ECAI 2014
T. Schaub et al. (Eds.)

doi:10.3233/978-1-61499-419-0-987

A deductive approach to the identification and description of clusters in Linked Open Data

Simona Colucci[1] and **Silvia Giannini** [2] and **Francesco M. Donini**[1] and **Eugenio Di Sciascio**[2]

Abstract. We propose an approach for inferring clusters in collections of **RDF** resources based on the features shared by their descriptions. The approach grounds on an algorithm for computing Common Subsumers in **RDF** proposed in a previous research work. The clustering service introduced here returns not only a possible partition of resources in a collection, but also a description of the knowledge shared within each cluster, in terms of (generalized) **RDF** triples.

1 Introduction

The Web of Data [7] is nowadays a fact, as testified by the huge amount of data available in machine-understandable and inter-operable formats, like **RDF**[3]. The Linked Open Data (LOD)[4] initiative has in fact been joined by several organizations, that chose to publish their data following the **RDF** standard notation. As a consequence, a significantly rich informative content becomes available, opening new challenges to be addressed through reasoning.

The proposed approach aims at finding commonalities in LOD by exploiting a specifically developed reasoning service, Common Subsumer(CS) of pairs of **RDF** resources [2], which copes with the difficulties arising from the attempts of reasoning over **RDF** [4]. As the service name may suggest, CS is defined in analogy with a specific DLs inference: Least Common Subsumer (LCS) [1]. Differently from LCS, CS computation gives up subsumption minimality and searches for knowledge pieces which may be inferred by both input resources. The proposed approach shows how such knowledge, although not subsumption-minimal, is still useful to deduce descriptions of clusters of **RDF** resources in a given domain. In particular, we chose LOD by Chamber of Deputies of Italian Parliament[5] as case study.

In the next section we describe the main features of our approach, together with some reference to its implementation and some preliminary results. Section 3 closes the paper.

2 The Approach

We aim at automatically clustering collections of **RDF** resources according to a fully semantic-based classification. In particular, **RDF** descriptions are investigated to infer non-overlapping clusters of resources entailing the same sets of **RDF** triples, in order to provide a description of the informative content shared within each cluster.

The originality of the proposal lays in the choice of adopting deductive services to learn[6] clusters description from examples represented in **RDF**. In fact, although clustering is a thoroughly investigated task in machine learning literature, approaches solving it usually adopt induction to identify clusters according to some—sometimes semantic-based—distance between elements in the same cluster. On the other hand, we propose to cluster a target collection through a deductive and fully semantic-based approach, which relies on the iteration of two steps: i) the CS of two randomly selected **RDF** resources is computed; ii) the rest of the collection is queried in order to find other items entailing the same CS. The subcollection made up by the two initial resources and those returned by step ii) is one cluster of the collection. In [2] an anytime algorithm has been proposed to compute a CS of pairs of **RDF** resources. In order to ensure correctness and computability, the algorithm computes the CS always referring to a customized representation of **RDF** resources, which we call *r-graph*: a portion of the Web of Data we consider relevant for the description of each input resource.

The LOD by Chamber of Deputies of Italian Parliament[5] is organized in about thirty different interlinked RDF datasets, accessible through an OpenLink Virtuoso SPARQL-protocol endpoint. For the current experimental evaluation, we cluster only resources contained in the dataset `deputato.rdf`, even though their descriptions span multiple datasets.

The reader may find in Figure 1-a) two example r-graphs describing deputies Nilde Iotti (`ocd:d3140_10`) and Tina Anselmi (`ocd:d270_10`) of the 10^{th} legislature of the Italian Republic. In the most general case, we consider relevant to describe a resource only triples having the resource itself as subject. Here, we adopt a more restrictive strategy to compute the r-graphs and we consider not relevant for a resource r also triples as `r p o .` such that `p` $\in$ {`dc:date`, `dc:title`, `foaf:depiction`, `foaf:firstName`, `foaf:nick`, `foaf:surname`, `ocd:endDate`, `ocd:file`, `ocd:startDate`, `ods:modified`, `rdfs:comment`, `rdfs:label`, `terms:isReferencedBy`}.

In a nutshell, the algorithm for computing the CS of two resources t and s starts by computing the r-graphs corresponding to t and s according to a flexible criterion and returns their CS as a pair $\langle x, T\rangle$, made up by a blank node x and a set of triples T, entailed by the r-graphs of both input resources. In Figure 1-b), it is shown a CS derived from the r-graphs in Figure 1-a)). It is worth noticing that, differently from other

[1] DISUCOM, Università della Tuscia, Viterbo, Italy
[2] DEI, Politecnico di Bari, Bari, Italy
[3] `http://www.w3.org/RDF/`
[4] `http://linkeddata.org/`
[5] http://dati.camera.it/data/en/

[6] In Machine Learning this is called *unsupervised learning*

Table 1. Clustering Results adopting a randomly selected initial seed pair

Seed's URIs	ocd:rif_mandatoCamera	ocd:membro	ocd:aderisce	foaf:gender	dc:description	$\|P\|$
($d19990\_1, d20060\_1$)	_:x1	_:x2	_:x3	"male"	"Laurea in giurisprudenza; avvocato."@it	127
($d3140\_1, d14290\_1$)	_:x1	_:x2	_:x3	"female"	"Laurea in lettere; insegnante."@it	9
($d12560\_1, d13120\_1$)	_:x1	_:x2	_:x3	"male"	_:x4	431
($d26000\_1, d10090\_1$)	_:x1	_:x2	_:x3	"female"	_:x5	35
($d10800\_1, d25610\_1$)	_:x1	_:x2	_:x3	"male"		9
($d12140\_1, d8520\_1$)	_:x1		_:x2	_:x3		2

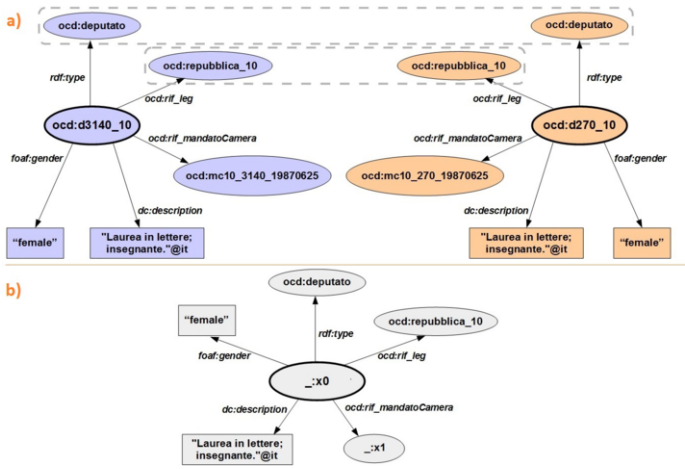

Figure 1. R-graphs (a) and CS (b) of N. Iotti and T. Anselmi.

proposed solutions [3], the algorithm also visits r-graphs of resources standing as triples predicate. As an example, let `t p b .` and `s q d .` be two triples in T_t and T_s, respectively; our algorithm explores also the r-graphs rooted in p and q, searching for triples entailed by both of them. This higher-order feature of **RDF** that we consider makes our approach more general than methods for computing CS in Description Logics, in particular $\mathcal{EL}$ [5].

Then, the set T of triples is used to model a SPARQL [6] query, which returns a subset P of the target collection R, such that the **RDF** description of each item in P entails all triples in T. To this aim, we need a *query compilation* phase in which we translate conditions expressed in the set T in an equivalent list of queries, such that the intersection of results of each query yields the desired cluster. As an example, consider the following set of triples: T ={`_:x0 rdf:type ocd:deputato.`, `_:x0 foaf:gender "female".`, `_:x0 ocd:aderisce _:x1.`, `_:x0 ocd:aderisce _:x2.`, `_:x1 rdfs:label "COMUNISTA".`} The compiled queries are shown in the following:

```
1. SELECT DISTINCT ?d WHERE {?d a <http://dati.camera.it/ocd/deputato>.}
2. SELECT DISTINCT ?d WHERE {?d <http://xmlns.com/foaf/0.1/gender> "female".}
3. SELECT DISTINCT ?d WHERE{?d <http://dati.camera.it/ocd/aderisce> ?s1.
                            ?s1 <http://www.w3.org/2000/01/rdf-schema#label> "COMUNISTA".}
4. SELECT DISTINCT ?d WHERE {?d <http://dati.camera.it/ocd/aderisce> ?s1.}
```

We show clustering results with reference to a set R of 613 resources corresponding to deputies of the first legislature of the Italian Republic. It means that every pair of resources (t, s) randomly selected from R returns a CS described, at least, by the following triples: `_:x rdf:type ocd:deputato .` and `_:x ocd:rif_leg ocd:repubblica_01 .` , where `_:x` stands for the blank node associated to the CS of t and s.

Tab. 1 reports a clustering proposal for R, in which the seed's URIs have been randomly selected. By looking at the first row in Table 1, one can notice how the algorithm aggregates 127 resources –including the seed pair– in R that (please follow the columns order): received an open mandate to the Chamber of Deputies; were members of a committee, joined a parliamentary group, are of male gender; worked as a lawyer, after obtaining a law degree.

3 Conclusion

We proposed a new, deductive strategy for clustering collections of **RDF** resources on the basis of the informative content shared by their descriptions expressed in form of generalized **RDF** triples. The clustering mechanism relies on the computation of a CS [2] of pairs of resources, which we describe by selecting a relevant portion of the Web of Data. The evaluated execution time of the whole clustering approach, together with the clustering results in terms of provided informative content, seem to support the effort spent in designing and implementing the clustering strategy and suggests new efforts in our future work. Future work will be devoted to the extension of CS definition and computation to other entailment regimes and to the investigation on general criteria for the selection of relevant triples.

Acknoledgements

We acknowledge support of project "A Knowledge based Holistic Integrated Research Approach" (KHIRA - PON 02_00563_3446857).

REFERENCES

[1] W. Cohen, A. Borgida, and H. Hirsh, 'Computing Least Common Subsumers in Description Logics', in *Proc. of AAAI'92*, pp. 754–761. AAAI Press, (1992).
[2] S. Colucci, F. M. Donini, and E. Di Sciascio, 'Common Subsumers in RDF', in *Proc. of AI*IA 2013*. Springer, (2013).
[3] J. Lehmann and L. Bühmann, 'AutoSPARQL: Let Users Query Your Knowledge Base', in *The Semantic Web: Research and Applications*, volume 6643 of *LNCS*, 63–79, Springer, (2011).
[4] P. F. Patel-Schneider, 'Reasoning in RDFS is Inherently Serial, At Least in The Worst Case', in *Proc. of ISWC'12 (Demos & Posters)*, volume 914 of *CEUR WP*, (2012).
[5] R. Penaloza and A.-Y. Turhan, 'A practical approach for computing generalization inferences in $\mathcal{EL}$', in *Proc. of ESWC 2011*, pp. 410–423. Springer, (2011).
[6] J. Pérez, M. Arenas, and C. Gutierrez, 'Semantics and complexity of SPARQL', *ACM Trans. Database Syst.*, **34**(3), (2009).
[7] N. Shadbolt, W. Hall, and T. Berners-Lee, 'The Semantic Web Revisited', *Intelligent Systems, IEEE*, **21**(3), (2006).