Computer Methods and Programs in Biomedicine 157 (2018) 225–235

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Data mart construction based on semantic annotation of scientific articles: A case study for the prioritization of drug targets

Marlon Amaro Coelho Teixeira[a,d,*], Kele Teixeira Belloze[b], Maria Cláudia Cavalcanti[c,*], Floriano P. Silva-Junior[a,*]

[a] Oswaldo Cruz Institute (IOC), Oswaldo Cruz Foundation (FIOCRUZ), Av. Brasil 4365, Manguinhos, Rio de Janeiro 21040-360, Rio de Janeiro, Brazil
[b] Federal Center for Technological Education Celso Suckow da Fonseca, CEFET/RJ, Av. Maracanã 229, Rio de Janeiro, Rio de Janeiro, Brazil
[c] Computer Engineering Department, Military Institute of Engineering (IME), Praça General Tibúrcio 80, Urca, Rio de Janeiro 20271-064, Rio de Janeiro, Brazil
[d] Acre Federal Institute of Education and Science and Technology (IFAC), Av. Brasil 920, Xavier Maia, Rio Branco 69.903-068, Acre, Brazil

ARTICLE INFO

Article history:
Received 21 July 2017
Revised 21 December 2017
Accepted 10 January 2018

Keywords:
Semantic annotation
Decision support systems
Drug target prioritization

ABSTRACT

Background and objectives: Semantic text annotation enables the association of semantic information (ontology concepts) to text expressions (terms), which are readable by software agents. In the scientific scenario, this is particularly useful because it reveals a lot of scientific discoveries that are hidden within academic articles. The Biomedical area has more than 300 ontologies, most of them composed of over 500 concepts. These ontologies can be used to annotate scientific papers and thus, facilitate data extraction. However, in the context of a scientific research, a simple keyword-based query using the interface of a digital scientific texts library can return more than a thousand hits. The analysis of such a large set of texts, annotated with such numerous and large ontologies, is not an easy task. Therefore, the main objective of this work is to provide a method that could facilitate this task.
Methods: This work describes a method called Text and Ontology ETL (TOETL), to build an analytical view over such texts. First, a corpus of selected papers is semantically annotated using distinct ontologies. Then, the annotation data is extracted, organized and aggregated into the dimensional schema of a data mart.
Results: Besides the TOETL method, this work illustrates its application through the development of the TaP DM (Target Prioritization data mart). This data mart has focus on the research of gene essentiality, a key concept to be considered when searching for genes showing potential as anti-infective drug targets.
Conclusions: This work reveals that the proposed approach is a relevant tool to support decision making in the prioritization of new drug targets, being more efficient than the keyword-based traditional tools.

1. Introduction

Everyday new discoveries arise in the biomedical area and many of these advances are related to new techniques and new equipments used in high throughput experiments. An increasing volume of structured data has become available as a result from these experiments. Still, textual repositories are rich sources from which important information can be extracted by biomedical researchers. One of the most important digital repositories is PubMed[1], which accounts for approximately 26 million scientific texts. A typical scientific paper covers topics from distinct domains within the same area. Digital libraries classify and index large sets of scientific papers according to these topics, facilitating the scientist to find the papers of interest for his/her research interest. However, it is usual that a scientist may be interested in many combinations of distinct topics.

Text annotation allows the identification of the occurrence of such multidimensional combination of topics, and therefore makes it possible to rank these articles according to the scientists interest. However, an annotation should be well defined, not ambiguous and easy to understand by domain specialists, in a way that it could be useful for the information retrieval process [1]. Semantic annotation is one of the main efforts towards associating text content to its well-defined meaning. There are already many semantic annotation systems [2], which provide mechanisms to bring semantic to documents through text annotation. It means to handle and associate metadata or ontology concepts with text content.

* Corresponding author at: Acre Federal Institute of Education and Science and Technology (IFAC), Av. Brasil, Xavier Maia 69.903-068, Rio Branco, Acre 920, Brazil.
E-mail address: marlon.teixeira@ifac.edu.br (M.A.C. Teixeira).

[1] https://www.ncbi.nlm.nih.gov/pubmed.

https://doi.org/10.1016/j.cmpb.2018.01.010
0169-2607/© 2018 Elsevier B.V. All rights reserved.

An ontology is a model that represents a domain of reality, i.e., a formal description of concepts and relationships [3]. The use of ontologies is recommended not only to maintain annotations based on a uniform vocabulary, but also to benefit from the richness of the ontological representation. Through ontologies it is possible to make inferences about the annotations, getting information that is not always explicit to the user, and possibly, enriching annotations.

Nowadays, there are more than 500 ontologies on the biomedical domains [4,5]. On the other hand, taking into account that an ontology is typically designed to cover a single domain, in order to cover a scientific text, typically multi-domain, it is required the use of multiple ontologies while semantically annotating them. The multi-ontology annotation of such texts can be especially valuable for the biomedical scientist. Despite many initiatives on semantic annotation [2], none of them handle an analytical view of such annotations, such as the co-occurrence of a set of terms, representing each one a specific aspect of the scientist interest. Moreover, to the best of our knowledge, there are no previous reports on a method to bring an analytical view of a database of scientific text annotations.

Since the 90s a powerful approach for analytical view and decision support, known as data warehousing (DW), has been largely used. DW are non-volatile thematic driven databases, capable of integrating multiple data sources. In the context of DW arises the concept of data mart (DM). DM is usually defined as a subset of a DW, with a specific focus, or with a reduced number of dimensions [6,7]. There are many initiatives on the development and usage of DW on biomedical data [8,9]. These DW aim to provide a multidimensional view of the data, to answer complex analytical queries, such as “*what is the mortality rate for postmenopausal female patients admitted from the general ward with fever?*” (extracted from a Clinical Decision Support System [10]). Note that to answer such query, the DW needs to keep information about the temperature for each patient, hospital unit, time and patient status. Typically, for maintaining historical data, a DW is a high volume database. Moreover, since complex queries involve the manipulation of a wide set of records from multiple tables and the use of common joins and aggregations, performance and response time is a challenge in DW environments [6,7].

The model behind the DW design is based on facts and dimensions, and aims to address those performance issues. The fact is what needs to be observed, such as the temperature. Each observed fact is described or characterized by aspects or dimensions, such as: patient, hospital unit, time, etc. In order to facilitate the user in expressing such complex queries, the DW database is usually handled through On-Line Analytical Processing (OLAP) tools. The use of such tools allows users to perform operations like aggregation, detailing of hierarchical levels, selection, projection and reorientation of multidimensional view along multiple dimensions, enabling better insight of historical data and heterogeneous sources [11]. Using DW and OLAP systems in scientific research can help in the sense that it enables to *observe* scientific texts annotations and to correlate informations about different organisms.

In order to build and maintain data in a DW or DM, a process called ETL (Extract, Transform and Load) should be designed and implemented. Traditional methods of DM design are well consolidated, and usually apply generic techniques for structured data sources. However, there are just a few studies exploring specific methodologies to build DMs from unstructured data for analysis and decision support [12–14]. Some of them [15–18] propose ontology-based approaches, that they claim to be useful for dealing with textual data sources. On the other hand, none of them details how to handle and select large amounts of textual data. As mentioned before, this feature is fundamental to support scientific research, specially in the biomedical field.

Therefore, new approaches are needed to address scientific research, as for example in the search for new drugs in fight against neglected diseases. Neglected diseases are diseases caused by protozoa and they reach the poorest populations of third world countries. For these reasons, experimental data on drug targets from these organisms are still very scarce. The correlation of protozoans information with relevant data from other well studied (model) organisms can direct the researchers’ experiments, making the searches less costly and obtaining relevant results in a shorter time [19].

The objective of this work is to present a design methodology of a data mart (DM), usually defined as a smaller DW, for textual data analysis by means of ontologies. Thus, the idea is to provide a systematic way to process a large set of scientific articles and support the researcher in better decision making with respect to his/her specific research interests. We also present a case study on the design and load of a data mart with focus on gene essentiality for the following five protozoa: *Entamoeba histolytica, Leishmania major, Plasmodium falciparum, Trypanosoma brucei* and *Trypanosoma cruzi*. At the end, useful queries illustrate the benefits of this approach in the search for new drug targets.

2. Methods

Usually, the existing traditional ETL process [6,7] consists of three steps, as shown in Fig. 1(i). The *Extraction* step receives data from a variety of data sources, including structured data and text data sources, and stores them in an intermediary database, also known as Data Staging area (DS). The next step applies transformations on the collected data and prepares them for the final step, which is responsible to load them according to the dimensional model (dimensions and facts).

The proposed method, named TOETL (Text and Ontology ETL), was designed based on the traditional ETL process, to address the scientific texts annotation context. The first step of this method was described in previous works [19,20] and summarized in Section 2.1. The following steps, *Transform* and *Load* steps, which are the contributions of the present work, are detailed in Sections 2.2 and 2.3.

Fig. 1(ii) provides a brief view of the adaptation of the traditional ETL process, in order to address the specificity of a Scientific text annotation DM. Note that the sources are mainly text-based sources and that the *Extract* step is focused on the Semantic Annotation, using a set of ontologies as input. In this step, the scientist, which is the main user, defines the set of articles of his/her interest, and also the set of ontologies that are more suitable for annotating those articles.

2.1. Annotation step (extraction)

The authors have carried out a previous work [19,20], where the semantic annotation is organized as a set of steps (illustrated in Fig. 2) and this step is briefly described here.

It begins with the *understanding the research theme* step, which involves the study of the classic literature in the area, as well as interviews with the users. A set of terms, expressions and their synonyms are raised in order to identify in the literature a significant amount of scientific texts (corpus). This set should cover all domains across the research theme.

The set of terms is then used as input to the *queries definition* step to compose keyword search queries on a digital library. Both generic and specific queries should be formed, in order not to miss important literature items while building the corpus. The same set of terms is used at the *digital library selection* step, in order to choose libraries (one or more) that cover most of the domains involved in the research theme. Once the queries expressions are