MANCHESTER 1824

The University of Manchester

NaCTeM

The National Centre for Text Mining

The National Centre for Text Mining

Providing Text Mining Services to the UK

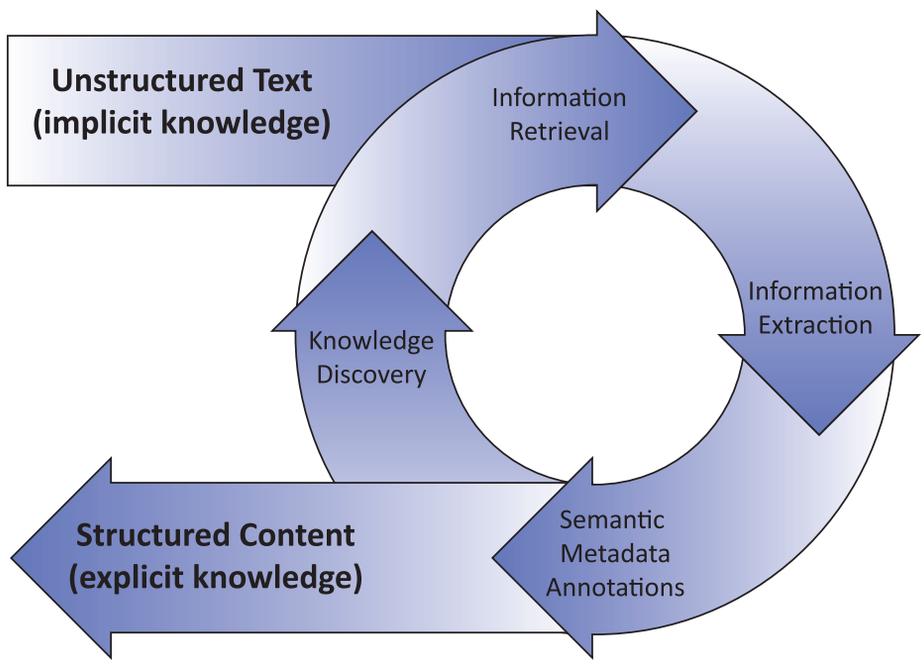
www.nactem.ac.uk

What is text mining?

Text mining is the process of discovering and extracting knowledge from unstructured data. This comprises three main activities:

- Information retrieval to gather relevant texts
- Information Extraction to identify and extract entities, facts and relationships between them
- Data mining to find associations among the pieces of information extracted from many different texts

In short, text mining can help make the implicit information in your documents more explicit, saving your organisation time and effort. Take a look at the sample text mining applications in this booklet and we hope you will soon see that there is more to text mining than simple search.



What is NaCTeM?

The National Centre of Text Mining started out as the world's first publicly funded centre for the development of tools and services to support the UK academic community. Building upon what is now the largest active text mining research group in the UK, NaCTeM draws upon decades worth of experience and commitment to provide dedicated infrastructure and support to its core communities. Recently NaCTeM has opened up these opportunities to the industrial market, leveraging our world class research and dedication to quality service provision. Don't just take our word for it; take a look at our list of clients on the next page.

Major Clients and Funders:

Arthritis Research Campaign
British Heart Foundation
Cancer Research UK
Chief Scientist Office
Department of Health- NIHR
Medical Research Council
Wellcome Trust
Astrazeneca
EPPI - Centre
Manchester Biomedical Research Centre
Nature Publishing Group
Pfizer

"Over the last couple of years, scientists at Pfizer's UK research site in Sandwich have been making use of the text mining tools and services developed by NaCTeM. One such tool, which has proven to be valuable, is TerMine, an automatic multi-word term recognition tool that has been used at Pfizer to enrich the labour intensive process of building dictionaries used for text mining. […]

Pfizer and NaCTeM have also been collaborating on a project called DECA (Disease Extraction with Concept Association) to extract associations between concepts in the biomedical domain such as diseases and symptoms from collections of biomedical texts (e.g. Medline). The aim of this project is to combine the strengths of the NaCTeM text mining tools, Kleio and FACTA to create an efficient search for associations between biomedical concepts. Also, a considerable amount of research is being applied to the challenge of lexical disambiguation of the biomedical terms. Pfizer values highly the world-class quality of the linguistic and semantic extraction skills and methodologies being developed and practised at NaCTeM which is located in the highly appropriate setting of the Manchester Centre for Integrative Systems Biology."

Ian Harrow, Senior Principal Scientist, Pfizer

JISC

MANCHESTER 1824
The University of Manchester

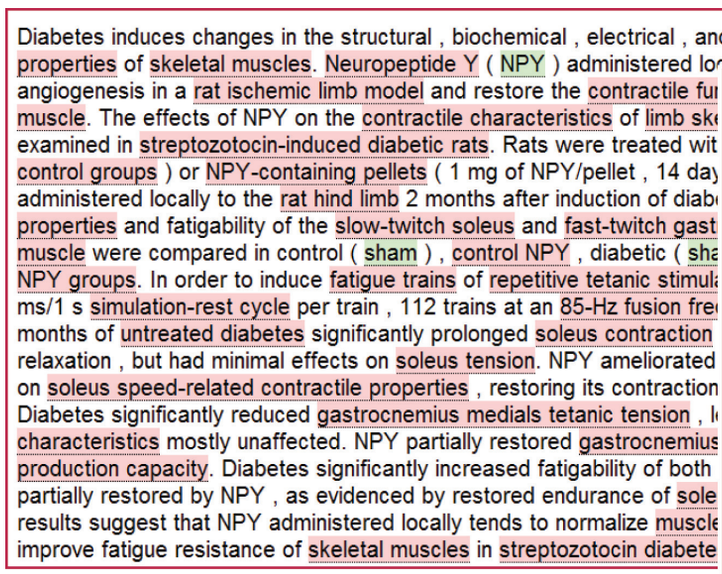
bbsrc
biotechnology and biological sciences research council

EPSRC
Engineering and Physical Sciences Research Council

For more quotes from our user communities visit http://www.nactem.ac.uk/comments

TerMine - Keyword Discovery Service

TerMine is a tool that automatically detects and extracts multi-word technical terms from text. The statistical techniques used allow for the scoring of these terms by importance in the document letting you create a ranked list. This is useful for applications such as key concept detection for enhancing metadata, building ontologies, enhancing other text mining components and even for the creation of an enhanced version of tag clouds.



Diabetes induces changes in the structural , biochemical , electrical , and contractile properties of skeletal muscles. Neuropeptide Y (NPY) administered loc
angiogenesis in a rat ischemic limb model and restore the contractile fu
muscle. The effects of NPY on the contractile characteristics of limb sk
examined in streptozotocin-induced diabetic rats. Rats were treated wit
control groups) or NPY-containing pellets (1 mg of NPY/pellet , 14 da
administered locally to the rat hind limb 2 months after induction of diab
properties and fatigability of the slow-twitch soleus and fast-twitch gast
muscle were compared in control (sham) , control NPY , diabetic (sha
NPY groups. In order to induce fatigue trains of repetitive tetanic stimul
ms/1 s simulation-rest cycle per train , 112 trains at an 85-Hz fusion fre
months of untreated diabetes significantly prolonged soleus contraction
relaxation , but had minimal effects on soleus tension. NPY ameliorated
on soleus speed-related contractile properties , restoring its contraction
Diabetes significantly reduced gastrocnemius medials tetanic tension , l
characteristics mostly unaffected. NPY partially restored gastrocnemius
production capacity. Diabetes significantly increased fatigability of both
partially restored by NPY , as evidenced by restored endurance of sole
results suggest that NPY administered locally tends to normalize muscle
improve fatigue resistance of skeletal muscles in streptozotocin diabete

Rank	Term	Score
1	contractile property	3
2	gastrocnemius tetanic tension production capacity	2.321928
3	contractile characteristic	2
3	rat ischemic limb model	2
3	soleus speed-related contractile property	2
3	fast-twitch gastrocnemius medial muscle	2
3	skeletal muscle	2
3	gastrocnemius medial tetanic tension	2
9	repetitive tetanic stimulation	1.584962
9	muscle contractile property	1.584962
9	streptozotocin-induced diabetic rat	1.584962
9	rat hind limb	1.584962
9	diabetic npy group	1.584962
9	85-hz fusion frequency	1.584962
9	local npy action	1.584962

Screenshots from the TerMine web demonstrator, available at http://www.nactem.ac.uk/software/termine

Improve the Access to Your Content

Describe your documents or populate ontologies using important terms, automatically detected. Use new ways of searching your content and to filter out irrelevant material.

Enhance your Existing Search Strategy

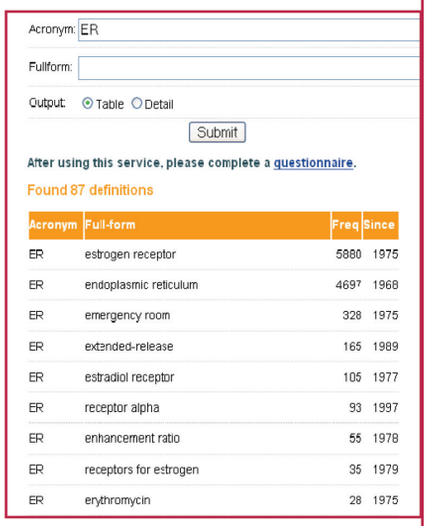
Automatically discover related concepts to your original search and use them to extend your query and identify important content you may have otherwise overlooked.

Profiling and Recommendation Engine

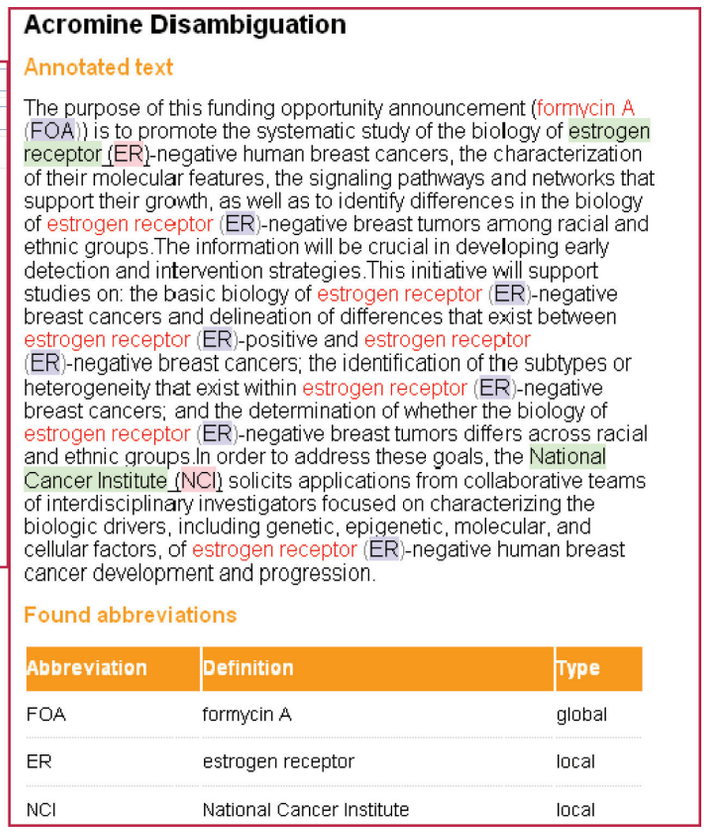
Automatically identify the interests of a reader based upon the concepts most significant to their choice of content. Present the reader with any new content they may otherwise have missed.

AcroMine - Acronym Disambiguation & Lookup

AcroMine finds expanded forms of acronyms from a database of those previously used by authors in a designated repository. The suite of tools comprises the acronym discovery component, an acronym expansion tool to list all known complete forms and a disambiguation tool for predicting the correct form based upon the surrounding context.



Screenshots from the AcroMine web demonstrator, available at http://www.nactem.ac.uk/software/acromine



Expand Context Sensitive Acronyms

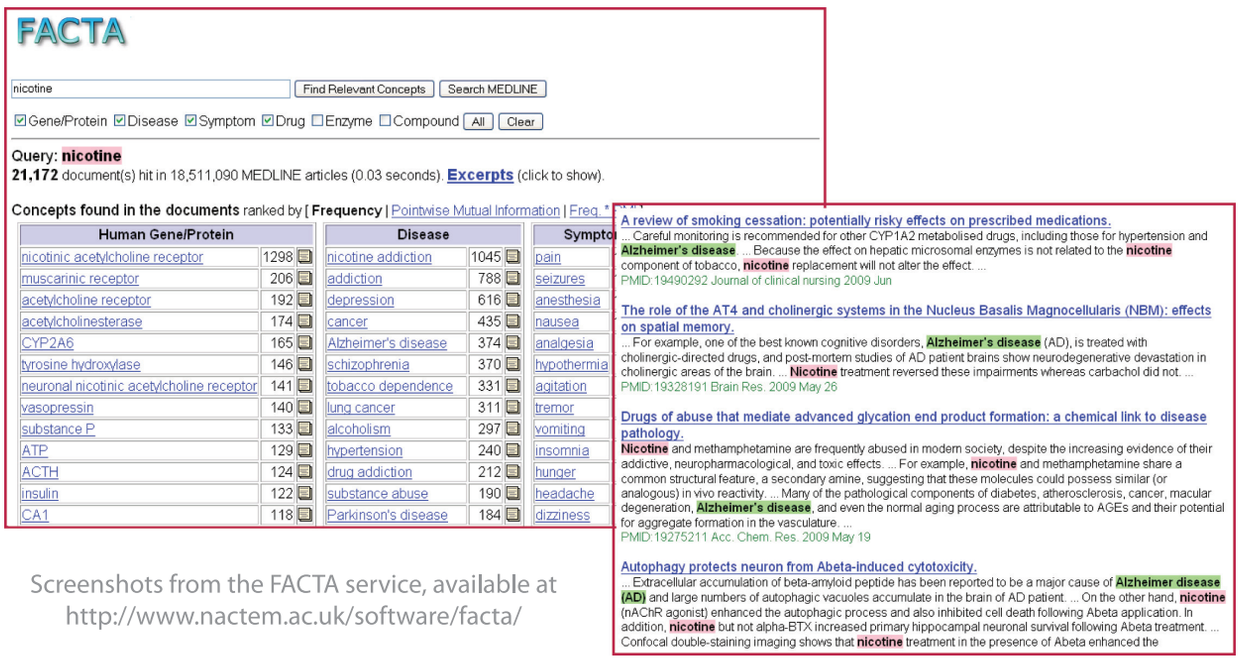
Acronyms are often shared by multiple concepts. AcroMine uses the context around an acronym to predict the most likely expanded form, making it ideal for working in interdisciplinary collections.

Enhance Tool Performance

Achieve better search results in other tools by automatically searching across multiple forms of an acronym with a single query. Filter out documents containing the acronym with alternate meanings.

FACTA - Find Associated Concepts with Text Analysis

FACTA is a tool that helps discover associations between biomedical concepts contained in MEDLINE articles. It allows the user to provide a flexible query (e.g. keywords or boolean combinations of concepts) and retrievesthe documents that match and the associations between the query term and concepts in a highly interactive manner. Snippets of information from MEDLINE provide evidence of these associations providing an additional level of context to your results.



Screenshots from the FACTA service, available at http://www.nactem.ac.uk/software/facta/

Explore an Overview of Your Content

Concepts closely correlated with your query appear in tabular form, categorized into semantically typed columns including proteins, diseases and symptoms, allowing you to gain a quick overview of the associations.

Intuitively Navigate to Related Concepts

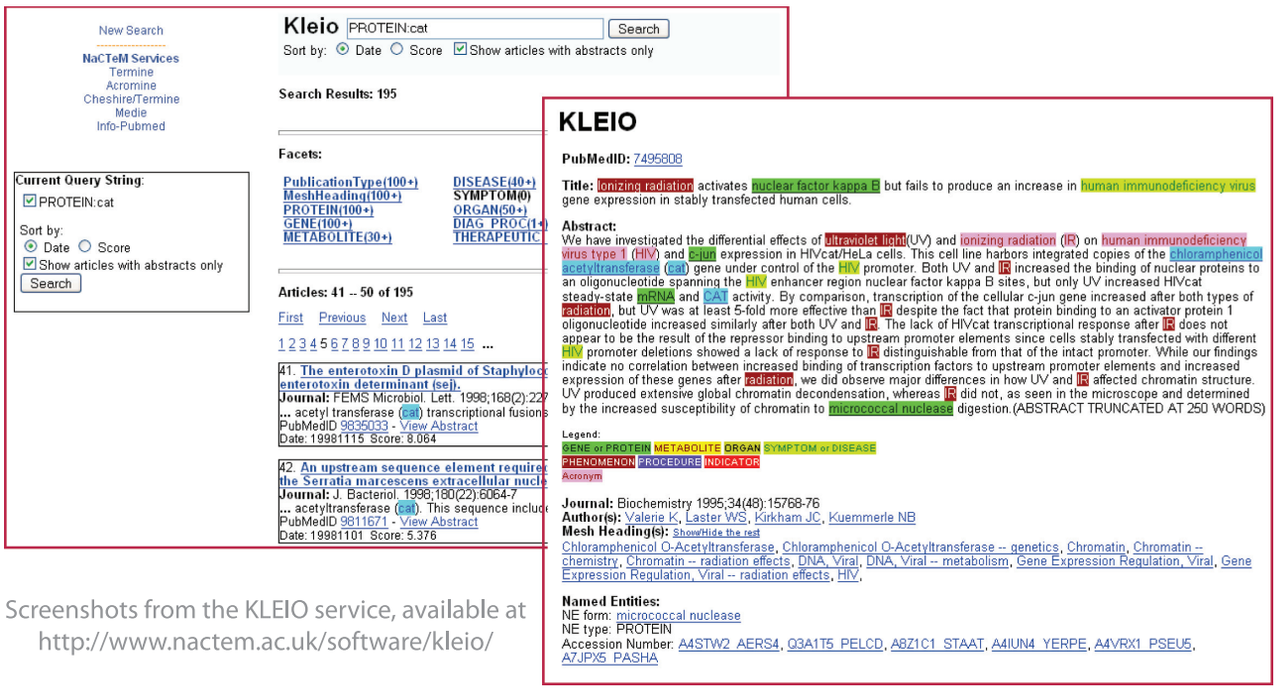
Users can navigate through the listed concepts to either enhance and filter their original query or to examine sections of the documents that show the associations. This flexibity allows users to control the query, whilst being supported by the system.

Easily Uncover Indirect Associations

One of the latest updates to FACTA is the ability to investigate potential indirect links between concepts. Ideal for hypothesis generation the new technology provides confidence scores for the detected relationships.

KLEIO - Semantic Search of Repositories

KLEIO offers a more convenient and intuitive tool for browsing document repositories by using textual and metadata search. It integrates various text mining components to offer textual and metadata search and enhanced functionality in terms of acronym expansion and interactive ranking. KLEIO draws upon a number of core technologies from NaCTEM's tool and resource collection to further enhance the automated detection and markup of semantic entities such as proteins and genes.



Screenshots from the KLEIO service, available at http://www.nactem.ac.uk/software/kleio/

Reduce Noisy Results with Entity Search

With each entity having a semantic type associated with it searching becomes much easier, particularly for ambiguous examples. A search for 'CAT' returns ~66,000 articles, whereas a search for 'PROTEIN:CAT' returns the much more manageable results (e.g. 195).

Navigate and Link to External Databases

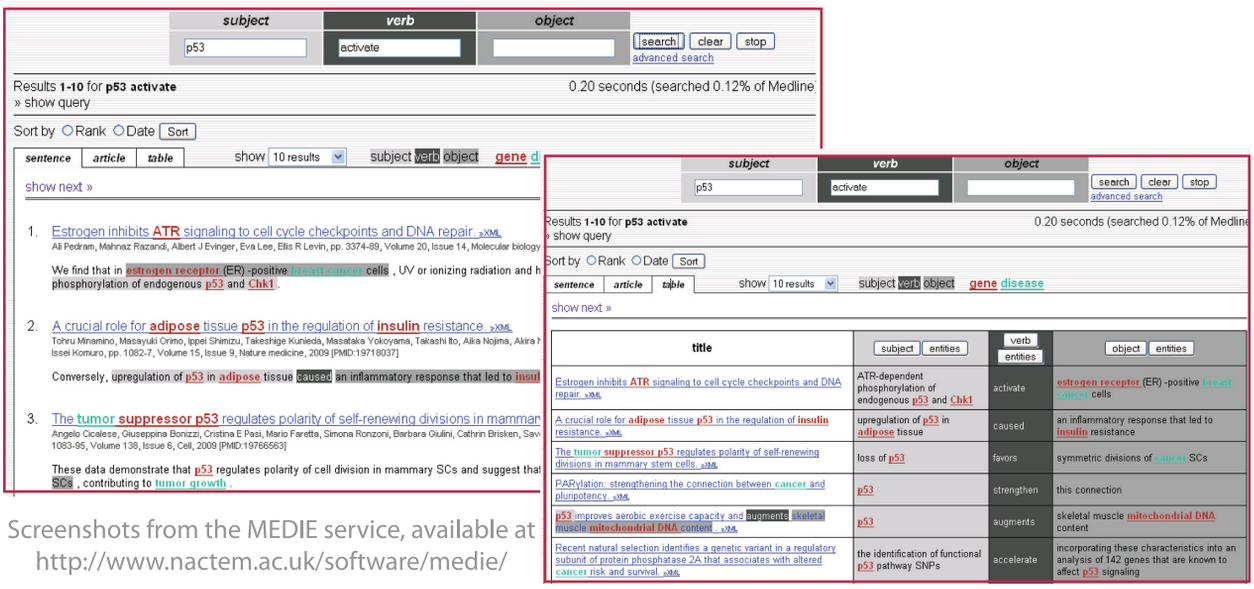
KLEIO draws upon information from a number of external resources, including SwissProt, KEGG, DrugBank amongst others. This enables fast lookup by known identifiers and efficient linking to external resources.

Filter Results with Faceted Search

Build your query by selecting semantic entities and narrow down the results with every new query. This saves you time by prompting the user with suggested means of navigating the results based upon the context that interests them most.

MEDIE - Semantic Fact Retrieval

MEDIE is an intelligent search engine for finding biological events from MEDLINE and discovering interactions between biomedical entities. The service uses advanced natural language processing technologies and a novel search engine for the accurate retrieval of concepts. It identifies sentence level biological relationships and map relevant facts to queries presented by the user. MEDIE moves away from simple correlations between objects to discover the underlying relationship regardless of the descriptive text surrounding the core fact.



Screenshots from the MEDIE service, available at http://www.nactem.ac.uk/software/medie/

Discover Direct Relationships in Content

Search for evidence in the literature, not just documents sharing keywords. The simple search interface allows for a combination of subject, verb and object e.g. 'p53 activate' (see above) and presents the user with documents detailing interactions where P53 activates another entity.

Filter Results by Document Section

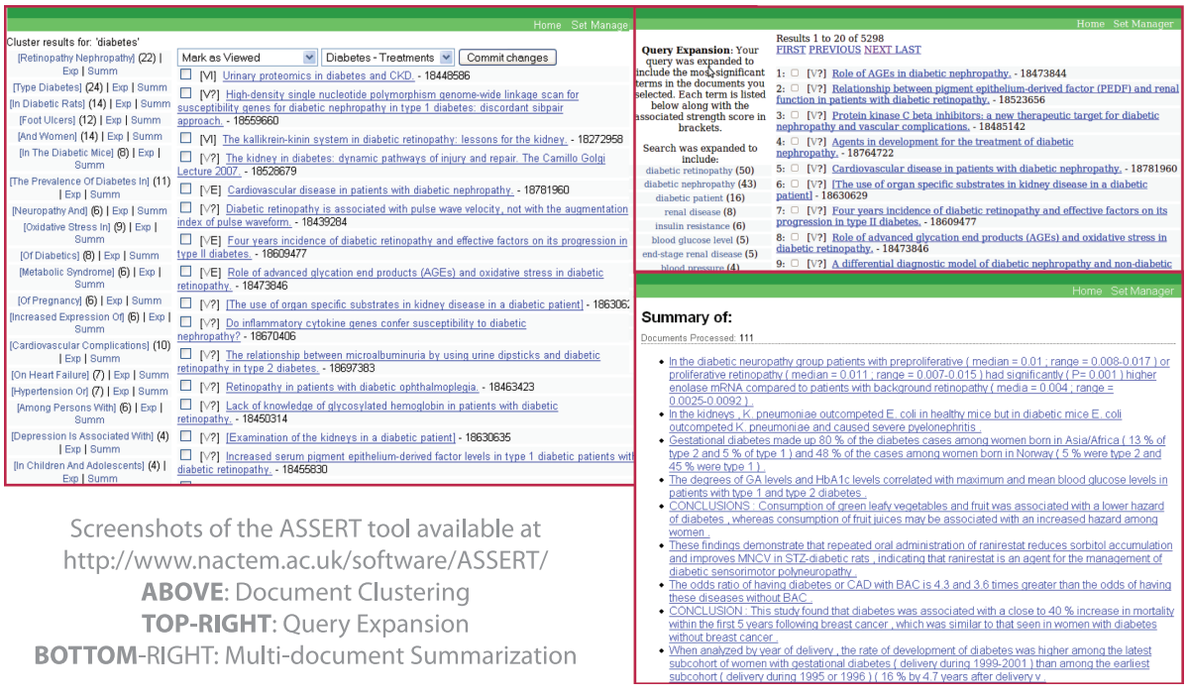
View the original evidence as sentences or in a tabular form sorting your results. Filter your results further by only showing those appearing in specific sections of the document such as results or methodology.

Broaden your Search Horizons

Automatically expand your query by using our built-in ontology. For example activate as a verb can be expanded to find relationships defined as induce, enhance, promote or up-regulate.

ASSERT - Supporting Systematic Reviews

Literature surveys have become the first port of call in most modern research, however, increasing problems of information overload and overlook due to the amount of literature can often lead to incomplete surveys. ASSERT solves this problem through a combination of text mining techniques designed to assist each stage of the formal systematic review, helping users to find the information they need, even in cases where it is difficult to define the scope at the outset.



Screenshots of the ASSERT tool available at
http://www.nactem.ac.uk/software/ASSERT/
ABOVE: Document Clustering
TOP-RIGHT: Query Expansion
BOTTOM-RIGHT: Multi-document Summarization

Move Towards Systematic Reviews

Save time and effort by using the guidelines for systematic review, as used by EPPI-Centre. Store results for later use, or remove them from the future queries entirely as part of this integrated framework for performing reviews.

Explore Contextual Clusters of Documents

Make handling large result sets easier by browsing the automatically generated clusters of relevant content with human readable, descriptive labels. Focus in on content of interest or branch out to find more content containing similar topics.

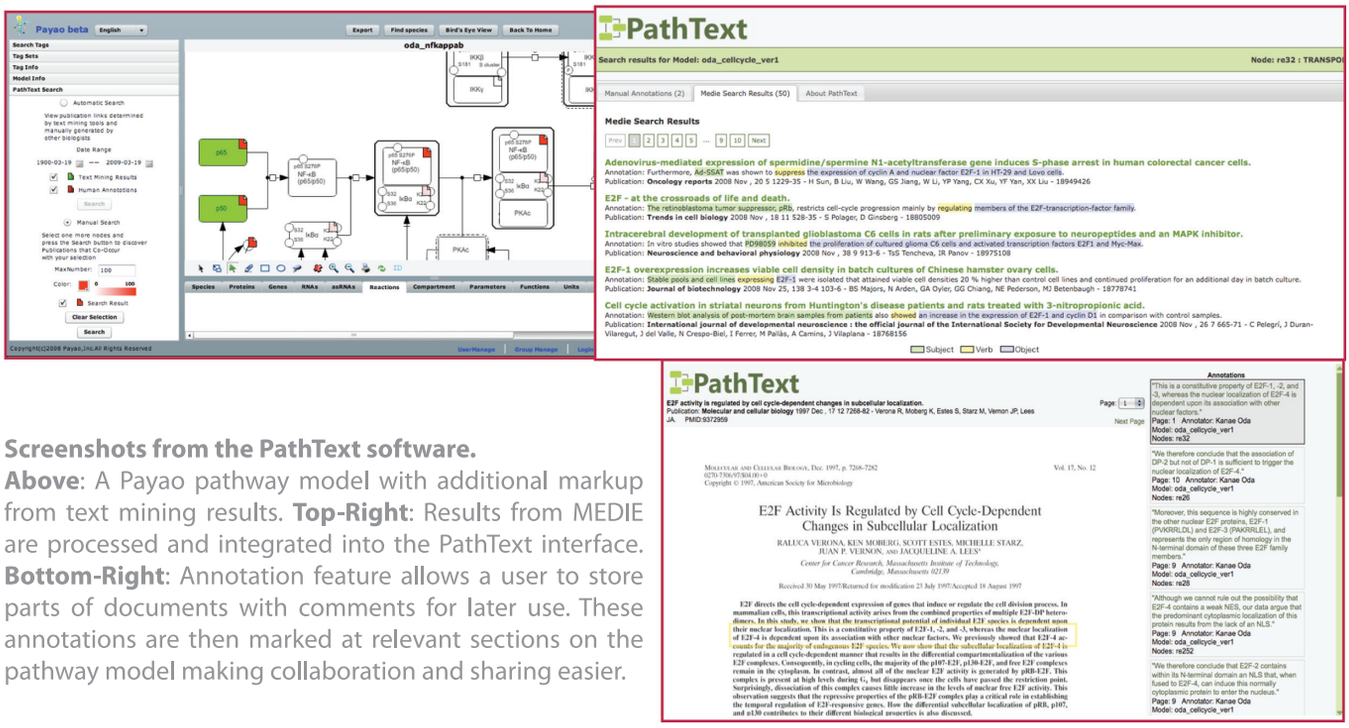
Expand or Summarise Your Search Results

Truly explore a large collection of literature by zooming in and out the content through our query expansion and summarization tools. Allow your search to shift its focus to bring in all of the relevant literature, whilst keeping a clear sight of the overall aims of your review.

PathText - Enriching Pathway Models with Evidence

To allow knowledge discovery, we integrate text mining techniques with pathways which store biological knowledge. Pathway construction is currently manual and relies on literature. PathText bridges pathway editors, visualisation interfaces and text mining systems to provide evidence from literature and to facilitate knowledge discovery. This case study combines the MEDIE and KLEIO services in addition to technologies such as the Systems Biology Markup Language, graphical notations and the CellDesigner program with transparent, underlying interfaces to the text mining tools at NaCTeM and Tokyo, to bridge the gap between two previously incompatible forms.

This case study shows how our text mining services can be integrated into your software, providing transparent solutions to meet the technology support needs of the information age.



Screenshots from the PathText software.
Above: A Payao pathway model with additional markup from text mining results. **Top-Right**: Results from MEDIE are processed and integrated into the PathText interface. **Bottom-Right**: Annotation feature allows a user to store parts of documents with comments for later use. These annotations are then marked at relevant sections on the pathway model making collaboration and sharing easier.

Filter Literature Using Pathway Diagrams

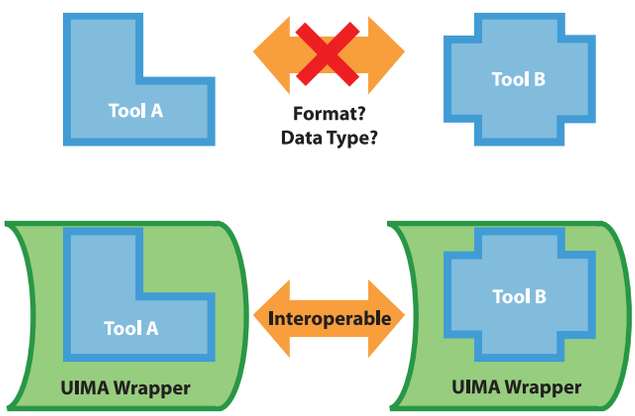
Simply by clicking on relevant parts of the pathway model PathText will perform complex searches for specific material of interest, without you having to learn a new query language. PathText automatically directs you to the appropriate tool to find results that are relevant to what you are looking for.

Find Information Not Documents

Rather than view a long list of documents, the PathText interface takes you direct to the information of interest. For identified relationships this reports the evidence in the form of the sentence that mentions the relationship and the immediate context surrounding it.

U-Compare - Workflows of Text Mining Components

U-Compare is an integrated text mining/natural language processing system based on the UIMA Framework which provides access to a large collection of ready-to-use, interoperable, natural language processing components. U-Compare allows users to build complex NLP workflows from these components via an easy drag-and-drop interface, and makes visualization and comparison of the outputs of these workflows simple.



Design custom workflows from TM building blocks

By using the drag-and-drop interface users can construct their own workflows tailored to an individual task. The vast collection of modules continues to grow and is comprised of open source tools through to full commercial applications.

Identify the best components for your tasks

A flexible interface allows users to compare different combinations of components on sample documents and automatically evaluate the best combination for achieving your goals.

U-Compare is a joint R&D project between the University of Tokyo, the Center for Computational Pharmacology (CCP) at the University of Colorado Health Science Center, and NaCTeM.

http://u-compare.org

BioLexicon - Language Resources to Support Analysis

Technical terms are a major barrier to bio-text processing. A huge number of biological, chemical and medical terms appear in the literature and new terms are coined every day. Furthermore, there are many spelling and semantic variants of these terms. Also, text mining tools need to know about the formal linguistic behaviour of domain terms, particularly domain verbs.

The BioLexicon is a unique linguistic resource that has been created to address such issues. Although primarily intended to support text mining and information retrieval in the biomedical domain, its standards-based structure and rich content make it a valuable resource for many other kinds of application.

Available now at ELRA
http://catalog.elra.info/

Maintain a competitive advantage

The BioLexicon's 2.2M entries include recently established terms not yet in other resources. As well as domain nouns, adjectives and adverbs, it contains detailed formal descriptions of the linguistic behaviour of domain verbs and nominalisations, describing and linking related forms such as retroregulate and retroregulation.

Enhance the performance of your tools

The BioLexicon's extensive coverage and sophisticated descriptions enable tools to perform richer search and more accurate analysis than is achievable with conventional resources.

The BioLexicon was designed and developed within the EC BOOTStrep project by 3 partners: NaCTeM, ILC-CNR (Pisa, Italy) and EMBL-EBI (Hinxton, UK). It is distributed by the European Language Resources Association (http://www.elra.info/) under both research and commercial use licences.

http://www.nactem.ac.uk/resources/biolexicon

http://www.nactem.ac.uk

NaCTeM

The National Centre for Text Mining

What We Can Offer Your Organisation

NaCTeM presents you the opportunity to build upon its expertise in research for text mining and natural language processing, combined with their development experience from working in a practical, user focussed and service orientated environment. Our tools are currently in active use in both the public and private sectors with an extensive list of satisfied clients.

NaCTeM's core expertise lies within the biomedical and life science domains, though we have active involvement in the social sciences, within scholarly communication and developing scalable solutions for large scale document collections. In addition to this the technologies used are applicable to a broad range of disciplines and we would welcome the chance to look into new areas. Further to the development of software products for academia we offer additional services to the wider community including:

Training Events and Consultancy

Commercial Technology & Resource Licenses

Customisation and Support Agreements

Collaborative R&D Programmes

Bespoke Software Development

Contacts and Future Collaborations

NaCTeM is always keen to hear from both industry and academics. If you are interested in working with us, or would like to know more about our work, then please contact:

John McNaught

Deputy Director of NaCTeM

email: john.mcnaught@manchester.ac.uk

tel: 0161 306 3098

The University of Manchester,
Oxford Road, Manchester M13 9PL
Royal Charter Number RC000797