Extracting Usage Patterns of Ontologies on the Web: a Case Study on GoodRelations Vocabulary in RDFa

Ewa Kowalczuk[1], Jedrzej Potoniec[1], and Agnieszka Ławrynowicz[1]

Institute of Computing Science, Poznań University of Technology, Poland

Abstract. The number of publicly available resources that re-use terms from various OWL ontologies has increased massively over last years, with the presence of Linked Open Data datasets and the growing number of websites that embed now structured data into HTML pages using markup languages such as RDFa, microdata and microformats. In this paper, we describe an approach to exploratory analysis of ontology usage patterns on the Web. We have conducted a case study on usage patterns extraction of GoodRelations ontology vocabulary from an RDFa dataset of the Web Data Commons corpus. For this purpose, we designed and ran experiments using a recently proposed pattern mining method for RDF(s) data: Fr-ONT-Qu. Rather than simple statistics or frequent term co-occurrences, we were able to discover more complex usage patterns of structured form of graph patterns, which express how GoodRelations vocabulary is used.

1 Introduction

A number of ontology development methodologies have been proposed such as METHONTOLOGY [1], NeON [2] or DiDOn [3]. Initially, they focused on creating a single ontology from scratch. Subsequently proposed methodologies promoted the use of available resources to create ontology networks by re-using existing ontologies, vocabularies, and design patterns [2, 4]. Now, the number of publicly available resources that re-use terms from OWL ontologies has increased massively, with the presence of Linked Open Data (LOD) datasets [5] and the growing number of websites that embed structured data into HTML pages using markup languages such as RDFa or microdata. Engineering Linked Data (LD) ontologies and vocabularies, and more generally LOD, is thus an urgent research problem. Despite existing studies in this direction [6, 7], far more work is needed on the topic of how to effectively use ontologies in LD and on the Web (cf. [8–10]).

This work deals with exploratory analysis of data published on the Web, where vocabulary from OWL ontologies is re-used. Our approach, using recently proposed pattern mining method Fr-ONT-Qu [11], aims not merely at computing simply statistics of an OWL ontology vocabulary re-use but for computing structured usage patterns of such vocabulary in RDF data. This allows us to study both: *which* vocabulary is used, and *how* it is used, i.e. the study of emerging design patterns and the vocabulary that instantiates them in practice.

We describe a case study on usage pattern extraction of the GoodRelations ontology vocabulary [12] (for product, price, and company data) from an RDFa dataset of the Web Data Commons corpus[1] consisting of over 2.6 billion quads. We designed and ran experiments using Fr-ONT-Qu aiming at finding more complex patterns of structured form of graph patterns rather than simple statistics or frequent term co-occurrences.

[1] http://webdatacommons.org

2 Related work

An early work that studied the usage of vocabulary on the Web of Data is described in [13]. It characterised structural properties and distributions of the raw data of over 1.5 million RDF Web documents with terms mainly from FOAF[2] and Dublin Core[3]. A recent survey [6] with 79 participants studied the most preferred vocabulary reuse strategies in LOD. In[14] the reuse of ontologies in LOD is analysed. Some other studies on LOD datasets, e.g. [15], were focused on investigating their conformance with best practices. Finally, the Linked Open Vocabulary index (LOV)[4] provides the information on most popular vocabularies.

Another line of related works deals particularly with syntactic properties of OWL ontologies on the Web. In [16] syntactic regularities in ontologies were studied. In [17] OWL DL restriction violations were investigated. The study presented in [18] described the extracted statistics related to the frequency of occurrences of OWL language constructs and the structure of ontology class hierarchies in the studied ontologies. Glimm et al. [19] analysed the uptake of OWL in Linked Data, concluding that the OWL fragment that is actually used on the Web of Data is likely a simplified profile based on OWL RL that was coined OWL LD.

3 Usage Pattern Mining

The workflow of our approach to usage pattern mining over high volume RDF data is depicted in Fig. 1. The input data (in our case RDFa dataset from Web Data Commons corpus) is loaded in chunks to several RDF repositories. Subsequently, *Recursive Concise Bounded Descriptions (Recursive CBD)* of objects belonging to chosen classes in the analysed dataset are calculated. The fragment of the dataset extracted via the Recursive CBD is loaded into a final repository over which a pattern mining algorithm Fr-ONT-Qu is run. Extraction of the fragment allows us to efficiently find patterns only from the area of interest and to list classes and properties as candidates for building blocks of SPARQL patterns (as a part of the declarative bias of Fr-ONT-Qu, described in Sect. 3.2). Next, we describe the notion of the Recursive CBD and Fr-ONT-Qu.

3.1 Recursive Concise Bounded Description

The notion of recursive CBD is straightforward and not entirely new, it was for example used in [20], though not explicitly called this way (the authors analyse the benefits of *increased property depth*). Our definition extends the definition of (asymmetric) CBD [21].

Definition 1 (Recursive Concise Bounded Description). *Recursive Concise Bounded Description of a chosen starting node of depth n is a subgraph of defined RDF graph calculated as follows:*

1. *Add to subgraph the CBD of starting node.*
2. *For each statement added in the previous step, add to the subgraph the CBD of statement object (this includes* `rdf:Statements`*, i.e. one needs to add the CBD for objects of triples with property* `rdf:object`*.)*
3. *Repeat step 2. $n-2$ more times.*

[2] http://www.foaf-project.org/

[3] http://dublincore.org/

[4] http://lov.okfn.org/dataset/lov/

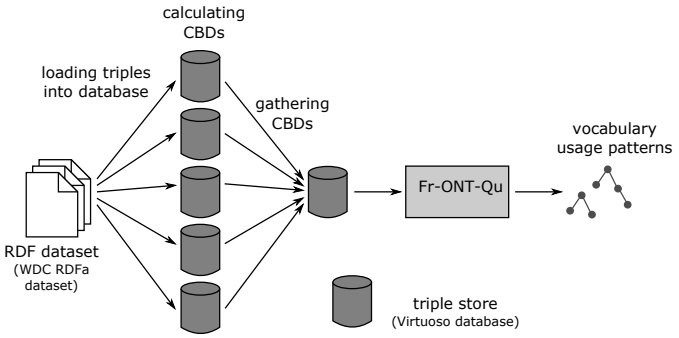

Fig. 1. Pattern calculation workflow.

3.2 Fr-ONT-Qu algorithm for pattern mining

The purpose of the Fr-ONT-Qu algorithm, described in details in [11], is to discover patterns in a given RDF graph. Each of these patterns is a SPARQL query with a single variable denoted in the query head. Every other variable occurring in the query is connected to the denoted variable by a property or chain of properties.

Input of the algorihtm is a declarative bias limiting search space (i.e. classes and properties to use), number of iterations (i.e. maximal number of triple patterns and filter expressions in a pattern – d), pattern quality measure (e.g. number of objects covered by a pattern), number of refinement iterations and number k of patterns selected in every iteration for futher expansion. Below, a sketch of the Fr-ONT-Qu algorithm with a small example (declarative bias containing classes `PassengerTrain`, `CargoTrain` and property `hasEngine`, $k = 2$) is presented:

1. For every pattern from the previous iteration, specialize it by adding a single constraint for a variable already existing in the query. E.g. consider pattern $\{$`?x a :Train.`$\}$, generated specialisations are 1) $\{$`?x a :Train, :PassengerTrain.`$\}$, 2) $\{$`?x a :Train, :CargoTrain`$\}$, 3) $\{$`?x a :Train; :hasEngine ?y`$\}$.
2. Measure quality of generated specialisations and remove all except k best ones. E.g. computed quality for 1) is 3, for 2) is 5 and for 3) is 7 and only 2 best patterns go to the next iteration, i.e. 3) and 2).
3. If the number of iterations does not exceed maximal number d, go to the step 1.

4 Case Study: GoodRelations Vocabulary in Web Data Commons

4.1 Materials

The experiment input data was RDFa dataset from the Web Data Commons [22] November 2013 crawl. The Web Data Commons RDFa dataset contains 2,636,964,693 triples [23]. Due to the large volume of data it was not possible to load it into a single triple store and extract CBDs efficiently. Instead, we divided the data into 5 chunks that were loaded into separate Virtuoso7 database instances (we found 550 million triples being a limit after the crossing of which the loading time increased significantly).

We extracted the fragment pertaining to the GoodRelations namespace by calculating the Recursive CBD of objects belonging to one of the most prominently used[5] classes: `gr:BusinessEntity` and `gr:Offering`. These complementary classes are used to describe two most important notions in the commercial world: an offer maker, such as company or shop, and the offer itself, such as product or service offered under certain conditions. The Recursive CBDs trees rooted in objects of these classes brought other notions from the GoodRelations vocabulary.

To facilitate data loading we also removed the content of all string-related literals, as we were not considering literals in our pattern extraction. One database instance was running at a time, with 48GB RAM available for its exclusive use. One chunk of data took approximately 4h to load, and it took on average 37min and 102min to calculate CBDs for `gr:BusinessEntity` and `gr:Offering` objects respectively. The results (making 9,964,299 triples in total) were subsequently gathered in a single database that was queried by Fr-ONT-Qu.

We run two pattern extraction processes. One included GoodRelations and OWL-related vocabulary (from namespaces `owl:`, `rdfs:` and `owl:`) with d=5 and k=20. The second one included all popular notions (classes with more than 200 and properties with more than 100 occurrences) present in the analysed fragment, with d=3 and k=30. The quality measure used was support on knowledge base (number of distinct values bound to the variable `?x`) minus penalty for pattern length (precisely number of triple patterns in a pattern divided by 100). The penalty factor was added to promote shorter patterns over longer ones. The first process took 46min to run, and the second 467min.

4.2 Results

Table 1. presents a selection of interesting patterns discovered, together with the number of occurrences. Complete results are publicly available[6]. We also include a summary of `owl:`, `rdfs:` and `rdf:` most frequent vocabulary occurences.

Looking at such patterns as 1 and 2, that describe offers either having a price specification that has eligible quantity or a price specification that applies to a particular delivery method, one may notice that they express the relations that are not expressible by the simple frequent itemset representation of term co-occurence.

Many of the computed patterns demonstrate integration of GoodRelations vocabulary with other prevalent namespaces, like `vcard:` for defininig location of businesses and `foaf:` for linking to product depictions (patterns 5 and 7). This seamless integration is possible due to the fact that RDFa allows to use different vocabularies at the same time. Data publishers can even use competitive e-commerce vocabularies at the same time, in order to ensure extensive compatibility. However, as we can see in pattern 6, this is often done in arbitrary manner leading to incoherent typing. In this case price spefication is defined to be a `dv:Product`[7]. Similar problem can be observed for pattern 5[8]. The reason for such inaccuracies might be misunderstanding of RDF data

[5] They are the two most popular (counting per number of pay-level domains) GoodRelations classes in analysed dataset [23].

[6] http://semantic.cs.put.poznan.pl/%7Eekowalczuk/OWLandGR/

[7] In Data-Vocabulary the class characterized by price is `dv:Offer`: http://rdf.data-vocabulary.org/rdf.xml

[8] In this case `?d` would have to be both `foaf:Image` and `gr:Offering`: http://xmlns.com/foaf/spec/#term%5Fdepiction.

Table 1. Selected discovered patterns.

	Pattern	No. of distinct ?x
1.	`select distinct ?x where { ?x gr:hasPriceSpecification> ?s . ?x gr:validThrough ?v . ?s hasEligibleQuantity ?q }`	35,521
2.	`select distinct ?x where { ?x gr:hasPriceSpecification ?s . ?s gr:appliesToDeliveryMethod ?d }`	390
3.	`select distinct ?x where { ?x gr:typeOfGood ?g . ?g rdf:type ?t . ?t a owl:Class }`	2,012
4.	`select distinct ?x where { ?x rdf:type ?d . ?d rdfs:subClassOf ?c }`	2,233
5.	`select distinct ?x where { ?x gr:hasBusinessFunction ?f . ?x foaf:depiction ?d . ?d gr:acceptedPaymentMethods ?m }`	5,480
6.	`select distinct ?x where { ?x gr:acceptedPaymentMethods ?m . ?x gr:hasPriceSpecification ?s . ?s a dv:Product }`	14,115
7.	`select distinct ?x where { ?x a gr:BusinessEntity . ?x vcard:adr ?a }`	36,031
8.	`select distinct ?x where { ?x gr:hasBusinessFunction ?f . ?f rdfs:isDefinedBy ?o . ?o a owl:Ontology . ?o owl:imports ?i . ?o rdfs:seeAlso ?s }`	243,342

model behind RDFa-annotated tags, especially when extensively nested. These issues might lead to the confusion of search tools, highly undesired by the publishers.

One of the remedies to the described problem might be the increase in usage of schema-related information (and it subsequent verification). The RDFa websites recognise the fact that they are ontologies as examplified by pattern 8. They also import ontologies (such as GoodRelations) and declare some of the features used to be defined by them (using `rdfs:isDefinedBy`). They sometimes use OWL-related features for creating hierarchy of categories (pattern 3 and 4) but more advanced features are very uncommon. The hope for increase in schema-related data lies in the fact that `owl:` basic features are as easily integrated in RDFa as any other common vocabularies. If there exists a clear commercial benefit for establishing an expressive schema layer on the top of the product data, such as being understood by advanced semantic information driven search tools, e-commerce owners will certainly do that, as they undoubtedly seized the opportunities RDFa gave them.

5 Conclusions

In our work, we demonstrated how Fr-ONT-Qu algorithm can be used to discover common usage patterns of combined vocabularies (GoodRelations, OWL and other). We also described how the domain of an interest can be extracted from large, heterogeneous volume of RDF data using the Recursive Concise Bounded Descriptions. Through the analysis of the extracted patterns we found which vocabularies are commonly used together with GoodRelations data. We analysed the current utilisation of OWL-related features and discussed possible benefits from their extended usage in RDFa.

Acknowledgments Agnieszka Lawrynowicz and Jedrzej Potoniec acknowledge the support from the PARENT-BRIDGE programme of Foundation for Polish Science, cofinanced from European Union, Regional Development Fund (Grant No POMOST/2013-7/8).

References

1. Fernandez, M., Gomez-Perez, A., Pazos, A., Pazos, J.: Building a Chemical Ontology using METHONTOLOGY and the Ontology Design Environment. IEEE Intelligent Systems **14**(1) (1999) 37–46
2. Suárez-Figueroa, M.C., Gómez-Pérez, A., Fernández-López, M.: The NeOn Methodology for Ontology Engineering. In: Ontology Engineering in a Networked World. (2012) 9–34
3. Keet, C.M.: Transforming Semi-structured Life Science Diagrams into Meaningful Domain Ontologies with DiDOn. Journal of Biomedical Informatics **45** (2012) 482–494
4. Presutti, V., Blomqvist, E., Daga, E., Gangemi, A.: Pattern-Based Ontology Design. In: Ontology Engineering in a Networked World. (2012) 35–64
5. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. Int. J. Semantic Web Inf. Syst. **5**(3) (2009) 1–22
6. Schaible, J., Gottron, T., Scherp, A.: Survey on common strategies of vocabulary reuse in linked open data modeling. In: ESWC. (2014) 457–472
7. Poveda-Villalón, M.: A reuse-based lightweight method for developing linked data ontologies and vocabularies. In: ESWC. (2012) 833–837
8. Jain, P., Hitzler, P., Yeh, P.Z., Verma, K., Sheth, A.P.: Linked Data Is Merely More Data. In: AAAI Spring Symposium: Linked Data Meets Artificial Intelligence, AAAI (2010)
9. Poveda-Villalón, M., Vatant, B., Suárez-Figueroa, M.C., Gómez-Pérez, A.: Detecting Good Practices and Pitfalls when Publishing Vocabularies on the Web. In: WOP, CEUR (2013)
10. Janowicz, K., Hitzler, P., Adams, B., Kolas, D., Vardeman, C.: Five stars of Linked Data vocabulary use. Semantic Web **5**(3) (2014) 173–176
11. Ławrynowicz, A., Potoniec, J.: Pattern Based Feature Construction in Semantic Data Mining. IJSWIS **10**(1) (2014)
12. Hepp, M.: GoodRelations: An Ontology for Describing Products and Services Offers on the Web. In: EKAW. (2008) 329–346
13. Ding, L., Finin, T.: Characterizing the semantic web on the web. In: ISWC. (2006) 242–257
14. Poveda Villalón, M., Suárez-Figueroa, M.C., Gómez-Pérez, A.: The landscape of ontology reuse in linked data. (2012)
15. Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A., Decker, S.: An Empirical Survey of Linked Data Conformance. Web Semant. **14** (July 2012) 14–44
16. Mikroyannidi, E., Stevens, R., Iannone, L.: Tradeoffs in Measuring Entity Similarity for Pattern Detection in OWL Ontologies. In: OWLED. (2013)
17. Bechhofer, S., Volz, R.: Patching syntax in owl ontologies. In: ISWC. (2004) 668–682
18. Wang, T.D., Parsia, B., Hendler, J.A.: A survey of the web ontology landscape. In: ISWC. (2006) 682–694
19. Glimm, B., Hogan, A., Krötzsch, M., Polleres, A.: OWL: Yet to arrive on the Web of Data? In: LDOW2012. (2012)
20. Hellmann, S., Lehmann, J., Auer, S.: Learning of OWL Class Descriptions on Very Large Knowledge Bases. IJSWIS **5**(2) (2009) 25–48
21. Stickler, P.: CBD - Concise Bounded Description. `http://www.w3.org/Submission/CBD` (2005)
22. Mühleisen, H., Bizer, C.: Web Data Commons - Extracting Structured Data from Two Large Web Corpora. In: LDOW2012. (2012)
23. Bizer, C., Mühleisen, H., Harth, A., Stadtmüller, S., Meusel, R., Schuhmacher, M., Völker, J., Eckert, K., Petrovski, P.: Web Data Commons - RDFa, Microdata, and Microformats Data Sets - November 2013. `http://webdatacommons.org/structureddata/2013-11/stats/stats.html` (2013)