Finding Commonalities in Linked Open Data

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

[1] DEI, Politecnico di Bari, Bari, Italy
[2] DISUCOM, Università della Tuscia, Viterbo, Italy

Abstract. The availability of a data source as huge, open, accessible and machine–understandable as the Web of Data asks for new and sophisticated inferences to be implemented in order to deeply exploit such a rich informative content. Towards this direction, the paper proposes an approach for inferring clusters in collections of **RDF** resources on the basis of the features shared by their descriptions. The approach grounds on an algorithm for Common Subsumers computation proposed in a previous work of some of the authors. The clustering service introduced here returns not only different cluster proposals for a given collection, but also a description of the informative content shared by the **RDF** resources within the clusters, in terms of (generalized) **RDF** triples.

1 Introduction

The Web of Data [7], born as a research challenge supporting the Semantic Web [1] initiative, is nowadays a fact, as testified by the huge amount of data available in machine-understandable and inter-operable formats, like the Resource Description Framework[3] (**RDF**). 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. Once such an open, continuously enriched and up to date data-source is at hand and accessible, a significantly rich informative content becomes available, opening the way to new challenges to be addressed through reasoning.

The approach proposed in this paper aims at finding commonalities in LOD by exploiting a specifically developed reasoning service, Common Subsumer (CS) of pairs of **RDF** resources [3], 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) [2]. Differently from LCS, CS computation gives up subsumption minimality and searches for knowledge pieces which may be inferred by both **RDF** input resources. In this paper we show how such information, although not subsumption-minimal, is still useful to deduce descriptions of clusters of **RDF** resources in

[3] `http://www.w3.org/RDF/`
[4] `http://linkeddata.org/`

a knowledge domain. In particular, we chose LOD by Chamber of Deputies of Italian Parliament[5] as case study.

The paper is organized as follows: in the next section we describe the main features of the proposed approach, together with some details on its implementation. In Section 3, some preliminary results are shown, before closing the paper.

2 The Approach

The approach we propose here aims at automatically clustering a target collection of **RDF** resources according to a fully semantic-based classification. In particular, **RDF** descriptions are investigated to infer 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 ([6], [8]).

We propose to solve clustering through a deductive and fully semantic-based approach, which relies on the iteration of the following two steps: i) the CS of two randomly selected **RDF** resources (in the following referred as *seed-resources*) is computed; ii) the rest of the target collection is queried in order to find other items entailing the same CS. The sub-collection made up by the two initial resources and those returned by step ii) is one cluster of the collection. Therefore, it is subtracted from the initial target collection, and the two steps are iterated until there are no more resources to be clustered.

An anytime algorithm to compute a CS of pairs of **RDF** resources has been proposed in [3] by some of the authors. In order to ensure correctness and computability, the CS computation refers 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. In a nutshell, the algorithm for CS computation starts by computing the r-graphs corresponding to the input resources t and s, and returns their CS as a pair $\langle x, T\rangle$, made up by a blank node x (*i.e.,* the CS of t and s itself) and a set of (generalized) **RDF** triples T, entailed by the r-graphs of both input resources[7].

Then, the set T of triples is used to model a SPARQL [5] 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.

In order to exemplify our clustering approach, we adopt the LOD by Chamber of Deputies of Italian Parliament as use case. Such an informative source

[5] `http://dati.camera.it/data/en/`

[6] In Machine Learning vocabulary, what we do is called *unsupervised learning*

[7] Due to space limits, the CS extraction algorithm [3] is only sketched through an example in the sequel.

is organized in about thirty different interlinked RDF datasets[8] (last update on the 5th November, 2012), accessible through a public SPARQL endpoint[9]. Each dataset contains the metadata describing a resource (by properties `dc:date`, `dc:description`, `dc:title`, and `rdfs:label`), and the statements about possible relations between that resource and other domain-related or web ones. For the current experimental evaluation, we cluster only resources contained in the dataset `deputato.rdf`, even though their descriptions span multiple datasets.

In Figure 1, the reader may find two example r-graphs describing a pair of resources (`ocd:d3140_10` and `ocd:d270_10`) in our reference dataset. In the most general case, all triples of datasets of interest having the seed-resource as subject are considered relevant for the description of the resource itself. Here, we adopt a more restrictive strategy driven by specific knowledge of the domain, and discard as not relevant for the description of a resource r also triples $<<r\ p\ o>>$ 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`$\}$.

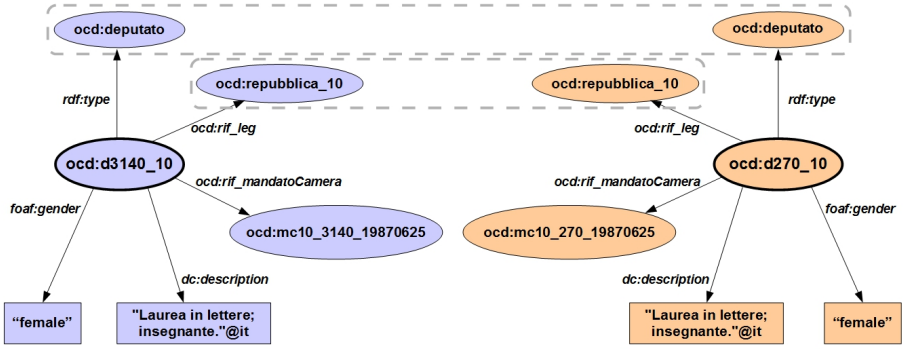

Fig. 1. Two possible r-graphs for **RDF** resources `ocd:d3140_10` and `ocd:d270_10` corresponding, respectively, to deputies Nilde Iotti and Tina Anselmi of the 10^{th} legislature of the Italian Republic. For the sake of clarity, resources `ocd:deputato` and `ocd:repubblica_10`, common to both r-graphs, are depicted as distinct nodes and surrounded by a smoothed dashed-line rectangle

Figure 2 shows a CS of resources `ocd:d3140_10` and `ocd:d270_10` whose two possible r-graphs are those in Figure 1.

[8] `http://data.camera.it/data/en/datasets/`
[9] `http://dati.camera.it/sparql`

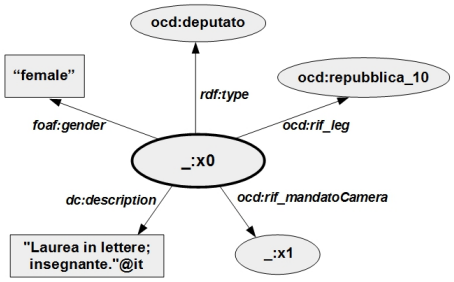

Fig. 2. A graphical representation of a CS $\langle x0, T \rangle$ of the **RDF** resources in Fig. 1.

3 Results

In Table 1, we report a clustering proposal for all deputies of the 10th legislature, where resources `ocd:d3140_10` and `ocd:d270_10` of the above example[10] were forcibly selected as seed pair for the extraction of the first cluster (P_1). By looking at the first row, one can notice how no other resource in the collection shares the features of the CS originated by the first seed-pair, *i.e.*, $|P_1| = 2$. All subsequent clusters have been extracted with a random selection of seed's URIs.

Table 1. Clustering results adopting `ocd:d3140_10` and `ocd:d270_10` as first seed pair.

#P	Seed's URIs	ocd:rif_mandatoCamera	ocd:membro	ocd:aderisce	foaf:gender	dc:description	ocd:rif_ufficioParlamentare	$\|P\|$
P_1	($d3140\_10, d270\_10$)	_:x1	_:x2	_:x3	"female"	"Laurea in lettere; insegnante."@it		2
P_2	($d200023\_10, d22710\_10$)	_:x1	_:x2	_:x3	"female"			81
P_3	($d30010\_10, d17060\_10$)	_:x1	_:x2	_:x3	"male"	"Laurea in giurisprudenza; avvocato"@it		44
P_4	($d20910\_10, d30570\_10$)	_:x1	_:x2	_:x3	"male"		_:x4	148
P_5	($d30140\_10, d60499\_10$)	_:x1	_:x2	_:x3	"male"			398
P_6	($d24780\_10, d31040\_10$)	_:x1		_:x2	"male"			7

[10] Notice that, although the resources are the same as in the above example, we here adopt a different criterion for selecting relevant triples.

Table 2. A clustering result adopting randomly selected initial seed pairs.

#P	Seed's URIs	ocd:rif_mandatoCamera	ocd:membro	ocd:aderisce	foaf:gender	dc:description	$\|P\|$
P_1	($d19990\_1, d20060\_1$)	_:x1	_:x2	_:x3	"male"	"Laurea in giurisprudenza; avvocato."@it	127
P_2	($d3140\_1, d14290\_1$)	_:x1	_:x2	_:x3	"female"	"Laurea in lettere; insegnante."@it	9
P_3	($d12560\_1, d13120\_1$)	_:x1	_:x2	_:x3	"male"	_:x4	431
P_4	($d26000\_1, d10090\_1$)	_:x1	_:x2	_:x3	"female"	_:x5	35
P_5	($d10800\_1, d25610\_1$)	_:x1	_:x2	_:x3	"male"		9
P_6	($d12140\_1, d8520\_1$)	_:x1		_:x2	_:x3		2

Table 2, instead, shows a clustering result obtained for the target collection R of 613 resources corresponding to deputies of the first legislature of the Italian Republic. It means that every pair of seed's URIs (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 (for improving readability, these triples are not reported in the table).

The six listed partitions have been obtained in 14.138 s. By looking at the first row as an example, one can notice how the algorithm aggregates all resources 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.

4 Conclusion

This paper proposed a new and 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 the CS [3] of pairs of resources used as seed. In order for such a computation to be finite, we select a relevant portion of the Web of Data to describe the seed resources, according to a characteristic function to be determined on the basis of domain-dependent criteria.

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.

Part of the future work will be devoted to the extension of CS definition and computation to other entailment regimes, to the investigation on (as general as possible) criteria for the selection of relevant triples, aimed at the combined optimization of both description expressiveness and computational complexity, and to a comparative experimental evaluation involving the definition of a metric for clusters quality assessment.

Acknoledgements

We acknowledge support of projects "A Knowledge based Holistic Integrated Research Approach" (KHIRA - PON 02_00563_3446857) and "Enhance Risk Management through Extended Sensors" (ERMES - PON01_03113/F3).

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American 248(4) (2001), (34-43)
2. Cohen, W., Borgida, A., Hirsh, H.: Computing Least Common Subsumers in Description Logics. In: Rosenbloom, P., Szolovits, P. (eds.) Proc. of AAAI'92. pp. 754–761. AAAI Press (1992)
3. Colucci, S., Donini, F.M., Di Sciascio, E.: Common Subsumers in RDF. In: Proc. of AI*IA 2013. LNAI, Springer (2013)
4. Patel-Schneider, P.F.: Reasoning in RDFS is Inherently Serial, At Least in The Worst Case. In: Glimm, B., Huynh, D. (eds.) Proc. of ISWC'12 (Demos & Posters). CEUR Workshop Proceedings, vol. 914. CEUR-WS.org (2012)
5. Pérez, J., Arenas, M., Gutierrez, C.: Semantics and complexity of SPARQL. ACM Trans. Database Syst. 34(3), 16:1–16:45 (Sep 2009)
6. Qi, L., Lin, H.T., Honavar, V.: Clustering remote RDF data using SPARQL update queries. In: Data Engineering Workshops (ICDEW), 2013 IEEE 29th International Conference on. pp. 236–242. IEEE (2013)
7. Shadbolt, N., Hall, W., Berners-Lee, T.: The Semantic Web Revisited. Intelligent Systems, IEEE 21(3), 96–101 (2006)
8. Zhang, X., Zhao, C., Wang, P., Zhou, F.: Mining link patterns in Linked Data. In: Web-Age Information Management, pp. 83–94. Springer (2012)