Methods 74 (2015) 97–106

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Application of text mining in the biomedical domain

Wilco W.M. Fleuren [a,b], Wynand Alkema [a,c,*]

[a] Computational Discovery & Design Group CMBI, Radboudumc, Nijmegen, The Netherlands
[b] Netherlands eScience Center, The Netherlands
[c] NIZO Food Research BV, Ede, The Netherlands

ARTICLE INFO

Article history:
Received 8 May 2014
Received in revised form 21 January 2015
Accepted 23 January 2015
Available online 30 January 2015

Keywords:
Text mining
Ontology
Natural language processing
Drug discovery
Biomedical research
Automatic information extraction

ABSTRACT

In recent years the amount of experimental data that is produced in biomedical research and the number of papers that are being published in this field have grown rapidly. In order to keep up to date with developments in their field of interest and to interpret the outcome of experiments in light of all available literature, researchers turn more and more to the use of automated literature mining. As a consequence, text mining tools have evolved considerably in number and quality and nowadays can be used to address a variety of research questions ranging from *de novo* drug target discovery to enhanced biological interpretation of the results from high throughput experiments. In this paper we introduce the most important techniques that are used for a text mining and give an overview of the text mining tools that are currently being used and the type of problems they are typically applied for.

1. Introduction

The scientific literature provides a wealth of information to researchers. It may serve as a starting point for assessing the state of the art in a particular field, or as a source of information that can be used for building research hypotheses that subsequently can be experimentally validated. Additionally this knowledgebase may serve as a source for interpretation of experimental results.

A large number of bibliographic databases is available in the life sciences domain, and have been reviewed by Masic and Milinovic [1]. One of the most important entry points to scientific literature sources for biomedical research is PubMed which gives access to more than 24 million scientific literature citations from MEDLINE, life science journals, and online books [2].

The number of articles that are added to the literature databases is growing fast. Fig. 1 shows the results of a PubMed search using terms that describe diseases, drugs and model organisms. In all cases, the number of papers that have been published on these subjects has increased exponentially. In addition to the exponential growth of the literature databases, the rate at which experimental data are produced has increased as well. For example in high throughput gene expression profiling or proteomics experiments, regulation of hundreds or thousands of genes and proteins is measured under multiple experimental conditions.

Retrieval of relevant information from literature databases and combining this information with experimental output is time consuming and requires careful selection of keywords and drafting of queries. This is often a biased and time consuming process, resulting in incomplete search results, preventing the realization of the full potential that these databases can offer [3].

Automated processing and analysis of text (referred to as text mining (TM)) can assist researchers in evaluating the scientific literature. Nowadays TM is applied to answer many different research questions, ranging from the discovery of drug targets and biomarkers from high throughput experiments [4–9] to drug repositioning, the creation of a state-of-the art overview of a certain disease or therapeutic area and for the creation of domain specific databases [10–15].

Due to the heterogeneous nature of written resources, the automated extraction of relevant biological knowledge is not trivial. As a consequence TM has evolved into a sophisticated and specialized field in the biomedical sciences where text processing and machine learning techniques are combined with mining of biological pathways and gene expression databases.

A number of reviews exists about TM in the biomedical domain that often emphasize the technical aspects of TM and the available tools or focus on gene and protein oriented information and less on the applications and real life research questions that even go beyond gene and protein research [16–19].

Here we give a state of the art overview of the use of TM for the biomedical domain and drug discovery. First we give a general

* Corresponding author at: NIZO Food Research BV, Ede, The Netherlands.
E-mail address: wynand.alkema@nizo.com (W. Alkema).

http://dx.doi.org/10.1016/j.ymeth.2015.01.015
1046-2023/© 2015 Elsevier Inc. All rights reserved.

description of TM, the different steps involved and the types of techniques that are used and describe some publicly available systems for TM. Subsequently we discuss a number of examples in which TM approaches have been applied to solve actual research questions. Finally we present an outlook in which we highlight the opportunities that TM can offer in the near future and the challenges that need to be addressed.

2. Text mining

A widely accepted definition of text mining has been provided by Marti Hearst, as “the discovery by computer of new, previously unknown information, by automatically extracting and relating information from different written resources, to reveal otherwise ‘hidden’ meanings” [20]. New hypothesis or facts that are the results of TM can subsequently be validated by experiments.

TM analysis typically involves a number of distinct phases, reviewed among others in [17,18,21,22], which are shown in Fig. 2 and described in detail below:

In the last decade a large number of applications have been developed (Table 1) that perform TM at various levels, and implement one or more steps from the scheme shown in Fig. 2. Each application has its own flavor of implementation, often driven by the exact question and the type of answer that is required and the intended user group.

2.1. Information retrieval

The first step in TM is to retrieve relevant textual resources for a given subject of interest. This process is referred to as information retrieval (IR) and is typically done by querying bibliographic databases with a set of keywords. The most used IR system by researchers in the biomedical domain is PubMed [9] that gives access to open source full text articles and abstracts of the MEDLINE database. Besides scientific literature, other literature resources such as patents, medical records, FDAs Medwatch reports, EudraVigilance reports, biomedical related blogs and websites are relevant text resources for biomedical research [1,49–51].

Notably, a lot of TM applications are built on the MEDLINE database, because it is freely available, features a rich applied programming interface and provides annotated abstracts with Medical Subject Heading (MeSH) Terms.

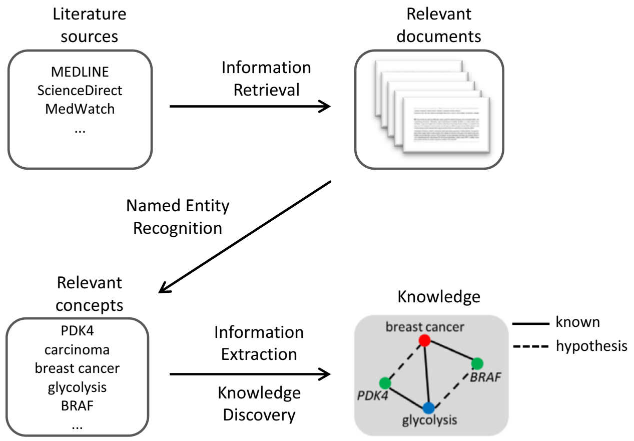

Fig. 2. Overview of a typical TM workflow. A typical TM workflow starts with information retrieval (IR) to get relevant documents for a given subject of interest. Using named entity recognition (NER) these documents will be analyzed for the occurrence of specific keywords. Information extraction (IE) is about detecting links between the found keywords and can be done in a number of ways (see text for more details). During knowledge discovery (KD) links between keywords can be used to infer new relations, so called hidden relations that can be seen as ‘true’ new knowledge.

A number of TM solutions offer enhanced IR by expanding the queries of the user by organizing similar keywords such as synonyms and alternative names into one concept based on a controlled vocabulary and subsequently incorporating all keywords of the same concept into the query. These tools are marked in Table 1 with the input type ‘Concepts’. The use of these extended queries may yield more comprehensive and more specific results. Next to enhancing the IR process, a number of TM tools analyze the results of the query and classify the retrieved documents based on their content or the occurrence of specific keywords in the documents. Sentence extraction, in which only those parts of the document are shown in which specific keywords occur assists the user in focusing on the relevant part of the IR output.

2.2. Named entity recognition

After IR, the resulting document set can be analyzed by search algorithms for the occurrence of specific keywords of interest

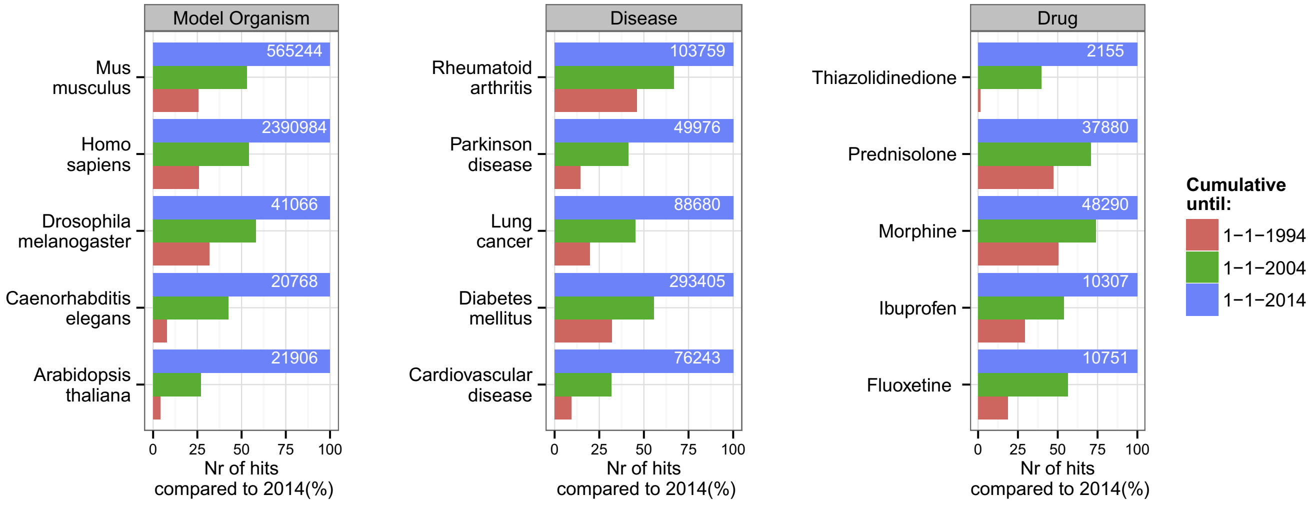

Fig. 1. Growth of the PubMed database. A number of queries for diseases, drugs and model organisms, indicated on the *y*-axis of each graph, were executed. The queries were run with varying cut-off values for the publication date indicated in the legend. The relative number of abstracts retrieved relative to the query with 01-01-2014 as the cut-off date is shown on the *x*-axis. The absolute number of abstracts for the query with 01-01-2014 as the cut-off date is shown on the right side of each group of bars.