

# National Centre for Text Mining

## An introduction to tools for researchers

Briefing Paper

September 2008

With an overwhelming amount of knowledge recorded in texts, it has become imperative to use automated techniques that can identify, extract, manage, integrate and exploit this knowledge for research and education, efficiently and systematically. Text mining exploits these techniques. The National Centre for Text Mining (NaCTeM) offers text mining services to UK researchers that enable semantic searching of text – that is, searches based on the meanings of words, phrases or terms in different contexts – thus improving access to information and increasing the efficiency of new research methodologies and techniques based on advanced information and communication technologies (e-science and e-research).

The following tools are now available. Most were developed for the life sciences, the focus of NaCTeM's first phase, but are now being applied to other domains, in particular the social sciences.

*The National Centre for Text Mining (NaCTeM) offers text mining services to UK researchers that enable semantic searching of text and the discovery of new knowledge*

**TerMine** identifies the most important terms in a document, ranking them according to their significance, where a term is a word or compound word denoting a specialised concept in a subject domain. TerMine is thus a quick way for a reader to pick out articles of potential interest from a large body of text.

Humans and language processors struggle with terminology because of term variation, when a concept is expressed in several different ways, and term ambiguity, when the same term is

used to refer to multiple concepts. Term variation and ambiguity may cause irrelevant information to be retrieved (low precision) and relevant information to be overlooked (low recall). TerMine overcomes this problem.

TerMine has also been used to build controlled vocabularies and ontologies, that is collections of words and phrases common to a subject area, by extracting candidate terms from a body of text.

[www.nactem.ac.uk/software/termine](http://www.nactem.ac.uk/software/termine)

**AcroMine** is based on a novel approach to the recognition of acronym definitions in a text collection. Often we retrieve more documents by using an acronym as a query term rather than its expanded form. For example, a study in 2005 reported that PubMed, the database of biomedical publications, could retrieve 5477 documents for the search term JNK, but only 3773 documents for the expansion, c-jun N-terminal kinase. AcroMine finds expanded forms of acronyms from a database created from the whole of MEDLINE, the biomedical bibliographic database, and removes ambiguity based upon the context in which they appear. This allows users to expand their queries, when searching large document collections, to include synonymous acronyms without losing the specificity of the original query.

[www.nactem.ac.uk/software/acromine](http://www.nactem.ac.uk/software/acromine)

**KLEIO** deals with the variety of names (terms) for the same concept, for example IL2, IL-2 and Interleukin-2. It is able to learn term variation patterns automatically and discover ambiguous and variant terms in a body of text. For example,

**ASSERT** provides text mining services to facilitate the process of producing systematic reviews, especially in the social sciences although the techniques are applicable to other domains. Key features include interactive query expansion to maximise coverage of an unbiased search, document clustering and visualisation to aid the overview of search results, and the ability to summarise multiple documents automatically and hence create reviews.

[www.nactem.ac.uk/assert](http://www.nactem.ac.uk/assert)

## Text Mining Tools

September 2008

---

a search for 'cat' in the whole of MEDLINE retrieves 60,711 documents, but by specifying 'protein cat' in KLEIO only 237 documents are retrieved. All of them are about the protein named 'cat', as opposed to a set of documents containing 'protein' and 'cat' retrieved by conventional Boolean query.

[www.nactem.ac.uk/software/kleio](http://www.nactem.ac.uk/software/kleio)

**MEDIE** is an advanced search engine for the life sciences, retrieving biomedical knowledge from MEDLINE. As with several of NaCTeM's services, it relies on text mining having already been applied to the document collection. This pre-analysis discovers and marks named entities (eg names of genes) and constructs a complex ('deep') semantic representation for every sentence. These representations and named entity annotations are then stored and indexed for rapid user web-based access.

User requests are converted on the fly into semantic patterns which are then matched against the stored semantic representations, and the corresponding sentences returned for inspection. The user can inspect the context of interesting sentences within the wider document and can launch more sophisticated queries. The main advantage of this service is that precise facts are returned, representing instances of the relationship between concepts expressed in the user's query, where these relationships may be expressed very differently from text to text.

For example, responses to the query 'what is reduced by the use of thalidomide' include the following sentences from different documents:

- These findings indicate that targeted delivery of thalidomide using APA capsules could facilitate its usage in reducing the inflammation associated with chronic conditions such as Crohn's disease and ulcerative colitis.
- Thalidomide downregulates angiogenic genes in bone marrow endothelial cells of patients with active multiple myeloma.

In the second result, ontological knowledge built into MEDIE recognises that 'reduce' is linked semantically to 'downregulate'.

This service has been developed by NaCTeM's partner, the University of Tokyo.

[www-tsujii.is.s.u-tokyo.ac.jp/medie](http://www-tsujii.is.s.u-tokyo.ac.jp/medie)

**FACTA** is a fast and interactive semantic search engine for the life sciences. It automatically finds associations between genes, diseases, drugs, compounds, enzymes and symptoms and their associated documents from the whole of MEDLINE, based again on pre-analysis to annotate text semantically and then store various kinds of entity. As well as accepting a

concept as a query, it also accepts an arbitrary combination of keywords, enabling the user to express a concept that cannot be captured by a single word. FACTA displays snippets of text containing the query term and its textual variants, highlighting any entities. The user can link from a snippet to the document it comes from.

[www.nactem.ac.uk/software/facta](http://www.nactem.ac.uk/software/facta)

**Info-PubMed** searches the whole of MEDLINE to find information about biomedical entities such as genes, proteins and the interactions between them. It is supplemented by a term dictionary containing more than 200,000 protein and gene names, which leads to a fine level of analysis and increases the number and types of terms that users can query. This service has been developed by NaCTeM's partner, the University of Tokyo.

[www-tsujii.is.s.u-tokyo.ac.jp/info-pubmed](http://www-tsujii.is.s.u-tokyo.ac.jp/info-pubmed)

**NaCTeM's** text mining tools and services are enabling users to reduce the time and effort taken to find and link pertinent information from large bodies of text and, through the use of tools provided as web services, to develop their own customised solutions in semantic data analysis and knowledge management. Where existing services do not meet specific user requirements, NaCTeM can develop customised text mining solutions, for example by becoming a partner in a research grant proposal or through the cost of such customisation being written into a proposal by the investigators.

As NaCTeM enters its second phase, it is aiming to bridge the gap between digital library and semantic grid initiatives through an improved facility for constructing semantic metadata descriptions from text using the various services, tools and techniques mentioned above.

This briefing document was written by Sophia Ananiadou, Director, National Centre for Text Mining, University of Manchester, and edited by the JISC Communications team.

Alternative formats of the briefing paper can be found at:

[www.jisc.ac.uk/publications](http://www.jisc.ac.uk/publications)

### Further Information and Resources

[www.nactem.ac.uk](http://www.nactem.ac.uk)

[1] Ananiadou, S. and McNaught, J. (eds) (2006) *Text Mining for Biology and Biomedicine*. Artech House Publishers, ISBN 1-58053-984-X, 302pp.

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