

What Text Mining Can Do

An alert reader will make connections between seemingly unrelated facts to generate new ideas or hypotheses. However, the burgeoning of published text means that even the most avid reader cannot hope to keep up with all the reading in a field, let alone adjacent fields. Nuggets of insight or new knowledge are at risk of languishing undiscovered in the literature.

Text mining offers a solution to this problem by replacing or supplementing the human reader with automatic systems undeterred by the text explosion. It involves analysing a large collection of documents to discover previously unknown information. The information might be relationships or patterns that are buried in the document collection and which would otherwise be extremely difficult, if not impossible, to discover. Text mining can be used to analyse natural language documents about any subject, although much of the interest at present is coming from the biological sciences.

Take interactions between proteins, for example. This area of research is important for the development of drugs to modify protein interactions that are linked to disease. Text mining can not only extract information on protein interactions from

documents, but it can also go one step further to discover patterns in the extracted interactions. Information may be discovered that would have been extremely difficult to find, even if it had been possible to read all the documents. This information could help to answer existing research questions or suggest new avenues to explore.

How Text Mining Works

Text mining involves the application of techniques from areas such as information retrieval, natural language processing, information extraction and data mining. These various stages of a text-mining process can be combined into a single workflow. We will now look in more detail at each of these areas and how, together, they form a text-mining pipeline.

Information retrieval (IR) systems identify the documents in a collection which match a user's query. The most well known IR systems are search engines such as Google™, which identify those documents on the WWW that are relevant to a set of given words. IR systems are often used in libraries, where the documents are typically not the books themselves but digital records containing information about the books. This is however changing with the advent of digital libraries, where the documents being retrieved are digital versions of books and journals.

IR systems allow us to narrow down the set of documents that are relevant to a particular problem. As text mining involves applying very computationally intensive algorithms to large document collections, IR can speed up the analysis considerably by reducing the number of documents for analysis. For example, if we are interested in mining information only about protein interactions, we might restrict our analysis to documents that contain the name of a protein, or some form of the verb 'to interact' or one of its synonyms.

Natural language processing (NLP) is one of the oldest and most difficult problems in the field of artificial intelligence. It is the analysis of human language so that computers can understand natural languages as humans do. Although this goal is still some way off, NLP can perform some types of analysis with a high degree of success. For example:

- *Part-of-speech tagging* classifies words into categories such as noun, verb or adjective

Text mining suggests new uses for thalidomide

Marc Weeber and colleagues used automated text mining tools to infer that the drug thalidomide could treat several diseases it had not been associated with before. Thalidomide was taken off the market 40 years ago, but is still the subject of research because it seems to benefit leprosy patients via their immune systems. Weeber and Grietje Molema, an immunologist, used text mining tools to search the literature for papers on thalidomide and then pick out those containing concepts related to immunology. One concept, concerning thalidomide's ability to inhibit Interleukin-12 (IL-12), a chemical involved in the launch of an immune response, struck Molema as particularly interesting. A second automated search for diseases that improve when the action of IL-12 is blocked revealed several not previously linked with thalidomide, including chronic hepatitis, myasthenia gravis and a type of gastritis.

'Type in thalidomide and you get 2-3000 hits. Type in disease and you get 40,000 hits. With automated text mining tools we only had to read 100-200 abstracts and 20 or 30 full papers. We've created hypotheses for others to follow up,' says Weeber.

Weeber et al. *J Am Med Inform Assoc.* 2003 10 252-259

Text Mining

Version 2: September 2008

- *Word sense disambiguation* identifies the meaning of a word, given its usage, from the multiple meanings that the word may have
- *Parsing* performs a grammatical analysis of a sentence. Shallow parsers identify only the main grammatical elements in a sentence, such as noun phrases and verb phrases, whereas deep parsers generate a complete representation of the grammatical structure of a sentence

The role of NLP in text mining is to provide the systems in the information extraction phase (see below) with linguistic data that they need to perform their task. Often this is done by annotating documents with information such as sentence boundaries, part-of-speech tags and parsing results, which can then be read by the information extraction tools.

Information extraction (IE) is the process of automatically obtaining structured data from an unstructured natural language document. Often this involves defining the general form of the information that we are interested in as one or more templates, which are then used to guide the extraction process. IE systems rely heavily on the data generated by NLP systems. Tasks that IE systems can perform include:

- *Term analysis*, which identifies the terms in a document, where a term may consist of one or more words. This is especially useful for documents that contain many complex multi-word terms, such as scientific research papers
- *Named-entity recognition*, which identifies the names in a document, such as the names of people or organisations. Some systems are also able to recognise dates and expressions of time, quantities and associated units, percentages, and so on
- *Fact extraction*, which identifies and extracts complex facts from documents. Such facts could be relationships between entities or events

A very simplified example of the form of a template and how it might be filled from a sentence is shown in **Figure 1**. Here, the IE system must be able to identify that 'bind' is a kind of interaction, and that 'myosin' and 'actin' are the names of proteins. This kind of information might be stored in a dictionary or an ontology, which defines the terms in a particular field and their relationship to each other. The data generated during IE are normally stored in a database ready for analysis in the final stage, data mining.

Data mining (DM) (often also known as knowledge discovery) is the process of identifying patterns in large sets of data. The aim is to uncover previously unknown, useful

The National Centre for Text Mining (NaCTeM)

NaCTeM (www.nactem.ac.uk) provides text-mining services to the UK academic community, including:

- **Software tools and services.** These enable researchers to apply text-mining techniques to problems in their areas of interest. Many are offered to users as web services, free for academic use and under commercial licence for industry. Initial interest was largely from the biological sciences. However, NaCTeM is also providing services to other sciences, including the social sciences, and medicine
- **Customised solutions.** NaCTeM also offers its expertise in developing customised solutions in close collaboration with both academic and industrial partners
- **Support and advice.** The Centre provides support and advice, organises seminars, hosts workshops and tutorials and provides access to document collections and text-mining resources

NaCTeM is operated by the University of Manchester's School of Computer Science. It is funded by the Joint Information Systems Committee (JISC), with additional funding from the Biotechnology and Biological Sciences Research Council (BBSRC).

knowledge. When used in text mining, DM is applied to the facts generated by the information extraction phase. Continuing with our protein interaction example, we may have extracted a large number of protein interactions from a document collection and stored these interactions as facts in a database. By applying DM to this database, we may be able to identify patterns in the facts. This may lead to new discoveries about the types of interactions that can or cannot occur, or the relationship between types of interactions and particular diseases and so on.

We put the results of our DM process into another database that can be queried by the end-user via a suitable graphical interface. The data generated by such queries can also be represented visually, for example, as a network of protein interactions.

Further Information

Ananiadou, Sophia (2007), *The National Centre for Text Mining: a Vision for the Future*, Ariadne (53) www.ariadne.ac.uk/issue53

Ananiadou, Sophia and McNaught, John (eds) (2006), *Text Mining for Biology and Biomedicine*, Artech House Publishers, ISBN 1-58053-984-X, 302pp

For more information please go to www.nactem.ac.uk or contact: nactem@manchester.ac.uk

This paper has been written by members of the National Centre for Text Mining and produced and edited by Judy Redfearn and the JISC Communications team.

Alternative formats of this briefing paper can be found at: www.jisc.ac.uk/publications

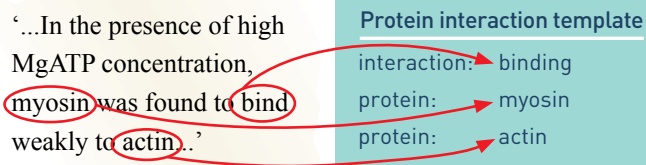


Figure 1: Template-based Information Extraction