WILEY

TIMELY. PRACTICAL. RELIABLE.

Data Mining Techniques

Third Edition

For Marketing, Sales, and Customer Relationship Management

Gordon S. Linoff
Michael J. A. Berry

Several thousand words have been categorized into these categories. The basic set of words consists of those occurring more than four times out of a million, in a large corpus of published documents.

The psycho-social dictionary can be used as a basis for mining sentiment in documents. The simplest measure would simply be the number of positive words minus the number of negative words in the document. A greater number of positive words implies a more positive sentiment.

This is only the beginning. The analysis can incorporate information from other categories as well, so “strong” positive would have a bigger influence than “weak” positive. This information can even be combined into more refined measures than just the number of words. By assigning a “positivity” score to each word that varies between zero and one, you can create a model set appropriate for using naïve Bayesian to combine the words into an overall sentiment value for the document. Naïve Bayesian is definitely not the only approach to opinion mining, but it is a common approach.

Some interesting successes have been achieved using sentiment analysis, both from the academic perspective and from the business perspective. For instance, several academic studies have shown that stock prices are correlated with the sentiment of articles written about them during earlier time periods. This information has also proven useful for investors. Unfortunately, companies that use sentiment analysis for investment are not willing to share case studies.

DIRECTV: A Case Study in Customer Service

DIRECTV is the leading provider in the United States of direct broadcast satellite service that distributes video, music, data, and more via satellite. It has been a leader in using data mining to understand its business, particularly what its customers are saying the hundreds of thousands of times that they call customer service every day.

This case study is not about a single project that used text mining for a specific purpose. Instead, it shows the power of data mining in transforming the customer service side of the business — a power made possible by text mining. The authors are immensely grateful to Dirk De Roos, leader of the customer care analytic team at DIRECTV, and John Wallace of Business Researchers, who worked with Dirk on text mining efforts. The work primarily uses SAS Text Miner and custom code developed by Business Researchers.

Background

DIRECTV competes against other multi-channel video programming providers, including satellite television as well as cable. By using satellites, DIRECTV is able to provide service throughout the United States with a minimum of blackout areas, including to remote areas that are “off the grid.”

The customer service department receives hundreds of thousands of calls every day that are generally in one of four categories:

- Information requests, particularly about the cost of channels: "How much does Showtime and HBO cost?"
- Billing requests: "What is this charge?" "Change the bill to Spanish." "Change address."
- Technical requests: "Customer needs to have slimline dish installed."
- Retention requests: "I want to stop my service."

With a growing business and so many calls, the customer care department is concerned about three things:

1. Gaining a general understanding of what customers are calling about. In particular, DIRECTV wants to identify whether new problems are arising in the business.
2. The average duration of calls. Each second of talk time — on average — costs hundreds of thousands or millions of dollars per year.
3. Understanding and predicting call volumes, particularly for staffing purposes.

The Call Center Interface

The call center software utilized by DIRECTV was originally developed to facilitate customer relationship management. The goal of the software was twofold. First, the software needed to help agents effectively complete calls, record what happened on the call, and take effective action. Second, the software needed to help agents finish the calls in as short a time as possible.

The software was also designed to gather information for business purposes. For instance, there would be a way to track whether the call was a retention call, trying to retain customers who might leave. It also had a way to track upgrade calls, technical service calls, and so on. A key goal of the CRM system was to resolve problems on the first call — resulting in better service for the customer and more efficient use of the call center.

The software was designed to be efficient as well, with drop-down menus for the agents to describe the call. High-level drop-down menus led to lower level ones, and lower level ones to even lower ones, reminiscent of the proverb: "Big fleas have little fleas upon their backs to bite 'em; little fleas have lesser fleas, and so on ad infinitum."

To actually do their work, agents would identify the problem and then access other systems to schedule appointments, updating billing information, and so on. Agents could also type in comments, providing even more detailed information.

What Happened Over Time

The call center interface became the primary source of information about calls. To facilitate marketing efforts, drop-down menus became more and more important. Other departments would add items to the menus to better understand particular offers and products. These "add-on" items to the menus often survived long after the original need was gone.

The end result was confusion. As the menus became more complex, the agents focused on fewer and fewer menu items. In fact, just a handful of items out of several hundred accounted for the majority of calls. Not surprisingly, these items were also the top choices of the menus.

Having most calls assigned default codes greatly reduced the effectiveness of the information. The items on the menus were intended to provide insight into the calls. In the end, they did not.

Applying Text Mining

At this point, Dirk De Roos, contacted John Wallace to develop a system that would help DIRECTV better understand what was happening on the calls. The idea was to use the comments typed in by the agents to categorize the calls.

Using undirected data mining, the system would find clusters in the text. (More technical detail is provided later in this case study.) Although clustering is undirected, there were "directed" goals for the clusters. The resulting clusters needed to make sense, in one of three ways:

- Customer segment
- Root cause of call
- Sentiment

All of these are important parts of the customer interaction. All are presumably available in the information on the call.

Determining the Usefulness of the RV Segment

The initial effort resulted in some interesting clusters. For instance, one of the clusters was the Recreational Vehicle (RV) cluster. It could be readily identified by words in the cluster such as Winnebago, camper, or RV. Satellite TV is popular for such "roving living rooms" because cable and broadcast TV are not available everywhere the campers might roam. Presumably their calls to customer service represented the needs of this particular customer segment.

As a customer segment, the RV group is of interest. Of course, DIRECTV had already identified this group as an important customer segment and had in fact been directing advertising and promotions to this group.

From the perspective of customer service, the RV group was less interesting. This group did not provide insight into improving customer service or agent efficiency or even in understanding why these particular customers were calling. The RV segment was as likely to call about setting up the satellite dish, as about billing problems (because the customers were often away from home), seasonal rate plans, and other issues.

The presence of clusters such as the RV cluster led to the next conclusion. Text mining was producing interesting, but not actionable, results. In researching the problem, the analysts realized that the comments themselves were not providing enough detail about the calls. Without the detail, separating comments into useful clusters was simply not possible.

The second problem was the goal of the analysis itself. Understanding sentiment, customer segment, gathering marketing information, and assigning the root cause for the call are all worthwhile tasks. But trying to attack all of them at the same time did not lead to actionable results.

Acting on the Results

The first round of data mining demonstrated that clusters based on comments could be interesting. The real understanding was that the project needed to be more focused:

- Agent comments needed to be more complete and assigned to more calls.
- The target chosen for the analysis was the determination of the "root cause" of customer calls.

Effecting these changes required modifications both in the call center and in the analytic environment.

WARNING **Trying to accomplish too much in one data mining effort leads to disappointment. Instead, focus on a task that is feasible and develop a culture of constant improvement.**

The call center interface was modified to encourage agents to include comments. For one thing, the comments became mandatory. Another was that the menu items were replaced with actions that occurred on the call. So if the agent changed a customer's address, then "change address" would be added onto the call record automatically.

These modifications turned out to be hugely successful. After the company retrained the agents to focus on typing in comments rather than using the menus, the average call duration went down by almost 5 percent, as shown in Figure 21-7.

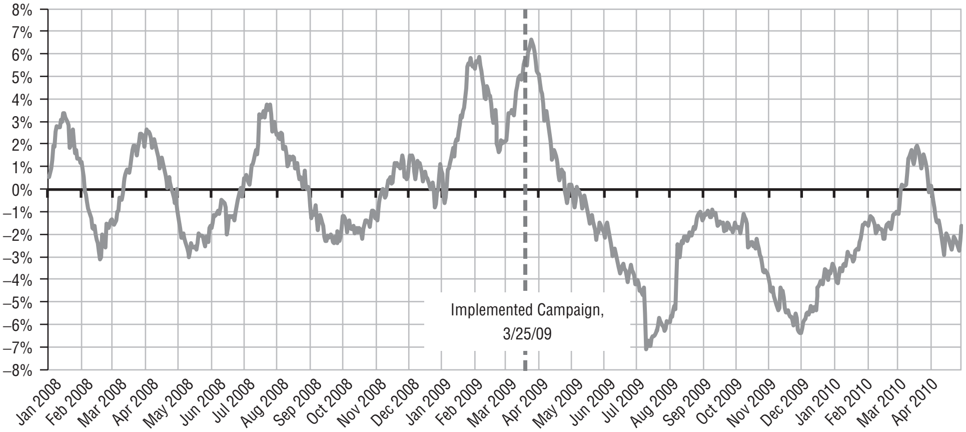


Figure 21-7: The implementation of the new call center interface, inspired by text mining efforts, reduced the average call duration by a noticeable amount.

To someone not familiar with call centers, this may not seem like a large number. Other efforts to reduce the duration of calls generally did not produce any long-lasting decrease at all. Furthermore, this decrease resulted in more complete data, because call center agents included more extensive comments. The decrease in call duration was an early and unexpected win for the text mining efforts.

Continuing Clustering

After working with the agents to clean up the comments, better data was available. The first task for using the improved data was to focus on root-cause clusters. Focusing on one type of clustering would allow DIRECTV to get more actionable results.

Better data and focus did indeed lead to better results. The clustering effort worked in tandem with cleaning and preparing the data — helping to determine what synonym lists to use, which stop words to include in the stop words list, and so on. The goal was to find a set of clusters that defined that "root cause" of the call. The end result was more than 100 clusters for different root causes.

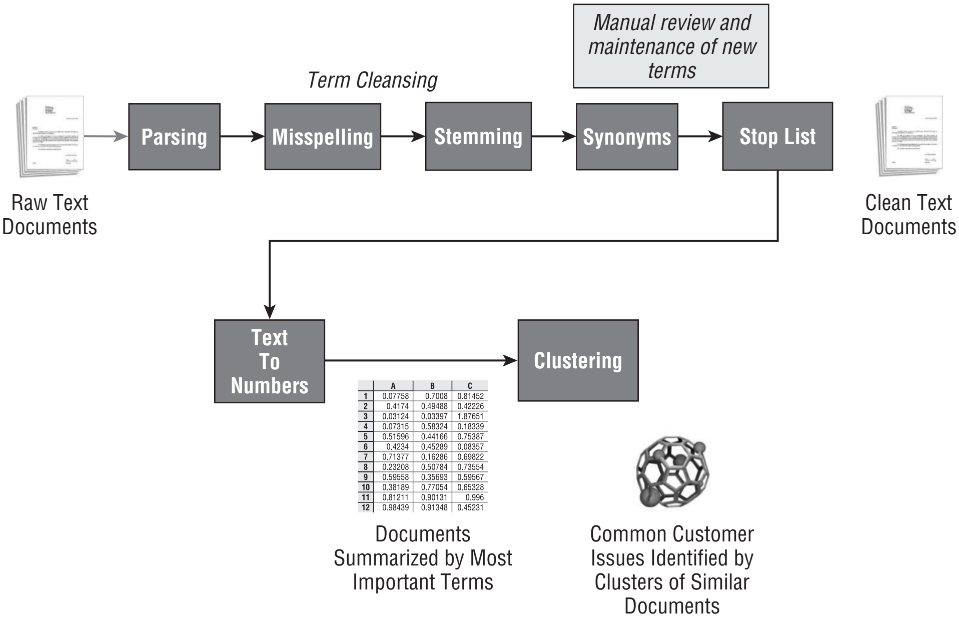
One cluster, for instance, is about changing billing information, such as changing an address or phone number. It turns out that all comments in this cluster are about calls during which such a change was made, even though the clustering only uses information in the comment fields.

This does not mean that everyone who changed billing information is in the cluster, because other things might happen during the call. Perhaps the call is predominantly about satellite dish issues, and requires a service call. During the course of the service call, the customer says something like, "Oh, that's my landline phone number. Use my cell phone instead." The change of billing information is not the root cause.

In the end, more than 90 percent of the comments were able to be assigned confidently to a root cause cluster.

Taking the Technical Approach

Figure 21-8 shows the overall process for creating the clusters as described by Business Researchers. This process has several steps, almost all of them related to managing the lexicon and massaging the text. Only the last step uses data mining algorithms.


Manual review and maintenance of new terms
Term Cleansing
Parsing
Misspelling
Stemming
Synonyms
Stop List
Raw Text Documents
Clean Text Documents
Text To Numbers
Clustering
Documents Summarized by Most Important Terms
Common Customer Issues Identified by Clusters of Similar Documents

	A	B	C
1	0.07758	0.7008	0.81452
2	0.4174	0.49488	0.42226
3	0.03124	0.03397	1.87651
4	0.07315	0.58324	0.18339
5	0.51596	0.44166	0.75387
6	0.4234	0.45289	0.08357
7	0.71377	0.16286	0.69822
8	0.23208	0.50784	0.73554
9	0.59558	0.35693	0.59567
10	0.38189	0.77054	0.65328
11	0.81211	0.90131	0.996
12	0.98439	0.91348	0.45231

Figure 21-8: The process for building document clusters involves many steps to transform the data into a structure usable for analysis.

Parsing the Comments

The first step of the process to create clusters was to identify a set of comments for analysis and to parse them into terms — creating the model set. In general, parsing works by replacing punctuation with spaces, and then taking all terms between spaces. The bag-of-words approach works well for understanding agent comments, because the agents are not trying to write grammatically correct sentences. They are actually trying to quickly capture important ideas in the customer interaction.

There are some tricks in the parsing phase, because punctuation can be used in unusual ways. For instance, one agent might type in “Customer wants HBO and Showtime.” Another might type in “Customer wants HBO/Showtime.” In this case, the slash looks like punctuation. However, in other cases, a slash might be part of a word, such as “w/out.” The trick is to handle these special cases by using special lists of terms that accept punctuation in the middle.

For the DIRECTV work, a random sample of about 400,000 comments was used, containing a total of about 7 million terms. These 7 million terms represented only 76,930 different unique terms, of which more than half (44,826) appeared only one time. Terms that are so rare that they appear only once are not useful for analysis.

Fixing Misspellings

Customer comments are typed in by real people with minimal or no editing. This means that misspellings are quite common. The problem of fixing misspelled words has been investigated almost since the dawn of the computer age. The sidebar “Spelling Suggestions Using Levenshtein Distance” discusses the basic method.

The automated task is simply a matter of constructing a valid dictionary and choosing the closest term. This is an iterative task, where you start with a dictionary and find the closest word to each word not in the dictionary. Some words are quite close (real misspellings) and some are quite far away (suggestions for new words to add into the dictionary).

In the case of DIRECTV, a standard dictionary is augmented with business-specific terms, such as the names of channels (“ESPN,” “Showtime,” and so on). Of the 32,104 non-unique terms in the sample comments, more than half were misspellings (or abbreviations). After taking this into account, the number of unique terms fell to about 22,000.

Stemming

Stemming transforms words into their root forms. For example, one comment might contain “Customer paid too much on last bill; money refunded.” Another might say “Refunding overpayment.” These two comments have no words in common, yet they are saying essentially the same thing. The stemming algorithm recognizes that “overpayment” and “paid” both have the root of “pay.” Similarly, “refunded” and “refund” both have the root of “refund.”

Stemming turns the two comments into “Customer pay too much on last bill; money refund” and “Refund pay.” After stemming, the comments are not grammatically correct. On the other hand, comments similar to each other are much more likely to contain similar terms.

SPELLING SUGGESTIONS USING LEVENSHTEIN DISTANCE

Spell checking has become such a common feature of word-processing software that most people expect to see squiggly red lines underlining misspellings or alternate suggestions whenever they are typing text. These tools work very well for interactive spelling checkers.

The problem of finding misspellings is easier than the problem of suggesting alternatives. Finding misspellings is simply a matter of looking up words in a dictionary, while taking into account proper nouns and grammar (so "sister's" might not have to be in the dictionary).

For text mining purposes, automated spell checking is necessary. In addition, the dictionary of correct spellings is different from a standard dictionary. For instance, it might contain such useful phrases as HD, HBO, and NFL for the DIRECTV example.

This problem of finding alternative suggestions was first investigated in the 1960s by Vladimir Levenshtein when he was working at Moscow State University. He investigated a "distance" metric between two words. The idea is to count the minimum number of "edits" required to go from one word to another. An edit is one of three operations:

- **Substitute one letter for another in the same position.**
- **Delete one of the letters.**
- **Add a new letter.**

Sometimes, a fourth is added: transposing two letters (that is, going from "teh" to "the").

For instance, to transform "man" to "woman" requires two edits, adding a "w" in the first position and then adding an "o" in the second position. Of course, there are other possibilities, but two edits is the shortest transformation between the words.

Levenshtein also devised an algorithm for finding the fewest transforms. This algorithm is efficient enough that it can be run on large documents against a large dictionary in a short amount of time (computers have improved a lot since the 1960s).

The approach has also been refined, such as by giving different weighting to the first letter, breaking words into smaller pieces, and using context to determine the best alternative word.

Applying Synonym Lists

Synonym lists are words and phrases that are recognized in the text and replaced by a common synonym. These serve several purposes, including fixing misspellings and finding word phrases. The synonym lists used by DIRECTV also combined similar details into a higher level idea. For example, "Change address" and "Change phone number" both turn into "Change account info."

The lists can also be used to fix misspellings. For example, these might all be synonyms for Showtime (one of the channels offered by DIRECTV):

- Showtime
- Show time
- Show-time
- ST
- Showt
- Shwotme

The synonym lists turn these into a single word (in this case "Showtime"). Notice that the last term is simply a misspelling. Looking up the term in a dictionary of common words would come up with "snowmen," illustrating why common dictionaries must be augmented by domain-specific terms.

Using a Stop List

The stop list contains words that have minimal meaning, as discussed earlier in this chapter. Stop words can also be meaningful terms that simply do not distinguish between comments. The word *customer* occurs in many comments, but doesn't provide information about the root cause. Similarly, *television* and *TV* are not useful, because all DIRECTV customers are, presumably, calling about something related to their television. Almost all customers have high-definition TV, so *HD* is a stop word as well.

TIP The stop word list should remove words that might seem meaningful but are not likely to distinguish between documents. In the DIRECTV example, this includes *customer, television*, and *HD*.

The purpose of the stop word list is to remove words that do not distinguish between different comments, even when these words might seem meaningful. The synonym and stop list used by DIRECTV did not significantly reduce the number of terms in the document. The size of the vocabulary for the comments ended up having about 22,000 distinct terms.

Converting Text to Numbers

Figure 21-8 shows a step called "Text to Numbers." This chapter has already described term-document matrixes and the use of singular value decomposition (SVD) to transform documents into numbers. This technique was used in this case, using SAS's Text Miner software to calculate 50 singular value vectors for the clustering.

Clustering

The final step is clustering. As described in Chapters 13 and 14, many different clustering algorithms can find clusters in data. The clustering method used in this case was Gaussian mixture models (GMM), also known as expectation-maximization clustering. In fact, several different methods were tried, notably k-means as well as GMM. The clusters generated by the GMM algorithm tended to be better, based on subjective judgment of the resulting clusters. As described in Chapter 14, GMM does a better job of distinguishing between clusters that have oblong shapes.

Not an Iterative Process

Describing the process as a set of seven steps makes it sound like one of Julia Child's recipes. Like the recipes for fine French food, the steps of data mining are all interrelated and affect each other. Unlike the steps in such recipes (at least when Julia followed them), steps often need to be repeated and in different sequences.

The choice of dictionary for correcting misspellings has a big impact on the choice of synonyms later. The validity of the clusters, in turn, has an impact on the choice of stop words and synonyms. In fact, looking at preliminary clusters made it obvious that words like *customer* and *HD* should be part of the stop word list.

Continuing to Benefit

DIRECTV did not stop its text mining effort with a single set of clusters used for analyzing customer comments. On a daily basis, it "scores" new customer calls with the root-cause clusters, provided by data mining. Every six to twelve months, it dives deeper into the clusters, to determine whether the clusters are still working and whether new ones are needed.

The root-cause clusters have become more and more useful over time. In one case, DIRECTV was assessing the clusters by looking at average call duration over time. Normally, this would provide insight into where the customer service agents were doing a good job and where room for improvement might reduce the average talk time. It might also suggest new "hard" problems that are taking the agents longer to solve.

In this case, though, the clusters at the top of the list for increased call duration were all similar and related to billing. Changing a billing address had an increase in talk time. Providing a refund had an increase in talk time. Changing the rate plan had an increase in talk time. Most other clusters did not show an increase in talk time.

After investigating more, analysts found that the increase started in one particular month, rather abruptly. Further discussions revealed that there had been an IT "upgrade" during this period, an upgrade that slowed down the billing system. Without the benefit of root-cause clusters, such a change would probably not have been noticed — and the cost of the extra talk time would have continued until the next upgrade would (presumably) fix the problem.

Lessons Learned

Text mining is the application of data mining to text data, which can come from many different sources. Typical sources used in business include news articles, blog entries, comments on customer service calls, and e-mails sent in by customers.

The purpose of text mining may be to understand the documents, summarize them in some way, or to cluster them into similarity groups. Identifying a specific characteristic (such as finding which customers stopped because of a boycott) might be sufficient. In other cases, features extracted from text are combined with other data to build more traditional data mining models. Newer applications are being developed in the area of sentiment analysis and opinion mining, determining the attitude of the writer to what he or she is writing about.

The most challenging aspect of text mining is managing the text itself. At the extreme are two different methodologies. The first is the bag-of-words approach, which treats documents as unordered lists of their words. At the other extreme is the natural language processing approach, which takes into account the meanings and grammatical features of the language. There are many variations of both methods.

Processing the text for the bag-of-words approach usually involves removing stop words, common words, fixing misspellings, stemming, and replacing words with synonyms. The next step is to define the term-document matrix, which is a giant matrix with a row for each document and a column for each term. The cells generally contain the inverse term frequency of each term in the set of all documents.

The next step is the text equivalent of principal components, called singular value decomposition. This reduces the text to points in a multidimensional space, where more traditional data mining techniques can be used. Alternatively, naïve Bayesian models are often a very effective modeling tool.

As the examples in this chapter show, text mining has a very broad range of applications. The DIRECTV example provides an excellent illustration not only of text mining, but also of the virtuous cycle of data mining introduced in Chapter 1.