

CICAG Meeting : Scientific Text and Data Mining

20th.May 2009 Burlington House, London

Introduction

Chair : Alan Tonge

The last 20 years has seen the creation of a vast quantity of scientific documentation in electronic form: journals, patents, word-processed documents, e-theses and web pages. These have come in both structured (formatted) and unstructured (e.g. natural language) form, and text- and data-mining techniques have developed to automatically extract information and look for trends and hidden patterns in such documents and their data.

Speakers today represent practitioners from both academic and commercial areas.

Annotating Chemical Names

Peter Corbett, Unilever Centre for Molecular Informatics, Cambridge

Automatic named chemical entity (NCE) recognition. There is a significant problem arising from disagreement between experts about manual NCE assignments: inter-annotator agreement is only of the order ~90%, therefore any computer program which claims better success than this must be treated with suspicion. Automatic annotation has additional problems arising from: punctuation, whitespace, acronyms & abbreviations, ambiguity (e.g. compound vs. compound class).

OSCAR3 is an Open Source application, available online, which attempts to optimize NCE recognition by the use of: chemical dictionary look-up; character 4-grams of recognisable motifs (e.g. using regular expressions with wildcards: "one\$", "^ace") and Markov entropy rule filtering. Latter used for confidence estimates of extracted terms. Balance between precision (what proportion of identified terms are correct?) and recall (what proportion of relevant terms are recovered?).

Semantic Searching on Digital Repositories Based on Text Mining

Sophia Ananiadou, NaCTeM, Manchester

The challenges of annotating text: converting unstructured text (with implicit knowledge) into structured (with explicit knowledge). Searching will be semantically based - i.e. beyond the use of simple indexed terms. Requires ontologies (e.g. ChEBI) and lexicons prepared by domain experts to recognise chemical terms. How do you train automated methods to do this, test the results? Future development: ChETA (Chemistry using Text Annotations) with Unilever Centre.

Text Mining with Pipeline Pilot: a Bibliography Platform Example

Stephane Vellay, Accelrys

Data pipelining: customizable components for data filtering, processing and visualization. 'Chemical Text Mining' component processes e.g. PubMed articles identify documents (authors, titles, abstracts), extract descriptors and evaluate document similarity.

From High Throughput to High Frustration: Data Mining in Drug Discovery

Darren Green, GSK

Massive investment in the past decade by the pharmaceutical industry to develop automatic (robotic) methods for compound synthesis and testing - perhaps 100,000 primary screen plate pots per day. A number of new methodologies developed to interpret the results:

- Predictive ADMET attempts to take a lead candidate (potency-based) in order to predict the required drug-like attributes (toxicity, absorption, solubility)
- Enterprise Decision Management: operational decisions based on automated analysis
- Data mining ADMET (e.g. using applications such as ClearForest). Challenge: understanding the nature of the data extracted.

Analysis and Visualization Tools

Jeanette Eldridge, AstraZeneca

Patent submissions do not need to be machine readable (e.g. graphical images of structures) and there are no fixed standards for either chemical names (trivial or systematic), structures (explicit connection tables, Markush, CAS registry numbers) or their substituents (generic/homology, formulae).

There are a range of tools for searching and tabulating: SciFinder, STN Express, OmniViz/BioWisdom, Temis Luxid; Excel, Spotfire, STN AnaVist. International Patent Classification IPC codes are a potentially powerful.

Mining Scientific Information: Pitfalls and Possibilities

Neil Stutchbury, Neil Stutchbury Consultancy

The Name-2-Structure Challenge: chemical documents use inconsistent naming conventions (historical/trivial vs. systematic (IUPAC/CAS)) with additional problems caused by e.g. spelling mistakes. There are a number of commercially available applications which address the problem of automatic chemical name to structure (i.e. connection table) conversion: CambridgeSoft Name=Chem, LexiChem TK, ACD/Name.

Vision: use of Open standards, markup language(s) and domain ontologies to enable rapid document searching either by chemical name or structure/substructure. Will require author annotation or some automated markup process.

Mining and Meaning: Applying Standards to Chemical Content

Richard Kidd, RSC

Promotion and development of standards (e.g. identifiers (InChI) and ontologies (ChEBI)) for:

1. Finding chemistry on the Web
2. Enabling links between chemistry and biology

RSC developments:

1. TOAST (Tiny ontologies all strung together) - subject classifications, starting with named reactions and molecular processes
2. ChemSpider - InChI resolver standardization of data deposit to largest online database of chemical structures

Selected presentations are available on the CICAG website:

<http://www.rsc.org/Membership/Networking/InterestGroups/CICAG/meetings.asp>

Alan Tonge (August 2009)