

Text Mining: Getting more value from literature resources

Roger Hale, Chief Operating Officer, Linguamatics Ltd

Scientists in the pharmaceutical and biotech industries need to review vast quantities of literature during the discovery and development of a new drug. Pinpointing the most relevant information is hard with current search tools. New developments in text mining are proving effective in getting better value from available literature resources.

Literature-based research: current challenges

The past forty years have seen a 5-fold increase in the number of abstracts published annually in Medline. There are now more than 12 million, and most are also available online as full-text articles. In addition to this there are patents, internal reports, and other potentially valuable in-house and public sources. While a small proportion of this information is in a structured form that can be managed using database systems, around 80% is unstructured and written in a natural language. This is far too much to review or keep up-to-date with manually.

Literature-based research is critical in turning an idea into a marketable therapeutic product 12-15 years later. Industry estimates suggest that 90% of drug targets are derived from the literature [1]. In fact, most background research for drug discovery is publicly available through published articles and abstracts, patent application data, and other sources. Strategic decisions during R&D need to be supported by information extracted from a wide range of text content, sometimes about unrelated or unfamiliar topics. In particular, it is usually important to extract and understand relationships, for example between genes and diseases or compounds and side effects. Typical questions include:

- *Which proteins interact with my target protein?*
- *Who is patenting on what targets, compounds and diseases?*
- *Will my lead compound fail due to toxicity?*

In the case of the last question, surveys suggest that about 50% of all potentially therapeutic compounds undergo attrition due to safety concerns and that about 50% of them had some indication in the literature already [1].

To answer questions like those above from unstructured information sources using conventional search technologies requires a large amount of human intervention and curation. Studies by IDC estimate that an enterprise employing 1,000 knowledge workers wastes nearly \$2.5 million per year due to an inability to locate and retrieve information, resulting in lost opportunities and diminished competitiveness [2].

Information Retrieval

Information Retrieval tools based on traditional keyword search, such as Google or PubMed, are familiar and highly tuned to the task of finding the most relevant documents containing certain keywords, but they have limitations on both the kinds of query that the user can make and the kinds of output returned. Input is usually a set of keywords (or MESH terms), and users get documents back instead of answers. Various enhancements to the basic search include *query expansion*, where your keywords are

augmented by related keywords to broaden the search, and *categorisation*, where results are clustered for easier browsing, but the basic proposition is still to input words and return documents.

Shortcomings of keyword search include:

- Although the documents returned may contain relevant information, a laborious and often prohibitively time-consuming manual review is needed to extract that information.
- It is ill suited to finding relationships, which is often the whole purpose, for example when finding the mechanism of interaction between two proteins.
- It does not cope well with complexity and ambiguity in the source material, for instance where synonyms and homonyms are common.

Information Extraction and Text Mining

Information Extraction systems are better suited to finding relationships, using precise queries that look for patterns of concepts and words, e.g. "*protein* interacts with *protein*" to find pairs of proteins that interact [3,4]. The result of a search is structured output, e.g. spreadsheet or database, in which the sought-after information is readily accessible.

Information Extraction is based on an understanding of the structure and meaning of the natural language in which documents are written. The ability to search for items occurring within the same sentence or section already gives a greater chance of a relationship between them than if they occur only in the same document. Recognition of grammatical syntax enables nouns and verbs to be distinguished. Entities, such as proteins and diseases, correspond to noun groups, whereas the relationships between them appear as verb groups. This enable specific searches for direct relationships.

The major advantage of information extraction systems is the precision of the queries and the clarity of the output, which can be efficiently reviewed, entered into a database or displayed visually. The term *text mining* has come to apply to such systems by analogy with conventional data mining systems.

An alternative to text mining using natural language processing is to use statistical co-occurrence methods. Although the occurrence of a particular gene in the same document as a particular disease may not tell us much, the co-occurrence of the same gene and disease in many documents suggests some kind of relationship. Although Information Extraction and statistical co-occurrence are sometimes seen as competitors, the two can usefully be combined. Information Extraction provides the direction of relationships, and can find information containing negative relationships. Co-occurrence may suggest indirect relationships that cannot easily be found by precise linguistic patterns.

Over the past few years, text mining systems have begun to move from academic research into the market place. A number of factors have driven this move, including market pull due to competitive pressure, the increasing availability of electronic documents, and the falling cost of computing power and storage.

We **find** that **p42mapk** **phosphorylates** **c-Myb** on **serine** and **threonine**, but not on **tyrosine**.

12E Query Results
16 results in 0.26 seconds processing time.

mitogen-activated pr..	relation: verbal	protein	sentence	Link	Score
p42mapk	phosphorylates	c-Myb	we find that p42mapk phosphorylates c-Myb on serine and threonine, but not on tyrosine.	source cache	100
ERK2	is positioned to phosphorylate	normal tau	since ERK2 was detected in neurofibrillary tangles and senile plaque neurites in the AD hippocampus, ERK2 is positioned to phosphorylate normal tau and could play a role in the generation of PHEs in AD.	source cache	100
MEKK1	phosphorylates	ME			
purified recombinant p42 MAPK	was found to phosphorylate	recombinant Wee1	purified recombinant p42 MAPK was found to phosphorylate recombinant Wee1 in vitro at sites that are phosphorylated in extracts.		
MEKK1	binds	raf-1	components:	cache	
ERK2	directly phosphorylated	GRASP55	furthermore, ERK2 directly phosphorylated GRASP55 on the same residues that generated the MPM2 phospho-epitope.	source cache	100
nuclear ERK2	phosphorylates	p53	in this study, we showed that nuclear ERK2 phosphorylates p53 at Thr55 in response to doxorubicin.	source cache	100
NIK	binds to and divergently activates	the plasma membrane Na(+)-H(+) exchanger NHE1	we now show that NIK binds to and divergently activates the plasma membrane Na(+)-H(+) exchanger NHE1	source cache	100

We **now show** that **NIK** **binds to and divergently activates** **the plasma membrane Na(+)-H(+) exchanger NHE1**.

An example output from Linguamatics' interactive text-mining system showing use of linguistics in extracted sentences.

Early text mining systems were aimed at information specialists. They typically require a combination of domain and informatics expertise to configure. Once set up for a particular task they can be run repeatedly over different sets of documents, producing high quality results, but changing to a new task or domain can take weeks and a lot of specialist linguistic expertise. Typical uses include analysis of news feeds and constructing pathway databases for subsequent access by end users.

Recent Advances: Ontologies and Usability

The challenge is to make such powerful tools more readily deployable and accessible. Two recent advances have made an important step forward in achieving this. The addition of *ontologies* enables more powerful searches, better-defined results, and easier customisation to new domains. The development of *Interactive Information Extraction* enables answers to be found in seconds rather than hours, and for the first time makes information extraction available on demand.

Early text mining systems typically have a few concepts (such as *company* or *protein*) built in, but the incorporation of ontologies enables semantic search for many thousands of specific concepts (such as *protein kinase*), and semantic output with standardised naming of concepts (such as the *LocusLink ID*), which enables results to be sorted and used directly in a database. An ontology is a way of organising the relationships and entities in a domain. A taxonomy is a simple ontology in which entities or relationships are organised hierarchically so that, for example, *MEK* is a *protein kinase*, and a search for all *protein kinases* would also identify all instances of *MEK*. Similarly, in an ontology of relationships, *up-regulation* may be classified as a direct relationship between proteins. Ontologies may contain synonyms for each entity, again powerful when querying.

Interactive Information Extraction combines the attractive features of information retrieval and extraction by providing the ease-of-use, scalability and interactivity of information retrieval with the precision and output of information extraction [5]. End users can move naturally and seamlessly from keyword-style searches to the full power and precision of information extraction. Co-occurrence within a document or within a sentence is also possible. Information scientists can now make ad hoc searches and get their results back in seconds.

Case Study: Mining Nuclear Receptor Cofactors

Validation is a key driver for the adoption of new technologies, and text mining is no exception. While much good work is being done towards community-accepted validations [6], case studies involving use by real customers can be very persuasive. In a recent case study [7], Linguamatics' I2E software was used with Biowisdom ontologies at AstraZeneca to mine MEDLINE and EMBASE for nuclear receptor cofactors.

Typically, 100 abstracts can be manually analysed for cofactors in one person-day whereas in just 2 person-hours with I2E, cofactors were extracted with a recall of 91% from around 8000 abstracts, rivalling the quality of hand analysis. This represented an order of magnitude speed-up over existing best practice. Further, the system discovered new cofactors that were not in the standard set.

Summary

In summary, text mining systems:

- Allow more specific queries than keyword search
- Provide structured output for easy review and analysis

Some systems can also:

- Use linguistics to interpret human language and identify relationships
- Use concepts and ontologies to provide domain knowledge
- Mine vast quantities of text interactively

The main benefits of using text mining are:

- Getting to decision points more quickly
- At least 10x speedup over previous methods
- Finding information that wouldn't otherwise be found

References

- 1
Fickett, J. and Hayes, W. (2004) Text Mining for Drug Discovery. *European Pharmaceutical Contractor*. Autumn 2004.
- 2
Feldman, S. and Sherman, C. (2001) The High Cost of not Finding Information. *IDC White Paper*.
- 3
Thomas, J. *et al.* (2000) Automatic Extraction of Protein Interactions from Scientific Abstracts. *Proceedings of the Pacific Symposium on Biocomputing 2000*, 4-9 January 2000, Honolulu, Hawaii. Altman, R.B. *et al.* (eds). World Scientific Publishing Co., Singapore. 541-552
- 4
Blaschke C, Hirschman L and Valencia A. (2002) *Information extraction in Molecular Biology*, Briefings in Bioinformatics 3: 154-165
- 5
Milward, D. and Thomas, J. (2000) From Information Retrieval to Information Extraction. *Proceedings of the ACL-2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval*, 8 October 2000, Hong Kong University of Science and Technology, Hong Kong, Klavans, J and Gonzalo, J. (eds). 85-97
- 6
Krallinger, M., Alonso-Allende Erhardt, R. and Valencia, A. (2005) Text Mining approaches in molecular biology and biomedicine. *In this issue*.
- 7
Milward, D., *et al.* (2005) Ontology-Based Interactive Information Extraction from Scientific Abstracts. Proceedings of the BioLINK SIG Text Mining Workshop, ISMB/ECCB 2004. To appear in *Comparative and Functional Genomics*.