What Is Text Mining?

Marti Hearst
SIMS,UC Berkeley
hearst@sims.berkeley.edu

October 17, 2003

I wrote this essay for people who are curious about the topic of text mining after having read the New York Times article by Lisa Guernsey (10/16/2003) or heard my Future Tense interview with Jon Gordon (10/20/2003).

What is text mining? What are its potential applications and limitations?

Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation.

Text mining is different from what we're familiar with in web search. In search, the user is typically looking for something that is already known and has been written by someone else. The problem is pushing aside all the material that currently isn't relevant to your needs in order to find the relevant information.

In text mining, the goal is to discover heretofore unknown information, something that no one yet knows and so could not have yet written down.

Text mining is a variation on a field called data mining, that tries to find interesting patterns from large databases. A typical example in data mining is using consumer purchasing patterns to predict which products to place close together on shelves, or to offer coupons for, and so on. For example, if you buy a flashlight, you are likely to buy batteries along with it. A related application is automatic detection of fraud, such as in credit card usage. Analysts look across huge numbers of credit card records to find deviations from normal spending patterns. A classic example is the use of a credit card to buy a small amount of gasoline followed by an overseas plane flight. The claim is that the first purchase tests the card to be sure it is active.

The difference between regular data mining and text mining is that in text mining the patterns are extracted from natural language text rather than from structured databases of facts. Databases are designed for programs to process automatically; text is written for people to read. We do not have programs that can "read" text and will not have such for the forseeable future. Many researchers think it will require a full simulation of how the mind works before we can write programs that read the way people do.

However, there is a field called computational linguistics (also known as natural language processing) which is making a lot of progress in doing small subtasks in text analysis. For example, it is relatively easy to write a program to extract phrases from an article or book that, when shown to a human reader, seem to summarize its contents. (The most frequent words and phrases in this article, minus the really common words like "the" are: *text mining, information, programs,* and *example*, which is not a bad five-word summary of its contents.)

There are programs that can, with reasonable accuracy, extract information from text with somewhat regularized structure. For example, programs that read in resumes and extract out people's names, addresses, job skills, and so on, can get accuracies in the high 80 percents.

I don't consider this to be text mining; rather it falls into an area called information extraction. However, I am a bit of a purist when it comes to defining what text mining is. I distinguish between what I call "real" text mining, that discovers new pieces of knowledge, from approaches that find overall trends in textual data.

An analogy I like to use comes from the realm of crime fighting. I think discovering new knowledge vs. showing trends is like the difference between a detective following clues to find the criminal vs. analysts looking at crime statistics to assess overall trends in car theft.

People are using the output of such programs to try to link together information in interesting ways. For example, one can extract all the names of people and companies that occur in news text surrounding the topic of wireless technology to try to infer who the players are in that field. There are a number of companies that are investigating this kind of application.

One problem with these approaches is that it is difficult to recognize which of the many relations that are shown are truly interesting. You'll immediately see who the big players are, but anyone who knows the business will already be aware of this. You'll also see many, many weak links between various players, hundreds or thousands of such links, and you can't tell which are the really interesting ones that you should pay attention to.

The most active, and I think promising, application area for text mining is in the biosciences. The best known example is Don Swanson's work on hypothesizing causes of rare diseases by looking for indirect links in different subsets of the bioscience literature.

As another example, one of the big current questions in genomics is which proteins interact with which other proteins. There has been notable success in looking at which words co-occur in articles that discuss the proteins in order to predict such interactions. The key is to not look for direct mentions of pairs, but to look for articles that mention individual protein names, keep track of which other words occur in those articles, and then look for other articles containing the same sets of words. This very simple method can yield surprisingly good results, even though the meaning of the texts are not being discerned by the programs. Rather, the text is treated like a "bag of words".

To get farther though we need more sophisticated language analysis. A number of us are working on statistical techniques that try to assign semantics, or meaning, to parts of the text. We break off pieces of the problem of analysis, targetted towards particular applications, rather than trying to "read" the articles as a whole. This goal is especially promising in the biosciences due to the nature of the text itself. In some ways it is easier to process automatically than ordinary text. It is less ambiguous and the processes it describes are somewhat mechanical, and so representable in a computer.

The fundamental limitations of text mining are first, that we will not be able to write programs that fully interpret text for a very long time, and second, that the information one needs is often not recorded in textual form. If I tried to write a program that detected when a where a new word came into existence and how it spread by analyzing web pages, I would miss important clues relating to usage in spoken conversations, email, on the radio and TV, and so on. Similarly, If I tried to write a program that processes published documents in order to guess what will happen to a bill in Washington DC, I would fail because most of the action still happens in negotiations behind closed doors.

For more information, see:

Untangling Text Data Mining, Marti Hearst, *ACL'99*.