INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 77 (2008) 354–362

journal homepage: www.intl.elsevierhealth.com/journals/ijmi

Literature-based concept profiles for gene annotation: The issue of weighting

Rob Jelier, Martijn J. Schuemie, Peter-Jan Roes, Erik M. van Mulligen, Jan A. Kors*

Erasmus University Medical Centre, Department of Medical Informatics, P.O. Box 2040, 3000 CA Rotterdam, The Netherlands

ARTICLE INFO

Article history:
Received 30 November 2006
Received in revised form
17 April 2007
Accepted 24 July 2007

Keywords:
Text-mining
Vector space model
Gene Ontology

ABSTRACT

Background: Text-mining has been used to link biomedical concepts, such as genes or biological processes, to each other for annotation purposes or the generation of new hypotheses. To relate two concepts to each other several authors have used the vector space model, as vectors can be compared efficiently and transparently. Using this model, a concept is characterized by a list of associated concepts, together with weights that indicate the strength of the association. The associated concepts in the vectors and their weights are derived from a set of documents linked to the concept of interest. An important issue with this approach is the determination of the weights of the associated concepts. Various schemes have been proposed to determine these weights, but no comparative studies of the different approaches are available. Here we compare several weighting approaches in a large scale classification experiment.

Methods: Three different techniques were evaluated: (1) weighting based on averaging, an empirical approach; (2) the log likelihood ratio, a test-based measure; (3) the uncertainty coefficient, an information-theory based measure. The weighting schemes were applied in a system that annotates genes with Gene Ontology codes. As the gold standard for our study we used the annotations provided by the Gene Ontology Annotation project. Classification performance was evaluated by means of the receiver operating characteristics (ROC) curve using the area under the curve (AUC) as the measure of performance.

Results and discussion: All methods performed well with median AUC scores greater than 0.84, and scored considerably higher than a binary approach without any weighting. Especially for the more specific Gene Ontology codes excellent performance was observed. The differences between the methods were small when considering the whole experiment. However, the number of documents that were linked to a concept proved to be an important variable. When larger amounts of texts were available for the generation of the concepts' vectors, the performance of the methods diverged considerably, with the uncertainty coefficient then outperforming the two other methods.

* *Corresponding author.*Tel.: +31 10 4087045.
E-mail addresses: r.jelier@erasmusmc.nl (R. Jelier), m.schuemie@erasmusmc.nl (M.J. Schuemie), peterjanroes@charta.org (P.-J. Roes), e.vanmulligen@erasmusmc.nl (E.M. van Mulligen), j.kors@erasmusmc.nl (J.A. Kors).
1386-5056/$ – see front matter © 2007 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2007.07.004

1. Introduction

The number of scientific publications is increasing exponentially. In the fields of molecular biology and the biomedical sciences, scientists find themselves unable to read every publication of interest. Additionally, high-throughput experiments on genes and proteins, such as with DNA microarrays, have become common practice in these fields, causing a true information overload. The need for computational support to attempt to manage this information overload has become widely recognized and has spawned a lively area of research.

However, much of the knowledge on genes and proteins is locked in unstructured free text and cannot be used directly in computational systems. To save this several databases have become available that offer structured information on genes and proteins. These databases are either public, e.g. the databases offered by the Gene Ontology Annotation project [1] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) project [2], or commercial, e.g. as offered by GeneGO (http://www.genego.com) and Ingenuity (http://www.ingenuity.com). For a large part, these databases are filled with manually encoded information, generated by experts reading the scientific literature. Manual encoding is generally considered a reliable method for extracting information from the literature, but due to its labor-intensive nature, it is necessarily limited in scope and flexibility. Complementary to manual encoding, currently much research effort is spent in the field of text-mining: the development of computerized algorithms for extracting information from the scientific literature [3]. Automated methods have the advantage of speed and adaptability, though it is more difficult to achieve high precision and recall.

In text-mining two main approaches can be distinguished. One approach focuses on the extraction of precise relationships between genes and other biomedical concepts, using techniques varying from the detection of simple patterns such as "protein A - action X - protein B" [4,5], to the complete parsing of whole sentences [6]. The second approach uses the occurrence and co-occurrence statistics of terms from a thesaurus or lexical features, such as words or bi-grams, in a set of documents.

Here we focus on the use of occurrence and co-occurrence information in text-mining. Despite its conceptual simplicity, the approach has proven quite effective in the field of information retrieval and information extraction in the biomedical domain. For example, several authors [7–10] demonstrated the value of (co-)occurrence based systems for the analysis of DNA microarray data, and Stapley et al. [11] used weighted word counts to predict the sub-cellular location of proteins. The field of literature-based discovery, where the objective is to generate new hypotheses about relationships between concepts, makes ample use of occurrence and co-occurrence statistics (e.g. [12–14]). The approach has also been used to combine textual information with other types of information, typically to achieve specific tasks. For example, Xie et al. [15] use textual information together with sequence homology and information on protein domains to automatically assign Gene Ontology (GO) codes to proteins. Others combine gene expression data with text mining to identify disease genes [16].

In a number of text-mining approaches, concepts are represented by a set of texts related to the concept. Subsequently, concepts are related to each other by comparing the linked sets of texts. To make the comparison of two sets of texts, several authors [8,11,12,17] have used the so-called vector space model to characterize a set of texts. Using this model, a concept is represented by a concept vector: a list of associated concepts, together with weights that indicate the strength of the association. The associated concepts in the vectors and their weights are derived from the set of documents linked to the concept of interest. These concept-associated vectors, which we will call *concept profiles*, can be used to easily and transparently compare concepts based on underlying literature. Furthermore, patterns of similarity in a set of vectors can efficiently be found, for instance with clustering approaches. However, when using this approach, the determination of the weights in the concept profiles is an issue. Various weighting schemes have been proposed, with a wide range of motivations and statistical properties (see e.g. [7,8,18,19]), but a comparative study of these weighting schemes is lacking. Here we compare three weighting schemes for generating concept profiles:

(1) Weighting by averaging, an empirical approach. In this approach each document is characterized by a document vector, a (weighted) list of concepts found in the document. Glennison et al. [8] generated concept profiles by averaging document vectors.
(2) The log likelihood ratio, a test-based measure. The log likelihood ratio has been used in statistical natural language processing for collocation discovery [20] and has recently been applied in text-mining [17].
(3) The uncertainty coefficient, an information-theory based measure. The uncertainty coefficient is a normalized version of the mutual information measure, which is commonly used to measure stochastic dependence. An adapted mutual information measure was used by Wren [19] for his knowledge discovery system.

To compare the weighting schemes, they were applied in a system that annotates genes with GO codes, a task used before as a benchmark for text-mining systems (e.g. [21–23]). The Gene Ontology was designed to annotate gene products with their associated biological processes, cellular components and molecular functions in a species-independent manner [24]. As the gold standard for our study we used annotations of genes with GO codes as provided by the Gene Ontology Annotation project [1].

2. Methods

2.1. *Corpus and thesaurus*

The corpus of literature for our experiments consisted of 3,072,396 MEDLINE abstracts, selected with the PubMed query "(protein OR gene) AND mammals". We used titles, MeSH headings, and abstracts. Stop words were removed and words were stemmed to their uninflected form by means of the normalizer of the lexical variant generator [25].