Chapter I
Text Mining for Biomedicine

Sophia Ananiadou
University of Manchester, National Centre for Text Mining, UK

ABSTRACT

Text mining provides the automated means to manage information overload and overlook. By adding meaning to text, text mining techniques produce a much more structured analysis of textual knowledge than do simple word searches, and can provide powerful tools for knowledge discovery in biomedicine. In this chapter, the author focus on the text mining services for biomedicine offered by the United Kingdom National Centre for Text Mining.

INTRODUCTION

Text mining covers a broad spectrum of activities and a battery of processes, but essentially the goal is to help users deal with information overload and information overlook (Ananiadou and McNaught, 2006). Key aspects are to discover unsuspected, new knowledge hidden in the vast scientific literature, to support data driven hypothesis discovery and to derive meaning from the rich language of specialists as expressed in the plethora of textual reports, articles, etc. With the overwhelming amount of information (~80%) in textual unstructured form and the growing number of publications, an estimate of about 2.5 million articles published per year (Harnad, Brody, Vallieres, Carr, Hitchcock, Gingras, Oppenheim, Stamerjohanns, and Hilf, 2004) it is not surprising that valuable new sources of research data typically remain underexploited and nuggets of insight or new knowledge are often never discovered in the sea of literature. Scientists are unable to keep abreast of developments in their fields and to make connections between seemingly unrelated facts to generate new ideas and hypotheses. Fortunately, text mining offers a solution to this problem by replacing or supplementing the human with automated means to turn unstructured text and implicit knowledge into structured data and thus explicit knowledge (Cohen, and Hunter,

2008; Hirschman, Park, Tsujii, and Wong, 2002; (McNaught and Black, 2006)(Jensen, Saric, and Bork, 2006; Hearst, 1999).

Text mining includes the following processes: information retrieval, information extraction and data mining.

Information Retrieval (IR) finds documents that answer an information need, with the aid of indexes. IR or 'search engines' such as Google™ and PubMed© typically classify a document as relevant or non relevant to a user's query. To successfully find an item relevant to a search implies that this item has been sufficiently well characterised, indexed and classified such that relevance to a search query can be ascertained. Unfortunately, conventional information retrieval technology, while very good at handling large scale collections, remains at a rough granular level. Moreover, such technology typically focuses on finding sets of individual items, leaving it up to the user to somehow integrate and synthesise the knowledge contained in and across individual items. Thus, the content of documents is largely lost in conventional indexing approaches. To address this problem, we have improved the search strategy by placing more emphasis on terms in a collection of documents. In Biomedicine new terms are constantly created creating a severe obstacle to text mining and other natural language processing applications. In addition, term variation and ambiguity exacerbate the problem. We extract the most significant words in a collection of documents by using NaCTeM's TerMine service.[a] TerMine extracts and automatically ranks technical terms based on our hybrid term extraction technique, C-value (Frantzi, Ananiadou, and Mima, 2000). The C-value scores are combined with the indexing capabilities of Lucene 2.2 for full text indexing and searching.

Based on the assumption that documents sharing similar words mention similar topics, the extracted terms can be used for subsequent associative search. The output of associative searching is a ranked list of documents similar to the original document. This allows us to link similar documents based on their content. Another enhancement of the search strategy is query expansion. One of the major criticisms with current search engines is that queries are effective only when well crafted. A desirable feature is automatic query expansion according to the users' interests, but most search engines do not support this beyond mapping selective query terms to ontology headings (e.g. PubMed[b]). Therefore, there are inevitable limitations of coverage. To address this, we have used term-based automatic query expansion drawing upon weights given to terms discovered across different sized document sets. Query expansion embedded in searching allows the user to explore the wider collection, focusing on documents with similar significance and to discover potentially unknown documents

Information Extraction (IE) is characterized as the process of taking a natural language text from a document source, and extracting the essential facts about one or more predefined fact types. We then represent each fact as a template whose slots are filled on the basis of what is found from the text.

A template is a "form" which, when filled, conveys a fact. Each form has a label which characterises the type of fact it represents, whilst the slots identify the attributes that make up the fact. An example of a simple fact is:

James Smith, Chief Research Scientist of XYZ Co.

Examples of events are:

XYZ Co. announced the appointment of James Smith as Chief Research Scientist on 4th August 2005.

We hypothesized that retinoic acid receptor (RAR) would activate this gene.