

October 16, 2003

Digging for Nuggets of Wisdom

By Lisa Guernsey

Digging Deeper

Text-mining software can examine thousands of documents, pulling relevant information into categories and drawing connections between the categories.

Mining unstructured data

While data-mining software pulls items from structured databases like inventories, text-mining software "reads" unstructured items like news articles.

(c) 2001, Chicago Tribune.
By Stephen J. Hedges and
Cam Simpson

The **Finbury Park Mosque** is the center of radical Muslim activism in England. Its doors have passed at least three of the men now held on suspicion of terrorist activity in France, England and Belgium, as well as one Algerian man in prison in the United States.

"The mosque's chief cleric, **Abu Hamza al-Masri** lost two hands fighting the Soviet Union in Afghanistan and he advocates the elimination of Western influence from Muslim countries. He was arrested in London in 1999 for his involvement in a Yemen plot, but was set free because the Yemen failed to produce evidence to have him extradited."

The software reads through one of many documents to begin to draw links between them on the subject of terrorism.

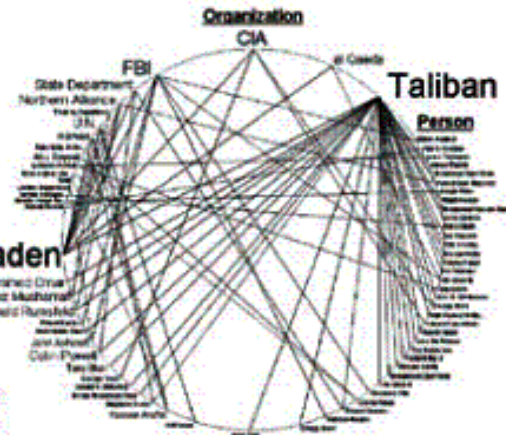
Here, it determines that **Abu Hamza al-Masri** is the name of a person. It then determines his position or title by referring to the first part of the sentence and the previous paragraph.

```
<Person|Position|Organization>
<OFFLEN OFFSET="3576" LENGTH="45">
<Person>Abu Hamza al-Masri</Person>
<Position>chief cleric</Position>
<Organization>Finbury Park Mosque</Organization>
<Person|Position|Organization>
```

Source: ClearForest

Making connections

The software can create links between documents and draw visual maps.

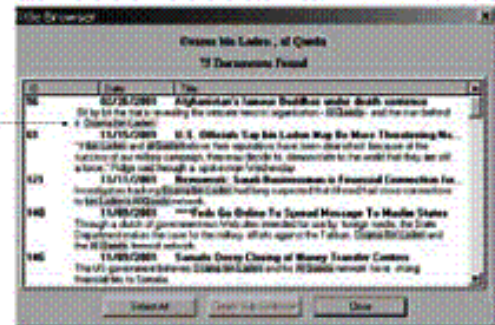


Osama bin Laden

After reviewing several thousand documents, it ties together names that appear to be linked. The larger words have stronger associations.

Retrieving the results

In some cases, the user can click on certain linkages to get a list of all the documents that include references to both words or concepts.



MICHAEL N. LIEBMAN knows his limitations. Even with a Ph.D. and a long career in medical research, he cannot keep up with all the developments in his area of interest, breast cancer. Medline, the database that already houses more than 10 million abstracts for journal articles, is adding 7,000 to 8,000 abstracts per week. Only a fraction of these are about cancer, but the volume of information is daunting nonetheless. "There is just too much literature to be able to go through it all," said Dr. Liebman, the director of biomedical informatics at the Abramson Family Cancer Research Institute at the University of Pennsylvania. Yet Dr. Liebman is convinced that new cures could someday emerge for breast cancer if only someone could read all the literature and synthesize it. So he has found a solution: enlisting a computer program to read the articles for him.

"The software is not going to get tired," he said. It also happens to be a speed reader: The product he is using, from a Chicago-based software company called SPSS, can zip through 250,000 pages an hour. Another product, from the text-mining company ClearForest, boasts a speed of 15,000 pages an hour, still far surpassing the human rate of a mere 60 pages.

Of course, no one, Dr. Liebman included, is arguing that these products are actually reading anything. What they are engaged in is "text mining," a technique that academics have been experimenting with for years but for which tools have only recently become commercially available. The prospect of rapidly scanning through reams of documents is stirring interest among researchers and analysts faced with more material than they can handle.

To the uninitiated, it may seem that Google and other Web search engines do something similar, since they also pore through reams of documents in split-second intervals. But, as experts note, search engines are merely retrieving information, displaying lists of documents that contain certain keywords.

Text-mining programs go further, categorizing information, making links between otherwise unconnected documents and providing visual maps (some look like tree branches or spokes on a wheel) to lead users down new pathways that they might not have been aware of.

Currently these programs are used by academic researchers and companies, but information scientists expect that to change. Lower-cost text-mining tools eventually will be offered to ordinary people who want to dig into medical or political issues using public documents. Madan Pandit, an expert in text analysis in Bangalore, India, who runs a Web site called K-Praxis (k-praxis.com), has suggested that text mining could help people make sense of voluminous documents that are already on the Web, like the 858-page report on the congressional inquiry into intelligence failures regarding the 9/11 terrorist attacks.

"There is a need to make these technologies available for publicly available information," he wrote at his site.

In most cases, text-mining software is built upon the foundations of data mining, which uses statistical analysis to pull information out of structured databases like product inventories and customer demographics. But text mining starts with information that doesn't come in neat rows and columns. It works on unstructured data - e-mail messages, news articles, internal reports, transcripts of phone calls and the like.

To make sense of what it is reading, the software uses algorithms to examine the context behind words. If someone is doing research on computer modeling, for example, it not only knows to discard documents about fashion models but can also extract important phrases, terms, names and locations. It can then categorize them and draw connections among the categories.

How well computers truly make sense of what they are reading is, of course, highly questionable, and most of those who use text-mining software say that it works best when guided by smart people with knowledge of the particular subject.

"I was an F.B.I. agent for 20 years," said Randall S. Murch, now a researcher at the Institute for Defense Analyses, which works for the Office of the Defense Secretary and other government agencies. "And I have yet to see anyone who is able to model the way an agent thinks and works through an investigation."

Text-mining software also can stumble when trying to parse the nuances of language. In other words, hold the sarcasm: If you send an e-mail complaint with references to "oh-so-helpful salesmen who clearly know their customers," text-mining software might eventually categorize your note as a compliment.

But advocates say that when the software is used on niche sets of text, it can make a difference. Intelligence agencies, for example, can start to find connections between seemingly unconnected individuals and organizations. People responsible for keeping up with developments in an industry can use the software to scan, categorize and even summarize thousands of articles at a time. Computer makers can better analyze the masses of e-mail messages that pour into technical support centers.

As much as 80 percent of a company's knowledge base may reside in documents that might have been considered unusable, industry analysts say. With text mining, they say, that text can become part of the stream of data flowing through a company's analytic systems.

"Now it's not just about what is easily encoded in a medical claim record," said Dan Sullivan, president of the Ballston Group, an information-management consulting firm. "Words matter, and words will become accessible again."

So what exactly has text mining discovered already?

Take the Fireman's Fund Insurance Company, which in 1999 started seeking a way to spot fraud without always having to rely on its staff to pick up on suspicious activity.

The company had already tried to solve the problem by crunching just the hard data, like the coded numbers signifying types of claims. But the computers kept flagging so many false positives that Marty Ellingsworth, director of operations research, decided to take a look at how the company's human investigators did their work. He found that they gleaned a lot from the free-form notes typed in by claims adjusters - about a client's behavior, say, or offhand comments.

An adjuster examining an accident between two vehicles, for example, may have noted that the front driver slammed on his brakes in light traffic, a potential tipoff to a staged rear-ender. "When we looked over their shoulders and saw them reading the text, we saw that we had to do that too," he said. Now the company uses software to process those notes along with hard data.

The company uses similar methods to determine which insurance cases might be fruitful for subrogation, the practice of extracting payments from other insurance agencies for damages that appear to be caused by the other agency's clients. Since it started mining for that purpose, it has collected \$1.4 million that it would have otherwise missed.

The best-known anecdote about text mining involves Don R. Swanson, a professor emeritus of information science at the University of Chicago who in the 1980's decided to take a deep look at medical literature on migraines. Starting only with the word "migraine," he downloaded abstracts from 2,500 articles from Medline and looked closely at the titles. When certain concepts caught his eye, he conducted new searches to see whether that concept existed in the full texts of other articles related to migraines.

In one instance, a reference to a neural phenomenon called "spreading depression" caused him to look for articles with that term in their titles. Reading those pieces, he found that magnesium was often mentioned as preventing this spreading depression. Other connections to magnesium deficiencies started to appear, so he dug further. In a 1988 paper on his research, he wrote, "One is led to the conjecture that magnesium deficiency might be a causal factor in migraine."

Today, Dr. Swanson's work is considered significant both for migraine studies and for text mining. The link between the headaches and magnesium deficiency was soon backed up by actual experiments.

Information scientists say his 1988 discovery is a perfect example of the unexpected connections that can reveal themselves among scattered text fragments - revelations that may surface even more quickly with the help of powerful software scanning thousands of pages an hour.

Dr. Swanson produced that work before the days of the Web, with the help of very rudimentary programs that organize data, and did most of the connecting of concepts and terms himself. But even today's more sophisticated text-mining programs - which can cost corporate clients thousands of dollars - are not yet designed for Web searches.

And even the most ardent fans of text mining warn that the software is useless without human brain power.

Marti Hearst, an associate professor of information systems at the University of California at Berkeley, said that text-mining analysts can suffer from overload. The visual maps that present unexpected links in data "can turn into spaghetti," Dr. Hearst said. "You have a million links. Which one is important?"

Before the software goes to work, it requires a human's expert input. To prepare for his text-mining project on breast cancer articles, for example, Dr. Liebman spent months building a framework of knowledge on the disease so that the software could categorize articles and concepts in a meaningful way. Over the summer he began to feed that framework into the software, along with the reams of articles that publishers made

available.

At his desktop, he can now pull up the fruits of the software's work: a visual map of extracted concepts all tied together in sometimes unexpected ways. Terms like breast cancer are linked to others like obesity or puberty, which when clicked on call up details about the articles that mention such concepts, even if they are deep within their texts.

"If you collect weak observations over a large number of journals, they become strong," Dr. Liebman said.

"We're using text mining to take advantage of what other people have seen but may not necessarily have felt to be significant observations."

After the software has yielded its stunning map, he develops hunches about which connections are worth exploring further. That takes him down paths requiring more computer analysis and more brain power.

Dr. Liebman said that while it was too early to know if he has uncovered anything extraordinary, he and his staff had become intrigued with the connections - or, rather, the lack of them - that appear about diseases like cancer, and women who deliver babies later in life.

"There have been studies that show there is no impact on the pregnancy or the child," he said, "but one of the things we're trying to pull out is, does late pregnancy have an impact on postmenopausal disease?" The question is not one that they had ever studied, he said.

Still, the question was not posed by the software. It sprang from the minds of the researchers, people with human curiosity and years of personal experience. "This is about identifying the right question, not just synthesizing data," Dr. Liebman said. "If you don't have the right question, it doesn't matter how much data you are looking at."