Artificial Intelligence in Medicine 63 (2015) 153–170

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Recognizing lexical and semantic change patterns in evolving life science ontologies to inform mapping adaptation

Julio Cesar Dos Reis [a,b,*], Duy Dinh [b], Marcos Da Silveira [b], Cédric Pruski [b], Chantal Reynaud-Delaître [c]

[a] Faculty of Campo Limpo Paulista, Rua Guatemala, 167, 13231-230 Campo Limpo Paulista, SP, Brazil
[b] Luxembourg Institute of Science and Technology, 29 Avenue John F. Kennedy, L-1855 Luxembourg, Luxembourg
[c] Laboratoire de Recherche en Informatique, University of Paris-Sud, Bâtiment 650 (Ada Lovelace), 91405 Orsay, France

ARTICLE INFO

Article history:
Received 30 June 2014
Received in revised form
27 November 2014
Accepted 30 November 2014

Keywords:
Biomedical ontologies
Ontology evolution
Ontology versions
Ontology changes
Mapping evolution
Mapping adaptation
Mapping maintenance

ABSTRACT

Background: Mappings established between life science ontologies require significant efforts to maintain them up to date due to the size and frequent evolution of these ontologies. In consequence, automatic methods for applying modifications on mappings are highly demanded. The accuracy of such methods relies on the available description about the evolution of ontologies, especially regarding concepts involved in mappings. However, from one ontology version to another, a further understanding of ontology changes relevant for supporting mapping adaptation is typically lacking.
Methods: This research work defines a set of *change patterns* at the level of concept attributes, and proposes original methods to automatically recognize instances of these patterns based on the similarity between attributes denoting the evolving concepts. This investigation evaluates the benefits of the proposed methods and the influence of the recognized change patterns to select the strategies for mapping adaptation.
Results: The summary of the findings is as follows: (1) the *Precision* (>60%) and *Recall* (>35%) achieved by comparing manually identified change patterns with the automatic ones; (2) a set of potential impact of recognized change patterns on the way mappings is adapted. We found that the detected correlations cover ∼66% of the mapping adaptation actions with a positive impact; and (3) the influence of the similarity coefficient calculated between concept attributes on the performance of the recognition algorithms.
Conclusions: The experimental evaluations conducted with real life science ontologies showed the effectiveness of our approach to accurately characterize ontology evolution at the level of concept attributes. This investigation confirmed the relevance of the proposed change patterns to support decisions on mapping adaptation.

1. Introduction

Mappings interconnect entities of domain-related ontologies through semantic relations [1]. They play a key role for enabling the interoperability between information systems, allowing software applications to semantically interpret annotated data using different ontologies [2]. In particular, recent work has explored the definition of mappings between medical terminologies for healthcare interoperability reasons [3]. This demands that mappings remain up to date despite the continuous evolution of ontologies. Besides errors at ontology alignment time, changes affecting ontology entities are one of the main reasons for invalidating mappings [4].

This leads to the mapping adaptation problem, referring to the way existing mappings are modified according to changes affecting ontology entities, to keep them semantically valid and complete over time [5,6]. Mapping adaptation remains challenging for many reasons. While knowledge engineers can manually accomplish this complex task on small ontologies, those of some specific domains, *e.g.*, the biomedical domain, require automatic software applications by virtue of their volume, the dynamic nature and the significant quantity of mappings impacted by ontology evolution [4]. Although sophisticated ontology changes can be discovered by modern tools (*e.g.*, OBO-Edit [7], OBO2OWL [8], *Conto-Diff* [9]), it is still difficult to exploit them for the adaptation of mappings impacted by ontology evolution. The reason for this is a lack of

* Corresponding author. Tel.: +55 (19) 3887 59 75.
E-mail address: juliocesardosreis@gmail.com (J.C. Dos Reis).

http://dx.doi.org/10.1016/j.artmed.2014.11.002
0933-3657/

adequate characterization of ontology evolution that would be suitable for supporting mapping adaptation. Indeed, we have demonstrated that even if concepts are interrelated in their entirety via mappings, only some partial information about concepts (*e.g.*, a set of most significant attributes) is useful for defining mappings [10,11]. In consequence, better describe the evolution of specific ontological entities will serve to inform the mapping adaptation task more effectively.

An example of a real situation observed in our studies refers to the transfer of information between concepts. We found cases where textual statements, which consist in values of attributes describing concepts, are completely transferred from one concept to sibling concepts. This has affected the associated mappings, since their definition relies on such textual information. We observed this case with the concept "560.39" of the ICD-9-CM[1] (ICD for short). Such concept contains three attributes and one of them has as value "Fecal impaction" (release 2009). Five mappings are defined having this concept as domain, and one of these mappings has a range called "Fecal impaction (disorder)", in SNOMED CT[2] (SCT for short). After evolution (*i.e.*, ICD release 2010), the attribute value "Fecal impaction" is no longer associated with the ICD concept 560.39 and the previously mentioned mapping has been removed. Moreover, the concept "Fecal impaction" has been newly created in ICD (release 2010) and is reconnected to "Fecal impaction (disorder)" of SCT. The example clearly highlights the need of methods for the characterization of the evolution of ontology concepts, by analysing changes in the attributes values to use this information to adapt invalid mappings.

Although significant research efforts have dealt with issues related to ontology evolution, understanding how this evolution affects dependent artefacts, such as mappings, has received very little attention [6]. Recent research work on ontology evolution mainly concerns internal logical inconsistencies of an ontology [12]. Even though existing researches investigate the evolution of specific life science ontologies like SCT [13,14], they fail to address the impact of the evolution on established mappings. For instance, Gonçalves et al. [13] analyze the changes between two ontology versions and use SCT to show the applicability of their approach. Spackman [14] investigates rates of change in large clinical terminologies using SCT as an object of study, although mappings are not taken into consideration. Groß et al. [15] study how mappings in life sciences ontologies change. They empirically analyze which ontology changes can lead to an addition or deletion of correspondences between concepts. Using a computed dataset of mappings, the study demonstrates how ontology changes can impact mapping changes, and the ratio of changes in mappings according to three general categories of ontology changes. The findings of the aforementioned studies have motivated a new study to gain a more in-depth understanding of how a more fine-grained classification of ontology changes would affect mapping adaptation, considering also real-world sets of mappings in the experiments.

In this article, we propose a formal definition of *ontological change patterns* that are relevant for supporting the automatic mapping adaptation. More specifically, the objectives are two-fold:

- We introduce a novel approach to automatically recognize instances of the proposed ontology change patterns by comparing successive ontology versions. This research examines whether techniques based on linguistic characteristics of attribute values combined with similarity measures play a role in supporting change pattern identification at the level of concepts' attributes.
- We experimentally assess the outcomes of the proposed methods and their ability to support mapping adaptation on realistic case studies using real mappings between large life science ontologies. We specifically explore, through our experiments, whether identified change patterns may influence different mapping adaptation actions.

Table 1
Notations. This table shows the relevant notations for this study and their descriptions.

Notation	Description
a_i	Attribute a_i denoted by a string
$a_i.name$	Attribute name (string)
$a_i.value$	Attribute value (string)
a_i^j	Attribute a_i at time j
w_{ki}^j	Single word/token w_k from attribute value $a_i.value$ at time j
$Ord(w_{ki})$	Position of word w_k from attribute value $a_i.value$
$A(c_k^j)$	Set of attributes of the concept c_k at time j
$R(c_k^j)$	Set of relationships of the concept c_k at time j
$C(O_x^j)$	Set of concepts of the ontology O_x at time j
$CT(c_k^j)$	Context of the concept c_k at time j
c_{cand}	Candidate concept
$W(a_i^j)$	Set of words/tokens from attribute value $a_i.value$ at time j
$Ch(w_{ki})$	Set of characters in word w_{ki}
$\|\|X\|\|$	Size of the given set X
$sim(a_i, a_j)$	Similarity between attributes a_i and a_j
M_{ST}^j	Set of mappings between ontology O_S and O_T at time j

We organize the remainder of this article as follows: Section 2 introduces the notations used throughout the article as well as the research problem. Section 3 presents the related work on change pattern identification and gives an overview of our original contributions. Section 4 reports on our approach to recognize lexical and semantic change patterns. Section 5 describes our experimental objectives and used materials. Sections 6–8 present the obtained results. We then discuss several aspects of our change pattern recognition approach in Section 9, prior to concluding this article and outlining directions for future work.

2. Background

We introduce the notations and definitions used in this article (Section 2.1) and describe the research problem (Section 2.2).

2.1. Notations and definitions

We present the notations used in this article in Table 1.

Ontology. An *ontology O* explicitly specifies a conceptualization [16]. More specifically, it describes a conceptualization of a domain by means of *concepts*, *attributes* and *relationships*. The definition adopted in this article considers an ontology O as a set of concepts interrelated by various relationships, *e.g.*, "*is-a*", "*part-of*", "*related-to*". The set of concepts of an ontology is defined as $C(O_x)=\{c_1, c_2, \ldots, c_n\}$. Each concept, characterized by a set of attributes, has a unique identifier and a set of relationships with other concepts of the same ontology. Given a concept c_k in ontology O_x, we denote O_x^j as ontology O_x at time $j \in \mathbb{N}$ and thus $c_k^j \in C(O_x^j)$.

We define the set of attributes defining a concept c as $A(c)=\{a_1, a_2, \ldots, a_n\}$ (*e.g.*, name, definition, synonym, *etc.*). The attributes can differ from one ontology to another, but in general an attribute describing a concept has a *name* and an associated *value*, *e.g.*, "*name:cardio_vascular_diseases*". Each attribute has particular semantics. For instance, the attribute "name" is used for denoting concept labels or the attribute "definition" is used for giving the meaning of a concept in a particular context. In some situations, we can have equivalent attributes in terms of meanings, *e.g.*, attribute "*name:hypotension*" is equivalent to "*synonym:low_blood_pressure*"

[1] http://www.cdc.gov/nchs/icd/icd9cm.htm (accessed 26.11.14).
[2] http://www.ihtsdo.org/snomed-ct (accessed 26.11.14).