1

Introduction

People are talking about your business every day. Are you listening?

Your customers are talking. They're talking about you to your face and behind your back. They're saying how much they like you, and how much they hate you. They're describing what they wish you would do for them, and what the competition is already doing for them. They are writing emails to you, posting blogs about you, and discussing you endlessly in public forums. Are you listening?

Other businesses and organizations are talking too. Researchers talk about new technologies and approaches you might be interested in. Other businesses describe innovations you could leverage in your products. Your competitors are revealing technical approaches and broadcasting their strategies in various publications. They talk about what they are working on and what they think is important. Are you listening?

Your employees are also talking. They are producing great ideas that are languishing for lack of the right context to apply them. They are looking for the right partners to help them innovate and create the next big thing for your company. They reveal new ways to improve your internal processes and even change the entire vision for your company. Are you listening?

All of this talk is going on out there now, even as you read these pages. And you can listen—if you know how. This book is about how we learned to listen to the talk and to turn it into valuable business insights for our company and for our customers. Now we would like to share that knowledge with you.

A Short Story…"The Contest"

Writing this book has been a project that beckoned for many years. We had started and stopped multiple times. We knew we wanted to write the book, but we had trouble convincing ourselves that anyone would want to read it. At a gut level, we knew that what we were doing was important and unique. However, there were a lot of competing methods and products, with more added every day, and we could not spend all of our time evaluating each of them to determine if our approach was measurably superior. Then, in May 2006, an event happened that in one day demonstrated convincingly that our approach was significantly better than all the other alternatives in our field. The results of this day would energize us to go ahead and complete this book.

It began when a potential client was considering a large unstructured data mining project. Like most companies, they had a huge collection of documents describing customer interactions. They wanted to automatically classify these documents to route them to the correct business process. They questioned whether or not this was even feasible, and if so, how expensive would it be. Rather than invite all the vendors in this space to present proposals, they wanted to understand how effective each technical approach was on their data. To this end, they set up the following "contest."

They took a sample of 5,000 documents that had been scanned and converted to text and divided them manually into 50 categories of around 100 documents each. They then invited seven of the leading vendors with products in this space to spend one week with the data using whatever tools and techniques they wished to model these 50 categories. When they were done, they would be asked to classify another unseen set of 25,000 documents. The different vendors' products would be compared based on speed, accuracy of classification, and ease of use during training. The results would be shared with all concerned.

That was it. The "contest" had no prize. There was no promise of anything more on the client's part after it was over. No money would change hands. Nothing would be published about the incident. There was no guarantee that anything would come of it. I was dead set against participating in this activity for three very good reasons: 1) I thought that the chances it would lead to eventual business were small; 2) I didn't think the problem they were proposing was well formed since we would have no chance to talk to them up front to identify business objectives, and from these to design a set of categories that truly reflected the needs of the business as well as the actual state of the data; and 3) I was already scheduled to be in London that week working with a *paying* customer.

I explained all of these reasons to Jeff, and he listened patiently and said, "You could get back a day early from London and be there on Friday."

"So I would have one day while the other vendors had five! No way!"

"You won't need more than one day. You'll do it in half a day." I didn't respond to that—I recognize rank flattery when I hear it. Then Jeff said, "I guess you really don't want to do this."

That stopped me a moment. The truth was I did want to do it. I had always been curious to know how our methods stacked up against the competition in an unbiased comparison, and here was an opportunity to find out. "OK. I'll go," I found myself saying.

As planned, I arrived at the designated testing location on Friday morning at 9AM. A representative of the client showed me to an empty cubicle where sat a PC that contained the training data sample. On the way, he questioned me about whether or not I would want to work until late in the day (this was the Friday before Memorial Day weekend). I assured him that this would not be the case. He showed me where on the hard drive the data was located and then left. I installed our software[1] on the PC and got to work.

About an hour later, he stopped by to see how I was coming along. "Well, I'm essentially done modeling your data," I said. He laughed, assuming I was making a joke. "No, seriously, take a look." We spent about an hour perusing his data in the tool. I spent some time showing him the strengths and weaknesses of the classification scheme they had set up, showing him exactly which categories were well defined and which were not, and identifying outliers in the training set that might have a negative influence on classifier performance. He was quite impressed.

"So, can you classify the test set now?" he asked.

"Sure, I'll go ahead and start that up." I kicked off the process that classified the 25,000 test documents based on the model generated from the training set categories.

We watched it run together for a few seconds. Then he asked me how long it would take. I tried to calculate in my head how long it should take based on the type of model I was using and the size of the document collection. I prevaricated just long enough before answering. Before I could give my best guess, the classification had completed. It took about one minute.

"So that's it? You're done?" he asked, clearly bemused.

"Yes. We can try some other classification models to see if they do any better, but I think this will probably be the best we can come up with. You seem surprised."

He lowered his voice to barely a whisper. "I shouldn't be telling you this, but most of the other vendors are still here, and some of them still haven't come up with a result. None of them finished in less than three days. You did it all in less than two hours? Is your software really that much better than theirs? How does your accuracy stack up?"

"I don't know for sure," I answered truthfully, "but based on the noise I see in your training set, and the accuracy levels our models predict, I doubt they will do any better than we just did." (Two weeks later, when the results were tabulated for all vendors, our accuracy rate was almost exactly as predicted, and it turned out to be better than any of the other participating vendors.)

"So why is your stuff so much better than theirs?" he asked.

"That's not an easy question to answer. Let's go to lunch, and I'll tell you about it."

What I told the client over lunch is the story of how and why our methodology evolved and what made it unique. I explained to him how every other unstructured mining approach on the market was based on the idea that "the best algorithm wins." In other words, researchers had picked a few sets of "representative" text data, often items culled from news articles or research abstracts, and then each created their own approaches to classifying these sets of articles in the most accurate fashion. They honed the approaches against each other and tuned them to perform with optimum speed and accuracy on one type of unstructured data. Then these algorithms eventually became products, turned loose on a world that looked nothing like the lab environment in which they were optimally designed to succeed.

Our approach was very different. It assumed very little about the kind of unstructured data that would be given as input. It also didn't assume any one "correct" classification scheme, but observed that the classification of these documents might vary depending on the business context. These assumptions about the vast variability inherent in both business data and classification schemes for that data, led us to an approach that was orders of magnitude more flexible and generic than anything else available on the market. It was this flexibility and adaptability that allowed me to go into a new situation and, without ever having seen the data or the classification scheme ahead of time, quickly model the key aspects of the domain and produce an automated classifier of high accuracy and performance.

In the Beginning…

In 1998, a group from IBM's Services organization came to our Research group with a problem. IBM Global Services manages the computer helpdesk operations of hundreds of companies. In doing so, they document millions of *problem tickets*—records of each call that are typed in by the helpdesk operator each time an operator has an interaction with a customer. Here is what a typical problem ticket looks like:

```
1836853 User calling in with WORD BASIC error when opening files in word. Had
user delete NORMAL.DOT and had her reenter Word, she was fine at that point.
00:04:17 ducar May 2:07:05:656PM
```

Imagine millions of these sitting in databases. There they could be indexed, searched, sorted, and counted. But this vast data collection could not be used to answer the following simple question: What kinds of problems are we seeing at the helpdesk this month? If the data could be leveraged to do this analysis, then some of the more frequent tasks could potentially be automated, thus significantly reducing costs.

So why was it so hard to answer this question with the data they had? The reason is that the data is *unstructured*. There is no set vocabulary or language of fixed terms used to

describe each problem. Instead, the operator describes the customer issue in ordinary everyday language…as they would describe it to a peer at the helpdesk operations center. As in normal conversation, there is no consistency of word choice or sentence structure or grammar or punctuation or spelling in describing problems. So the same problem called in on different days to different operators might result in a very different problem ticket description. This kind of unstructured information in free-form text is what we refer to as "talk." It is simply the way humans have been communicating with each other for thousands of years, and it's the most prevalent kind of data to be found in the world. Potentially, it's also the most valuable, because hidden inside the talk is little bits and pieces of important information that, if aggregated and summarized, could communicate actionable intelligence about how any business is running, how its customers and employees perceive it, what is going right and what is going wrong, and possibly solutions to the most pressing problems the business faces. These are examples of the gold that is waiting to be discovered if we can only "Mine the Talk."

And so with this challenge began the journey that culminated in this book.

The Thesis

The purpose of this book is to share with you the insights and knowledge that we have gained from the journey we have been on for nearly a decade. We are applied researchers and software engineers that have been developing technologies to address real-world business problems. We have implemented and experimented with variations of most approaches and algorithms that are espoused in the literature, as well as quite a few new techniques of our own. Through trial and error, insight, and sometimes good luck, we have come up with an approach, supported by technology, that we think will revolutionize the definition of business intelligence and how businesses leverage information analytics into the future.

Our work can be summarized in one simple thesis:

> A methodology centered around developing taxonomies that capture both domain knowledge and business objectives is necessary to successfully unlock the business value in all kinds of unstructured information.

In this introduction, we will take you through the thinking that has led us to this conclusion, and outline the methodology we use to *Mine the Talk* to create lasting business value.

The Business Context for Unstructured Information Mining

In parallel to, and in collaboration with, technological progress, businesses have adapted and evolved to better leverage structured and unstructured information analytical

capabilities. This is part of a larger phenomenon sometimes referred to as the co-evolution of business and technology.

The Enterprise Ecosystem

Gone is the day of the monolithic enterprise. To create efficiencies, most industries have disaggregated into their component parts. For example, in the automotive industry, the manufacturers do not mine the ore to make the steel to fabricate the parts that make up their products. All of these tasks are performed by specialized companies for the automotive industry, and as many other potential industries and customers as possible, in order to maximize their return on investment. Similarly, as a consumer, you do not actually buy or service your vehicle from the manufacturer, but there is a selection of dealers and service providers to choose from. Today's enterprise exists in a complex network of suppliers, vendors, business partners, competitors, and industries, which we call the enterprise ecosystem (see Figure 1-1).

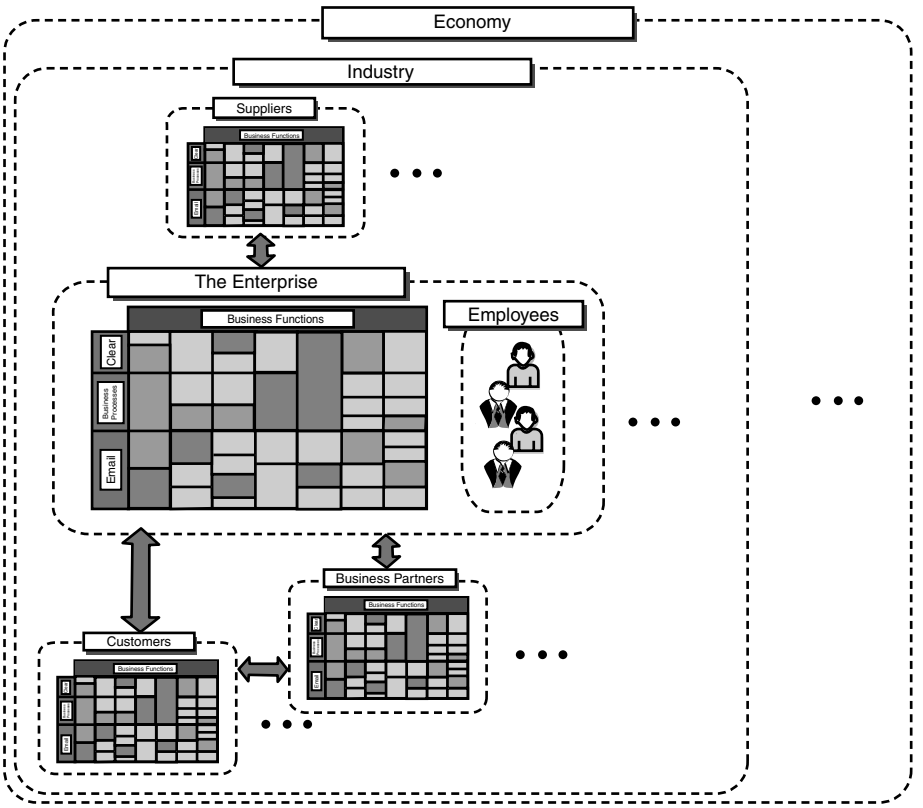

Figure 1-1 The enterprise ecosystem

An increasingly important aspect of the enterprise ecosystem that creates an integrating backdrop is the information space. This consists of all human knowledge accumulated in its various forms: written, spoken, and otherwise. This is manifested in communication mediums such as newspaper, magazines, books, television, and the Web. With the growth of the internet and electronic media, the rate of growth in the information space is increasing exponentially. But more information is not always better. Figuring out what is important is becoming an increasingly difficult problem.

Services Industry Growth

A second trend in the evolution of business is the significant growth of the services sector. Over the last century, the world's labor force has migrated from agriculture to goods manufacturing to services (see Figure 1-2). This largely has to do with efficiency gains in agricultural and manufacturing processes through technological innovation and labor optimization. There is no reason to believe that this won't occur in the services sector as well, so the advantage will go to those who lead in this process, not those who ignore it or resist it.

Top Ten Nations by Labor Force Size
(about 50% of world labor in just 10 nations)
A = Agriculture, G = Goods, S = Services

Nation	% ww labor	% A	% G	% S	25 yr % delta S
China	21.0	50	15	35	191
India	17.0	60	17	23	28
U.S.	4.8	3	27	70	21
Indonesia	3.9	45	16	39	35
Brazil	3.0	23	24	53	20
Russia	2.5	12	23	65	38
Japan	2.4	5	25	70	40
Nigeria	2.2	70	10	20	30
Banglad	2.2	63	11	26	30
Germany	1.4	3	33	64	44

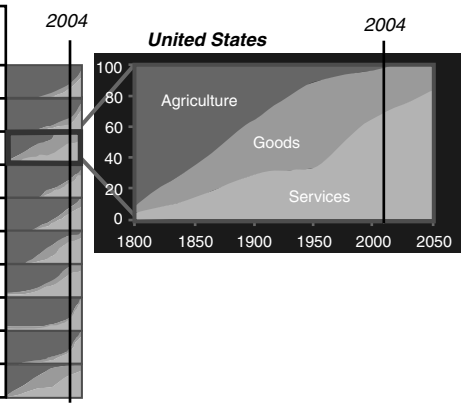

Figure 1-2 Services industry growth

So why is this important in the context of *Mining the Talk*? Services, by their nature, are more unstructured. Services involve an interaction between the service provider and the customer. Customers are integral to the process, because they are involved with the co-production of value between a provider and a consumer. This makes each transaction inherently unique. When modeling or capturing this process, this uniqueness or variability lends itself to a more unstructured representation. With the explosion of the

services sector and its affinity to unstructured information, the ability to extract value from this information will only grow in importance.

One of the techniques to create efficiencies in agriculture and manufacturing was process standardization. This is made somewhat more difficult in services with customers in the loop, because in many cases, they may not care about your process efficiencies. We can all relate to being annoyed by complex phone navigation trees to try to get to the right person or service to handle our concern. It is clear that business process standardization will continue and is necessary, but those who do it most effectively without alienating customers will have a distinct advantage.

From Transaction to Interactions and Relationships

It is no longer sufficient to only look at the attributes of a transaction with customers at the boundary of the enterprise to understand your business. Businesses need to look at the lifecycle of their interactions within the business ecosystem (see Figure 1-3). This means all of the interactions that you have with your customers, suppliers, business partners, employees, and competitors, as well as the industries and economies in which you compete. On its face, this can seem either obvious, insurmountable, or both. However, there is hope, and *Mining the Talk* will help get you there.

How many times have you heard the sage business advice, "listen to your customers"? Although this is good advice, it is necessary but not sufficient. You will also need to listen to your employees, because they are closest to the action. Your vendors, suppliers, and business partners are also a source of incredible information. In addition, the information space, what is being said in the press, on the Web, and in the technical literature, are other sources that can be utilized. All of this requires constant attention and monitoring. These are not transactions to be counted and sorted, but they are ongoing interactions and relationships that need to be understood and leveraged.

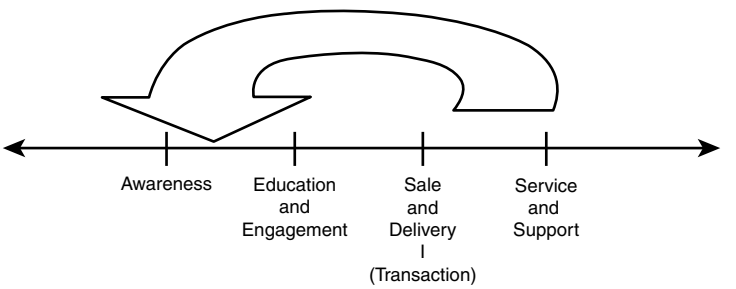

Figure 1-3 Interaction lifecycle

Capturing Business Objectives and Domain Expertise

In our journey of discovery, we have seen one mistake made repeatedly. We have seen static business models and static data models try to be used to model inherently dynamic business processes, particularly at the point of interaction. For example, virtually every customer relationship management system we have come across has a manual classification scheme (or taxonomy) that is meant to be used by the service agent to classify the nature of the customer interaction. This approach has two major flaws. First, as soon as the classification scheme is published, it is out of date, because interactions with your customers are unpredictable and continually changing. Second, even if the classification scheme was representative of your customer interactions, it is unreasonable to expect any number of service agents to classify their interactions with their customers in a consistent way and with high quality. This very often makes such classification data completely useless, or, more dangerously, misleading. This issue is true throughout the business ecosystem where unstructured information exists.

An adaptable data model is critical when incorporating unstructured information mining. One problem commonly encountered is that the analysis typically leads to more questions. In business intelligence or data mining, if the data model is not designed to handle the new question, the data model must be modified and the data manipulated and reloaded, which is often a difficult and cumbersome process, many times taking months. This problem is compounded even more with unstructured information because of its very nature. It is important to be able to add or enhance existing taxonomies, classifications, or extracted data as the information leads you through the discovery process.

Additionally, it is important to combine the right mix of algorithmic assistance with domain expertise. We have found that most people naively want to push a button and magically receive the answer. However, we have come to the conclusion that, particularly with unstructured information, the analytical process cannot be fully automated. Even for structured information, it typically needs an analyst in the loop to interpret the results. For unstructured information, an analyst is required in the loop to help guide the process with a combination of algorithmic assistance, a useful set of metrics to assist in the interpretation, and the analyst's domain expertise. The key is to make efficient use of this expensive and scarce resource, not to eliminate it entirely.

A Common Analytical Methodology

We struggled for some time and went through multiple iterations before we settled on a common analytical methodology that spans domains, business objectives, and information sources. Our method has three major phases: *Explore*, *Understand*, and *Analyze*.

Each of these phases leverages different capabilities that build on each other. However, it is not always necessary to use every phase or capability. For very large information collections, we have developed the capabilities in *Explore*, and we will discuss this in Chapter 5, "Mining to Improve Innovation." Most of this book will concentrate on the *Understand* and *Analyze* phases, which are the unique differentiators and the key to unlocking the business value of unstructured information in *Mining the Talk*. In this section, we introduce each of these concepts and describe what they entail.

Explore

Many times we are dealing with very large repositories of information. Depending on our information source and our business objective, not all of the information will be relevant. For example, if you are interested in analyzing the Web to understand issues around a specific brand, you only need the portion of the Web that pertains to that brand. If you are analyzing patents related to a technology area, you only need the patents that are relevant to that area. We use a combination of techniques to locate the relevant set of information from a larger set. With structured information, we can use various queries; for unstructured information, we can use search, and we can combine them in different ways using various set operations. We call this process *Explore*.

Queries

We use *query* as the term to describe how we use structured fields in a database to select the subset of information that is of interest. For example, we can select customers based upon their location, the product they have purchased, or the time frame that we wish to investigate. We can select patents based upon the assignee, the inventor, or the classification code. These are all typically structured fields that are stored in a database. These types of queries are quite simple to perform using the standard SQL query language to find the sub-collections of interest. This technique is very powerful and effective, given you have the appropriate attributes in the database and know which of their values will select the subset that is relevant to address the issue being analyzed.

Search

Search is the process of finding those documents that contain specific words or phrases in their unstructured text. We use search as the means to find collections of documents that have concepts of interest within them, rather than to find individual documents. Although it is a valuable tool, search is not the solution to all problems. The use of language does not always lend itself to easily disambiguate concepts. Some words have more than one meaning, known as homonyms. For example, using "shell" as a query will likely return information on sea shells, Shell Oil, Unix shell, egg shells, and many others. Disambiguation is one problem, but coverage is another. Some meanings can be

described with more than one word, known as synonyms. For example, we have found that valium has more than 150 unique names—have fun typing that query.

Set Operations

Because in many cases no single query or search is sufficient to get to the optimally desired collection for deeper analysis, we have found it necessary to be able to perform set operation on collections. The most commonly used operations are join and intersect. Joins are useful in combining multiple searches for synonyms. Intersection is useful when you are looking for the subset that has two attributes that could be from either the structured or unstructured fields. In some cases, when the result of a combination of queries and searches is still too large to effectively analyze in a reasonable time, sampling techniques may be used to select a statistically valid subset.

Recursion and Expansion

Results of queries and searches can be used as input to subsequent Explore operations. This allows us to refine the subject of our mining study incrementally as we learn more about the data. Also, we can use query expansion to take the results of a query done on a subset of the data and apply it to the entire data collection.

Understand

The result of the *Explore* phase is a collection of information that covers the topic of interest. The *Understand* phase is about discovering what the information contains. We have developed a unique method of creating structure from unstructured information through the process of taxonomy generation and refinement. We use a combination of practical steps, statistical techniques, algorithms, and a methodology for editing taxonomies that allows for the flexible capture of domain expertise and business objectives. We call this process *Understand*. The Understand process works in two directions: the analyst understands the underlying structure inherent in the unstructured information, and the models captured as a result of the analyst's edits represent an understanding of domain knowledge and business objectives.

Statistics are fundamental to our *Understand* process. We are all familiar with the idea of summarizing numerical data with statistical techniques. For example, a grade point average is a way to summarize your overall academic performance. It doesn't tell you everything, but most people have agreed that it is a pretty good indicator. What about something more complex, like a sporting event? Pick your favorite sport—whether it is baseball, basketball, tennis, or football—and there are usually various ways to summarize the game or match that allow you to understand the essence of what transpired. Such summaries are no substitute for watching the game, but they can convey a lot of information about the game in a very small space.

Partitioning

If you have a large body of text, there is probably one or more natural ways to partition it into smaller sections. A book naturally falls into chapters, and each chapter into paragraphs, and each paragraph into sentences, just as a baseball game has innings and innings have outs. Breaking a large document into smaller entities makes it much easier to summarize the message of the text as a whole, because it makes statistics possible. If we try to summarize a baseball game without breaking it down by innings and outs, we are left with only the final score. But if we can break down the game into innings or at-bats and measure what happened during each of these smaller units (e.g., hits, walks, outs), then we can create meaningful statistics such as *Earned Run Average* or *On Base Percentage*.

There may be many suitable ways to do partitioning, each with its own advantage. However, the best methods for partitioning are those that produce a section that talks about only one concept with respect to the questions we want answered. The level of granularity should roughly match that of the desired business result. The analogy in baseball is that we measure innings for pitchers and at-bats for batters. The different levels of granularity make sense for different kinds of outcomes that need to be measured.

Similarly, if we want to understand the issues for which customers are calling into a call center, then individual problem records, which may span multiple calls, are the right partitioning. On the other hand, if we wish to understand better what affects customer satisfaction, we may decide to analyze each individual call record. A customer might be both satisfied and dissatisfied during the course of resolving an issue, and we want to isolate the interactions in order to analyze the underlying causes.

Feature Selection

Once the partitioning granularity is properly adjusted, we need to decide what events we are going to measure and what statistics we will keep. In a baseball box score, we don't measure everything about the game. For example, we don't know the average number of swings each batter took, or the number of pitches each pitcher threw. We could measure these things if they were important to us, but that level of detail is not interesting to the average baseball fan. Similarly in statistical analysis of text, we could measure the average number of times each letter of the alphabet occurs. We could measure the average word length, or the number of words in sentences. In fact, such statistics are used as a means for roughly measuring how "readable" a section of text will be for readers of various grade levels.[2] However, these kinds of statistics are not helpful to answer typical business questions, such as "What are my customers most unhappy about?"

So what are the right things to measure about each text example? The answer is, it all depends. It depends on what we want to learn and what kind of text data we are dealing with. Word occurrence is a good place to start for most types of problems, especially those where you don't have much specific domain knowledge to draw upon and where

the language of the documents is fairly general. When the text is more technically dense or focused on a very specialized area, then it may make sense to also measure sequences of words, also known as phrases, to get a more precise kind of statistic.

We use word and phrase occurrence as the features of a document. However, we don't use every word and phrase, because there can be a very large number of them and they are not all meaningful or useful. We use a combination of techniques to reduce the feature space to a more manageable size. We eliminate non-content–bearing words, called stopwords, such as "and" and "the." We also remove repetitive or structural phrases (we call them stock phrases). If every document contains "Copyright IBM" or "IBM Confidential," then it can safely be removed. We also combine features using a synonym list. This can be done manually where deemed appropriate or automatically through a technique called stemming. Stemming allows "jump," "jumping," and "jumped" to be treated as one. There are also various domain-specific synonym lists that can be used where stemming will fall short. Finally, we remove features that occur infrequently in the document collection because these tend to have little value in creating meaningful categories. Once we have reduced the features to a manageable size, we can use this to create summary statistics for each document. We call the collection of all such statistics for every document in a collection the *vector space model*.

Clustering

We use clustering to quickly and easily seed the process of taxonomy generation. *Clustering* is an algorithmic attempt to automatically group documents into thematic categories. These thematic categories, which together constitute a taxonomy, give an overview of what information the document collection contains. There are many different clustering algorithms that could be used, and our approach could support them all. However, we have relied heavily on variations of the k-means algorithm, because it is fast and does a reasonable job. We have also developed our own algorithm, which we call *intuitive clustering*, that we also employ.

Taxonomy Editing

Clustering is a wonderful tool, but we rarely find it to be sufficient. No matter how good the algorithmic approach to clustering becomes, it cannot embed the nuance of business objectives and the variations of language from different information sources within an algorithm. This is the critical missing element that our method incorporates. We have developed a unique set of capabilities that allow for an analyst or domain expert to quickly assess the strengths and weaknesses of a taxonomy and easily make the changes necessary to align the taxonomy with business objectives.

Analyst knowledge about the purpose of the taxonomy trumps every other consideration. Thus, a category may be created by an analyst for reasons that have nothing to do

with text features. An example would be a category of "recent" documents—those created most recently out of all the documents in the corpus. Depending on the business analysis goals, such a category may be very important in helping to understand emerging trends and issues.

Ideally, the name of a category should describe exactly what makes the category unique. An analyst may decide to change a system-generated name to one that is better aligned with the analyst's view of what the category contains. This category renaming process thus becomes an important way that domain expertise is captured.

In addition to the name, a category can also be described by choosing examples that best summarize the overall content. We describe these as "Typical Examples" because they are selected by virtue of having all or most of the features that typify the documents in the category as a whole. Using the vector space model, it is possible to automatically compare examples and select those that have the most typical content. By reading and understanding typical examples, it is possible for the analyst to make sense of a large collection of documents in a relatively short period of time.

It is also important to measure the variation within a category of documents. If there is a statistically large variation among the documents within a category, this may indicate that the category needs to be split up, or subcategorized. We call the metric that measures within category variation *cohesion*. Additionally, it is important to measure the similarity between categories. We call this *distinctness*. Categories with low distinctness scores indicate a potential overlap with another category. This overlap may indicate the need to merge two or more categories together.

The categories created using clustering and summarized with various statistics can also be edited based on this understanding. This is where analysts adds their domain knowledge and awareness of the business problem to be solved to the results—creating categories that are more meaningful.

There are many kinds of editing that are typically employed, at all levels of the text categorization. Categories can be merged or deleted. They can be created wholesale from documents matching individual words, phrases, or features. Categories can be edited—splitting off subsets of a category to create new categories. Documents can be selectively removed from one category and placed in another.

The taxonomy editing process can be thought of as the human expert training the computer to understand concepts that are important to the business. There may be many different types of categorizations that can be created on the same set of data, each representing a different important aspect of the information to the enterprise.

Visualizations

The visual cortex occupies about one third of the surface of the cerebral cortex in humans. It would be a shame to waste all of that immense processing power during the *Understand* process. We employ visualizations of taxonomies to create pictures of the

information that the human brain can process in order to locate areas of special interest that contain patterns or relationships. There are many types of visualizations that can be used to show relationships in structured and unstructured information. Scatter plots, trees, bar graphs, and pie charts can all help in the process of understanding the information, and in modifying taxonomies to reflect business objectives.

The vector space model of feature occurrence in documents is the primary data source for automatically calculating visual representations of text. Using this representation, a document becomes not just words, but a position or point in high-dimensional space. Given this representation, a computer can "draw" a set of documents and allow a human analyst to explore the text space in much the same way an astronomer explores the galaxy of stars and planets.

Analyze

At the end of the *Understand* phase, we have one or more taxonomies that represent characteristics of the unstructured information, along with a feature set that describes the individual documents that make up each taxonomy. But a taxonomy by itself rarely achieves the business objectives of mining unstructured information. The final step is to take combinations of structured and unstructured information and look for trends, patterns, and relationships inherent in the data and use that to make better business decisions. We call this process *Analyze*.

Trending

Timing is everything in comedy, in life, and in business. Knowing how categories occurred in the data stream over time will often reveal something interesting about why that category occurs in the first place. Trend analysis is also useful for detecting spikes in categories as well as in predicting how categories will evolve in the future. Trending can be interesting from a historical perspective, but it is usually most valuable when used to detect emerging events. If you can detect a problem in your business before it costs you a lot of money, that goes straight to the bottom line. If you can spot a trend before your competition, you have a leg up.

Correlations

Taxonomies capture the concepts embedded in unstructured information. Co-occurrence analysis reveals hidden relationships between these concepts and other attributes or between categories of different taxonomies. For example, we can look for a relationship between technology areas and companies to see where our competition is investing. Or we can find a correlation between a specific factory and a certain kind of product defect.

A correlation is based on the simple idea that two different phenomenon have occurred together more than expected. For example, if 100 customers who talked to a specific call center representative ended up dissatisfied with their overall customer experience, then depending on the total percentage of unsatisfied callers and the total percentage of calls that particular representative took, we could calculate whether there was a correlation between dissatisfaction and talking to this representative. Keep in mind that, even if there is such a correlation, it doesn't mean that this representative is actually responsible for the poor customer satisfaction. It could be that this person only works during weekends and that people who call on the weekends are generally more dissatisfied. This example serves to show that correlations are not causes. They are simply indicators of potential explanations that should be explored further. Think of them as "leading indicators" of business insights.

Classification

One you have a taxonomy that models an important aspect of the information, it is important to be able to apply this classification scheme to new unstructured data. Many classification algorithms exist, and we have incorporated a large variety of them into our approach, allowing us to select the best algorithm for a given taxonomy and information collection. The specifics of how we do text classification is a more technical subject that is beyond the scope of this book. However, the general approach is to pick the algorithm that most accurately represents each category, based on a random sampling of the documents in that category being used to test the accuracy of each modeling approach.

Applications for *Mining the Talk*

Although the number of possible different applications for unstructured mining is virtually limitless, we will explore in-depth applications that fall into five major categories.

Customer Interaction

This is the mining of information coming from unstructured interactions between representatives of the business and customers. It is one of the most common forms of information that nearly every business possesses. This kind of information is good for providing insights into your current processes, what's working and what's not, and identifying areas of potential cost reduction or quality improvement.

Voice of the Customer

This type of mining also involves the customer, or potential customer, but the difference is there is no direct dialog with the business. This includes mining of customer monologues available from surveys, discussion forums, or web logs. This kind of information is useful for discovering what your customers think about you and about your industry. It can provide ideas for improving your products, burnishing your image, or inventing innovative new service offerings.

Voice of the Employee

In this area, the information comes from internal surveys, suggestion boxes, employee discussion events, or open employee forums. This data can provide the business with valuable insight into the collective consciousness of the organization. Such insight can help set a company's vision or generate new ideas for innovation. The results of this kind of mining may help bring disparate groups of the company together to collaborate on new projects and opportunities.

Improving Innovation

Mining the Talk to improve innovation involves looking at both internal and publicly available information sources to find potential ways to innovate through partnering with other businesses. The patent literature is one good source to look for potential opportunities for cross licensing or joint development programs. This data can also be mined to gauge the potential viability of new technologies or product offerings.

Seeing the Future

The ability to see a little bit further ahead than your competition is a crucial competitive advantage. Mining a wide spectrum of publicly available unstructured information sources over time can help the business spot important product and technology trends as they emerge. It can also help in gaining time to react to emerging external events before they become major business catastrophes.

Figure 1-4 shows how the application areas described in the remaining chapters of this book cover the business ecosystem.

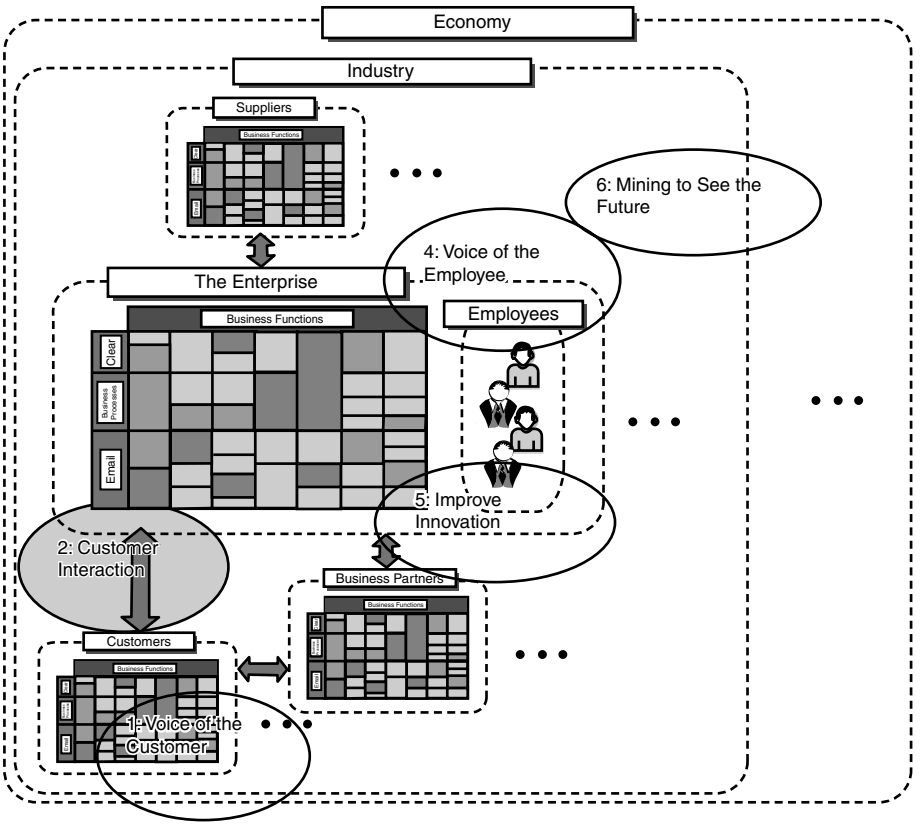

Figure 1-4 Application areas by chapter

The Transformation Process

As we wrote this book, we noticed that the documented examples and case studies followed a general pattern. Based on our experience of working within our own company and working with clients in many different industries, we have found that it is very difficult to get to the full scope of mining unstructured information's potential in one big step. We have found the process to be iterative. As the users and consumers of *Mining the Talk* become more familiar with the approach and build trust that it will achieve the results intended, they will embrace and expand the scope.

Each new potential customer of our approach believes that their situation is unique and their data is different. So typically, we go through a process of showing them the efficacy of our approach on a limited set of their data. This allows us to collaborate with

the customer in exploring the data and enables them to envision the potential of the endeavor. This is usually followed by an enhanced proof point in a major application area, but one still focused on a limited portion of the business. Once the business value is achieved on this proof point, we are on our way to a full-scale deployment and exploring additional potential opportunities in other parts of the business.

This breaks down roughly into five key steps for business transformation process enabled by *Mining the Talk*: 1) Identify the business drivers; 2) Identify the key information sources; 3) Identify what we can learn; 4) Set up an initial unstructured mining process; and 5) Establish a sustainable process that can continue to create value. Subsequent chapters will cover each of these steps in more detail, for each of the application areas.

Summary

We hope this introduction has motivated you to learn more about what this amazing method we call *Mining the Talk* can do. By this point, you should be convinced that there is a growing problem faced by businesses due to ever-increasing complexity of relationships and interactions, and that mining of unstructured information is a potential solution to this problem. We have also introduced the basic premise behind our method, which is that mining of unstructured information requires the capture of domain knowledge and business objectives if it is to consistently succeed. We trust the rest of this book will convince you that this premise is correct.

In subsequent chapters, we will take you on our journey through five application areas of our approach, giving you specific insights into how our approach can be used to address needs within the business. As we go through these application areas, the steps of the *Mining the Talk* methodology will be fleshed out. The next three chapters will address the informal talk that occurs with customers and employees. Subsequently, Chapters 5 and 6, "Mining to See the Future," will address more formal talk that is found in technical literature, patents, and the Web. Finally, in Chapter 7, "Future Applications," we discuss some potential future applications of *Mining the Talk*.

Endnotes

1. This software is called IBM Unstructured Information Modeler, and the appendix of this book describes it in detail.
2. There are many well-established formulas for calculating the readability and grade level of a text sample based on word length, syllables, sentence length, and so on. Here is a reference for more information on these approaches: http://www.readability.info/info.shtml.