Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Supporting biomedical ontology evolution by identifying outdated concepts and the required type of change

Silvio Domingos Cardoso[a,b,*], Cédric Pruski[a], Marcos Da Silveira[a]

[a] *LIST, Luxembourg Institute of Science and Technology, 5, avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg*
[b] *LRI, Univ. Paris-Sud, CNRS, Université Paris-Saclay, Rue Noetzlin, Bât. Ada Lovelace (650), 91405 Orsay, France*

ARTICLE INFO

Keywords:
Ontology evolution
Biomedical ontology
Semantic Web

ABSTRACT

The consistent evolution of ontologies is a major challenge for systems using semantically enriched data, for example, for annotating, indexing, or reasoning. The biomedical domain is a typical example where ontologies, expressed with different formalisms, have been used for a long time and whose dynamic nature requires the regular revision of underlying systems. However, the automatic identification of outdated concepts and proposition of revision actions to update them are still open research questions. Solutions to these problems are of great interest to organizations that manage huge and dynamic ontologies. In this paper, we present an approach for (i) identifying the concepts of an ontology that require revision and (ii) suggesting the type of revision. Our analysis is based on three aspects: structural information encoded in the ontology, relational information gained from external source of knowledge (i.e., PubMed and UMLS) and temporal information derived from the history of the ontology. Our approach aims to evaluate different methods and parameters used by supervised learning classifiers to identify both the set of concepts that need revision, and the type of revision. We applied our approach to four well-known biomedical ontologies/terminologies (ICD-9-CM, MeSH, NCIt and SNOMED CT) and compared our results to similar approaches. Our model shows accuracy ranging from 68% (for SNOMED CT) to 91% (for MeSH), and an average of 71% when considering all datasets together.

1. Introduction

The achievement of the Semantic Web vision, as initially described by Berners-Lee [2], implies the use of ontologies to semantically enrich web data in order to make it machine-understandable. In this paper, we adopted the same definition of an ontology as that used in the BioPortal repository [33] and considered the OWL representation of four well-known resources of the biomedical domain: MeSH,[1] ICD-9-CM,[2] SNOMED CT[3] and NCIt.[4] However, we highlight that the logical part of these models was not used for inferring new knowledge for reasoning issues [27]. The models were used only to define the concepts (and their attributes) and regions related to them that can contribute to identifying concepts that will potentially evolve in the near future. Discussions on other uses or definitions of ontologies are out of the scope of this paper. In [4], Bodenreider enumerates a set of applications where ontologies play a central role in the biomedical domain. They are, among other things, used for indexing large document collections such as MEDLINE,[5] and to support information retrieval, by associating terms provided by the MeSH terminology with scientific publications. In a clinical environment, ontologies are used to encode medical reports and facilitate access to patient data or for public health issues [25]. As a result, physicians are relieved of the tedious task of researching information, thus enabling them to focus on the patient and define personalized treatments. Ontologies are also important for companies that provide data curation services. In this case, ontologies are used to annotate biomedical or omics data and semantically bind pieces of information together [30]. Semantic annotations generated using ontologies then allow advanced data analytic tools to identify correlations between distributed information, which leads to the definition of new knowledge and the development of new drugs for the pharmaceutical

* Corresponding author at: LIST, Luxembourg Institute of Science and Technology, 5, avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg.
E-mail addresses: silvio.cardoso@list.lu (S.D. Cardoso), cedric.pruski@list.lu (C. Pruski), marcos.dasilveira@list.lu (M. Da Silveira).

[1] https://meshb.nlm.nih.gov/search.
[2] https://www.cdc.gov/nchs/icd/icd9cm.htm.
[3] https://www.snomed.org/snomed-ct/.
[4] https://ncit.nci.nih.gov/ncitbrowser/.
[5] https://www.nlm.nih.gov/bsd/pmresources.html.

https://doi.org/10.1016/j.jbi.2018.08.013
Received 14 May 2018; Received in revised form 16 August 2018; Accepted 28 August 2018
Available online 08 September 2018

industry. In this context, the relevance of the documents retrieved via semantic annotations clearly depends on the quality of these annotations.

In order to ensure the optimal performance of the systems relying on ontologies, these have to be fully aligned with the knowledge of the domain and the changes that occur must be propagated to dependent artefacts, including semantic annotations [6,7] and RDF datasets within the Linked Open Data cloud [1], and ontology mappings [10]. However, the size of biomedical ontologies makes the task of identifying outdated concepts (i.e., concepts needing revision) difficult for subject matter experts. Furthermore, performing ontological changes requires a significant processing effort due to the quantity of available medical information to be analysed.

In this paper, we propose a stochastic model implementing machine-learning techniques for identifying whether the content of an ontology needs to be revised in order for the domain to evolve. This also includes the identification of the type of non-logical changes that will affect a concept [16] (Extension, Removal, ChgDescription, Move – see Section 2). We consider the extension of the ontology, i.e. the addition of new concepts, modification of the description of a concept, e.g. modification of the label, removal of a concept, and the decision on whether a concept will move to another part of the ontology. We base our proposal on the state-of-the-art approaches of the field [8,24,31] and extend them in several ways by adding new features that were identified as playing a key role, by evaluating different techniques to deal with unbalanced datasets, and by analysing the impact of different machine-learning methods on different types (in terms of expressivity, size and dynamics) of ontologies. In addition to classical feature selection, mainly based on structural information (see Section 2) derived directly from the ontology, we used web information obtained by querying relevant scientific publications in the domain and the subset of information accessible through UMLS (Unified Medical Language System[6]), and also temporal information like the past evolution of the considered concept, as well as ontology region stability. Moreover, unlike existing work that clearly focuses on one dedicated ontology, i.e. Gene Ontology for Pesquita and Couto [24] and MeSH for Tsatsaronis et al. [31] and on the extension of the ontology, our method has been designed to cope with any existing ontology. We therefore propose an experimental validation of our model on four OWL versions of standards within the biomedical field with different sizes, levels of expressivity and evolution frequencies: ICD-9-CM, MeSH, NCI thesaurus and SNOMED CT. Furthermore, we also compare our model to existing models when possible.

The remainder of the paper is structured as follows: Section 2 introduces relevant notions and presents related work from the field predicting ontology evolution. Section 3 presents the material and methods we used to design our approach. Section 4 shows the experimental results we obtained for the evolution of biomedical ontologies and Section 5 discusses them. Finally, Section 6 concludes the paper and outlines future work.

2. Background

2.1. Problem statement

The main problem addressed in our work is the identification of needs for the evolution of the non-logical part of an ontology. We divided this problem into:

1. The identification of the set of concepts that need to be revised (associated with the function $Evolv_K$, defined below),
2. The recommendation of the type of revision that need to be implemented to update the concept considered (associated with the function $IdentTypeOfChange_K$, defined below).

In our context, $O^t = (C^t, R^t, A^t)$ represents version t of an ontology where C^t denotes the set of concepts, R^t the set of relationships between the concepts and A^t the set of axioms. Following the definition provided by Wang et al. [32], we define the meaning $M(c^t)$ of a concept $c^t \in C^t$ as a triple

$$M(c^t) = (label(c^t), int(c^t), ext(c^t))$$

In this definition, $label(c^t)$ represents the label of c^t, $int(c^t)$ is a set of properties *e.g.* object and datatype properties in OWL, or more generally speaking concept attributes, and $ext(c^t)$ is the extension of c^t (the set of individuals).

By

$$K = Struct(c^t) \cup Temp(c^t) \cup Rel(c^t)$$

we denote the context for our work. $Struct(c^t)$ represents the structural characteristics of c^t. It includes the intrinsic characteristic of a concept *e.g.*, the number of attributes defining a concept, or the number of siblings, superconcepts and subconcepts. $Temp(c^t)$ denotes the temporal characteristics of c^t, which includes aspects dealing with the history of a concept. In this work, we considered (i) the stability of c^t obtained by measuring the elapsed time between t and the version l, with $0 < l < t$ and $M(c^t) \neq M(c^l)$ and (ii) the stability of the neighbourhood of c^t (see Table 2). $Rel(c^t)$ considers the relational aspect of c^t acquired from external sources of information from the Web (see Section 3.3). Given one concept $c^t \in C^t$, our goal was to identify whether the meaning of c^t was still up-to-date at time $t + 1$ in a given context K. Therefore, regarding this problem, the function $Evolv_K$ is defined as follows:

$$Evolve_K: C^t \to \{0, 1\}$$
$$c^t \to \begin{cases} 0 & if \quad M(c^t) = M(c^{t+1}) \\ 1 & otherwise \end{cases}$$

The first challenge of this work was to find an alternative to correctly execute this function when $M(c^{t+1})$ is unknown. In a detailed analysis on the evolution process, we observed that a concept could evolve in different ways. Complementary to the previous problem, knowing that a concept will evolve, we aimed to detect the type of revision required to update c^t and obtain c^{t+1}. We assumed that four types of revisions were possible

$$RevType = \{Extension, Removal, ChgDescription, Move\}$$

where *Extension* refers to new concepts to be added as subconcepts of c^t at time $t + 1$. This type of revision was shown as relevant in [24]. *Removal* refers to the complete removal of c^t at time $t + 1$. *ChgDescription* denotes the modification in the label as well as in the attributes structure and attribute values of c^t at time $t + 1$. *Move* refers to changes in at least one superconcept of c^t at time $t + 1$ (*i.e.* the set of superconcepts of c^t is different from the set of superconcepts of c^{t+1}, implying a move of c^t to another part of O^t). These revision categories regroup the ontological modifications identified by the literature from the field ontology evolution [20,16,29,13,14]. As explained in Section 1, we focused on the non-logical part of the ontologies. To cover the logical part, we invite you to read [11,17]. To detect the revisions we were interested in, we used the COnto-Diff tool [13], but other *diff* tools such as PROMPT-Diff [23] may also be used. The inputs into the tool were the two versions of the ontology, and the output is the set of concepts and the revision actions associated with them.

Knowing that a concept had evolved, without having any other information about c^{t+1}, the second challenge of our work was to determine what type of revision was applied to the concept. In other words, in the perspective of the "identification of revision needs for a concept", we wanted to provide complementary information about what type of revision (from *RevType*) would be appropriated to keep the concept up-to-date. We associated this problem with the following function:

[6] https://www.nlm.nih.gov/research/umls/.