4

Integrating Unstructured Text into the Structured Environment

The world of computing has grown from a small, unsophisticated world in the early 1960s to a world today of massive size and sophistication. Nearly every person worldwide—in one way or the other—is affected by or directly uses computation on a daily basis. Nothing less than national productivity from the 1960s to the present has been profoundly and positively affected by the widespread growth of the use of the computer.

The growth of computing can be measured in two ways: growth in structured systems and growth in unstructured systems.

Possibilities of Unstructured Systems

Structured systems are those for which the activity on the computer is predetermined and structured. Structured systems are designed by, built by, and operated by the IT department. ATM transactions, airline reservations, manufacturing inventory control systems, and point-of-sale systems are all forms of structured systems.

Structured systems are tied closely with the day-to-day operational activities of the corporation. Because of this affinity, structured systems grew

quickly. Cost justification and return on investment for structured systems came easily because of the close tie-in with the day-to-day business of the corporation. The growth of the structured environment was fueled by the desire of the business world to be competitive and streamlined.

Unstructured systems are those that have no predetermined form or structure and are full of textual data. Typical unstructured systems include emails, reports, contracts, transcripted telephone conversations, and other communications. When a person does an activity in an unstructured environment, he is free to do or say whatever he wants. The person doing the communication can structure the message in whatever form is desired, using any language. In an unstructured environment, the communication can range from a proposal of marriage to a notification of a layoff to the announcement of the birth of a baby, and everything in between. There simply are no rules for the content of unstructured systems.

The growth of the unstructured environment has been fostered by the needs for communications, informal analysis (such as that found on a spreadsheet), and personal analysis (of finances, personal goals, personal plans). There was (and is) a different set of motivations for the growth of the two environments. Figure 4.1 shows the different environments.

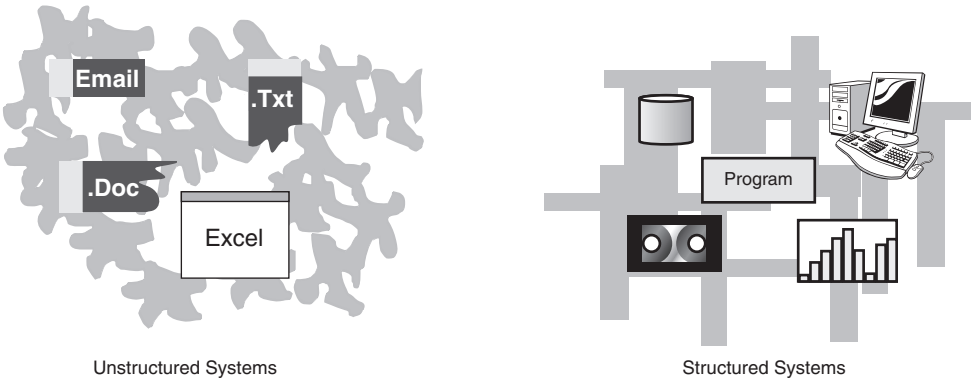

Figure 4-1 The two basic forms of data

From the beginning, the worlds of structured systems and unstructured systems have grown separately and apart and yet—at the same time—parallel with each other. It is no surprise that today each environment is separate from the other environment in many ways:

- Technologically
- Organizationally

- Structurally
- Historically
- Functionally

In truth, there is little overlap or connection between the two worlds.

Imagine what the world would look like if, indeed, there was overlap (or intersection) between the two environments. Imagine the possibilities if the two worlds could connect in an effective and meaningful way, the new types of systems that could be built, the new opportunities for the usage of computation, and the enhancements to existing systems in ways that are not possible using technology. When one accepts the limitations of today's technology and today's environment, there are only so many things that can be done. Imagine what would happen if those limitations suddenly disappeared.

If a bridge is to be built between the two environments, it makes sense to bring the unstructured text to the structured environment. In doing so, the decision support analyst can take advantage of the analytical processing capabilities that exist in the structured environment.

In most organizations an analytical infrastructure exists in the structured environment. This environment consists of things such as a database management system (DBMS), Business Intelligence (BI) software, hardware, and storage. Organizations have already invested millions in their analytical environment. The existing analytical infrastructure serves only structured systems, however. Data has to be put in a structure and a format that is particular and disciplined. Despite the particulars of the existing analytical infrastructure environment, it is less expensive to bring the unstructured data to the existing analytical infrastructure environment than it is to reconstruct the analytical infrastructure in the unstructured environment. By bringing unstructured data to the existing analytical infrastructure environment, the organization can leverage the training and the investment that has already been made in the existing analytical infrastructure environment.

When the gap between unstructured data and structured data is bridged, an entirely new world of possibilities and opportunity for information systems opens up. Figure 4.2 shows that a bridge between the structured and unstructured environments has many benefits.

The possibilities for new systems blossom when the gap between unstructured data and structured data is crossed. There are enormous and new opportunities that arise when the two types of data are merged.

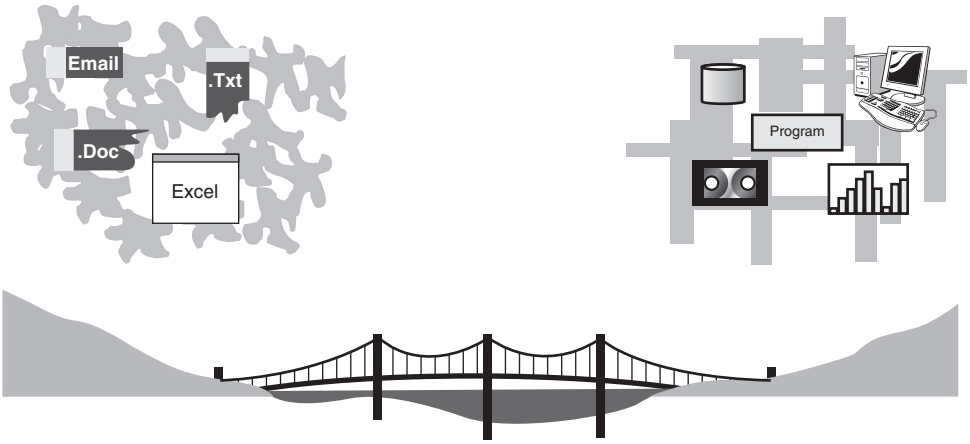

Figure 4-2 Forming the bridge between the structured and unstructured world

Integrating Unstructured Textual Data

In second generation textual analytics, the key to crossing the bridge between the two worlds is the integration of unstructured text before it is sent to the structured environment. Raw unstructured text cannot simply be placed into the structured world and still be meaningful and useful. Stated differently, unstructured text placed directly into a structured environment creates a mess. There is too much data—data that has different meanings and is recorded as a single name, alternate spellings, extraneous words, and documents that have no bearing on business. All these limitations of unstructured text become manifested when unstructured data is moved whole cloth into the structured environment.

To be effective, unstructured text must be integrated before it can be moved into the structured environment. By integrating unstructured text, the bridge between structured and unstructured data is created, and the stage is set for textual analytics.

Reading the Unstructured Textual Data

The first step in the integration of unstructured text is the physical reading of the text. To be integrated, raw text must first be read or "ingested."

In some cases, the text first appears in a paper format. In this case, the text on the paper must be read—scanned—and the text converted to an electronic format. This process is typically done in optical character recognition (OCR). There are quite a few challenges to this process of lifting text from a paper foundation:

- Sometimes the paper is old and brittle and is destroyed by the process of trying to read it. In this case, the analyst must not count on reading the paper more than once.
- Sometimes the print font on the paper is not easily recognizable by the scanner. In this case, there are a lot of manual corrections.
- Sometimes the scan process reads and interprets the words incorrectly.

As a rule, the process of converting from paper to electronics is one that involves a manual scan and correction after the electronic scan is done, if for no other purpose than to make sure the electronic scan is successful. In many cases, manual corrections must be made when the scanning and conversion process has made an error or the electronic scan process has made assumptions about what is read that are not true.

However it is done, the text needs to be lifted from the paper media and converted into an electronic format.

Then there is the case of voice recordings. Like data found on paper, voice data likewise needs to be lifted from the media in which it was stored and reset into an electronic format that is intelligible to a program that reads and analyzes text. Voice recordings can be converted to an electronic format by means of voice character recognition (VCR). Text can be lifted from VCR as well. The issues of quality and reliability for VCR are similar to OCR considerations.

Choosing a File Type

When the text is in an electronic format, the format and structure of the text needs to be taken into account. Some of the typical formats for the reading of electronic text follow:

- .pdf
- .txt
- .doc
- .ppt
- .xls
- .txt compatible
- Lotus
- Outlook

Often the vendor supplies software to read these file types. However, often the vendor does not guarantee a 100% successful reading. For this reason, third-party vendors supply software and software interfaces that are more efficient and more reliable than those supplied by the vendor. It is true that you have to pay for third-party solutions.

However, the third-party solutions are more reliable and more efficient than the vendor-supplied solutions. Also, the third-party vendor has the responsibility of keeping up with the different releases of the base software as new releases are made.

ALTERING THE ORIGINAL SOURCE

One of the issues faced by the systems programmer is whether to allow the original source text to be altered. In some cases, the software reading a source file wants to add data to or otherwise alter the source text. In other cases, the source text is never altered. It is read, but not altered.

By far, the safest policy is never to alter the source text, even at the expense of having redundant copies of data lying around.

Reading Unstructured Data from Voice Recordings

In some cases, where the text does not reside on paper, the text resides on tapes. Typical of this usage of tapes are telephone conversations that are taped and then transcribed. In this case, the tapes must be converted into an electronic format, much like scanning data, except the scan is not text. Typical software in this case includes VCR. VCR technology has many liabilities associated with it. VCR is subject to being fooled by accents, by people talking too softly, and other issues. As a rule, if a transcription can be done with 95% accuracy, that is considered to be good.

It is an interesting point that humans do not hear and understand 100% of the words that are spoken. Our brains "fill in the blanks" frequently. So it is not unreasonable that VCR does not do a 100% job of accurate transcription.

However it is accomplished, the original source text must be read and entered into the component that will begin the process of textual integration.

After the source text has been read, the next step is to actually integrate the text.

The purpose of textual integration is to prepare the data for textual analytics. It is true that raw text can be subjected to textual analytics. However, the reading, integration, and preconditioning of the raw source text sets the stage for effective textual analytics. Stated differently, textual analytics can be done on raw textual data, but not effectively. The data itself defeats much of the purpose of textual analytics. To be effective, textual analytics must operate on textual data that has been integrated and preconditioned.

The Importance of Integration

It is not always obvious why raw text needs to be integrated and preconditioned before it is useful and most effective for textual analytics. The following cases make the point of why integration of text is a necessary precursor to effective textual analytics.

Simple Search

A simple search is to be conducted on the name "Osama Bin Laden." Operating on unintegrated data, the search fails to find references when the name "Usama Bin Laden" appears or the name "Osama Ben Laden" appears. If textual integration had been done properly, the search for "Osama Bin Laden" would have turned up all occurrences of all spellings of his name.

Indirect Search of Alternate Terms

Suppose an analyst wants to find all places where there is a mention of a broken bone. If the analyst searches for "broken bone," the analyst finds all the places where there are permutations of the term. However, if data is integrated first, an indirect search for "broken bone" turns up the many terms that also mean "broken bone." Operating on integrated data, an indirect search on broken bone finds "fractured radius," "lacerated tibia," "oblique fractured ulna," and so forth.

Indirect Search of Related Terms

In addition to looking for alternate terms, related terms can also be accessed by the textual analyst. Consider the term "Sarbanes Oxley." If a direct simple search is made on the term "Sarbanes Oxley," the search will turn up the many places where that term is found. Consider what happens when raw textual data is integrated before the search is done. An indirect search can discover the many terms that are related to Sarbanes Oxley. For example, when the raw text is integrated and a search is done on related terms, an indirect search on "Sarbanes Oxley" finds items such as the following:

- Contingency sale
- Revenue recognition
- Promise to deliver

Permutations of Words

Another interesting aspect of integrating text is the recognition of the roots of words. When raw unintegrated text is searched for the phrase "moving the needle," if that phrase is used anywhere, the search finds it. When raw text is integrated, permutations of the base word are recognized as well. For example, when a search is made for "moving the needle" on integrated text where the stems of words have been recognized, the results find the following:

- Moves the needle
- Moved the needle
- Move the needle

From these simple examples of analysis of text against raw textual data and integrated textual data, it becomes obvious that if you are going to do effective textual analytics, the data that will be operated on must first be integrated.

The Issues of Textual Integration

The kinds of issues that must be addressed in the integration of unstructured text into the structured environment include the following:

- **Determining if the unstructured document has any relevance to the business**—If the unstructured document is not relevant to the business conducted, the unstructured document does not belong in the structured environment as a candidate for textual analytics. Figure 4-3 shows that raw unstructured data is fed to the integration component. The integration component then screens the data based on business relevance. For example, an email that said "I love you, darling" would not be deemed to have business relevance and would not be placed in the textual analytical database.
- **Removing stop words from the unstructured environment which are extraneous to the meaning of the text**—Typical stop words are "a," "and," "the," "is," "was," and "which." Stop words are used to lubricate language, but add little or nothing to the subjects that are discussed. Figure 4-4 shows that stop words need to be filtered out so that they don't get in the way of arriving at the heart of the matter when it comes to analyzing unstructured text.

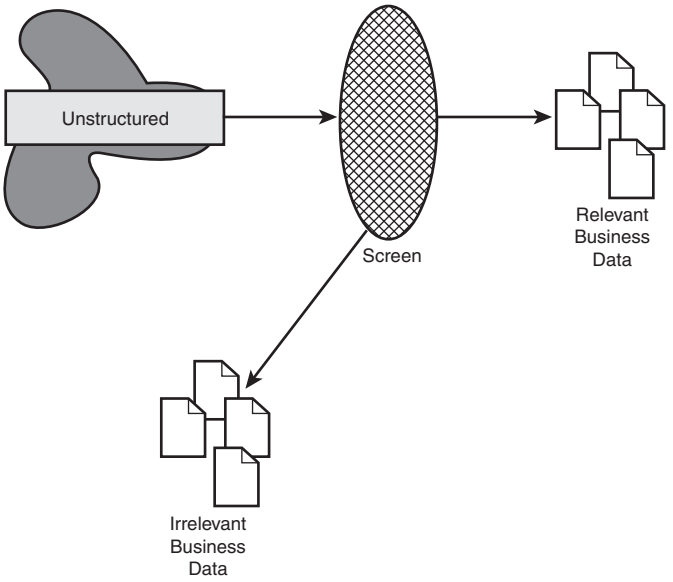

Figure 4-3 Relevant business data needs to be screened from irrelevant business data.

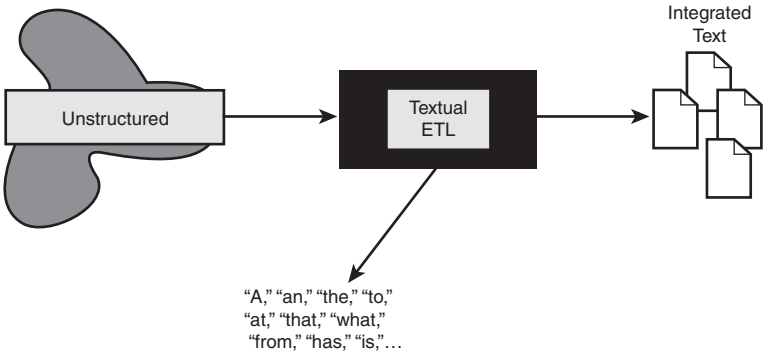

Figure 4-4 Stop words are removed.

- **Reducing words to their Greek or Latin stems**—By reducing the words found in unstructured text to a common stem, the commonality of words can be recognized when the words are literally not the same. Figure 4-5 shows that several ordinary words have a common Latin stem. The figure shows that the words "Moving," "Move," "Moved," and "Mover," all have a common stem—"Mov."

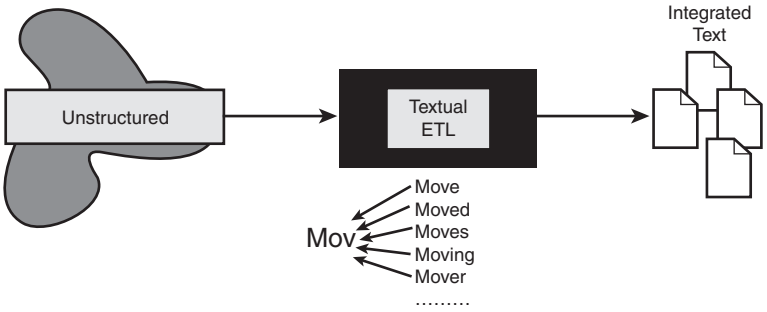

Figure 4-5 Words are reduced to a common stem.

- **Resolving synonyms**—Where there are synonyms, the reduction of the synonym to a common foundation allows for the possibility of a common vocabulary. It is only through the establishment of a common vocabulary that meaningful searches can be done. There are two basic ways for synonyms to be resolved. One of those ways is through synonym replacement. With synonym replacement, when a synonym is recognized, it is replaced by the more common (or more general) form of the word, as shown in Figure 4-6. The problem with synonym replacement is that after the replacement is made, a search for the original synonym cannot be made. After a word is replaced, it cannot be found by a search.

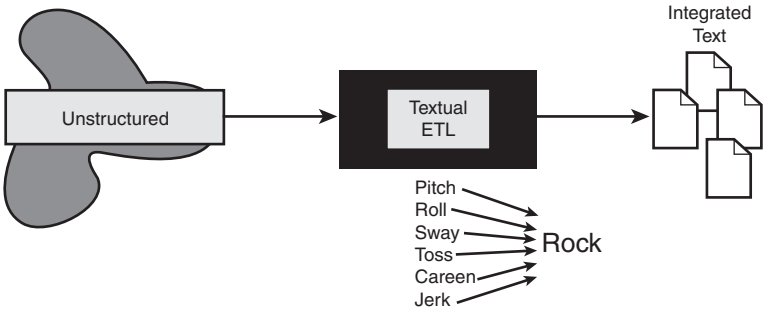

Figure 4-6 Synonym replacement is another activity that can be done to precondition unstructured data.

SYNONYM CONCATENATION

The other way for synonyms to be resolved is through synonym concatenation. In synonym concatenation, synonyms are not deleted. Instead, synonyms are concatenated with their original word, as shown in Figure 4-7. By using synonym concatenation, either the specific word or the synonym can be accessed and analyzed.

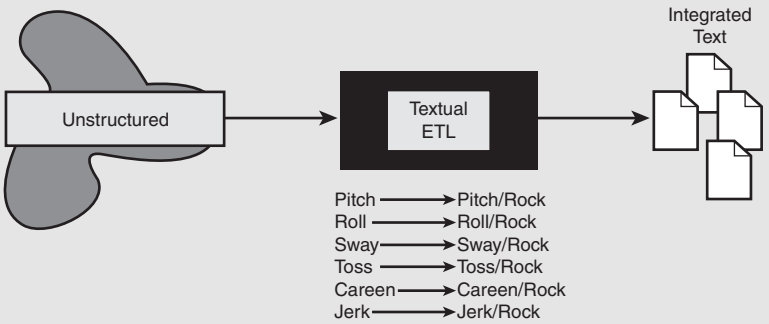

Figure 4-7 Synonym concatenation is another option for preconditioning unstructured data as part of the integration process.

The problem with synonym concatenation is that the texture of the original English sentence is destroyed. Usually, this doesn't matter. However, if there is a need to preserve the original texture of the English sentence, synonym concatenation is not useful.

As a rule, synonym concatenation is the best choice for managing synonyms.

- **Resolving homographs**—In the case of words that have multiple meanings, the correct and unique term replaces the nonunique common term. This is the second ingredient needed for the establishment of a common vocabulary in raw text. In many regards, homographic resolution is the reverse of synonym resolution. In Figure 4-8, the term "ha" is replaced with different medical terms based on the person who originally wrote the term.

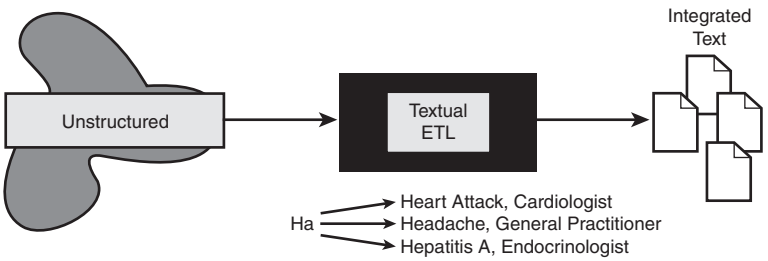

Figure 4-8 Homographs are expanded to a more precise meaning.

- **The capability to handle both words and phrases**—It is not sufficient to support textual analytic processing by using just words. Phrases need to be supported as well, as shown in Figure 4-9.

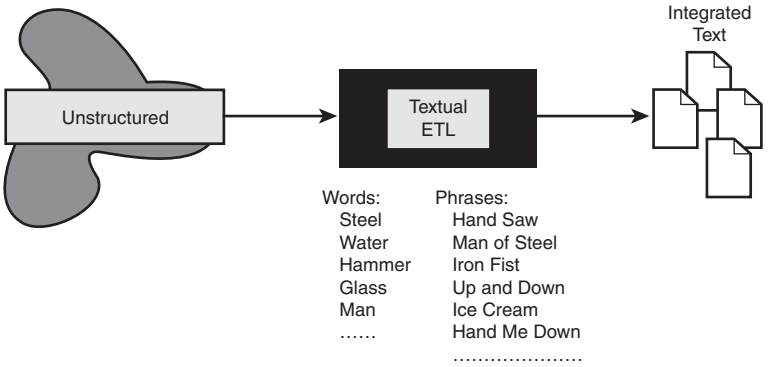

Figure 4-9 Both words and phrases need to be handled.

- **Allowing for multiple spellings of the same name or word**—Some names and words can be spelled in many different ways. Common misspellings need to be included as well. In Figure 4-10, some of the many common misspellings of "Osama Bin Laden" are incorporated into the integrated text. In doing so, the textual analyst is sure to find the references even if they are not spelled correctly or spelled as the person initiating the search thinks they should be spelled.

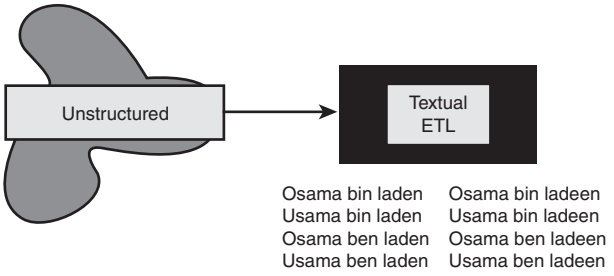

Figure 4-10 Alternate spellings should be handled as well.

- **Negativity exclusion**—In the case of negativity exclusion, where there is a negative, the words that follow the negative expression are removed from any indexing or other reference. In Figure 4-11, cancer is not included in the indexing process because it is preceded by a "not."
- **Punctuation and case-sensitivity**—Punctuation and case-sensitivity need to be removed as a consideration for searching. In Figure 4-12, the term "asher lev" can be found even though the term is written "Asher Lev" in the unstructured text. Punctuation and case are eliminated as a basis for finding a match between search argument and the text operated on.

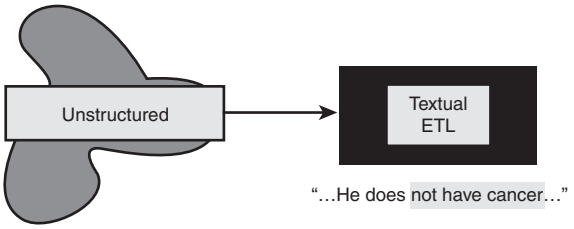

Figure 4-11 Negativity exclusion is another aspect of textual integration.

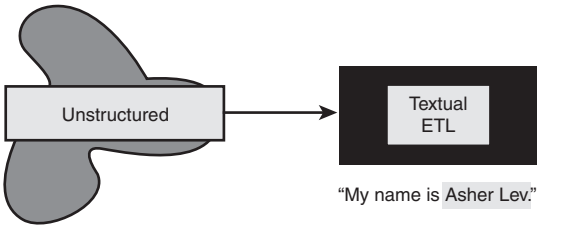

Figure 4-12 Punctuation and case-sensitivity should not be a factor in doing textual analytics.

- **Document consolidation**—On occasion, document consolidation is a useful aspect of textual integration. When textual consolidation is done, documents that hold like information are logically consolidated into a single document, as shown in Figure 4-13. The grouping of like documents can have the effect of enhancing the manageability of the process of textual integration.

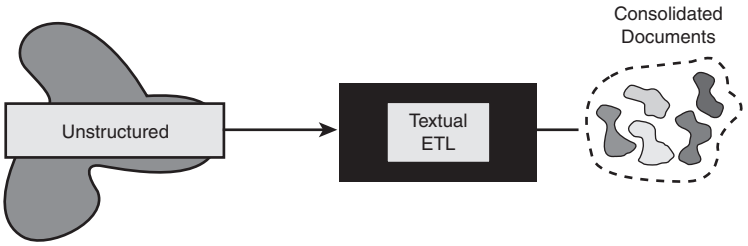

Figure 4-13 Document consolidation is sometimes a good thing to do as part of the textual integration process.

- **Themes of data**—Another important aspect of textual integration is that of determining basic themes of data. The themes can be discovered in a document or in the text that has been gleaned from multiple documents. In Figure 4-14, data is clustered around water and steel.

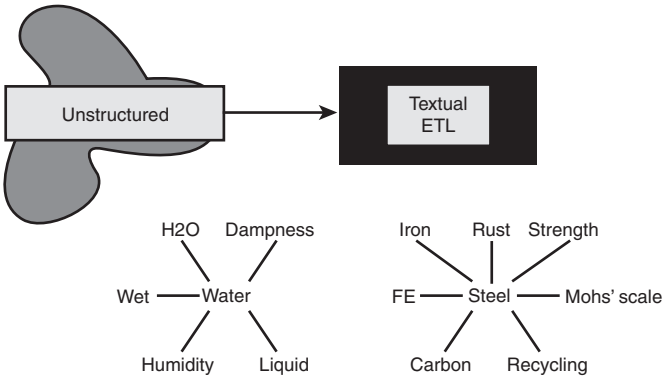

Figure 4-14 Creating themes for documents and groups of documents is another aspect of textual integration.

These basic activities of integrating unstructured text are the minimum subset of processes that need to occur to provide a sound foundation in the preparation of text for textual analytics. Many other related processes can be applied to unstructured text as it is prepared for movement to the structured environment.

External Categorization

An important process that needs to occur for the preparation for the movement of unstructured data into the structured environment is that of the creation of external categorizations (or indirect indexes).

The starting point for external categorization occurs when a topic is associated with multiple words. When the topic is associated with multiple words and phrases, the unstructured text is examined with the associated words and phrases. As a simple example, consider the topic "Sarbanes Oxley:"

Topic—Sarbanes Oxley
  Associated words and phrases:
    Promise to deliver
    Contingency sale
    Revenue recognition
    Contract terms

The unstructured data is examined, and wherever a word or phrase is found that is indirectly related to Sarbanes Oxley, a reference is made from that document to Sarbanes Oxley. This is an example of an indirect index. Now, when an indirect search of unstructured data is made on Sarbanes Oxley, all the references to terms that relate to Sarbanes Oxley are found.

Contrast an indirect search to a direct search. In a direct search, if the search were made on Sarbanes Oxley, only text where the term Sarbanes Oxley is found would be referenced. Such a direct reference has limited usefulness.

As data is placed in the structured environment from the unstructured environment, not only is the unstructured text integrated, but external categorization of the data is done as well. By integrating the unstructured text and creating indirect indexes, the gap between the two environments is bridged.

Simple Integration Applications

The first step in creating an integrated textual environment is accessing and gathering the unstructured documents to be processed. If an organization has many unstructured documents lying around in many different places, the ability to find, access, and gather those documents into a single location is an important feature.

After the raw documents have been gathered, the next step is to integrate them. The simplest kind of application that can be created from integrated unstructured text is one where data is simply integrated into the structured environment and then accessed and analyzed. When text has been integrated, there are many uses of the data. A standard BI tool can be used to create queries against the textual data.

As an example of a simple application, suppose there is a toxic chemical that has just been unearthed as a threat. The integrated text can be used as a basis for a search. It can take a matter of seconds to find out what information there is about this newly uncovered chemical.

A second simple use of integrated textual data in the structured environment is doing an indirect search. For example, suppose there is a need to do an indirect search on "toxic chemical." All the different kinds of toxic chemicals and all the information about the toxic chemicals can be accessed and organized quickly.

A third valuable use of unstructured data that has been integrated and placed in the structured environment is the ability to link that integrated text to other structured data. As a simple example, suppose there is text about "nitroglycerine." This text can be connected to and related to other occurrences of "nitroglycerine" in the structured environment, forming a robust query.

There is widespread applicability of simple unstructured integration. All companies that have unstructured data in any form need this capability.

Examining the Contents of Existing Unstructured Data

Another example of the use of integrating text and placing it in the structured environment is looking at large repositories of unstructured text. Many organizations collect unstructured data in the form of emails. Other corporations collect unstructured text in

software such as Documentum. Over time, these collections of unstructured text grow large.

The problem is that as these collections of data grow large, the content becomes unintelligible. Stated differently, there is so much content and the content is so scattered and disparate that it becomes impossible to find anything in the files of unstructured data.

By integrating the large volumes of unstructured text and then bringing the text over to the structured environment, the unstructured text is meaningfully read and examined.

There is a widespread need for the ability to look into large volumes of unstructured data. All corporations that have large stores of unstructured data—emails, documents—need this capability.

Enterprise Metadata Repository

There is a great need for the ability to find and integrate and gather enterprisewide metadata. Metadata is a special class of unstructured text. Metadata refers to the classes of data, not to the data itself. Metadata—wherever it is found—is nothing but unstructured text. The same tools that can enter the unstructured environment and integrate content of text can also be used for enterprisewide metadata gathering. When gathered, the enterprisewide metadata can be used for the creation and population of a repository.

All organizations that have metadata need this capability.

Customer Communications

One of the most important sources of information is the communications found in the corporation. Communications can be between employees of the corporation or between the employees, customers, and prospects of the corporation. Typically, communications are stored in emails or transcripted telephone conversations. In truth, these communications can be found practically anywhere.

These communications can be integrated and brought into the structured environment and used to form the 360-degree view of the customer.

One of the most obvious ways that unstructured data can be used to enhance existing analytical systems is in terms of the creation of an unstructured customer contact file, which is a file of every contact or communication the customer has had with the corporation. This can include emails, letters, and other documents. The unstructured customer contact file is an index of the date of the contact, the nature of the contact, and to whom the contact was made.

There are many and powerful uses of the customer contact file. One of the most powerful uses is supplementing a customer relationship management (CRM) system. Figure 4.15 shows that a true 360-degree view of the customer is desirable for many businesses.

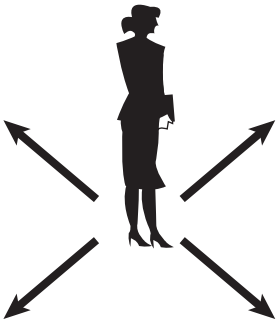

Figure 4-15 The 360-degree view of the customer

The essence of the CRM systems is to create a 360-degree view of the customer, thereby opening many avenues and opportunities. Some of the reasons why the 360-degree view of the customer is appealing follow:

- **Cross selling**—If you understand a lot about the customer in one arena, the opportunity to sell to the same customer in another arena will materialize.
- **Prospecting**—The more you understand about a customer, the better you can qualify a sales or sales prospect list.
- **Anticipation**—By understanding a lot about the customer, you can anticipate future needs.

One of the basic tenets of CRM is that it is much easier to sell into an established customer base than it is to bring in new customers. In this regard, creating a long-term and real relationship with the customer is certainly a worthwhile objective.

So exactly how is this relationship established? The basis of the relationship is the integrated knowledge about the customer. The integrated knowledge includes many different facets about the customer:

- Age
- Education
- Occupation
- Marital status
- Address
- Net worth
- Income
- Spending habits

- Children
- Car
- Cost of home

The idea behind creating the 360-degree view of the customer is to bring together data from many different places to integrate the data and achieve a truly cohesive and comprehensive view of the customer, as shown in Figure 4-16.

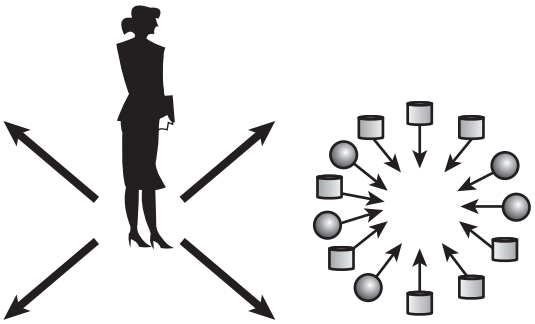

Figure 4-16 To achieve the 360-degree view of the customer, lots of different kinds of data are integrated together.

So exactly how does the unstructured contact file enhance the 360-degree view of the customer? The unstructured contact file adds the dimension of communication. Now, instead of just knowing odd facts about the customer, the corporation can know what the customer has been saying—what communications have transpired. If the customer has been irritated, the corporation can know that. If the customer has been especially pleased, the corporation can know that. If the customer has been trying to get through and talk to the corporation, the corporation can know that. In short, the unstructured contact file allows the corporation to know the recent state of mind of the customer. No 360-degree view of the customer comes anywhere close to knowing that kind of valuable information about the customer.

The Resulting Architecture

The resulting architecture that is created looks like Figure 4-17. By integrating unstructured text and organizing into the structure as shown, the following are possible:

- The unstructured data can be analyzed.
- The unstructured text can be accessed by direct or indirect searches.
- The unstructured text can be linked to structured databases and a composite query can be created.

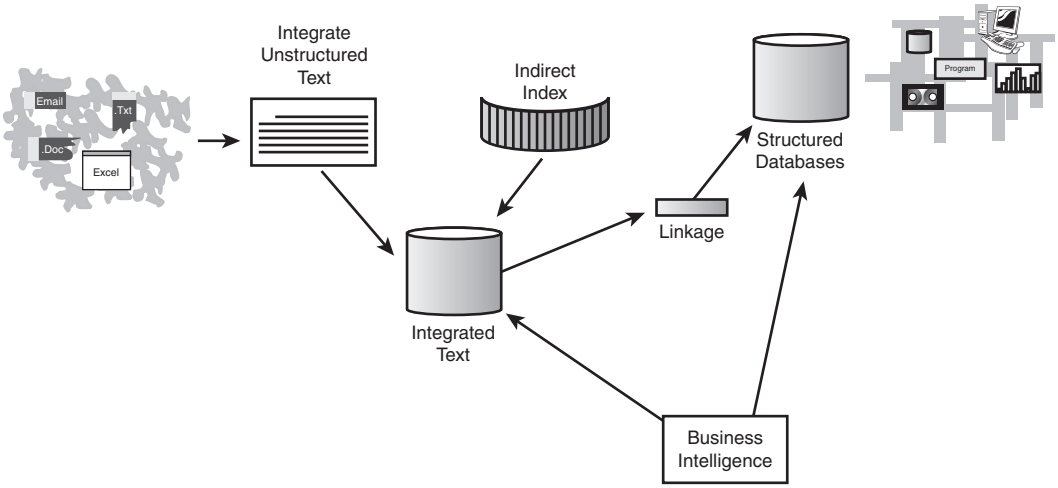

Figure 4-17 The architecture for textual analytics

Choosing the Best Types of Integration

There is no question that there is value to integration. Bringing together the many sources of unstructured data is no exception. When sources of unstructured data are integrated, it is possible to see a complete picture of whatever subject is captured. Different perspectives can be gathered together and compared and contrasted.

A lot of good things can happen when information is integrated. Figure 4-18 shows the integration of unstructured data.

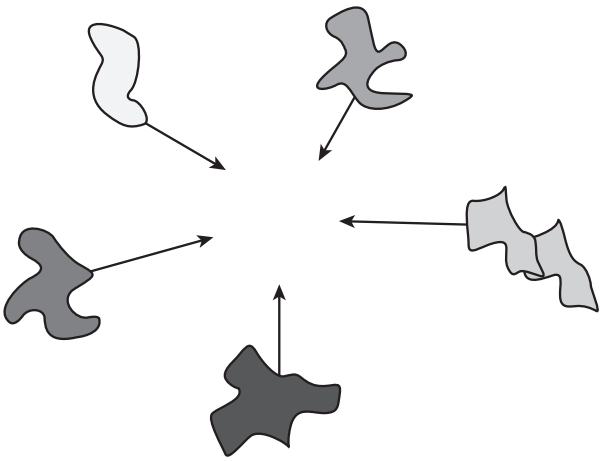

Figure 4-18 One of the values in the creation of the unstructured data warehouse is the integration of the data.

There are two basic forms of integration. One form of integration occurs when many different unstructured sources are integrated together. Another form occurs when unstructured data is integrated with structured data.

Figure 4-19 shows these two basic forms of integration.

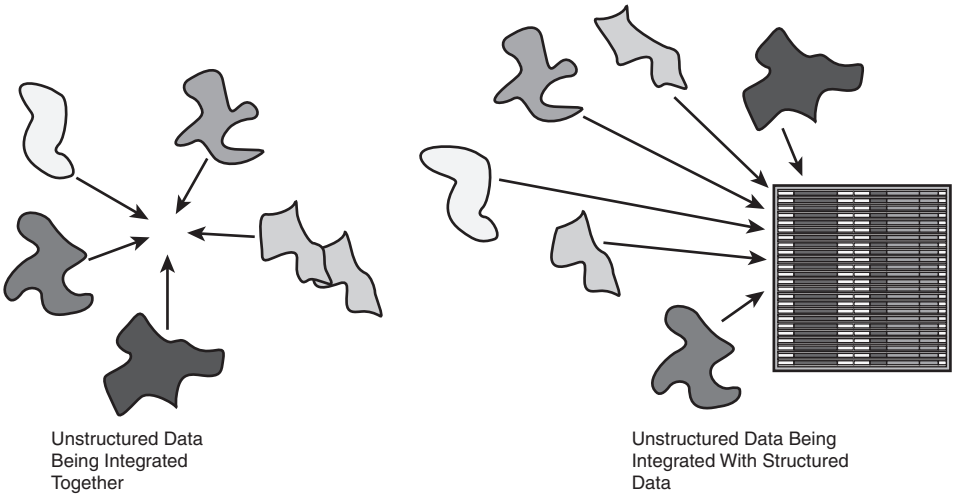

Figure 4-19 There are two basic forms of integrating text.

Both of these forms of integration are valuable; however, they essentially serve the same purpose.

Ways in Which Integration Occurs

There are at least two ways that integration can occur. One way is where data is accessed, gathered, and used to form a result. This can be called *access integration*. EII is a form of access integration. In access integration, the integration is done on-the-fly. The data is accessed, integrated, and then sent to the requesting party.

The other form of integration is *foundation integration*. With foundation integration, the sources of data are accessed in their entirety, the data is integrated in its entirety, and a foundation of integrated data is created. Then when it comes time to access integrated data, the foundation is accessed.

Figure 4-20 shows the two means by which data can be integrated. In general, when the data is integrated into a foundation, it is placed in a database.

Performance Limitations

There are two schools of thought as to which form of integration is better: the access-oriented integration or the building of a foundation.

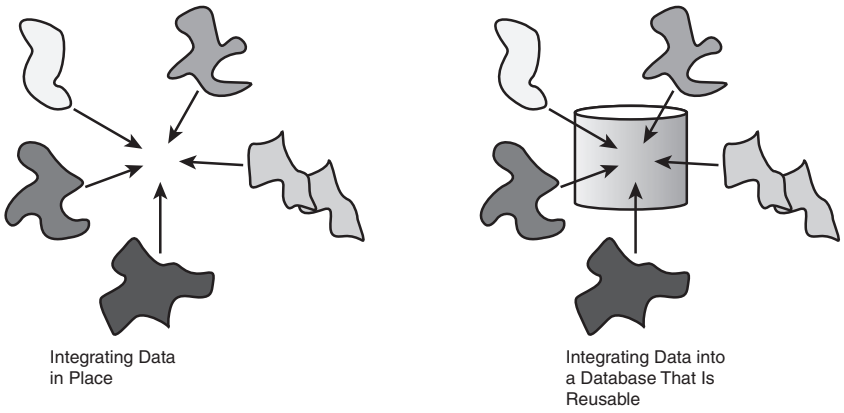

Figure 4-20 Two basic approaches to integration

At first glance it appears that access integration is superior because a lot of data doesn't have to be integrated. When a foundation is built, a lot of data has to be integrated. Furthermore, the data that is integrated does not have an analyst querying the data, so building a foundation is a large undertaking. Doing access integration takes far fewer resources, at least on the surface.

However, consider this: Every time a query is made to data, the same data has to be reaccessed. Figure 4-21 shows the many accesses to the same data.

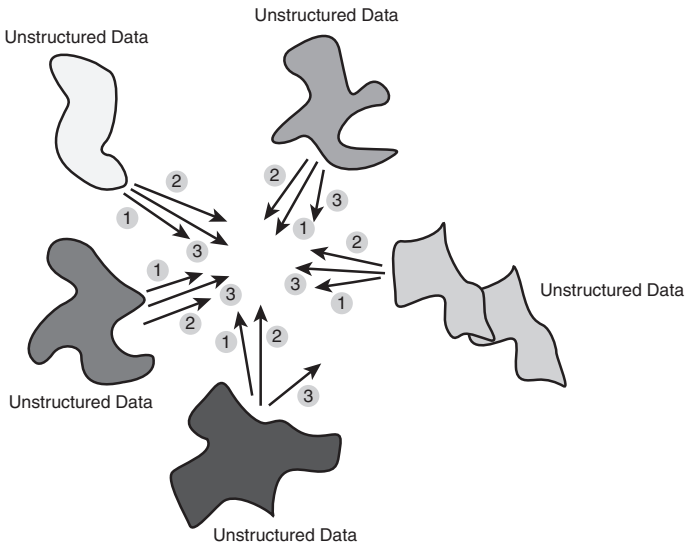

Figure 4-21 Every time there is a need for access, a whole new set of queries is made.

If only a few accesses/searches must be made, the work the system has to do, as shown in Figure 4-21, is not much. However, if there are many accesses/searches to be made, or if the accesses/searches actually go after the same data, the system must do an extraordinary amount of work to create access integration. The work the system has to do can result in poor performance and a need for large amounts of machine resources.

However, performance can be an issue for other reasons than a lot of queries being run. Another reason why performance can be a problem is that an integration access query runs at the speed of the slowest transaction. Consider the circumstances seen in Figure 4-22.

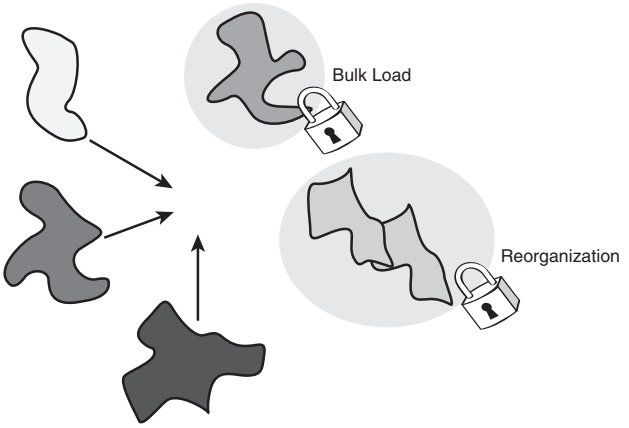

Figure 4-22 The speed of access is only as fast as the slowest source of data.

In Figure 4-22, a query has been submitted to access several sources of data. The query will not be complete until all other data is collected. The problem is that two of the sources of data are busy doing something else. One source of data is being reorganized while another source of data is busy with a bulk load. It might be hours before the data will be ready to be accessed. Performance is an issue that relates to more than just the total workload passing through the system.

However, there are other issues. In some cases, a heavy amount of processing must be done to integrate the data. The processing can take many different forms:

- Aggregation
- Summarization
- Conversion logic
- Reformatting
- Restructuring
- Addition of new fields

There are many activities that might have to occur when data is integrated.

Disadvantages to Integration

There are some problems. The first problem is that the same work must be done repeatedly. Every time the same data is accessed, the same integration activities have to occur. This is a waste of resources. Figure 4-23 shows this circumstance.

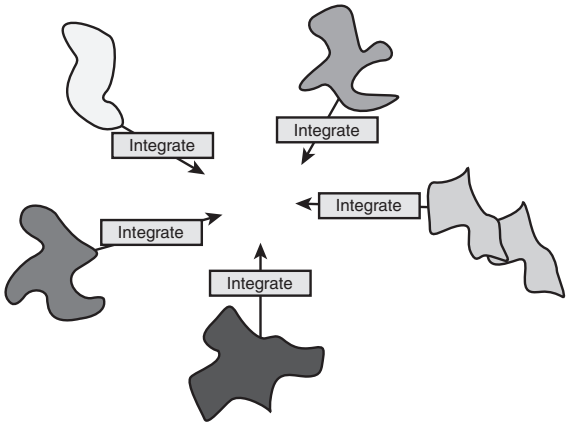

Figure 4-23 When integration must be done, it must be done repeatedly.

However, there is another related problem. If integration is done on-the-fly, what happens when the integration is done differently from one access of the data to the next? For example, at 10:41 AM the data is accessed and integrated. At 11:43 AM the same data is accessed, except this time the integration is done differently. If there is to be integrity of results, the same integration must be done each time. Otherwise, a query will yield one result one time and another result another time.

In the same vein is the circumstance where data is updated between different accesses of data. Figure 4-24 shows this instance.

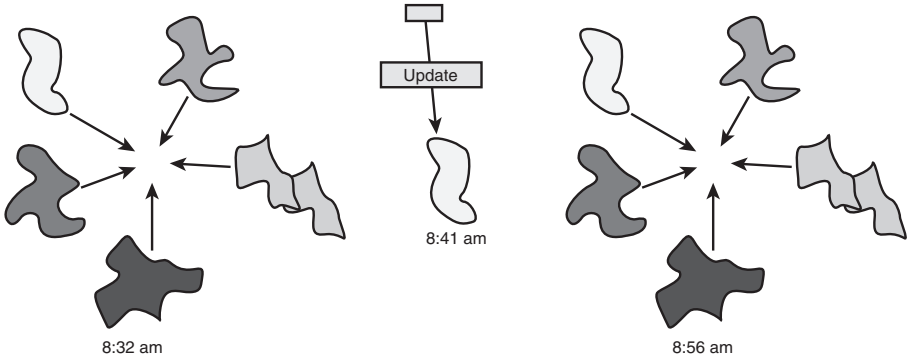

Figure 4-24 By accessing data from the source every time, if a change is made to the source data, the data has now become irreconcilable.

In Figure 4-24, you can see that data is accessed at 8:32 AM. An update is made to the source data at 8:41 AM. Then another access is made to the same data at 8:56 AM. The results at 8:56 AM will be different from the results of the access made at 8:32 AM.

If integrity of data is an issue, access data and then integrating it has a problem. However, when data is placed in a foundation database, as shown in Figure 4-20, the data does not have the problems that have been illustrated.

At first glance, it appears that building a foundation database of integrated data is an expensive, wasteful activity. However, when the limitations and disadvantages of access integration are considered, the foundation approach to integration starts to look really good.

Summary

Before textual analytics can be done properly, text must be integrated.

The first step in integration is the process of capturing the text in an electronic format. If the text is not in an electronic format, it must be lifted from the paper and produced electronically. OCR processing automates some of this process.

The next step is to physically read the data in whatever native format the data is in. Formats such as .doc, .txt, .ppt, .pdf are all common formats.

Next, the data is screened to separate the unneeded data from the relevant data. After the data is checked for relevancy, the next step is to read and analyze the text with the following criteria:

- Stemming
- Homographic resolution
- Synonym concatenation
- Synonym replacement
- Alternate spelling concatenation
- Negativity exclusion
- Punctuation, case, and font insensitivity
- Basic theming
- External categorization

After the text has been analyzed in accordance with these variables, the next step is to prepare the data for entry into a DBMS.

The different forms of search that can be done include the following:

- Simple direct searches
- Indirect searches
- Indirect searches based on external categorization
- Search on word permutations

As important as unstructured data, there are important considerations to semistructured data as well. Chapter 5, "Semistructured Data," addresses the "cousin" of unstructured data—semistructured data.

Prentice Hall — 0132360292