Discovering Subsumption Axioms with Concept Annotation

Pierre Monnin, Amedeo Napoli, Adrien Coulet

Discovering Subsumption Axioms with Concept Annotation

Découverte d'axiomes de subsomption avec l'annotation de concepts

Pierre Monnin
LORIA (CNRS, Inria Nancy-Grand Est, Université de Lorraine)
Campus Scientifique BP 239
Vandœuvre-Lès-Nancy, France 54506
pierre.monnin@loria.fr

Amedeo Napoli
LORIA (CNRS, Inria Nancy-Grand Est, Université de Lorraine)
Campus Scientifique BP 239
Vandœuvre-Lès-Nancy, France 54506
amedeo.napoli@loria.fr

Adrien Coulet
LORIA (CNRS, Inria Nancy-Grand Est, Université de Lorraine)
Campus Scientifique BP 239
Vandœuvre-Lès-Nancy, France 54506
adrien.coulet@loria.fr

CCS CONCEPTS

• **Information systems → Data mining**; **Resource Description Framework (RDF)**; **Ontologies**; • **Computing methodologies → Ontology engineering**;

KEYWORDS

Linked Open Data, Formal Concept Analysis, Concept Annotation, ontology engineering

1 INTRODUCTION

Linked Open Data (LOD) consist of a large and growing collection of inter-domain data sets represented using Semantic Web standards including the use of RDF (Resource Description Framework) and URIs (Uniform Resource Identifiers) [1]. In LOD, resources can represent entities of the real world (*e.g.*, persons, organizations, places) and are identified with a URI. RDF statements use predicates to link resources to other resources (from the same data set or from other data sets), to literals (*e.g.*, strings, integers) or to classes of an ontology (class instantiation). Here, an ontology is a formal representation of a particular domain that consist of classes and relationships between them [3]. Among these relationships, we focus our work on the subsumption relation, stating that a class is more specific than another.

Available ontologies and LOD data sets are of various quality and completeness. In this paper, we propose to complete an ontology by discovering subsumption axioms between classes based on the regularities in a data set whose resources are linked to classes of the ontology. Particularly, we present and extend preliminary results published by the authors [6]. Formal Concept Analysis, a mathematical framework, is used to build a hierarchical structure called lattice representing the regularities between RDF resources and the predicates they are subject of. Concept Annotation is then applied on this structure to introduce ontology classes. This two-step process allows to discover subsumption axioms considering only the regularities in the data set.

2 LINKED OPEN DATA AND ONTOLOGIES

LOD are represented in the form of graphs encoded using RDF. The atomic element of an RDF graph is a triple denoted by: $\langle subject, predicate, object \rangle \in (U \cup B) \times (U \cup B) \times (U \cup B \cup L)$ where U is the set of URIs, L is the set of literals and B represent blank nodes. To illustrate our approach, we consider in this short article an abstract RDF data set (Table 1) along with an abstract ontology O (Figure 1). We denote C_O the set of classes of O. We are interested in the discovery of subsumption axioms. The subsumption relation is a transitive relation denoted by $\sqsubseteq$, where $c \sqsubseteq d$ means that every instance of c is also an instance of d. For the sake of abstraction, O uses `abstract:subClassOf` to express subsumption relations and `abstract:type` to express instantiations.

Table 1: Example of RDF triples, represented with the Turtle syntax.

r_1	abstract:type	k_1, k_2 .	r_2	$pred_3$	o_5 .
r_1	$pred_1$	o_1 .	r_3	abstract:type	k_1, k_2 .
r_1	$pred_2$	o_2 .	r_3	$pred_1$	o_6 .
r_2	abstract:type	k_1, k_2, k_4, k_5 .	r_4	abstract:type	k_1, k_2, k_5 .
r_2	$pred_1$	o_3 .	r_4	$pred_2$	o_7 .
r_2	$pred_2$	o_4 .	r_4	$pred_3$	o_8 .

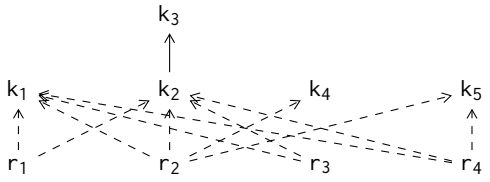

Figure 1: Example of ontology classes being instantiated by resources. Instantiations are represented using dotted arrows and subsumption relations using solid arrows. For example, r_1 instantiates k_1 and k_2, and k_2 is subsumed by k_3.

We define the *type* of a resource r as the set of classes of O that r instantiates. Formally, $\text{type}(r) = \{c \in C_O \mid \langle r, \texttt{abstract:type}, c \rangle\}$. Accordingly to the definition of the subsumption, we define the *extended type* of r as the whole set of superclasses of the classes of type(r). Formally, $\text{extdtype}(r) = \text{type}(r) \cup \{d \in C_O \mid \exists c \in \text{type}(r),\ c \sqsubseteq d\}$.

3 FORMAL CONCEPT ANALYSIS AND CONCEPT ANNOTATION

Formal Concept Analysis (FCA) is a mathematical framework whose basics can be found in [2]. It is used to build a hierarchical structure called lattice that denotes the regularities in a considered data set. Indeed, from the LOD triples in Table 1, we propose the formal

context in Table 2. It is noteworthy that only subjects and predicates are considered to build this context: a subject and a predicate are related in this context if and only if a RDF triple exists where the subject and the predicate appear together. Then, standard FCA may be applied to produce the lattice visible in Figure 2. Formal concepts (A, B) are formed by two sets: the extent A containing subjects and the intent B containing predicates. To link formal concepts with ontology classes, Concept Annotation [6] is applied to each of them. Considering a formal concept (A, B), its annotation is defined as $A^{\diamond} = \bigcap_{r \in A} \text{extdtype}(r)$. It represents the shared extended type of subjects in A. Given two formal concepts $(A_1, B_1) \leqslant (A_2, B_2)$, as $A_1 \subseteq A_2$, we have $A_2^{\diamond} \subseteq A_1^{\diamond}$. Therefore, the annotation can be depicted using an extension of the reduced labeling applied on lattices [2]. We call the *proper annotation* of a concept its annotation excluding classes appearing in the annotation of its superconcepts.

From the annotated lattice in Figure 2, subsumption axioms may then be discovered. Consider a concept (A, B), *e.g.*, concept 6, and one of its covering superconcept (E, F), *e.g.*, concept 4. As in Figure 2, we denote $A_A^{\diamond} = \{x_1, x_2, \ldots, x_p\}$ (here, $\{k_4\}$) and $E_A^{\diamond} = \{y_1, y_2, \ldots, y_q\}$ (here, $\{k_5\}$) the proper annotations of the two concepts. Then, consider two ontology classes $x_i \in A_A^{\diamond}$ and $y_j \in E_A^{\diamond}$. By definition, x_i appears in the extended type of all subjects in A and y_j appears in the extended type of all subjects in E. As $A \subseteq E$, y_j appears in the extended type of all subjects where x_i also appears. Moreover, y_j also appears in the extended type of other subjects. Thus, we consider this as the discovery of a subsumption axiom stating that x_i is subsumed by y_j (here k_4 should be subsumed by k_5). This axiom is then compared with axioms already defined in the ontology: *(i)* if $x_i \sqsubseteq y_j$ is already explicitly stated, this is a *redundant axiom*, *(ii)* if $x_i \sqsubseteq y_j$ is not already stated, but can be inferred, this is an *inferable axiom* and *(iii)* if $x_i \sqsubseteq y_j$ is neither already explicitly stated nor inferable, it is a n*ew subsumption axiom* that has been discovered. Here, the axiom $k_4 \sqsubseteq k_5$ is a new axiom.

If one considers only the covering concepts during the discovery process, this may lead to miss some interesting axioms. For example, in Figure 2, as the proper annotation of concept 2 is empty, we cannot discover axioms when considering concept 4 and its superconcept. Therefore, we propose here to discover subsumption axioms from the induced order on annotations from the lattice, obtained by considering only the annotations ordered by set inclusion following the order between formal concepts. This induced order is represented, using the reduced labeling, on the right of Figure 2. Set inclusion is read from top to bottom. Thus, subsumption axioms can be read from this reduced labeling from bottom to top, leading to discover more axioms: k_5 being subsumed by k_1, k_2 and k_3 and k_4 being subsumed by k_1, k_2 and k_3. Such axioms cannot be discovered by only considering the covering concepts.

4 PERSPECTIVES

We ran preliminary experiments on DBpedia [5], a LOD data set built from Wikipedia. It is necessary to evaluate and compare the obtained results with results of other methods. The annotation allows to take into account a third dimension (ontology classes) when describing data, in addition to the first two dimensions (RDF subjects and their predicates). Therefore, Triadic Concept Analysis [4] could be considered for comparison as it is an extension of FCA that also introduces a third dimension in the description of data. Finally, it could also be interesting to compare these two methods with results obtained using standard FCA. Indeed, because of its two-dimensional model for data description, predicates and ontology classes can only be mixed in one dimension.

Table 2: Formal context built from the RDF triples in Table 1. A cross in the table relates a subject and a predicate if and only if a RDF triple exists where the subject and the predicate appear together.

	abstract:type	$pred_1$	$pred_2$	$pred_3$
r_1	×	×	×	
r_2	×	×	×	×
r_3	×	×		
r_4	×		×	×

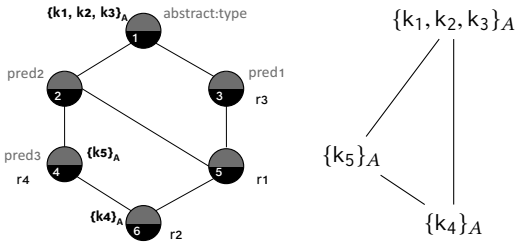

Figure 2: On the left: line diagram representing the annotated concept lattice built from the formal context in Table 2 and the ontology in Figure 1. On the right: line diagram representing the induced order on annotations. The lattice and the induced order are displayed using the reduced labeling extended to annotations. Subjects are depicted in black, predicates in grey and annotations are depicted by $\{\cdot\}_A$. Formal concepts are arbitrarily numbered from 1 to 6.

ACKNOWLEDGMENTS

This work is supported by the *PractiKPharma* project, founded by the French National Research Agency (ANR) under Grant No.: ANR-15-CE23-0028 (http://practikpharma.loria.fr/), and by the *Snowball* Inria Associate Team (http://snowflake.loria.fr/).

REFERENCES

[1] Christian Bizer, Tom Heath, and Tim Berners-Lee. 2009. Linked Data - The Story So Far. *Int. J. Semantic Web Inf. Syst.* 5, 3 (2009), 1–22.
[2] Bernhard Ganter and Rudolf Wille. 1999. *Formal Concept Analysis: Mathematical Foundations.* Springer.
[3] Thomas R Gruber et al. 1993. A translation approach to portable ontology specifications. *Knowledge acquisition* 5, 2 (1993), 199–220.
[4] Fritz Lehmann and Rudolf Wille. 1995. A triadic approach to formal concept analysis. *Conceptual structures: applications, implementation and theory* (1995), 32–43.
[5] Jens Lehmann and *et al.* 2015. DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web* 6, 2 (2015), 167–195. https://doi.org/10.3233/SW-140134
[6] Pierre Monnin, Mario Lezoche, Amedeo Napoli, and Adrien Coulet. 2017. Using Formal Concept Analysis for Checking the Structure of an Ontology in LOD: The Example of DBpedia. In *Foundations of Intelligent Systems - 23rd International Symposium, ISMIS 2017, Warsaw, Poland, June 26-29, 2017, Proceedings.* 674–683. https://doi.org/10.1007/978-3-319-60438-1_66