



**CENTRAL
CO-ORDINATION:**
Bente Maegaard
Center for Sprogteknologi
Njalsgade 80
DK - 2300 Copenhagen S
Tel: +45 35 32 90 90
Fax: +45 35 32 90 89
euomap@cst.dk
www.hltcentral.org/euomap

Mining for meaning: the joy of text

By Lucille Redmond

Sieving words for their meaning is what distinguishes humans from machines. But technology is catching up fast.

It started with data mining – software that could mill through billions of figures and analyse them for the statistics and trends apposite to your particular company. Or your army, or even your manor, if you were a police chief looking to know why there was suddenly a lot of crack turning up in the southwestern end of a suburb, beside where a new link road had just been built.

But four-fifths of the data on intranets and on the World Wide Web is not figures, but text. And the information in that text is invisible to the programs that trawl for data. So the race is on to develop software that can analyse text in the same way that data mining software analyses figures.

Huge profits dangle temptingly before the people racing to develop the technology – one figure being bandied around is an estimated €5 billion to be made in the next two years from text mining.

An early push for the technology came from a seminal event in the 1990s known as MUC – the Message Understanding Conference, founded by US defence research group DARPA.

It was a series of international trials of research systems. Groups working in text processing would be given a common task, and would have a few months to train their system to do it. Then they would be given some live data and have to run the data and try to extract knowledge – then the results would be compared and the systems ranked.

A decade later, two technologies are merging: data mining and text analysis. With the combination, suddenly the data miner has access to the data in text.

The data mining approach to text processing is to use information retrieval (IR) technology. It uses pattern matching, keyword matching or word frequency analysis, to discover what a document is about – essentially treating a document as if it were maths.

“This uses statistical techniques to process documents,” said University of Brighton text mining specialist Roger Evans. “You can use that to locate documents on a particular topic, or to try to route documents in your organisation.”

The linguistic approach, on the other hand, is text analysis – natural language processing (‘natural’ languages, to a linguist, being ones that humans speak – English, French, German, Japanese – rather than the ones like machine language that computers speak). This is cutting-edge stuff – in effect developing artificial intelligence.

“Although at a theoretical level we’re still not very good at it, we have developed it to a point where you could produce systems which actually try to understand a piece of text and pull out information, rather than just scanning it as a list of words,” said Evans.

Is my name in there?

“That technique is called information extraction. It will, for example, go through a lot of text trying to find references to companies.”

This is not as simple as it sounds. “It will look for the name of the company – but then often in a document, once you’ve mentioned the company you will refer to it in an abbreviated form.

“So if the software can detect words which are company names or personal names, and group them together into the same company, you can start counting how many references there were to that company.

“Then you can start analysing what the sentences say, and try to understand references to a company that might just be ‘it’ or ‘the company’ and see which company is being referred to – for instance if one company took over another. You start trying to identify events that occurred in a document – and you gradually build up more and more sophisticated linguistic techniques to try and build a still fairly simple-minded but more knowledge-intensive view of what the document is about,” said Evans.

“Once you’ve got into that information, you’ve got away from language, into a kind of symbolic state which ordinary data mining is happy with, to transform your document into something more like a data table. Then you can apply traditional data mining techniques.

“For instance, you might scan all the documents on a newswire to look for documents about company takeovers, and discover that one of your competitors is aggressively trying to take over small companies.

“Or you might look at newswire documents about the power industry and power stations, and discover trends – where the best markets are for nuclear power, for instance, or wave power.”

What do my customers say?

In customer relationship management, traditional data mining can analyse any market survey questionnaire – but only the answers in the check boxes. “There will be a big survey and what is collected is a mixture of multiple choice tick-boxes and customers’ comments about the products.

“They’ve never been able to do anything with those customers’ comments before on a large scale, because you have to process them. So these companies have a big interest – they can type the answers into their machines and be able to process them,” said Evans.

The traditional players in the market are companies like IBM, SPSS and SAS, all of which are already big in data mining. “The trend towards text mining is in the last two or three years, now that they’ve come alive to the fact that it’s possible, and that their data mining tools are exactly the right way to do it,” he said.

The leading edge of text mining research is into document understanding. For instance, on the World Wide Web there are billions of documents with headings and subheadings and hypertext links between them, and containing pictures. And the documents may be anywhere on the Web.

“The emphasis in document understanding is to try and cope with all that information. Pure text mining just thinks of documents as a stream of words, and ignores anything else that’s going on,” said Evans.

In a news article the story is in the headline, and this is expanded in the first paragraph, and further explained as the article progresses. So it is highly relevant to know whether you are in a headline – which contains the most important information, but uses the strange grammar of headline-writers – or in plain text, or in a hypertext link or a subheading.

Spooky science

Language technology is not yet widespread in the ordinary user market, but in specialised commercial areas it is finding a ready market for its tools, and for consultancy and training.

Yet it uses technologies that an ordinary Joe might find spookily futuristic. “Some of them use neural nets and genetic algorithms,” said Evans.

“Neural nets are computer programs which at some level make very simplified models of brains. So they process by having little neurons that are linked to each other with synapses.

“They have numbers related to them, to tell you when one fires relative to another. Then you train them by giving them lots of data for a particular problem, and that sets their probability weightings. Then you give them data for a new problem, and they settle down into a state where they give you an answer.”

Genetic algorithms are computer programs which learn in a way similar to evolution. “They try lots of solutions and work out which one is the best. They keep the best ones, then get them to breed with each other – solutions are changed by swapping bits of one solution with another so that they generate a new solution which is almost like the existing solution, but not quite. Then they make up thousands and thousands of those, and pick the best ones.”

The solutions get better and better, until you come up with a superb solution that you could never have written by hand, because the problem is too difficult.

An alternative, more traditional approach uses grammatical software that processes documents for its structure. Even this must use grammar in a relaxed way, looking for clues to the grammatical structure of a piece of text, rather than seeking exact structures.

One of the problems for computers is that they tend to lack a sense of humour – so far, anyway, no one has succeeded in programming funniness. “It’s virtually impossible to pick out if a document is being sarcastic, and is saying things positively which are actually meant negatively,” said Evans. “The number of contexts that can make what looks like a simple statement be negative is enormous – ‘You can scarcely believe that... X’ probably means ‘...not X’.”

Websearch plus infotech

Linguamatics is a one-year-old natural language processing company in Cambridge which specialises in information extraction and dialogue systems. It was founded by four Cambridge University buddies who had worked together in a US company which closed its Cambridge office – so they decided to set up in business themselves.

“We’re building an interactive information extraction system,” said founders Roger Hale and James Thomas. “Typically information extraction systems have been operated in a batch mode, and you’ve really needed to be a linguist or a domain expert or both to use them.”

Linguamatics wants to speed the process up, combining the technology used in something like a Web metasearch engine with the technology of information extraction, to get accuracy and speed, and using natural language technology to pinpoint exactly what a document is about.

“We’re looking at the biomedical domain,” said Thomas. “There are lots of people interested in finding out about proteins or genes, and there’s an absolutely enormous literature, but nobody’s got the time to read it all.”

For example, the medical database Medline has some 10 million abstracts in it. “Obviously if you’re a researcher interested in the interactions of a particular gene, there’s no way you could read all the abstracts – even if you could do an IR (information retrieval) query to find something that mentioned it.”

Currently there are several approaches, say the Linguamatics founders – information retrieval, question-answering systems, statistical clustering, document categorisation and summarisation. “They all offer useful services where you can retrieve pages of information.”

The multinational SPSS is at the cutting edge of this new technology. Tom Khabaza is programme manager of the advanced data mining group in its business intelligence division.

Originally Khabaza and his colleagues in a company called Integral Solutions developed a program called Clementine to streamline data mining processes. Clementine was a GUI – a ‘graphical user interface’ – in other words, its information was displayed as icons, like a Windows program.

Clementine pictured the connections between different pieces of data as icons connected by arrows, and allowed the users to play with the data by whizzing the information back and forth along the arrows.

When SPSS bought Integral Solutions it got Clementine, and Khabaza. "From the first days of Clementine, every so often we'd bump into somebody who'd say: 'I've got a lot of data, but it's in free text, so this isn't really going to help'," said Khabaza.

Industry and university

Around four years ago a gas company realised it needed to mine text as well as data, and asked Integral Solutions to try to write some software for it. Khabaza got together with Roger Evans of Brighton University to make a text mining demonstration program, to show what the technology might be able to do.

"It used information extraction technology – supplied by Roger and his team – to analyse documents and turn them into structured data, which we would then analyse using Clementine," said Khabaza.

The result was impressive, but it took a lot of natural language engineering to get to the point where the text could be analysed.

"The software that had been produced would really only analyse text about gas pipelines and suchlike, so it was fine for that particular application, but there aren't that many people who want to know that particular type of information," said Khabaza. "If you wanted to analyse a different type of text for a different purpose you'd have to get your engineers in again, to re-engineer the whole natural language part."

Then SPSS acquired a text mining company called Lexiquest, and integrated its Lexiquest Mine software with Clementine to produce software which can do both parts of the job – analysing the language in the text, and analysing the resulting data.

High end of the market

Text mining software is not – at this stage of the game, anyway – stack-'em-high-and-sell-'em-cheap ware. The customers are huge entities, and the prices are currently in tens of thousands – about €50,000 for a good data mining program, and up to €200,000 more for text mining capability. There are data and text mining workbenches that cost much more, and much less than this – it depends what functions you need how much you are prepared to pay.

Olivier Jouve, vice president of text mining in SPSS, says that today's customers are mainly large companies in the areas of customer relationship management (CRM), competitive intelligence or pharmaceuticals.

"The first customers were more involved in competitive intelligence – any big company has specific divisions looking at its competitors, finding out what they are doing, using external sources like the Web or databases," Jouve said.

What drives the price up in text mining is the factor of linguistics: analysing text involves using dictionaries, grammars, and processing the concepts of languages as different as English, French, Dutch and Japanese (data mining is big in Japan).

The price goes up as more functions are added – methods for accessing data from different sources, for dealing with graphics, for reporting.

This is a baby technology, in a place something like where XML was five years ago: everyone wants it, but no one is sure how it will turn out. But all the signs are that text mining is going to be huge, and that the technologies which it is developing are going to change the face of computing.