

*Text Mining: Science Digs Deeper

Many scientists trying to unearth nuggets of information from the vast online deposits have probably wished for an intelligent tool to automatically answer complex queries, such as “How does protein X affect disease Y”? Existing keyword searches largely ignore the relationships between words, so is this scenario just a pipe-dream? Perhaps not, as several research groups and companies are now pioneering innovative text-mining technologies that might ultimately allow personally relevant semantic searching.

Intelligent digging

Most electronic information is stored as unstructured text in the form of e-mails, news articles, academic papers, commercial reports and so on.¹ Keyword-based search engines, be they specialized for science like Entrez or more general like Google, pay little heed to the textual context or interrelationships between words.²

Text-mining programs aim to dig deeper, using natural-language processing to discover hidden patterns and associations, and providing visual maps to direct users down previously uncharted trails.¹ This essentially semantic approach uses statistical and linguistic rules, hand-crafted for specific types of text, to probe the underlying meanings. The result could be technologies capable of answering sophisticated questions and performing automated text searches with an element of intelligence.³

Labour-intensive manual text-mining approaches first surfaced in the mid-1980s, but technological advances have enabled the field to advance in leaps and bounds during the past decade.⁴ Electronic text-mining programs are now beginning to use artificial intelligence techniques to search text for entities (qualities or characteristics) and concepts (such as the relationship between two entities).⁴ The four key stages of this process are retrieving relevant documents, extracting lists of entities or relationships among entities, answering questions about the content, and delivering facts to the user in response to specific natural-language queries.

Fool's gold?

So just how realistic is this goal? It is true that language-processing software tools have been successfully applied to non-scientific text, such as news content.⁵ Yet the task facing science text-mining comes closer to the ultimate challenge of comprehending human languages.⁵

One of the major hurdles to be overcome is interpreting the dense layers of jargon that permeate science writing — this is known as the ‘ontology problem’.

⁵ Another obstacle to the progress of text mining is the

thorny issue of access; many journals do not make their full-text content publicly available, so search engines often scan abstracts alone. Further complications stem from the lack of standardization; few journals use the same format, and even within an article different sections (such as the methods and the discussion) might need to be assigned different weightings depending upon the precise nature of the search.



Images from Husar posted to Flickr:

<http://www.flickr.com/photos/husar/tags/mining/>

Lighting the way

Several groups of academics have begun addressing these concerns, and are developing text-mining software to scan open-access publications and meet the advanced needs of researchers in their particular fields. Examples from the life sciences include the Arrowsmith software and EBIMed retrieval engine, both of which perform sophisticated searches of Medline text focusing on the causes of disease and protein interactions, respectively.⁷ One of the newest such tools is Textpresso launched by Wormbase in February 2006 to serve the *Caenorhabditis elegans* research community; this resource relies on human ‘taggers’ to manually mark up its corpus of text, and outputs responses to complex queries in the form of citations, abstracts or paragraphs from relevant papers.⁵ An information resource with links to various similar projects is provided by Biomedical Literature and (text) Mining Publications (BLIMP).

An exciting new initiative to aid the academic onslaught against ‘data deluge’ was launched in March 2006.² The National Centre for Text Mining (NaCTeM) is a collaborative effort between the Universities of Manchester, Liverpool and Salford, funded by the Joint Information Sys-

tems Committee (JISC), the Biotechnology and Biological Research Council (BBSRC), and the Engineering and Physical Sciences Research Council (EPSRC). NaCTeM aims to provide tools, carry out research and offer advice to the academic community, with an initial focus on text mining in the biological and biomedical sciences.²



Images from Husar posted to Flickr:
<http://www.flickr.com/photos/husar/tags/mining/>

Pioneering publishers

Although science publishers will have a strong impact on the success of text-mining efforts, they have yet to develop a standard annotation of their content that will allow full-text access to computers.⁷ One route would be for each publisher to license its own electronic tools for searching its content, although such a system would still necessitate multiple separate searches. A more appealing prospect might be the adoption of a common format in which all publishers could issue content for text mining and indexing. Nature Publishing Group (NPG) recently proposed an initiative, known as the Open Text Mining Interface (OTMI), in an attempt to kick start the debate on how publishers should respond to requests for machine-readable copies of content.^{8,9} Their suggested approach is to establish a common format in which coded content could be made freely available for text mining and indexing while maintaining publishers' restrictions on human access. This aim would be achieved by labelling the different section of the paper and converting text into 'word vectors' and 'snippets', which give some indication of the content and structure of a piece. They propose that "If all publishers were to adopt this or some similar standard, the entire literature would become accessible for mining."⁷

Some argue that converting the text after publication is a mistaken way to approach the issue of adding meaning to new academic texts and that the process should begin much earlier, with the initial formatting of the journal article. This has been the approach taken by the National Institutes of Health (NIH) with their initiative to encourage publishers to adopt a common Journal Publishing Document Type Definition (DTD) to pro-

vide a standard method of xml tagging for journal content that could provide semantic meaning. Open access publishers, who do not face the problem of how far syntax can be retained while shielding meaning, also tend to take different approaches, discussed in BioMed Central's text mining page and the Open Archives Initiative.

Future prospects

Clearly, the ability of software to interpret text depends upon the knowledge and abilities of its human programmers and users. Recent document-retrieval studies have reported just 5–10% improvements in accuracy using existing text-mining technologies compared with standard keyword searches.³ Thus, while we could be poised on the threshold of the era of text mining, the technology is still in its infancy and much remains to be achieved. It is too early to predict whether text mining will ultimately strike gold, hit rock bottom, or be surpassed by newer forms of industry as semantic markup of scientific texts becomes the norm.

References

1. Guernsey, L. (2003) Digging for nuggets of wisdom. *New York Times* 16 October
<http://tech2.nytimes.com/mem/technology/techreview.html?res=950CE5DD173EF935A25753C1A9659C8B63>.
2. Joint Information Systems Committee (2005) Press Release: World's first text mining service to benefit British academics 17 March
http://www.jisc.ac.uk/index.cfm?name=pr_text_mining_170305.
3. Abrams, W. (2003) Text mining: the next gold rush. *Second Moment*
<http://www.seconddmoment.org/articles/textmining.php>.
4. Nightingale, J. (2006) Digging for data that can change our world. *The Guardian* 10 January
<http://education.guardian.co.uk/elearning/story/0,,1682496,00.html>.
5. Dickman, S. (2003) Tough mining: the challenges of searching the scientific literature. *PLoS Biol.* 1(2): e48
<http://biology.plosjournals.org/perlserv/?request=get-document&doi=10.1371%2Fjournal.pbio.0000048>.
6. Timmer, J. (2006) Mining scientific publishing. *Ars Technica* 3 May
<http://arstechnica.com/journals/science.ars/2006/5/3/3827>.
7. Editorial (2006) Machine readability. *Nature* 440: 1090
<http://www.nature.com/nature/journal/v440/n7088/full/4401090a.html>
8. Hannay, T. (2006) Open text mining interface. *Nascent* 24 April
http://blogs.nature.com/wp/nascent/2006/04/open_text_mining_interface_1.html.
9. Lynch, C. (2006) Open Computation: Beyond Human-Reader-Centric Views of Scholarly Literatures:
<http://www.cni.org/staff/cliffpubs/OpenComputation.htm>