KNOWLEDGE MANAGEMENT FOR BIOMEDICAL LITERATURE: THE FUNCTION OF TEXT-MINING TECHNOLOGIES IN LIFE-SCIENCE RESEARCH

Carmen Galvez
University of Granada
Granada /Spain
cgalvez@ugr.es

Abstract

Efficient information retrieval and extraction is a major challenge in life-science research. The Knowledge Management (KM) for biomedical literature aims to establish an environment, utilizing information technologies, to facilitate better acquisition, generation, codification, and transfer of knowledge. Knowledge Discovery in Text (KDT) is one of the goals in KM, so as to find hidden information in the literature by exploring the internal structure of knowledge network created by the textual information. Knowledge discovery could be major help in the discovery of indirect relationships, which might imply new scientific discoveries. Text-mining provides methods and technologies to retrieve and extract information contained in free-text automatically. Moreover, it enables analysis of large collections of unstructured documents for the purposes of extracting interesting and non-trivial patterns of knowledge. Biomedical text-mining is organized in stages classified into the following steps: identification of biological entities, identification of biological relations and classification of entity relations. Here, we discuss the challenges and function of biomedical text-mining in the KM for biomedical literature.

Keywords

Knowledge Management (KM), Knowledge Discovery in Text (KDT), Biomedical Tex-Mining, Ontology, Natural Language Processing (NLP)

1. INTRODUCTION

Recent discoveries in the domains of biomedicine have resulted in a huge volume of domain literature, which is constantly expanding both in size and thematic coverage. The healthcare knowledge base is expanding at an unprecedented rate. Although a great deal of crucial biomedical information is stored in factual databases, the most relevant and useful information is still represented in domain literature. Approximately 50,000 new records are added annually to the Melidne database alone. Open access publishers such as BioMed Central have growing collections of full-text scientific articles. High-throughput techniques are regularly used to capture thousands of data points in a single experiment, and many well-established low throughput experiments are performed in thousands of laboratories. The results of these experiments mostly end up in scientific databases and in scientific publications. Although there have been concerted efforts to capture more scientific data in specialist databases, it is generally acknowledged that only 20 per cent of biological knowledge and data is available in a structured format or a database. The remaining 80 per cent of biological information is hidden in the unstructured, free text of scientific publications [1]. The life-sciences, as a knowledge-driven discipline, currently produces more publications than any other scientific field. This means that, even for scientists who specialise in a specific subdiscipline, it is difficult to keep track of publications in their field of research. The main problem is the lack of a framework to manage the knowledge systematically. Without such a framework, it is not possible to efficiently deliver all the benefits of the knowledge that have already been discovered over the years.

The Knowledge Management (KM) for biomedical literature aims to establish an environment, utilizing information technologies, to facilitate better acquisition, generation, codification, and transfer of knowledge. A lot of effort has been focused on computerization, standardization, and automated

analysis of biomedical data. The information has to be interoperable among heterogeneous databases. Besides to interoperability, data standardization can also address semantic compatibility issues by mapping relevant text into standardized concepts in bio-ontologies, such as the *Unified Medical Language System* (UMLS) [2] or the *Gene Ontology*™ (GO) [3]. Ontologies help convey the semantics of textual information in a machine-understandable format so that data from different sources can be reliably integrated for more sophisticated analysis. Manual encoding or tagging of the entities to standardized concepts is labor intensive. Current expansion has heightened interest in: (a) *Information Retrieval* (IR), to gather, select, and filter documents that may prove useful; (b) *Natural Language Processing* (NLP) to automatically process the texts; and (c) *Information Extraction* (IE), a sub-area of NLP, to find relevant concepts, facts surrounding concepts, and relationships between relevant terms from the identified documents.

Techniques for literature mining are a requirement for effective knowledge discovery, management, maintenance and update in the long term. Processing biomedical literature faces many challenges, including both technical and linguistic. One of the main challenges in bio-text mining is the identification of biological terminology, which is a key factor for accessing the information stored in literature, as information across scientific articles is conveyed through the terms and their relationships. Thus, text-mining systems are indispensable tools to reduce the increasing flux of information in scientific literature to topics pertinent to a particular interest focus. Consequently, information processing systems must be applied to restrict the available information to that fraction which is pertinent to a particular topic or to a particular context within a topic. The challenge is to manage the increasing volume, complexity and specialization of knowledge expressed in the biomedical literature. Text-mining tools and methods can help researchers manage this affluence of information, and discovery facts, relationships and implications in literature that can be used to assist solve medical problems. Therefore, the objective of this work is to emphasize the potential that the text-mining techniques have in the broader methods of biomedical knowledge discovery and in the life-science research.

2. TEXT-MINING TECHNOLOGY IN THE BIOMEDICAL KNOWLEDGE MANAGEMENT

Text-mining has its origin from data-mining. The information in conventional data-mining is usually highly structured, containing mostly numbers and symbols. Data-mining is an analytical process entailing IR, NLP and IE, used to discover unsuspected associations; that is, combining or linking facts and events for the purpose of *Knowledge Discovery in Databases* (KDD). Data-mining methods can be generally grouped as: (a) *supervised methods*, to present documents according to predefined classes, such as techniques for inserting new documents into a previously existing ontology; and (b) *unsupervised methods*, such as clustering algorithms and visualization techniques, which gather texts on the basis of their similarity and thereby reduce the dimensionality of text representation. When data-mining processes are applied to texts in natural language, we speak of text-mining, also known as textual data-mining, intelligent text analysis, text data-mining, unstructured data management, or *Knowledge Discovery in Text* (KDT).

Knowledge discovery is one of the goals in KM. A large portion of biomedical information is available in electronic format. Because of the dynamic nature of biomedicine, the usage of available knowledge sources has to be combined with the dynamic management of terms and concepts encountered in texts. However, research articles, technical reports, clinical notes, and many other sources of information are stored as free-form text because of the flexibility it offers. Natural-language text often conveys rich concepts. It is difficult to extract new knowledge from these documents. KDT comprises three main tasks: IR, IE and text-mining. Biomedical knowledge in literature can be discovered through three basic procedures [4]:

(i) *Top-down approaches*, where researchers form hypotheses that lead to specific experiments, or create ontologies to describe the terminology and knowledge common to a given domain;

(ii) *Bottom-up approaches*, which try to discover interesting patterns or associations in existing data, in turn used to form new hypotheses (clustering techniques are used frequently for this purpose);

(iii) *Hybrid methods*, involving several techniques and knowledge sources in combination, such as information retrieval and term co-occurrence analysis, to arrive at complementary sets of documents that can help researchers articulate new hypotheses. In many cases implicit relationships are inferred simply by combining the principle of the co-occurrence of terms or concepts to some form of graphic association.

Biomedical text-mining, then, is the discovery by computer of previously unknown information, through the automatic extraction of information from different written resources. Most current systems address known relationships, and aim at the extraction of semantic or conceptual entities, properties of entities, and factual information involving identified entities. And this is one of the objectives of text-mining technologies, that is to identify non-trivial, implicit, previously unknown information in text. A key element is the linking together of this extracted information to form new facts or hypotheses that can be further explored using more conventional experimental means. Text-mining is the process of discovering and extracting knowledge from unstructured data, contrasting it with data-mining which discovers knowledge from structured data [5]. Text-mining enables analysis of large collections of unstructured documents for the purposes of extracting interesting and non-trivial patterns of knowledge. Biomedical text-mining is organized in stages classified into the following steps:

- **Identification of biological entities**. Biomedical texts are analyzed and stored in an internal representation form, after the elimination of stop-words, the exclusion of overly frequent terms, term standardization via stemming or lemmatisation, and the detection of noun phrases. Also text processing means tokenisation and then part-of-speech tagging, entity tagging or labelling and term recognition. Biomedical text-mining uses techniques from the field of data-mining but, because it deals with unstructured data, a major part of the text-mining process revolves around the crucial stage of pre-processing the document collections. NLP plays a major role in text-mining as it transforms text into structures that can be analyzed statistically.

- **Identification of biological relations.** Textual data can be analyzed using text-mining algorithms, that is, applying either unsupervised or supervised methods. Data analysis is dependant on the pre-processing. If a vector space representation has been chosen, the data can be analyzed using classic data-mining techniques, such as support vector machine. The vector is based on the bag-of-words model approach consisting of all words represented in the document. Clustering is an unsupervised learning problem where is necessary an automated way to organize this collection into documents relating to biomedical concepts. Text classification is a supervised learning problem where we know the labels of the documents (specified by domain experts) and train the corpus to effectively predict unknown future data in the right classes automatically.

- **Classification of entity relations**. The results are graphically represented, after constructing the biological entity-document index, this it is used to compute a network connecting graphically link between every pair of genes that co-occurred. The evaluation of extracted information or validation of results. Analyzing information from biomedical text is especially challenging because of the complexity of the field. Many text-mining techniques have incorporated ontologies to take advantage of the existing knowledge that they provides.

The process of text-mining needs a well-organized integration of these phases for knowledge discovery. Every phase of the text-mining process can be addressed with several different methods and technologies. Consequently, a large variety of method can be combined to solve the various aspects of literature mining. The combinatorial potentials are also reflected in the number of currently available tools for literature analysis. Then again, all of these tools address more or less the same tasks of identification and of gene relations.

2.1 Tex-mining technologies in the identification of biological entities

Biological entities — which here are names of genes, proteins, gene products, organisms, drugs or chemical compounds — are the means of scientific communication as they are used to refer to domain concepts: in order to understand the meaning of an article and to extract appropriate information, precise identification and association of terms is required. There are almost 300 biomedical databases containing terminological information. Many of such resources contain *descriptors* rather than terms as used in documents, which makes matching controlled sets of terms in literature difficult. Terminological processing (i.e. identification, classification and association of terms) has been recognised as the main bottleneck in biomedical text-mining [6], severely reducing the success rates of 'higher-level' text-mining processes which crucially depend on accurate identification and labelling of terms.

One of the inherent problems in automated biomedical literature mining using NLP is the difficulty of recognizing and resolving biomedical terms as certain named entities. Named entities in the biomedical literature can represent a variety of biological objects such as genes, proteins, diseases, drugs, organisms and other biological components. The caveat of biomedical literature is that naming schemes for these objects are very varied, and from one publication to another, gene names, protein names, and even disease names can significantly change in spelling, punctuation, abbreviation and even wording. Given a term in the context of one or two sentences, the task of correctly identifying it as a gene or protein or another biological object can be difficult even for a trained scientist who is familiar with the common nomenclature, let alone for an automated computer system. The problem intensifies when only the term is given, devoid of context. In such a case it is often impossible to classify the term correctly, because many proteins are named after the gene that produces them. Different biological objects may have the same exact name and can be differentiated only by the contextual words such as "protein" or "gene".

Irregular gene-naming arises in part because various researchers from different fields who are working on the same area of knowledge discover a large number of entities that need to be named. At present, some genes are denoted in publications under more than one name or symbol, and moreover, one symbol/name is sometimes used for several unrelated genes. Numerous hurdles in genomic information, then, are due to terminological variation and the complexity of names [/]. One major ensuing problem is ambiguity: among mouse gene names, variations accounted for 79% of the missing gene occurrences [8]. There is also a high correlation between the degree of term variation and the dynamic nature of biomedicine. As the use of gene symbols in publications can be confused or confusing, approved nomenclature is intended to enable scientists to access all data pertaining to a specific gene of interest, across species. And while nomenclature and ontological specifications are valuable for processing information, efforts toward systematic gene-naming have met with limited success.

There is large combined effort in the bioinformatics, NLP and biomedical communities to compile synonym lists of biomedical named entities. For example, the *National Library of Medicine* (NLM) along with *National Institutes of Health* (NIH) have a long established *Medical Subject Headings* (MeSH®) which define the "official" naming conventions for most of the biomedical terms in existence. However, this list of terms has to be manually curated, and the conventions are not universally recognized or followed. Trying to establish a universal naming scheme is an exercise in futility since there are no ways of effective enforcement of such standards in an open scientific community, and since new schemes are offered continuously.

2.2 Text-mining technologies in the identification of entity relations

The next goal after identification and disambiguation of biological entities is the detection of relations between the entities. The most simple approach for this task is to assume relations between entities based on co-occurrence in a text. The probability of an established relation between entities depends to some extent on the location of the entities within a text. The weakest hypothesis about a relation is due to a co-occurrence of entities anywhere in a text. If two entities co-occur within the same sentence, a true relation becomes more likely, while the coverage might decrease simultaneously. Sophisticated approaches try to further improve the analysis on sentence level by employing

dictionaries and rule-based analysis techniques. The dictionaries contain words related to the description of relations; the rules are designed for the analysis of sentence or phrase structures. These approaches lead to a better precision of results but decrease the recall owing to the restricted set of vocabulary and sentence structures. While all of these methods take the textual and sometimes grammatical context into consideration, none of them truly integrates the biological context.

The basis is that the co-occurrence of gene terms in the same sentence or the same document often implies real biological relationships between the named entities. Stapley and Benoit [9] tallied the number of co-occurrences of every pair of genes in Medline abstracts and used this data to calculate what they denote as '*BioBibliometric Distances*' between genes, so that the rarer the co-occurrence of two genes in the literature database, the larger the distance between them. The literature-derived gene-to-gene network may provide important information assigning a biological function to gene sequences and gene expression patterns. Therefore, *gene-to-gene co-citation networks* can be used to test new hypotheses, and new knowledge can be generated by reviewing these accumulated results in a concept-driven manner, linking them into testable chains and networks [10]. The nature of these relationships can be explored further using the *Medical Subject Headings* (MeSH®) index, or bio-ontologies. Considerable effort has likewise been centred on the construction of literature-based networks [11]. Novel approaches may also resort to the literature to establish functional relationships among genes—a methodology based on revealing coherent themes within literature through a similarity-based search in document space, after which the content relationships among abstracts are translated into functional connections among genes [12].

2.3 Text-mining technologies in the classification of entity relations

Relations between biological entities are not fixed but change according to the functional context in which an entity applies. The biological mechanisms and the environment in which the entity was observed generally specify the functional context of a biological entity. Consequently, the description of a functional context is usually distributed in multiple sentences and in-depth expert knowledge is required to decode the functional context from publications. Text-mining systems might support the identification of single aspects of a functional context such as a tissue type, but it is still impossible to automatically elucidate the complex dependencies between the components of a functional context. A relation might thus be described and correctly identified by steps 1 and 2 from a text, but the functional context in which the relation was observed might not correspond to the topic of interest.

The most common way to consider functional context in text-mining is the introduction of structured, hand-curated information about biological entities. Available sources such as GO assign biological entities to classes, such as biological functions and pathways. MeSH® assigns domains like diseases or anatomy to publications. These sources can be used to establish different 'lines of evidence' for an entity relation derived from text. A certain disease can be assigned to a relation if the paper in which the two entities were identified has been assigned to the disease via MeSH®. If both genes of an identified relation belong to the same class in GO, then this class is assigned to the relation. The natural internal consistency of biological facts and findings make such an approach possible.

All researchers would be able to associate certain related biological entities with others in the literature and databases. But it is difficult to know how this process is carried out. Attempts to impose standard names across the board are meeting stiff resistance, while approaches that would give genes unique numbers seem unlikely to take root [13]. Biomedical knowledge is particularly dependent on shared naming conventions: if researchers cannot clearly match a name to the underlying object (gene or structure), then some failure of communication is likely to occur [14]. Thus, this calls for improved text-mining tools of biological entity identification and better methods for visualizing information. Building such tools is critical for managing genomic information.

3. CONCLUSION

Text-mining promises to support many useful activities that are currently challenging to biologist, such as building models of biological systems as well as deriving novel hypotheses combining knowledge from different publications. Since textual format is a very flexible way to describe and store various

types of information, large amounts of information are stored and distributed as text. Moreover, the amount of accessible textual data has been increasing rapidly. Such data may potentially contain a great wealth of knowledge. However, analyzing huge amounts of textual data requires a tremendous amount of work in reading all of the text and organizing the content. The most important issue for this text-mining technology is how to represent the contents of textual data in order to apply statistical analysis. In terms of knowledge extraction, many kinds of knowledge can be extracted from textual data, such as linguistic knowledge for NLP and domain-specific lexical and semantic information that many be stored in a database. As a result, there are many useful applications for text-mining in the biomedical knowledge discovery and in the life-science research, particularly because of the huge amount of technical data and the relationships contained therein that are waiting to be identified and assembled.

4. REFERENCES

[1] Koehler J. Editorial. Briefings in Bioinformatics 2005; 6 (3): 220–221.

[2] Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The Unified Medical Language System: informatics research collaboration. J Am Med Inform Assoc 1998; 5(1): 1–11.

[3] Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nat Genet 2000; 25(1):25-9.

[4] Leroy G, Chen H. Genescene: an ontology-enhanced integration of linguistic and co-occurrence based relations in biomedical texts. J Am Soc Inf Sci Technol 2005; 56 (5): 457–68.

[5] Hersh W. Evaluation of biomedical text-mining systems: lessons learned from information retrieval. Briefings in Bioinformatics 2005; 6 (4): 344–56.

[6] Ananiadou S. Challenges of term extraction in biomedical texts. Available at: http://www.pdg.cnb.uam.es/BioLink/workshop_BioCreative_04/handout/

[7] Morgan AA, Hirschman L, Colosimo M, Yeh AS, Colombe JB. Gene name identification and normalization using a model organism database. J Biomed Informat 2004; 37: 396–410.

[8] Tuason O, Chen L, Liu H, Blake J, Friedman C. Biological nomenclatures: a source of lexical knowledge and ambiguity. In: Pac Symp Biocomput 2004. p. 238–49.

[9] Stapley BJ, Benoit G. Biobibliometrics: information retrieval and visualization from co-occurrence of gene names in Medline abstracts. In: Pac Symp Biocomput; 2000. p. 529–40.

[10] Jenssen T-K, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. Nat Genet 2001; 28(1): 21–8.

[11] Stephens M, Palakal M, Mukhopadhyay S, Raje R, Mostafa J. Detecting gene relations from MEDLINE abstracts. In: Pac Symp Biocomput 2001:483–496.

[12] Iliopoulos I, Enright AJ, Ouzounis CA. Textquest: document clustering of MEDLINE abstracts for concept discovery in molecular biology. In: Pac Symp Biocomput; 2001. p. 384–395.

[13] Pearson H. Biology's name game. Nature 2001;411:631–2.

[14] Hirschman L, Park C, Tsujii J, Wong L, Wu CH. Accomplishments and challenges in literature data mining for biology. Bioinformatics 2002; 18(12): 1553–61.