the information frontier—a key to future success in a knowledge-based economy.

4.1 WHAT IS TEXT MINING?

Text mining is the automated or partially automated processing of text. It involves imposing structure upon text and extracting relevant information from text.

People read documents. Computers read documents. People write and provide answers to questions. Computers print and provide answers to questions. In between the reading and the writing, we hope that there is human intelligence when people do the work and human-like intelligence when computers do the work.

Text mining deals with words. Words, millions of words transcribed and stored in electronic files, represent raw data for analysis. Internal documents, external publications, and the words of consumers and business buyers beg for analysis. Overwhelming quantities of text come to us unstructured. To make sense out of textual data, we must impose structure.

Classifying text documents, analyzing syntax, identifying relationships among documents, understanding a question expressed in natural language, extracting meaning from a message, summarizing the meaning—these are nontrivial tasks involving more than the mere matching of words in text. To do a thorough job of text mining requires technologies from computational linguistics and pattern recognition, as well as traditional and data-adaptive modeling tools.

Some suggest that text mining concerns itself with large quantities of textual data. Our definition of text mining, like our definition of data mining, is not dependent upon the size of the database or document collection. We view text mining as a process, an approach to doing research that begins with words rather than numbers. Text mining methods are relevant to small document collections as well as large document collections.

Major business applications of text mining fall under the general categories of text categorization, information retrieval, and measurement. Categorization has organization as its initial objective. Information retrieval relates to searching, finding the proverbial "needle in the haystack." Measurement and the definition of text measures involve converting textual information into numeric information, so that it may be analyzed like other business data.

Text mining has business value. Firms with efficient mechanisms of information retrieval have a competitive advantage. Text produced in the form of memos and reports in one area of a company may be relevant to other areas of the company. Business intelligence, both internal and external, is important to corporate survival. Many companies are in the process of building data or knowledge warehouses with the objective of making information more widely available to knowledge workers. Text mining can foster collaboration and the sharing of expertise within a firm.

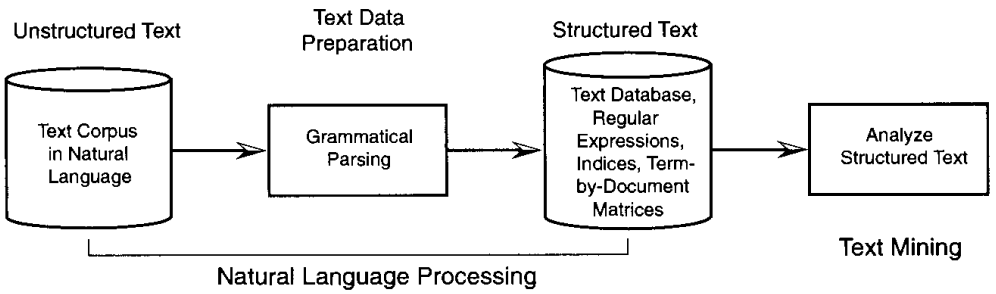
Measurement or text scoring, less well understood by the proponents of computational linguistics and text mining, flows from the work of text categorization. Research respondents listen to interviewer questions; they answer questions. Focus group participants talk about their experiences and feelings about products. User group participants type questions and comments into online message boards and chat rooms. Buyers send messages to sellers by electronic mail. To the extent that measures or scores can be extracted from these text sources, text mining holds promise as a business research tool.

4.2 WORKING WITH TEXT

> The ingredients of language are words and rules. Words in the sense of memorized links between sound and meaning: rules in the sense of operations that assemble the words into combinations whose meaning can be computed from the meanings of words and the way they are arranged.... We have digital minds in an analog world. More accurately, a part of our minds is digital. We remember familiar entities and their graded, crisscrossing traits, but we also generate novel mental products by reckoning with rules. It is surely no coincidence that the species that invented numbers, ranks, kinship terms, life stages, legal and illegal acts, and scientific theories also invented grammatical sentences and regular past-tense forms. Words and rules give rise to the vast expressive power of language, allowing us to share the fruits of the vast creative power of thought. (Pinker 1999, pp. 269, 287)

We have, as Pinker (1994) describes it, a "language instinct." The structure of sentences, the words we use—these are the observables of the speech act. Underlying the words and sentences is meaning—call it a latent message. It is the meaning or latent message that matters. "I have strong feelings for you. You touch my heart. You complete me. Deeply moved am I by the very sight of you. I can't bear to lose you. Let's stay together forever." Depending upon the context in which they are spoken, these sentences could be thought of as carrying the same meaning: "I love you." We may not be the best writers or speakers (or lovers), but we can understand the meaning of all manner

Figure 4.1 Natural Language Processing and Text Mining



of messages. We speak and listen with ease, recalling Juliet's words from Shakespeare: "What's in a name? That which we call a rose by any other name would smell as sweet."

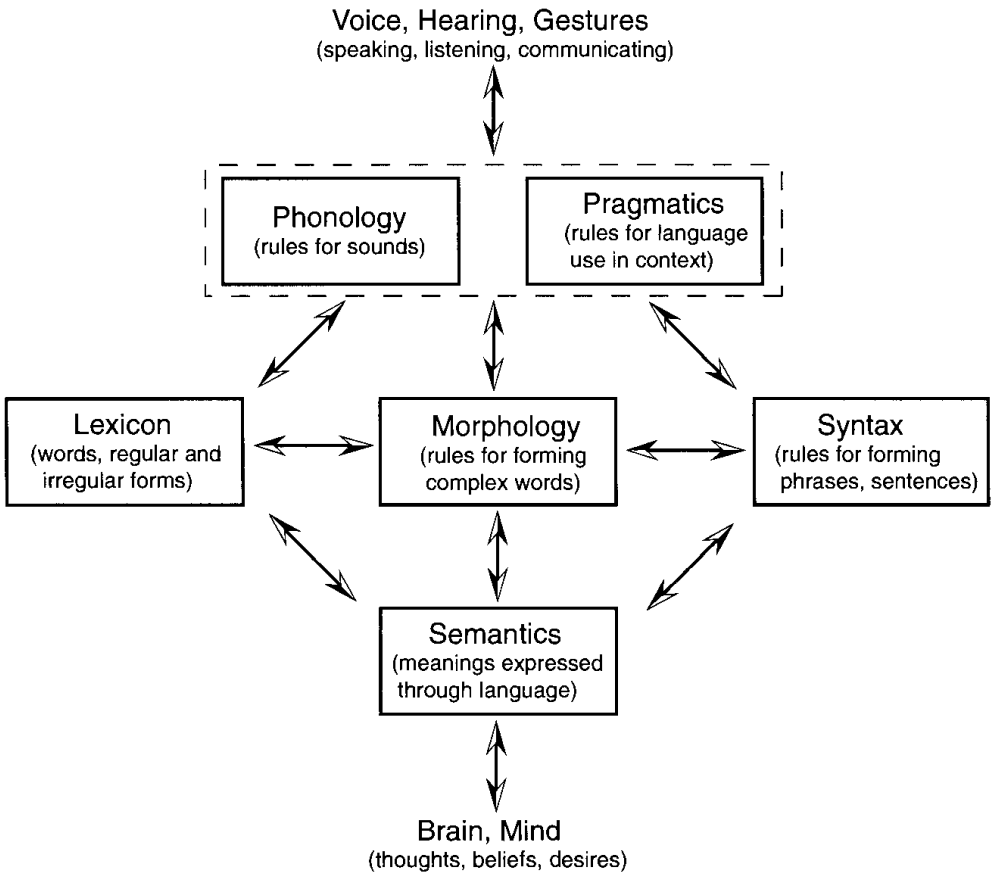
People are good at processing small quantities of unstructured text. Computers are good at processing large quantities of structured text. To do the work of text mining—to process large quantities of unstructured text—we must either give an enormous amount of time to people or a modicum of intelligence to computers. Because people's time is better spent in other ways, we try to give intelligence to computers. We find ways to structure text so that it can be understood by computers. This is the work of natural language processing, a preliminary to text mining. As illustrated in Figure 4.1, natural language processing involves parsing unstructured or grammatically structured natural language, creating regular expressions that are more easily analyzed by computer.

Text mining begins with a document collection, sometimes referred to as a "corpus." Compared with a traditional database, a document collection is unstructured, raw text. It is in the natural language, rather than a specialized computer language or set of codes. Documents contain paragraphs, paragraphs contain sentences, and sentences words. Natural language follows grammatical rules, with many ways of conveying the same idea and with many exceptions to rules. Words and the rules of grammar comprise the linguistic foundations of text mining as shown in Figure 4.2.

Linguists study natural language, the words and the rules that we use to form meaningful utterances. "Generative grammar" is a general term for the rules; "morphology," "syntax," and "semantics" are more specific terms. Computer programs for natural language processing use grammatical rules to mimic human communication and convert natural language into structured text for further analysis.

A corpus or document collection usually represents a particular content domain or publishing venue. All articles published in *The Wall Street Journal* in 2002 would be a corpus. The units of analysis, the documents, would be the individual articles. Each document has attributes—date of publication, newspaper section, page number, and length—that could be coded as "tags," alphanumeric data or search codes associated with the article. The meaning of a document, what it actually says, is reflected by its

Figure 4.2 Generative Grammar: Linguistic Foundations of Text Mining



Source: Adapted from Pinker (1999).

words. Discovering the meaning and coding that meaning are the hard parts of text mining.

The location of words in sentences is a key to understanding text. Words follow a sequence, with earlier words often more important than later words, and with early sentences and paragraphs often more important than later sentences and paragraphs. Words in the title of a document are especially important to understanding the meaning of a document. Some words occur with high frequency and help to define the meaning of a document. Other words, such as the definite article "the" and the indefinite articles "a" and "an," as well as many prepositions and pronouns, occur with high frequency but have little to do with defining the meaning of a document.

What makes text unstructured from the analyst's point of view is the fact that features of text must be defined. The features or attributes of text are often associated with terms—collections of words that mean something special. There are collections of words relating to the same concept or word stem. The words "marketer," "marketeer," and "marketing" build on the common word stem "market." There are syntactic structures to consider, such as adjectives followed by nouns and nouns followed by nouns. Most important to text mining are sequences of words that form terms. The words "New" and "York" have special meaning when combined to form the term "New York." The words "financial" and "analysis" have special meaning when combined to form the term "financial analysis." We talk of "stemming," which is the identification of word stems, dropping suffixes (and sometimes prefixes) from words. More generally, we are parsing natural language text to arrive at structured text.

Terms identified in the document collection act as variables in text mining. After coming up with a list of unique terms from the document collection, we can sort the list based upon frequency of occurrence. Additional structure may be imposed upon textual data by creating a term-by-document matrix (sometimes called a lexical table). The rows of this data matrix correspond to word stems from the document collection, and the columns correspond to the documents in the collection. Terms constitute a subset of the most important (or highest frequency) terms. The entry in each cell of a term-by-document matrix could be a binary indicator for the presence or absence of a term in a document, a frequency count of the number of times a term is used in a document, or (more likely) a weighted frequency indicating the importance of a term in a document. Exhibit 4.1 illustrates the process.

Typical text mining applications have many more terms than documents, resulting in sparse rectangular term-by-document matrices. To obtain meaningful results for text mining applications, analysts examine the distribution of terms across the document collection. Very low frequency terms, those used in few documents, may be dropped from the term-by-document matrix, reducing the number of rows in the matrix. After the term-by-document matrix has been refined, text mining turns into data mining. We have numbers suitable for analysis using a variety of traditional and data-adaptive methods, including principal components (singular value decomposition), cluster analysis, classification and regression trees, and support vector machines.

Exhibit 4.1 Creating a Term-by-Document Matrix

An initial step in text mining applications, such as information retrieval and text categorization, is the formation of a term-by-document matrix. The drawing below illustrates the process. The first document comes from Steven Pinker's *Words and Rules* (1999, p. 4), the second from Richard K. Belew's *Finding Out About* (2000, p. 73). Terms correspond to stems of words that appear in the documents. In this example, each matrix entry represents the number of times a term appears in a document. We treat nouns, verbs, and adjectives similarly in the definition of stems. The stem "combine" represents both the verb "combine" and the noun "combination." Likewise, "function" represents the verb, noun, and adjective form "functional." An alternative system might distinguish among parts of speech, permitting more sophisticated syntactic searches across the set of documents. Once created, the term-by-document matrix is like an index, a mapping of document identifiers and terms (keywords or stems). For information retrieval systems or search engines we might also retain information regarding the specific location of terms within documents.

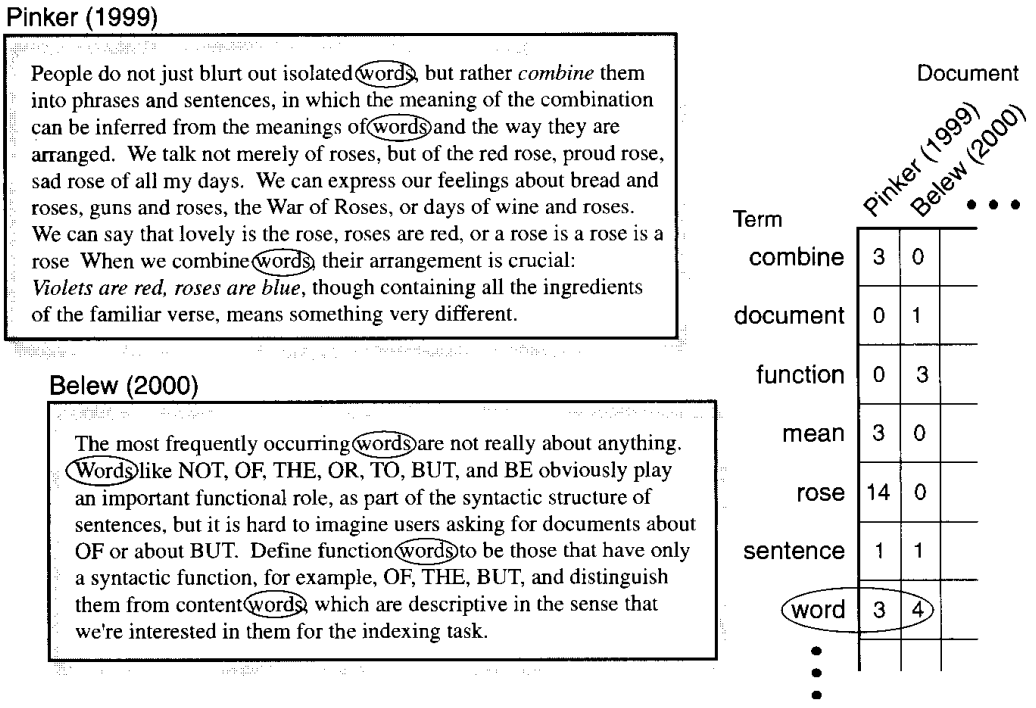
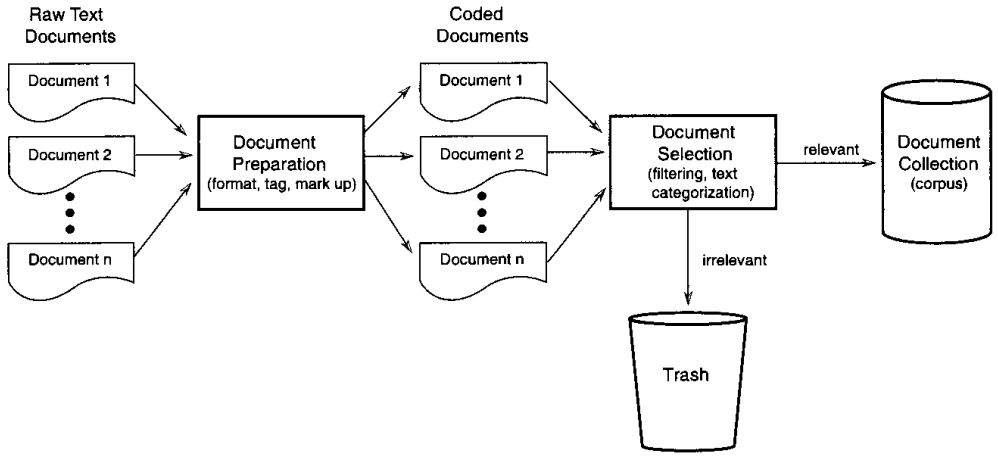


Figure 4.3 Selecting Documents for the Document Collection (Corpus)



Textual data, often overwhelming quantities of textual data, come to us unstructured. To make sense out of text, we impose structure. Forming the term-by-document matrix is often the beginning of the process of imposing structure. Note that the term-by-document matrix is specific to the document collection from which it is derived. When we add documents to the document collection, we must add columns to the matrix. When many documents are added, we should consider redefining the set of terms (the rows of the matrix). For large text mining systems, updating of the term-by-document matrix and subsequent analysis of the matrix may be carried out periodically as a batch operation.

4.3 TEXT CATEGORIZATION

Text categorization involves identifying common features across documents and organizing those documents into groups based upon the common features. Text categorization is an important component of many text mining applications, including those involving information retrieval and text measures. Text categorization is also a common first step in data mining. We categorize documents according to their relevance to the document collection or corpus. This is a filtering process, as shown in Figure 4.3. Relevant documents go into the document collection, irrelevant documents go into the trash.

4.3.1 Example: E-Mail Classification

A text categorization problem that most of us can relate to is the junk e-mail problem. For active e-mail users, thousands of messages are received each month. Separating the good messages from the junk can be a time-consuming process as users painstakingly review their in-boxes. Because the distinction between good and junk messages varies

from one user to the next, standard e-mail classification programs, such as those for identifying spam, fall short of providing the kind of facility that many users require.

We can use text categorization programs to automate the process of sorting e-mail messages into good and junk piles. The binary response, defined by an individual user, is the classification of the message as good or junk. Each message or document may be described by its sender, subject text/terms, message length, and message text/terms. The E-Mail Text Categorization from Appendix A illustrates the types of explanatory variables that could be used in such problems.

Suppose an individual user, Daniel, reviews one thousand e-mail messages. For each message, he decides whether it is good (he clicks on the message to read it) or junk (he ignores or deletes the message without reading it). A computer program records Daniel's behavior, noting the disposition of messages. Logs from the one thousand messages, with information about explanatory variables and responses, constitute a training set for the e-mail classification problem.

In going from e-mail text to continuous and categorical explanatory variables and a binary response, we have converted our text mining problem to a data mining problem. With Daniel's training set, we have a supervised learning problem. To build a classification model, we can use a variety of methods, including logistic regression, classification trees, neural network classifiers, and support vector machines. Subsequent observation of Daniel's behavior can provide a validation set for evaluating the performance of alternative methods and models. Further observation could provide a test set for evaluating the performance of the selected model.

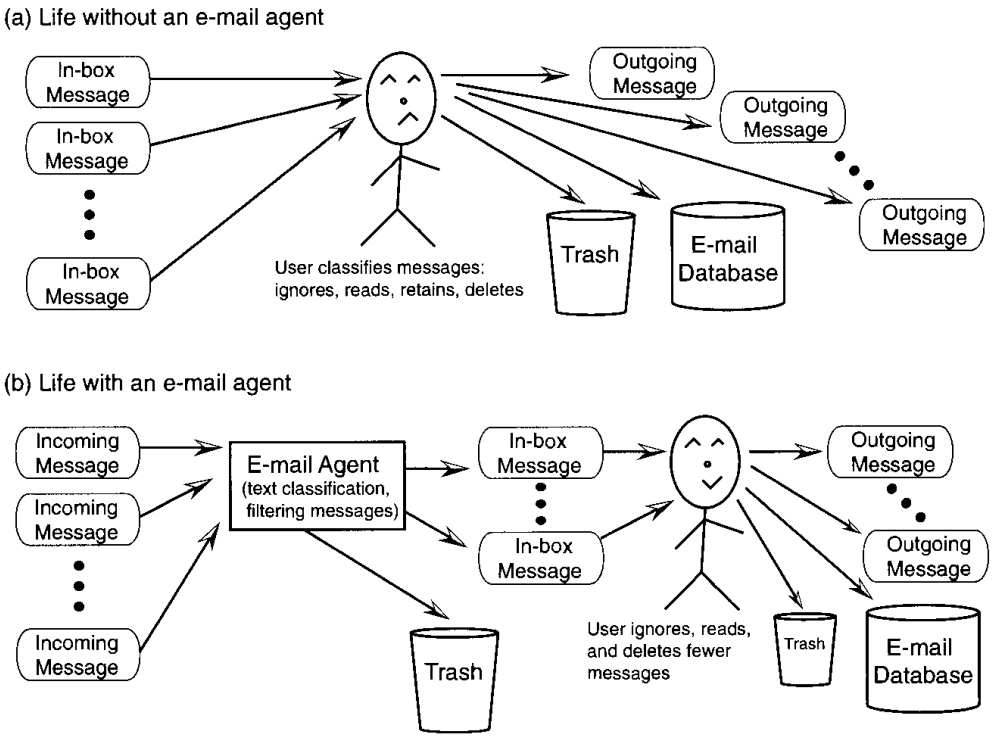
Why should we develop a text classification for e-mail? Because it can improve productivity of office workers. Deployment of the system could take the form of an intelligent e-mail agent customized for each office worker. Figure 4.4 depicts an office worker's life before and after the introduction of an e-mail agent.

4.3.2 Example: Electronic Business Library

Turning to the problem of business information management, we note that business libraries, like libraries in general, have changed with information technology. The physical library of books, periodicals, and corporate reports is being overtaken by a plethora of online information sources, both internal and external to the firm. As the amount of machine retrievable information grows daily, the need for efficient indexing and cataloging grows with it. Technologies for text categorization can help information specialists to build electronic libraries that serve user needs.

Recognize that the generic text categorization problem is a series of binary classification problems. That is, each document has the potential of being classified under many topical categories. Suppose the pharmaceutical firm Pfizer sends a letter to the drug retailer Walgreens. The letter reviews pricing policies as well as recommended pharmacist warning labels to be distributed to consumers of an ethical drug XYZ. Walgreens scans the letter and stores it as a document in its electronic library. The document could be classified under "Pfizer," "pricing," "pharmacist warnings," and "XYZ," among

Figure 4.4 An Electronic Mail Agent with Text Categorization



other indexing keys. Initial indexing and categorizing work is the job of clerical workers, information specialists, or business librarians; it is a manual process.

To automate the process of maintaining an electronic library, business information specialists must first devise an index or categorization scheme. They should rely upon an understanding of the information flows into and out of the organization and of the information needs of library users, including employees, suppliers, and customers. The index or categorization scheme should be comprehensive, so that every document may be assigned at least one key, and it should provide sufficient detail to accommodate the information retrieval needs of users.

After an index has been defined, information specialists or corporate librarians can build training, validation, and test sets of documents, assigning appropriate keys to each document. A letter from Pfizer gets the key "Pfizer" but not the key "Johnson & Johnson." It gets the key "XYZ" but not the key "aspirin." And so on. The set of training documents should be sufficiently large to have multiple instances of every indexing key.

Indexing keys serve the role of binary response variables. Automated text categorization utilizes classification models fit to the text data (represented by a term-by-document matrix, for example). Using the set of training documents, we fit a separate model for each of the keys in the index. Like the e-mail versus junk mail problem, supervised learning methods may be used for the classification models.

4.3.3 Example: Managing Information Overload

The text categorization examples of e-mail classification and business libraries presume supervised learning environments in which training documents are categorized by human experts. Now we consider a generic unsupervised learning problem.

Suppose that a support engineer for a software firm is asked to monitor the activity in a user chat room. In particular, she is asked to observe user discussions for a month and prepare a summary report for management identifying primary discussion threads or themes.

Hundreds, perhaps thousands, of chat room messages are logged each month as users, both satisfied and dissatisfied, express their opinions about the firm's software products. Reading the entire transcript from the chat room could take many hours. Identifying primary threads or themes could be a difficult task. The support engineer wonders whether it might be possible to use text mining tools to organize, explore, and summarize the domain of user opinion in the chat room.

Think of chat room messages as entries in a document collection or corpus. Construct a term-by-document matrix. Identify or define attributes that may be used to differentiate among documents, such as document length, word choice, or other text measures. Augment the term-by-document matrix by adding rows for these measures. The columns of the augmented term-by-document matrix represent numeric vectors describing the documents; these constitute input data for unsupervised learning methods.

Just as we can use multivariate methods like principal components and multidimensional scaling to develop product positioning maps from consumer surveys, we can use multivariate methods to develop semantic maps from chat room text. Berry and Browne

(1999), for example, describe how singular value decomposition may be used to analyze rectangular term-by-document matrices. Lebart (1998) shows how the multivariate technique of correspondence analysis converts frequency counts in a term-by-document matrix into maps of terms and documents, providing a visualization of textual data.

Just as we can use cluster analysis to identify product classes or consumer segments, we can use cluster analysis to identify semantic clusters or groups of documents with common themes. Messages within a cluster are more like one another than to messages in other clusters. If we can identify messages at the center of each cluster, we may be able to provide examples of prototypical chat room messages, those most descriptive of the major threads or themes of discussion.

The support engineer's job in this example—reading, organizing, and summarizing a large body of text—is not unlike the job of many knowledge workers. Unsupervised learning methods for text categorization, properly employed, help people to cope with the problem of information overload. Furthermore, knowledge workers can use unsupervised learning methods without preconceived notions about the nature of the information. They let the text define the categories and the structure of the summary.

Automatic text summarization is another area of research and development that can help with information management. Imagine a text analysis program with the ability to read each document in a collection and summarize it in a sentence or two, perhaps quoting from the document itself. To our hypothetical e-mail agent, categorizing messages as spam or normal messages, we add a component that reads the normal messages, summarizes them, and organizes them for review by management. Wishful thinking? Perhaps. But, given developments in the area of automatic text summarization (Mani and Maybury 1999), intelligent systems like this are becoming a reality.

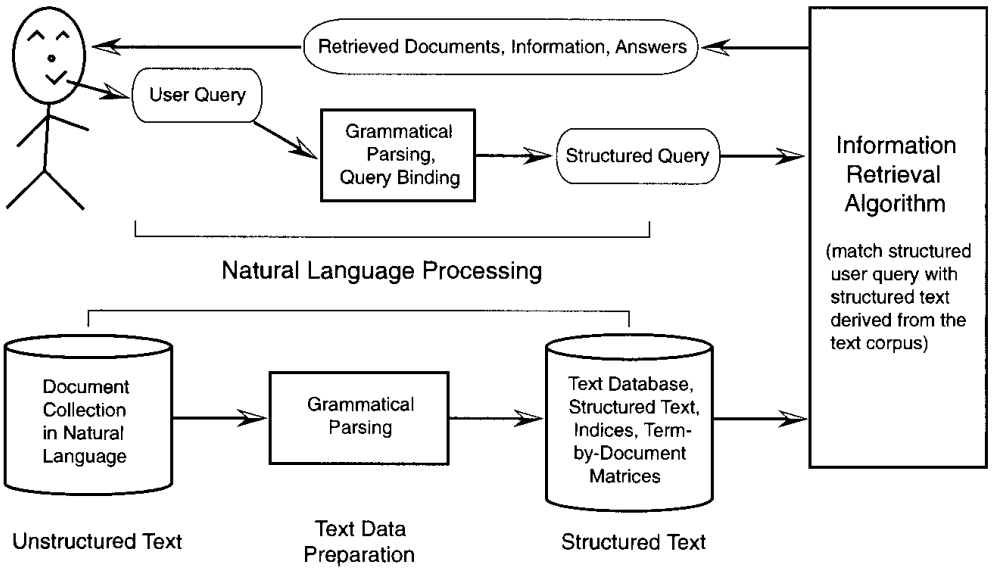
4.4 INFORMATION RETRIEVAL

Users of the Internet and World Wide Web are familiar with search engines that look across the expansive domain of the World Wide Web to find matches for user queries. This is an information retrieval operation.

Information retrieval systems are search engines. Figure 4.5 shows the typical structure of an information retrieval system. The fundamental task involves matching user queries with documents. If the user query is in natural language, the system must first convert the user query into a structured query; this is a grammatical parsing or "query binding" operation, as it is sometimes called. The information retrieval system also relies upon the prior text processing of the corpus, converting natural language documents into structured text.

We want search engines to have high recall, identifying all documents relevant to an information request. Search engines with high recall may also retrieve documents that are irrelevant or only indirectly related to the information request. A search engine that retrieves only documents that are relevant to an information request is said to have high precision. When using a high precision engine, however, we run the risk of missing relevant documents. Witten, Moffat, and Bell (1999) note that the tension between

Figure 4.5 Components of an Information Retrieval System



the dual objectives of high recall and high precision has been an enduring theme in information retrieval research.

Information retrieval systems or search engines perform a matching function; they attempt to match the words used in a user's query with words used in documents. Various methods may be used to accomplish the matching. Each user query may be thought of as a vector of search terms which may be compared with vectors of terms for documents being searched. A simple Boolean or logical method involves a matching of terms used in the user's query with terms used in documents. Some search engines complement Boolean matching with references to knowledge bases, dictionaries, or patterns in user queries. Vector space models and various text classification algorithms can facilitate the search.

What about the information retrieval or matching algorithm itself? Is is necessary to compare the user's structured query with the structured text of each and every document? No. We can build upon methods of text categorization as reviewed in the previous section to simplify the information retrieval and matching process.

Text categorization and methods that assess document affinities or similarities are useful preliminaries to information retrieval. When a user asks for information, her query, along with the documents of the corpus, may be parsed, categorized, or positioned within the semantic space of the corpus. Documents in the same text category as the user's query or close in space to the user's query are those most likely to satisfy the user's information needs. Retrieve one document in a class of documents, and the others can follow automatically. Retrieve one document, and nearby documents (in the semantic

space) can follow automatically. The matching algorithm, then, need not and should not entail a document-by-document comparison with the user query.

Another matching method is link or network analysis. Information about links can be incorporated in Web searches. Links show interrelationships among Web pages; they help us to find additional Web pages relevant to a user query. Hub and authority pages, for example, play special roles in Web searches. A hub page points to many other pages, and an authority page is pointed to by many other pages.

Information retrieval is an active area of research and development. We should expect more efficient search engines in the future, as natural language and text mining algorithms find their way into commercial offerings. Future systems are likely to employ interaction between search engines and user analysts. There will be natural language capabilities on the user side in the parsing of queries and on the database side in the organization and analysis of text. Insightful's InFact system, for example, permits the user to enter requests as natural language queries and employs technologies from computational linguistics.

Search engines like Google and AltaVista are based upon simple word matching operations, importance coding, or the linking of the documents or Web pages across the World Wide Web. We need to go beyond the capabilities of today's search engines. To quote Herbert Simon:

> Our search engines must become more intelligent so that they can select and filter from the forests of information the particular items our eyes need to see. They must not only be able to respond to our specific requests, but also to use broader knowledge of our needs to retrieve important information we have not asked for and may be surprised by. (Simon 2002, p. xvii)

In the future we will see intelligent applications and agents that perform syntactic processing and meaning-based Web searches. Syntactic modeling provides the ability to do relationship searches in addition to keyword searches. For most modern languages the order or configuration of words within a sentence is important to understanding the meaning of the sentence.

In English it is customary to place the subject before the verb and the object after the verb. In English verb tense is important. The sentence "Tom carries the Apple computer," can have the same meaning as the sentence "The Apple computer is carried by Tom." "Apple computer," the object of the active verb "carry" is the subject of the passive verb "is carried." Understanding that the two sentences mean the same thing is a first step in the direction of building intelligent search engines.

Programs with syntactic processing capabilities, such as Insightful's InFact, provide a glimpse of what search engines may become in the future. Such programs perform grammatical parsing with an understanding of the roles of subject, verb, object, and modifier. They know parts of speech (nouns, verbs, adjective, adverbs). And, using identified entities, representing people, places, things, and organizations, they can perform relationship searches. Working with a syntactic processing program and a business news database, for example, a user might ask to identify all activities (verb forms) relating

Coke and Pepsi. Retrieved documents could be quite useful as sources of competitive intelligence.

Facilitated by standards for information interchange like XML, the World Wide Web holds promise as a public information store, a vast repository of machine-retrievable and machine-understandable data. In the future we will be "spinning the Semantic Web," as described by Daconta, Obrst, and Smith (2003) and Fenzel et al. (2003). Text mining methods will be important in this information-rich environment, as we move from information retrieval to knowledge discovery.

4.4.1 Example: Searching a Document Archive

Text categorization and information retrieval often go hand in hand. The objective is to use the power of the computer to organize large bodies of text (text categorization) and to find answers from the text (information retrieval). In writing their book about the marketing research and information services industry, Miller and James (2004) had access to the electronic archive of *Inside Research*, a monthly newsletter about the marketing research industry written by Jack J. Honomichl and Laurence N. Gold. There were basic questions to be answered. How had the marketing research industry changed in recent years? What had been the history of mergers and acquisitions? What could be learned about factors of production? What about sales revenues and the demand for research and information services? The authors wondered whether answers could be found in the archive.

Searching the entire *Inside Research* archive for answers seemed a daunting task given the amount of detail and lack of structure in the original documents. The archive contains articles of interest to buyers and sellers of marketing research and information services. The archive available to Miller and James (2004) began with the first issue from January 1990 and continued through December 2003. Originally distributed as a set of Microsoft Word files, one for each year of publication, the *Inside Research* archive contained more than two thousand printed pages and a million words of text. A small body of text by text mining standards, *Inside Research* presented an interesting technical challenge because of its organization as many individual news items covering a wide range of content in each monthly issue. For the most part, documents were unstructured. Some paragraphs and sections of documents had titles; others did not. There was no comprehensive index for the archive.

The first step in analyzing the *Inside Research* archive was to organize the document collection into discrete units for analysis. Miller and James (2004) used text editors and a Perl program to convert the original Microsoft Word files, organized as newsletters, into ASCII text documents. Some documents represented long articles on selected marketing research topics. Others were short, such as one-sentence paragraphs about events affecting individuals and firms in the marketing research industry. Graphics were not part of the archive, but documents could include tables as well as paragraphs of text. Extensive text editing yielded a document collection of around seven thousand ASCII text documents.

How could one extract meaning from the *Inside Research* document collection? Faced with the task of reviewing large quantities of textual information, Miller and James (2004) turned to text categorization and information retrieval tools developed by Insightful Corporation. When the authors wanted to find documents related to pricing, they searched the document collection for "pricing." When they wanted to find news items about the ACNielsen company, they searched the collection for "ACNielsen." The text categorization system provided an efficient mechanism for finding and verifying facts about individuals and companies, as well as for finding feature articles on selected topics.

An interesting feature of the system employed by Miller and James (2004) was its ability to learn from user feedback. With each request for information, the system returns what its algorithms indicate are the most relevant documents. Then the analyst has the opportunity to rate documents according to her own judgment of their relevance. Analyst ratings are then available to the text categorization program, so they can be used in future information requests. Relevancy feedback, interaction between computer algorithms and human judgment, offers considerable promise for the future of text mining.

4.4.2 Example: Web Spiders for Competitive Intelligence

The Society of Competitive Intelligence Professionals (SCIP) provides this definition of competitive intelligence (CI):

> A systematic and ethical program for gathering, analyzing, and managing external information that can affect your company's plans, decisions, and operations. Put another way, CI is the process of enhancing marketplace competitiveness through a greater—yet unequivocally ethical—understanding of a firm's competitors and the competitive environment. (Society of Competitive Intelligence Professionals 2003)

Competitive intelligence professionals ask questions. Who is the competition, and what are they like? What is their financial position? What is their market share? What about business objectives? Do they have strategic alliances with other firms? What products and services do the competitors offer? What are their prices? Are there new competitors on the horizon? Competitive intelligence professionals answer these questions by gathering and analyzing information from public sources.

The Internet and World Wide Web have transformed the landscape of competitive intelligence by providing a vast store of publicly available information. Corporate Web sites and online reports, user chat rooms, news articles, and reports about products and prices from online shopping sites are sources of competitive intelligence. It is easy to retrieve information about competitors and competitive products. The challenge is to sort through mounds of retrieved information, identifying the most relevant documents for management.

Web spiders, also called crawlers, wanderers, Webbots, or bots, are software programs that search the Web for relevant information. They do this by following universal resource locator (URL) links from one Web page to another. Spiders usually work in

conjunction with scrapers, which gather information from Web pages. The general idea is to replace manual searching by analysts with automated searching by computer.

An exemplary Web spider for competitive intelligence is the CI Spider program developed by Chen, Chau, and Zeng (2002). To use the CI Spider, the researcher provides a starting list of Web addresses, such as home page URLs of known competitors. The researcher enters search terms, such as "products" and "prices," and constraints, such as limiting the search to commercial sites, excluding education and government sites.

Working from the initial list of Web addresses and employing a breadth-first algorithm that searches widely across sites before it searches deeply within sites, the CI Spider follows all possible links, finding Web pages that match the researcher's search terms and constraints. The search stops after the spider has identified a specified number of relevant Web pages. The result of the search is the list of URLs for these Web pages. The researcher can click on the URLs to view complete content of the pages.

In experiments with student users, Chen, Chau, and Zeng (2002) observed that the CI Spider program performed better than both Lycos searches constrained by Internet domain and analyst-driven, within-site searches. CI Spider is one of many programs for automatic Web searching (Rasmussen 2003; Zhong, Liu, and Yao 2003; Bar-Ilan 2004; Chen and Chau 2004; Hemenway and Calishain 2004). Sullivan (2000) discusses text mining methods for competitive intelligence research, referring to a number of commercially available programs. General reviews of the competitive intelligence landscape have been provided by Fuld (1994) and Bergeron and Hiller (2002).

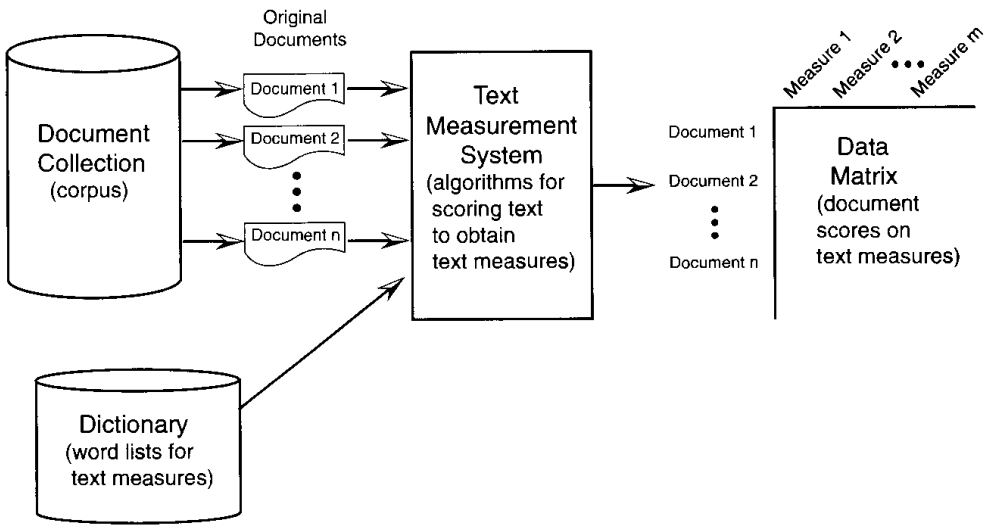
4.5 TEXT MEASURES

What are text measures? They are scores on attributes that describe text. Each document in a collection can be assigned scores. Measurement, in its most basic sense, is the assignment of numbers to attributes according to rules. Text measures can be used to assess personality, consumer preferences, and political opinions, just as survey instruments can. The difference between text measures and survey instruments is that text measures begin with unstructured text as their input data, rather than forced-choice questionnaire responses.

The term "text measures" may be new, but many examples of this type of text analysis exist, going under names such as content analysis, and thematic, semantic, and network text analysis (Roberts 1997; Popping 2000). Text analysis has seen a wide range of application within the social sciences, including the analysis of political discourse. West (2001) notes growing interest in the field of content analysis in recent years.

An early computer implementation of content analysis is found in the General Inquirer program (Stone et al. 1966; Stone 1997). Buvač and Stone (2001) describe a recent version of the program, which provides text measures based upon word counts across numerous semantic categories. Some of the more popular categories relate to bipolar dimensions identified in Charles Osgood's semantic differential research (Osgood, Suci, and Tannenbaum 1957; Osgood 1962), including positive–negative, strong–weak, and active–passive dimensions. A more recent example of computerized content analysis is

Figure 4.6 Components of a Text Measurement System



the DICTION program (Hart 2000b; Hart 2001) for analyzing the tone of text messages, as discussed in an example below.

Text measures involve a scoring of documents based upon predefined measurement categories or methods. Figure 4.6 shows a generic text measurement system. Text measures flow from a measurement model (algorithms for scoring) and a dictionary, both defined by the researcher or analyst. A dictionary in this context is not a traditional dictionary; it is not an alphabetized list of words and their definitions. Rather, the dictionary used to construct text measures is a repository of word lists, such as synonyms and antonyms, positive and negative words, strong and weak sounding words, bipolar adjectives, parts of speech, and so on. The lists come from expert judgment about the meaning of words. The text measurement system scores documents from the collection by using scoring algorithms and the dictionary.

Popular, but perhaps overly simplistic, examples of text measures are the "What sucks? What rocks?" programs for gauging public opinion through Web searches. Jon Orwant and Steve Lidie offered "What Languages Suck" programs, referring to programming languages. Don Marti developed what he called an "Operating System Sucks-Rules-O-Meter" that runs daily over the Web. A review of the logic behind these systems is provided by Dan Brian (2003).

"What sucks? What rocks?" systems utilize search engines like Google or AltaVista to gather text examples of the items being evaluated. Then they employ grammatical parsing programs, often written in Perl, to convert the natural language in the text examples into regularized expressions for analysis. Finally, text measures are employed; that is, the text examples for each item being evaluated are scored for the number of

hits associated with sets of bipolar verbs or adjectives, such as sucks-rules, bad-good, stupid-brilliant, and so on. The public-domain systems on the Web gather data on a regular, usually daily, basis and report summary data (usually frequency counts) for the items or objects being judged.

Nothing is to prevent us from using similar mechanisms for doing serious consumer and public opinion research. We can refine bipolar text scoring methods so the resulting measures are more trustworthy. We can use a series of bipolar adjectives to develop multivariate text measures or profiles for objects being judged. In addition, we can establish norms for text measures so that objects may be judged relative to other objects within the same category, and we can compare text measures across objects and time. The Web Text Mining case in Appendix A provides an exercise along these lines. Written in the Perl language, natural language and text scoring programs for the case are available in the public domain. With a little work these programs could be adapted to conduct meaningful business and market research.

A related area of text analysis (and a research stream with a long and storied history) concerns document word counts, literary styles, and authorship identification (Zipf 1949; Oakman 1980; Mosteller and Wallace 1984). While the objectives of historical and literary studies have little to do with contemporary business applications of text mining, methods of linguistic and literary analysis contribute to our general understanding of text mining methods. Characteristic of these studies is the scoring of documents, the creation of text measures.

4.5.1 Example: Text Measures for a Political Document Collection

To get a feeling for what text measures can provide, we examine the political research of Roderick P. Hart (Hart 2000a) in the book *Campaign Talk: Why Elections Are Good for Us*. To conduct this research, Hart had to prepare a representative text database, identifying political documents and storing them in electronic form. He also had to develop text measures, constructing dictionaries for the scoring of documents.

As with many data and text mining applications, the hard part, the time-consuming part, is data and text preparation. Preparing the document collection for Hart's (2000a) research was a monumental task. His research covered the thirteen U. S. presidential campaigns from 1948 to 1996. He collected text from press releases (7,309 documents) and broadcast media (1,219 documents), transcripts of campaign speeches (2,357 documents), political advertisements (553 documents), and letters to the editor from newspapers (6,126 documents). Documents were converted to electronic media and organized by time (election and campaign cycle), candidate, and political voice (people, politicians, and press).

Print media text came from *The New York Times*, *Washington Post*, *Christian Science Monitor*, *Atlanta Constitution*, *Chicago Tribune*, and *Los Angeles Times*, as well as from Associated Press and United Press International reports. Text for the broadcast media was gathered from photocopies of scripts and transcripts of audiotapes from the nightly news. News bureaus from ABC, CBS, NBC, CNN, and PBS were represented. Campaign speeches included all nationally broadcasted addresses during general elections as well

as selected campaign stump speeches. Political advertisements represented a sampling of advertisements across seven of the thirteen campaigns in the study. Finally, letters to the editor were drawn from newspapers representing twelve major cities across the United States.

The text measurement system for Hart's research draws upon the DICTION program and a dictionary of around ten thousand words arranged in thirty-three disjoint lists (Hart 2000a, 2000b, 2001). The scoring algorithm relies upon simple word counts. It assumes semantic independence, counting each word occurrence equally, regardless of context or grammatical role.

Five general text measures summarize the tone of documents along dimensions called certainty, optimism, activity, realism, and commonality (shared values). Each measure depends upon counting words that correspond to selected word lists. Some words receive positive weights, others negative weights. The optimism measure, for example, assigns positive weights to words relating to praise, satisfaction, and inspiration, while recording negative weights for words relating to blame, hardship, and denial. The realism measure draws upon the concept of familiarity. A text measure in its own right, familiarity is computed with reference to a dictionary of forty-four words that are regarded as the most common words in the English language.

Hart's research demonstrates how text measures may be used to understand large and diverse text sources. Reviewing the voices of the three main groups (politicians, press, and public), he observed increasing complexity (lower familiarity scores) over time. Across the entire period of the study, the voice of the press was decidedly negative in tone (low in optimism), compared with politicians and the public. Normalcy in speech, if not in policy, is rewarded in American politics. Among political candidates, centrist speakers, those who spoke in common parlance, were more successful than non-centrist candidates.

Judging from the text sources used in Hart's studies, there is great variability in the way people talk about politics. Hart (2000a) identified three general groups of political speakers: (1) "pundits," who focus upon campaign dynamics, (2) "traditionals" concerned about national values, and (3) "functionals," the largest group, who avoid both pundit and traditional language in order to focus upon community problems.

Hart's contribution to the analysis of political discourse lies in his reliance upon original text sources. Whereas many political analysts spend their time putting a "spin" on what people say, Hart lets people speak for themselves. While many analysts rely upon political polls to assess public opinion, Hart relies upon the actual words of the people. Reviewing his work with content analysis over the years, Hart (2001) writes,

> I have found virtually every stereotype about computerized content analysis to be untrue. It is alleged to be mechanical, but I have found it to be creative. It has been decried as oafish, but I am fascinated by its subtlety. It is said to be reliable but not valid, and yet I see its validity as its greatest strength. It is said to be reactive, colorless, and arcane; I have found it to be heuristic, exciting, and altogether normal. (Hart 2001, p. 43)

4.5.2 Text Measures and Qualitative Research

Qualitative research is currently a labor-intensive enterprise. Whether conducting in-depth interviews or focus group discussions, the moderator participates in a conversation. Talking and listening take time. Spoken words, recorded on video or audio tape, are often transcribed into text, another time-consuming process. The analysis of qualitative data is equally labor-intensive, with the researcher reviewing tapes and transcripts and turning the unstructured words of others into a coherent story, a report of research findings.

Many software tools designed for the qualitative data analyst assist in the organization process. They help the analyst to keep track of the details of text analysis and to restructure an analysis quickly when new text categories or theories become apparent. Tools such as Atlas/ti and NUD.IST are analyst-driven. They automate the clerical aspects of analysis, the tasks of data organization and record keeping, but are not text mining tools. Reviews of such tools have been provided by others (Kelle 1995; Weitzman and Miles 1995; Fielding and Lee 1998; Popping 2000; Weitzman 2000). Geisler (2004) discusses text data preparation and issues involved with the manual coding of transcripts from qualitative research.

Much qualitative research is small-sample research, which is not surprising given the time it takes to collect and analyze text. This situation could change as software tools emerge that automate the process of qualitative data collection and analysis. Miller and Walkowski (2004) suggest possibilities with focused conversations and qualitative research online. When we have large bodies of qualitative data focused upon the same subject, then we have an opportunity to explore additional possibilities of text mining.

Text measurement holds promise as a technology for understanding consumer opinion and markets. Just as political researchers a la Hart (2000a) can learn from the words of the public, press, and politicians, business researchers can learn from the words of consumers, press, and competitors. Commercial advertisements can be analyzed just as political advertisements. User group postings can be analyzed just as letters to the editor in newspapers.

The work of content analysts and qualitative researchers in the social sciences can serve as a model for business and marketing research. We are encouraged to understand consumers in their own words, to understand the competitive landscape in the words of competitors, and to use the power of the computer to do the work of text analysis.

Words and numbers—that's what we have. Ideas about the business world are expressed in words. Through measurement we find ways to convert words to numbers. Data analysis, traditional and data-adaptive, converts millions of numbers to thousands of numbers, and thousands to a few. At the end of analysis we look at summary numbers and models fit to the data, and we turn those numbers and models into words, interpreting results, telling a story to managers. This is, of course, an oversimplification of the research process. But it is interesting to think of research in terms of numbers and words. It helps to put data and text mining into perspective. Quantitative researchers go from words (theories and survey instruments) to numbers and from numbers to words (reports of results). Qualitative researchers, by contrast, deal almost entirely with words.

4.6 FURTHER READING

Data mining is quantitative research—a numbers game. Text mining is also a numbers game, but with words rather than numbers as the raw input. As computers, natural language processing, syntactic modeling, and text analysis algorithms become more capable, we will see new and exciting applications of text categorization, information retrieval, and text measures.

For those interested in learning more about text mining, reviews may be found in Trybula (1999), Witten, Moffat, and Bell (1999), Meadow, Boyce, and Kraft (2000), Sullivan (2001), Feldman (2002), and Sebastiani (2002). Hausser (2001) gives an account of generative grammar and computational linguistics. Statistical language learning and natural language processing are discussed by Charniak (1993) and Manning and Schütze (1999). The writings of Steven Pinker (1994, 1997, 1999) provide insight into grammar and psycholinguistics. Maybury (1997) reviews data preparation for text mining and the related tasks of source detection, translation and conversion, information extraction, and information exploitation. Detection relates to identifying relevant sources of information; conversion and translation involve converting from one medium or coding form to another.

Meadow, Boyce, and Kraft (2000) and the edited volume by Baeza-Yates and Ribeiro-Neto (1999) provide comprehensive reviews of computer technologies for information retrieval. Belew (2000) and Berry and Browne (1999) discuss technologies relevant to Web search engines, and Huberman (2001) reviews patterns of Web user activities. Merkl (2002) provides discussion of clustering techniques, which explore similarities among documents and group documents into classes. These can be quite useful to analysts and managers facing a glut of textual information. Dumais (2004) reviews latent semantic analysis and statistical approaches to extracting relationships among terms in a document collection.

The edited volume by Denzin and Lincoln (2000) provides an overview of qualitative research methods. Antecedents of text measures may be found in the text analysis methods of qualitative researchers, as described in Schrott and Lanoue (1994), Roberts (1997), and Silverman (2000, 2001). Also relevant are works dealing with applications and methods of content analysis, including computer-assisted methods (Popping 2000; Neuendorf 2002; West 2001).

Pfeifer (2003) reviews information retrieval systems based upon the public-domain language Perl. For those interested in programming for text preparation and text processing, Perl is often the language of choice. Introductory Perl references include Christiansen and Torkington (2003) and Schwartz and Phoenix (2001). For more advanced programming ideas, the reader can refer to Wall, Christiansen, and Orwant (2000), Conway (2000), Cross (2001), Quigley (2002), and Schwartz and Phoenix (2003). The online Perl community (`http://www.perl.org`) and *The Perl Journal* are excellent sources of Perl programs and lore. Edited volumes by Jon Orwant (2003a, 2003b) contain articles from *The Perl Journal.* Perl is often used for dynamic Web applications (Radcliff 2002), including those involving XML (Ray and McIntosh 2002; Riehl and Sterin 2003) and Web services (Ray and Kulchenko 2003).