The original publication is available at www.springerlink.com.
Riccardo Albertoni and Monica De Martino **Semantic Similarity and Selection of Resources Published According to Linked Data Best Practice**, On the Move to Meaningful Internet Systems: OTM 2010 Workshops

Book Series Lecture Notes in Computer Science
Publisher Springer Berlin / Heidelberg
ISSN 0302-9743
ISBN-10 3-642-16960-0 Springer Berlin Heidelberg NewYork
ISBN-13 978-3-642-16960-1 Springer Berlin Heidelberg NewYork
Volume 6428/2010
DOI 10.1007/978-3-642-16961-8_58
Pages 378-383

Semantic Similarity and Selection of Resources Published According to Linked Data Best Practice

Riccardo Albertoni, Monica De Martino

CNR-IMATI,
Via De Marini, 6 – Torre di Francia - 16149 Genova, Italy
{albertoni,demartino}@ge.imati.cnr.it

Abstract: The position paper aims at discussing the potential of exploiting linked data best practice to provide metadata documenting domain specific resources created through verbose acquisition-processing pipelines. It argues that resource selection, namely the process engaged to choose a set of resources suitable for a given analysis/design purpose, must be supported by a deep comparison of their metadata. The semantic similarity proposed in our previous works is discussed for this purpose and the main issues to make it scale up to the web of data are introduced. Discussed issues contribute beyond the re-engineering of our similarity since they largely apply to every tool which is going to exploit information made available as linked data. A research plan and an exploratory phase facing the presented issues are described remarking the lessons we have learnt so far.

1 Selecting Complex Resources

Effective sharing and reuse of data are still desiderata of many scientific and industrial domains, e.g., environmental monitoring and analysis, medicine and bioinformatics, CAD/CAE virtual product modelling and professional multimedia, where the selection of tailored and high-quality data is a necessary condition to provide successful and competitive services. For example, in the domain of environmental data, many data resources are usually obtained through complex acquisition-processing pipelines, which typically involve distinct specialized fields of competency. Oceanographers, biologists, geologists may provide heterogeneous data resources, which are encoded differently in text, tables, images, 2D and 3D digital terrain models.

Semantic web and in particular the emerging linked data best practice [1] provide a promising framework to encode, publish and share complex metadata of resources in these scientific and industrial domains. Enabling factors for establishing the web of data as preferred selling point for complex resources are: (i) linked data best practice relies on light-weighed ontologies encoded in Resource Description Framework (RDF) which can be exploited to provide ontology driven metadata. Such a kind of metadata takes advantage from the Open Word Assumption, enabling the adoption of

complex, domain specialized and independently developed metadata vocabularies, which are pivotal to document resources produced in complex and loosely coupled pipelines; (ii) linked data best practice relies on content negotiation exploiting the standard HTTP protocol, it is not proposing a brand new platform replacing the existing technologies. Rather, it can be placed side by side to domain specific protocol and standards (e.g., Open Geospatial Consortium specification for the geographic domain) making metadata available in human and machine consumable format; (iii) technological headways have brought to mature prototypes in order to expose resource as linked data (e.g., D2R and Pubby), to query them by appropriate query language (i.e., SPARQL), to retrieve their pertaining RDF fragments published around the web (e.g., Sindice), to reason, store and manipulate these fragments once there are retrieved (e.g., JENA API).

However, even supposing the linked data was massively adopted to share the metadata of complex resources, the **selection** of the most suitable datasets for complex domains like environmental analysis would still be an enervating task. A huge amount of resource features and their complex relations must be considered during the selection process.

Especially for assisting in this process, semantic similarity algorithms supporting a deep comparison of resource features are pivotal. The term "semantic similarity" has been used with different meanings in the literature. It sometimes refers to *ontology alignment,* where it enables the matching of distinct ontologies by comparing the names of the classes, attributes, relations, and instances [2]. Semantic similarity can also refer to *concept similarity* where it assesses the similarity among terms by considering their distinguishing features [3, 4]; their encoding in lexicographic databases [5,6,7,8]; their encoding in conceptual spaces [9].

In this position paper, however, semantic similarity is meant as *instance similarity* since this similarity is fundamental to support detailed **comparison**, **ranking** and **selection** of multidimensional data through its ontology driven metadata.

Different methods to assess instance similarity have been proposed. Some rely on description logics [10]; some have been applied in the context of web services [11]; some others have been applied to cluster ontology driven metadata [12, 13].

Surprisingly, none of these methods supports recognition in the case of those instances, albeit different, have effectively the same informative content: they lack of an explicit formalization of the role of *context* in the entity comparison, and they fail identifying and measuring if the informative content of one overlaps or is contained in the other. Thus, the similarity results are not easily interpretable in terms of gain and loss the users get adopting a resource in place of another. To address these problems, we have recently proposed an asymmetric and context dependent semantic similarity among ontology instances, which meets the aforementioned requirements. The results are shown to be very promising for fine-grained resource selection when operating on a local repository of resources [14]. Unfortunately, there are still many issues that have to be addressed to scale the instance similarity up to the web of data. In this position paper, we are going to discuss these issues.

2 Identified Issues

As more and more data resources are exposed on the web, semantic similarity should locate data on the fly on the web of data, considering multiple and possibly unknown sources. Extending instance similarity at such a scale forces to redesign the similarity addressing its invariance with respect to metadata varieties, which arise when independent stakeholders provide resources. In particular we have to deal with

(i) *non-authoritative metadata*, namely metadata published by actors who are neither the resource producers nor the owners, as it happens for metadata documenting resources that have been re-elaborated or reviewed by third parties;

(ii) *heterogeneous metadata*, i.e., metadata provided according to different, sometimes interlinked, more often overlapping metadata vocabularies, as it happens when the metadata for a resource are provided by stakeholders with different fields of competency;

(iii) *non-consistently identified metadata*, namely metadata occurring when the same resource has different identifiers in distinct metadata sets.

(iv) *efficiency and computational issue:* in a longer perspective an accurate similarity assessment might result computationally prohibitive as soon as the number of resources discovered and features considered increase.

3 Research Plan and Exploratory Phase

We propose a quite challenging research plan to fit the similarity into the web of data:

(i) *non-authoritative metadata* can be investigated considering how synergies with semantic web indexes (e.g., Sindice [15]) can be used to retrieve non authoritative features;

(ii) *heterogeneous metadata* can be addressed deploying schema and entity level consolidation using both explicit metadata statements and mining implicit equivalences through co-occurring resources annotations;

(iii) *non-consistently identified metadata* could be eased deploying reasoning techniques to be applied to web datasets, e.g., to smush fragments of distributed metadata, or developing specific scripts to interlink resources relying on a-priori knowledge about how datasets have been originated;

(iv) *efficiency and computational problems* can deploy strategies to speed up the assessment of semantic similarity, in particular, solutions based on the cashing of intermediate comparisons and techniques to prune the comparisons according to a specified application context might resolve the less severe cases. Moreover, algorithms for efficient parallelization can be studied, e.g., using the Map Reduce cluster-computing paradigm.

Before engaging in this challenging research plan, we have undertaken an exploratory phase analyzing real web data. The goal is to get a first-hand experience in varieties introduced by data providers publishing metadata. Although publishing metadata according linked data best practice has a huge potential for documenting resources produced in complex pipelines, it is not yet a common practice in the specialized

domains we have mentioned. For this reason, we have been forced to move on a simpler domain considering the scientific publications exposed as linked data by Semantic Web Dog Food-SWDF (http://data.semanticweb.org/) and DBLP in RDF (http://dblp.l3s.de/d2r). We aim at comparing a limited set of researchers considering the number of publications they wrote.

We have set up a first linked data enabled instance similarity redesigning the prototype developed in [14] in order to have a live test bed for experimenting and deepen the aforementioned issues. In particular, we have extended the notion of context making explicit to which namespaces properties belong to, so it is possible to build context considering properties from different RDF schemas. We have also updated the ontology model, which was previously based on Protégé-API, to a more linked data oriented module querying RDF models by SPARQL. Then we have started experimenting the new prototype to assess the semantic similarity among researchers whose metadata are available as linked data.

According to the linked data best practice, researchers are identified by URI, then our similarity prototype compares two researchers considering their URIs (i.e., http://dblp.l3s.de/d2r/resource/authors/giovanni_tummarello and http://dblp.l3s.de/d2r/resource/authors/Renaud_Delbru). The following context is provided to parameterize our instance similarity assessment:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
 [foaf:Person]->{{},{(foaf:made, Count)}}
```

According to this context the more two researchers are related through the foaf:made property to a similar number of entities, the more the researchers are considered similar. This is just a simple example of context, more complex cases can be easily considered as discussed in [14].

During the similarity assessment, researchers' URIs are dereferenced in order to get their authoritative RDF fragments. Researcher publications are either provided by DBLP or semantic web dog food, but dereferencing the DBLP researchers' URIs we get the publications from DBLP and not from semantic web dog food, which is in this case a non-authoritative info w.r.t. DBLP.

A first attempt to face with *Non-authoritative metadata* is then done considering Sindice. Given an URI, Sindice returns a ranked list of RDF fragments published all over the web and containing such a URI. Unfortunately, if you ask for http://dblp.l3s.de/d2r/resource/authors/giovanni_tummarello, Sindice returns just the fragments from DBLP, namely the authoritative fragment that corresponds to such an URI, and all the fragments that can be obtained dereferencing URIs contained in that authoritative fragment. We know SWDF RDF fragment pertaining to Tummarello provides metadata about his publications, but unfortunately it refers to other Tummarello's URIs. So these non authoritative info cannot be exploited during our similarity assessment. **First lesson: *Non-authoritative metadata* and *Non-consistently identified metadata* are tightly inter-related in the real practice. To effectively deal with the former issue often we have to care about the latter issue.**

Considering that we know a priori, semantic web dog food provides researcher's URI in the form http://data.semanticweb.org/person/name-[midlename]-[familyname], we

can add for each SWDF researcher the following owl:sameAs triples on the web to overcame the previous problem at least in this specific example.

```
<http://data.semanticweb.org/person/name-[midlename]-
[familyname]>  owl:sameAs
<http://dblp.l3s.de/d2r/resource/authors/name_[middle-
name]_familyname>
```

Assuming that each URI in the retrieved RDF fragments is dereferenced, we are then able to retrieve the non-authoritative RDF fragments from SWDF. The reasoner of JENA is exploited in the linked data enabled instance similarity to induce the symmetry and the transitivity of owl:sameAs and to exploit coherently the entities that have been already consolidated. In this simple case, the *heterogeneous metadata issue* does not appear, in fact both DBLP and SWDF use FOAF schema. We would have experienced this issue if one of the two datasets had used Dublin Core instead of FOAF. However, we experienced another sort of heterogeneous metadata: triples provided by DBLP relate publications to researchers by foaf:maker and not by its inverse property foaf:made specified in the context. The similarity ignores foaf:made is the inverse of foaf:maker unless that is specified by an ontology schema or a specific rule added a priori. **Second lesson: ontology/schema must be dereferenced as much as entity's URIs to make the semantics of properties exploitable.**
On the other hand, we must be careful dereferencing ontologies\schemas and adding rules otherwise we end up with huge RDF graph making even worst the efficiency and computational problems. Dereferencing schemata and URIs is extremely slow, and it adds to RDF graph plenty of info that is not exploited during the semantic similarity assessment (i.e., info not pertaining to specified context). Some kind of context driven crawling and local caching supporting by persistent RDF models has to be considered. **Third lesson: specific and context driven policies to dereference the URI and retrieve RDF fragments should be deployed in order to ease efficiency and computational problems.**
As soon as fragments are dereferenced, we can compare the researchers' publications. Some publications are provided twice, both by DBLP and SWDF, and of course they are provided with distinct URIs. If similarity considered them as distinct publications it would count twice some of the researchers' publications returning wrong results. **Fourth lesson: *Non-consistently identified metadata* is a recursive problem. Consolidating researcher without consolidating papers brings to wrong similarity results. We must be sure entities and properties in the similarity context have been properly consolidated before applying instance similarity.**

4 Conclusion

In this position paper, we discuss linked data best practice to make available metadata of resources produced throughout a complex pipeline. We claim our asymmetric and context dependent instance similarity as a tool for comparing complex metadata but some issues have to be faced. A research programme dealing with these issues is

drafted and an exploratory phase shows how some new sub-problems came up exploiting linked data even in very simple scenarios. We think relying on real data provided by third parties is pivotal in order to learn more about the metadata varieties. That time consuming practice is inspiring to make linked data consuming tools work effectively and to fully demonstrate the linked data potential in everyday business practices.

Acknowledgement. Part of the activity described in the paper has been carried out within the CNR Short Mobility programme granted to Riccardo Albertoni in 2009. We thank Dr. Bianca Falcidieno and Dr. Giovanni Tummarello for their precious suggestions, Dr. Renaud Delbru and Dr. Michael Hausenblas for their support during the short visit at DERI.

References

1. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. Int. J. Semantic Web Inf. Syst. 5(3): 1-22 (2009)
2. Euzenat, J., Shvaiko, P.: Ontology Matching, Springer Verlag, (2007).
3. Rodríguez, M. A., Egenhofer, M. J.: Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure. IJGIS 18(3): 229-256 (2004).
4. Janowicz, K., Keßler, C., Schwarz, M., Wilkes, M., Panov, I. Espeter, M. Bäumer, B.: Algorithm, Implementation and Application of the SIM-DL Similarity Server. GeoS 2007: 128-145, (2007).
5. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and Application of a Metric on Semantic Nets, IEEE Trans. on Systems, Man, and Cybernetics, 19, 1, pp. 17-30, (1989).
6. Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy, Proceedings of the 14th Int. Joint Conf. on Artificial Intelligence (IJCAI), (1995).
7. Lin, D.: An Information-Theoretic Definition of Similarity. Proc. of the Fifteenth InConference on Machine Learning. Morgan Kaufmann, 296-304, (1998) .
8. Pirrò, G.: A semantic similarity metric combining features and intrinsic information content. Data Knowl. Eng. 68(11): 1289-1308 (2009)
9. Schwering, A.: Hybrid Model for Semantic Similarity Measurement. ODBASE-OTM Conferences. LNCS Vol. 3761 Springer-Verlag 1449-1465, (2005).
10. D'Amato, C., Fanizzi, N., Esposito, F.: A dissimilarity measure for ALC concept descriptions. SAC 2006: 1695-1699, (2006).
11. Hau, J., Lee, W., and Darlington, J.: A Semantic Similarity Measure for Semantic Web Services. Web Service Semantics: Towards Dynamic Business Integration, workshop at WWW 05. (2005)
12. Maedche, A. and Zacharias, V.: Clustering Ontology Based Metadata in the Semantic Web. PKDD 2002. LNAI Vol. 2431 Springer-Verlag 348-360, (2002)
13. Grimnes, G. A. Edwards, P. Preece, A. D.: Instance Based Clustering of Semantic Web Resources. ESWC 2008: 303-317 (2008)
14. Albertoni, R., De Martino, M: Asymmetric and Context-Dependent Semantic Similarity among Ontology Instances. In Journal on Data Semantics X, LNCS 4900:1-30, (2008).
15. Oren, E., Delbru, R., Catasta, M., Cyganiak, R., Stenzhorn, H., Tummarello, G.: Sindice. com: A document-oriented lookup index for open linked data, International Journal of Metadata, Semantics and Ontologies, 3 (1), Inderscience 37--52, (2008)