Expert Systems with Applications 41 (2014) 5495–5508

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Assessing semantic annotation activities with formal concept analysis

Juan Cigarrán-Recuero [a], Joaquín Gayoso-Cabada [b], Miguel Rodríguez-Artacho [a], María-Dolores Romero-López [c], Antonio Sarasa-Cabezuelo [b], José-Luis Sierra [b,*]

[a] Escuela Técnica Superior de Ingeniería Informática, Universidad Nacional de Educación a Distancia, 28040 Madrid, Spain
[b] Facultad de Informática, Universidad Complutense de Madrid, 28040 Madrid, Spain
[c] Facultad de Filología, Universidad Complutense de Madrid, 28040 Madrid, Spain

ARTICLE INFO

Keywords:
Semantic annotation
Formal concept analysis
Ontology
Annotation tool

ABSTRACT

This paper describes an approach to assessing semantic annotation activities based on formal concept analysis (FCA). In this approach, annotators use taxonomical ontologies created by domain experts to annotate digital resources. Then, using FCA, domain experts are provided with concept lattices that graphically display how their ontologies were used during the semantic annotation process. In consequence, they can advise annotators on how to better use the ontologies, as well as how to refine these ontologies to better suit the needs of the semantic annotators. To illustrate the approach, we describe its implementation in @note, a Rich Internet Application (RIA) for the collaborative annotation of digitized literary texts, we exemplify its use with a case study, and we provide some evaluation results using the method.

1. Introduction

The enormous efforts to digitize physical resources (documents, books, museum exhibits, etc.), along with recent advances in information and communication technologies, have democratized access to a cultural, scientific and academic heritage previously available to only a few. Likewise, the current trend is to produce new resources in a digital format (e.g., in the context of social networks), which entails an in-depth paradigm shift in almost all the humanistic, social, scientific and technological fields. In particular, the field of the humanities is one which is going through a significant transformation as a result of these digitalization efforts and the paradigm shift associated with the digital age. Indeed, we are witnessing the emergence of a whole host of disciplines, those of Digital Humanities (Berry, 2012), which are closely dependent on the production and proper organization of digital collections.

As a result of the undoubted importance of digital collections in modern society, the search for effective and efficient methods to carry out the production, preservation and enhancement of such digital collections has become a key challenge in modern society (Calhoun, 2013). In particular, the annotation of resources with metadata that enables their proper cataloging, search, retrieval and use in different application scenarios is one of the key elements to ensuring the profitability of these collections of digital objects. While the cataloging and retrieval of resources (whether digital or non-digital) have been the object of study in library sciences for decades (Calhoun, 2013), modern applications require annotating resources in semantically richer and more flexible ways, in many cases allowing multiple alternative annotations in the same collection. In consequence, the tendency is to introduce the use of ontology-based semantic technologies, in addition to conventional metadata schemas (Keyser, 2012).

While in recent years we have witnessed significant advances in the automatic annotation of resources, in particular of those with heavy text content (see Section 6), there are multiple scenarios in which resource annotation cannot be inferred from the contents of these resources (e.g., scenarios involving resources in which the content is not directly related to the meta–information required). In these cases it is necessary to involve human annotators in the semantic annotation of the resources. The resulting activities are referred to as *semantic annotation activities* in this paper. Some examples of semantic annotation activities are the annotation of digital educational resources (e.g., *learning objects*) in the eLearning domain (Aroyo & Dicheva, 2004; Devedzic, Jovanovic, & Gasevic, 2007; Kurilovas, Kubilinskiene, & Dagiene, 2014; Tiropanis, Davis, Millard, & Weal, 2009), the annotation of media content in the multimedia domain (Hunter & Gerber, 2010; Labra, Ordónez, & Cueva-Lovelle, 2010; Mu, 2010; Šimko, Tvarožek, & Bieliková, 2013), or the one chosen as a case study in this paper: the annotation of digitized literary texts (Azouaou & Desmoulins, 2006; Donato et al., 2013; Gayoso, Sanz, & Sierra, 2013; Gayoso et al., 2012; Koivunen, 2005; Rocha, Willrich, Fileto, & Tazi, 2009; Schroeter, Hunter, Guerin, Khan, & Henderson, 2006; Tazi, Al-tawki, & Drira, 2003).

* Corresponding author. Tel.: +34 913947548.
E-mail address: jlsierra@fdi.ucm.es (J.-L. Sierra).

http://dx.doi.org/10.1016/j.eswa.2014.02.036
0957-4174/

The main objective of any semantic annotation activity should be to produce an annotation of the resources in the underlying digital collection that satisfies all the requirements of accuracy, completeness and adequacy posed by the intended uses of the collection. Therefore, being able to assess to what extent these requirements are accomplished is an obligation in order to guarantee the quality of the final annotation outcomes. On one hand, the result of this assessment could help annotators to make a better use of the semantic models (i.e., the *annotation ontologies*) during the annotation of the resources. On the other hand, it could also be useful to the creators of the ontologies (i.e., the experts in the domain), who could identify how their ontologies should be modified, augmented or refined on the basis of the actual use of these assets during the annotation process. However, for huge collections or dense and semantically-rich annotations, the accomplishment of this assessment by individual inspection of every single annotated resource can become a titanic task. Therefore, providing automatic or semi-automatic assistance in the assessment of semantic annotation activities is an overriding concern in guaranteeing the quality of the annotations performed.

This paper addresses the formulation of mechanisms that support the assessment of semantic annotation activities, in order to enable: (i) better guidance of annotators during the annotation process, and (ii) the iterative refinement of the annotation ontologies. For this purpose, it presents a method of assessing the use of ontologies in semantic annotation activities, based on formal concept analysis (FCA). In this approach, annotators are provided with ontologies specifically designed by domain experts, and they use these ontologies to annotate a collection of digital resources. Then, the annotated collections are automatically analyzed using FCA to allow domain experts access to a lattice-based graphical representation that summarizes the overall annotation activity. This representation is linked to the concepts in the ontology so that at a glance, domain experts can assess how the proposed ontology is being used by annotators. Along with other aspects, they can see which concepts are not being used, which concepts are always used together, and which concepts are used more often than others. As a result, they can provide guidance to the annotators, enabling them to better use the ontologies proposed, or they can find aspects of the ontology that can be improved (e.g., several concepts might be combined into a single concept or they could include new concepts made apparent from the concept lattice). Therefore, and under reasonable assumptions, FCA provides domain experts with the machinery necessary to address the assessment of semantic annotation activities, at least to a semi-automatic extent.

The approach proposed in this paper has been successfully used in @note, a Rich Internet Application (RIA) for the collaborative annotation of digitized literary texts for educational purposes. In @note, teams of annotators (students, in this case) must complete the annotation of digitized literary works with free-text notes, and they must catalogue these notes using concepts taken from an ontology provided by the domain experts (teachers, in this case). Once the annotation activity is complete, and according to the aforementioned approach, @note allows teachers to examine how students performed the annotation activity by showing them a concept lattice created by considering notes as objects and ontology concepts as attributes in a formal context.

The remainder of this paper is organized as follows. In Section 2, we describe the annotation assessment approach. In Section 3, we describe its implementation in @note. In Section 4, we present a case study, i.e., an annotation activity of a literary work (*The Library of Babel*, a short story authored by the Argentinian writer Jorge Luis Borges). In Section 5, we present some evaluation results. In Section 6, we describe some related works. Finally, in Section 7, we present the conclusions and directions for future work.

2. The assessment approach

This section describes our approach to the assessment of semantic annotation activities using FCA. In Subsection 2.1, we summarize the elements of FCA required in the approach. In Subsection 2.2, we present an overview of such an approach. In Subsection 2.3, we describe the nature of annotation ontologies. Finally, in Subsection 2.4, we present the use of FCA to facilitate the assessment of annotation activities by domain experts.

2.1. The elements of FCA

The annotation assessment approach proposed in this paper relies heavily on the construction of concept lattices from annotated digital resources. As mentioned earlier, we use the well-known FCA technique. FCA is a mathematical theory of concept formation derived from lattice and ordered set theories that provides a theoretical model for organizing information and revealing relationships (Wille, 1992; Ganter & Wille, 1999; Carpineto, & Romano, 2004; Wille, 2009). The main construct of the theory is the *formal concept,* which is derived from a *formal context*.

A *formal context* can be defined as a set of objects, a set of attributes and a set of *is-a* or *has-a* relationships between objects and attributes. A formal concept is a pair (A,B), where A is a set of objects (also known as the *extent* of the formal concept), and B is a set of attributes (also known as the *intent* of the formal concept). The extent and the intent of a formal concept are connected as follows:

- The extent A consists of all the objects that are related to all the attributes in the intent B.
- The intent B consists of all the attributes shared by the objects in the extent A.

Formal concepts can be ordered by their extents. More formally, $(A,B) \subseteq (C,D) \Leftrightarrow A \subseteq C$; in this case, (C,D) is called a *super-concept* of (A,B) and, conversely, (A,B) a *sub-concept* of (C,D). This order relation is a generalization-specialization, and it can be proven to be a *lattice* (i.e., a concept lattice) based on the basic theorem of FCA (Ganter & Wille, 1999; Wille, 1992).

In a concept lattice, two important types of formal concepts are *object concepts* and *attribute concepts*:

- The *object concept* associated with an object o is the most specific concept that includes o in its extent. The intent of an object concept is defined by all the attributes of o, whereas the extent contains not only object o but also all those objects related to all the attributes of o.
- The *attribute concept* associated with attribute a is the most generic concept that includes a in its intent. Its extent contains all the objects with attribute a, and its intent is defined by all the attributes shared by the objects belonging to the extent set.

Because concept lattices are ordered sets, they can be displayed naturally in terms of *Hasse diagrams* (Ganter & Wille, 1999). In a Hasse diagram: (a) there is exactly one node for each formal concept; (b) if, for concepts $C1$ and $C2$, $C1 \subseteq C2$ holds, then $C2$ is placed above $C1$; and (c) if $C1 \subseteq C2$ but there is no other concept $C3$ such that $C1 \subseteq C3 \subseteq C2$, there is a line joining $C1$ and $C2$.

Fig. 1(a) shows an example of a formal context, and Fig. 1(b) shows its associated concept lattice using a Hasse diagram.[1] This example illustrates that Hasse diagrams are particularly useful for visualizing concept lattices; thus they will be used in our approach as the primary means of presenting lattices to domain experts. The

[1] Concept lattices in section 2 have been generated with the ConExp application (http://conexp.sourceforge.net/).