Data Mining tools for Biomedical Discovery: a framework application in a Biotech Pharma Industry.

A case report of a working group.

Gualtieri F., Colace F., Garofalo P., Rovati L.
Rottapharm Biotech s.r.l., Monza (Italy)
Francesca Gualtieri, francesca.gualtieri@rottapharmbiotech.com

Abstract
Scientific literature provides a wealth of information to researchers to generate knowledge and assist in decision-making.
The number of articles in the literature databases is growing fast. Taking into consideration the huge amount of available information about diseases, drugs and targets is a time consuming process, it requires high-qualified search queries with keywords and synonyms or alternative names and care in items selection by professionals.
Automated process and analysis of text - Text Mining (TM) - may assist librarians and researchers in evaluating scientific literature in biomedical field.
Methods and Results
This study, while reviewing some Text Mining Tools, provides a practical example of their possible application in drug discovery and hypothesis generation in a biotech-pharma research team. In particular, we showed how KNIME (the Konstanz Information Miner, an open source data analytics) and R could be useful in improving the workflow of library search in Rheumatology.
Conclusion
TM represents an opportunity to automate literature search and has potential benefits such as time saving in the following activities: screening of relevant articles, decision making process, analysis of "big data" and ranking citations, integration of different sources as databases, biomedical and chemical libraries. Librarians and Information Specialist can play an important role in supporting, organizing the text mining tools and workflow.

Key words: Ontologies, Data Mining, Literature Searching, Text Mining Tools, Biomedical Discovery

Introduction

TM is generally defined as the automated processing of large amounts of digital data or textual content for purposes of information retrieval, extraction, interpretation, and analysis.(1) TM is a subfield of Data Mining that seeks to extract new valuable information from unstructured sources within documents, aggregates the parts of an entire collection of source documents, and reveals ”hidden” minings to readers (2).
The huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Data Mining provides the methodology and technology to transform these data into useful information for decision-making. (3) This automated process and analysis of text may assist librarians and researchers in evaluating the scientific literature in biomedical fields. Taking into consideration the huge amount of results and retrieving relevant information about diseases, drugs and targets is a time consuming process; it requires high-qualified search queries with keywords and synonyms or alternative names and care in items selection by professionals (4)

In fact, understanding large amount of text with the aid of computer tools is a harder job than simply installing or teaching grammar and dictionary to a computer. After term recognition and mapping (MeSH), it is important to include relations (taxonomies) and to bridge terms and resources.(5) TM can be used to analyse network of biological pathways (to understand molecular mechanism), gene prioritization and function prediction, biomarkers in prediction and disease progression, identification of new targets in drug discovery or repositioning (disease-specific drug-protein connectivity map), drug target identification using side effect similarity.
TM involves six common classes of tasks:
a) *Anomaly* detection – It is the identification of unusual data records that might be interesting or data errors that require further investigation.
b) *Association* rule learning (Dependency modelling) – It is the search for relationships between variables.
c) *Clustering* – It is the task of discovering groups and structures in the data. Groups of data that are in some way or another "similar".
d) *Classification* – it is the task of generalizing known structure to apply to new data.
e) *Regression* – it attempts to find a function, which models the data with the least error.
d) *Summarization* – it provides a more compact representation of the data set, including visualization and report generation.
The aim of our study is to compare two different TM tools and show their possible application in the drug discovery process and hypothesis generation in a biotech-pharma workflow in the rheumatology field.

Methods and Results
We considered **Knime** (the Konstanz Information Miner, an open source data analytics), an open source data analytics, reporting and integration platform, and **R,** a software environment for statistical computing and graphics widely used among statisticians and data miners for developing statistical software and data analysis.
We decided to work on these systems for several reasons, mainly because of their web-based open source availability and freely fee accessibility. The following search query was performed in both systems: *Osteoarthritis, Knee"[Mesh] AND ("humans"[MeSH Terms] AND pain*, to search for new valuable target(s) in knee osteoarthritis.
The process to extract concepts and their relationship from abstracts and papers is a very particular and ticklish; it needs the assistance of a controlled vocabulary (MeSH or UMLS). A correct identification of biomedical terms for disease, genes or gene/protein, described with a variety of descriptors, names and synonymous, is crucial in the disambiguation process of term recognition. A second step is to find co-occurrence concepts and analyze the nature of the relationship in a unified way. The aim of the process is to map semantically the extracted terms (words) and show them in a structured network or word cloud visualization to guide researchers in a fast and correct interpretation of results.
Via *word cloud* visualization it is possible to identify the most frequent occurrence (the largest font) of a term in a given topic. Instead, network analysis shows the relationships between subjects within the chosen topic. For instance, by network analysis it is possible to find an opinion leader in the osteoarthritis field or the best site where to conduct a clinical trial (keep me posted); to analyze direct or indirect interactions of protein-protein or gene-protein involved in a disease of interest (in osteoarthritis we extracted a network focused on matrix metallopeptidase 13) (5)(6).
We conclude that TM represents an opportunity to automate literature search and has potential benefits such as time saving in the following activities: screening of relevant articles, analysis of "big data" and ranking citations.

Biomedical Librarians can play an important and critical role in the TM process, providing researchers access to TM tools, taxonomy and vocabulary use or creation, and sharing content platform. They are actors in building the right conditions for collaborative computer-assisted process and analysis, supporting users by making data resources available as integrated data, bridging knowledge from literature.

REFERENCES (SUGGESTED READINGS)

1. Reilly BF. When Machines Do Research, Part 2: Text-Mining and Libraries. Charlest Advis. 2012;14(1):75–6.
2. Gonzalez GH, Tahsin T, Goodale BC, Greene AC, Greene CS. Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery. Brief Bioinform [Internet]. 2015;(August):bbv087. Available from: http://bib.oxfordjournals.org/lookup/doi/10.1093/bib/bbv087
3. Doms A, Schroeder M. GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Res [Internet]. 2005;33(Web Server):W783–6. Available from: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gki470
4. Fleuren WWM, Alkema W. Application of text mining in the biomedical domain. Methods [Internet]. Elsevier Inc.; 2015;74:97–106. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1046202315000274
5. Rodriguez-Esteban R. Biomedical text mining and its applications. PLoS Comput Biol. United States; 2009 Dec;5(12):e1000597.
6. Settles B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics. England; 2005 Jul;21(14):3191–2.