Ontology Search by Categorization Power

Ondřej Zamazal and Vojtěch Svátek

Department of Information and Knowledge Engineering,
University of Economics, W. Churchill Sq.4, 130 67 Prague 3, Czech Republic
{ondrej.zamazal|svatek}@vse.cz

Abstract. The demo paper presents novel functionality of ontology search based on categorization power as extension to the OOSP tool. We explain the approach and introduce two scenarios supported.

1 Introduction and Motivation

When searching for a suitable ontology, knowledge engineers commonly use keyword search over different ontology entities such as classes, properties and individuals.[1] This approach generally yields ontologies that are to some degree relevant to the submitted terms. However, there are also finer-grained cases of search. Ontologies are often used for entity categorization, where instances known to belong to a general class are to be partitioned to its finer subcategories. For example, an individual `John` known to be a `Person` can be further categorized as a `Man`, or as someone who has a wife (via the compound concept expression `hasWife some owl:Thing`), provided the respective entities are available in the ontology. We thus might be interested not only in ontologies having a class named `Person` but specifically in ontologies having `Person` as *focus class*: one for which such (preferably, rich) subcategorization can be expressed using the ontology's entities. The search should then reflect the *categorization power* of the ontology wrt. particular focus classes such as `Person`, expressed in terms of the number of available *categorization options* (COs).

While a formal model underpinning the notion of categorization power is forthcoming, this paper presents, in intuitive terms, preliminary support for the mentioned kind of search as implemented in the extension of the pre-existing 'Online Ontology Set Picker' (OOSP)[2] tool [5]. We demonstrate its functionality on two scenarios of recommending ontologies based on their focus classes:

- **Sc1**: For a given keyword, focus classes are found (by localname) and *lists of their COs* are shown for each ontology where they appear. This functionality supports the collection of explicit *relevance feedback* in order to gather training data for future recognition of 'meanigful' COs.
- **Sc2**: The overall number (focus) classes with *number of COs reaching a certain threshold* is calculated per ontology, together with *other ontology metrics* such as number of classes, taxonomy depth etc.; optionally, the number of COs for a given focus class can be computed as well.

[1] See, e.g., the Watson engine, `http://watson.kmi.open.ac.uk/`.

[2] `http://owl.vse.cz:8080/OOSP/`

In the remainder, Section 2 briefly explains the main notions; Section 3 introduces the new functionality of OOSP; Section 4 describes related work, and, finally, Section 5 wraps up the paper with conclusions and future work.

2 Categorization Power of Focus Classes

Most obvious *categorization options* (COs) in ontologies are named subclasses of the given focus class (FC). However, further COs may be constructed as compound concept expressions. Since the number of all COs under standard OWL expressiveness would often be infinite (e.g., considering arbitrary cardinality or unlimited recursion of properties), we currently restrict the scope of COs under consideration to those built using a few simple, common concept expression patterns, yielding a finite number of COs for every FC. In addition, only some COs are meaningful, reusable categories, which we denote as *ontologistic*[3] *categories* (OCs). While the OC status of COs may sometimes be context-dependent or subjective to some degree, we believe that it is worth accounting for. For example, let us consider[4] an ABox consisting of one class instantiation and three property assertions:

```
Individual: John
     Types: Man             Facts: bornIn UK
     Facts: hasWife Mary    Facts: insuranceCategory Enterpreneur
```

Let us examine the candidate COs for the Person FC built using different structural patterns with respect to being OCs. The named class Man (P1 pattern) is clearly an OC. The compound expression hasWife value Mary, i.e. having Mary (a specific person) as wife, is not an OC due to its low reusability; on the other hand, having a wife in general (hasWife some owl:Thing, P2 pattern) or, possibly, having a wife who is doctor (P3 pattern) is an OC. Being born in the UK (i.e. a specific country, P4 pattern) would be an OC in many contexts; being born in general (P2) should not (since it should be a functional property holding for every instance of Person. Finally, insuranceCategory with a specific value (again, P4 pattern) is clearly an OC; in contrast, insuranceCategory some owl:Thing (P2) may only be an OC under the assumption that being insured is not mandatory in the given context and that the fact of non-insurance is manifested by the absence of any insuranceCategory assertion.

In this preliminary work we attempt to approximate the *categorization power* for focus classes as computed from the ontology TBox (*focused categorization power*, FCP). We selected five variants of CO construction, of which four (1–4) simply follow the patterns P1 to P4, while the additional variant 5 is derived from P4 but the individual *i* is not part of the ontology itself but of an associated SKOS[5] codelist. For each variant we consider a specific formulation of the FCP

[3] While 'ontological category' might read as 'category expressible using an ontology' (possibly with a complex, unintuitive descriptions), 'ontologistic category' refers to presumed plausibility of the CO (as reusable domain concept) for human ontologists.

[4] In Manchester OWL syntax, http://www.w3.org/TR/owl2-manchester-syntax/.

[5] https://www.w3.org/2004/02/skos/

V.(P.)	fcp_{var}(FC)
1 (P1)	$\vert\{C;\ C\ SubClassOf:\ FC\}\vert$
2 (P2)	$\vert\{P;\ P\ Domain: FC\}\vert$
3 (P3)	$\vert\{(P,C);\ \exists D\ \ P\ Domain:\ FC \wedge P\ Range_a\ D\ \wedge\ C\ SubClassOf:\ D\}\vert$
4 (P4)	$\vert\{(P,i);\ \exists C,D\ \ P\ Domain:\ FC \wedge P\ Range_a\ D\ \wedge\ C\ SubClassOf:\ D\ \wedge\ i\ Types:\ C\vert$
5 (P4)	$\vert\{(P,i);\ \exists s\ \ P\ Domain:\ FC\ \wedge\ P\ Range_a\ skos: Concept \wedge$ $\wedge\ P\ Range_a\ value(skos: inScheme, s)\ \wedge\ i\ skos: inScheme\ s\ \wedge\ i\ Types:\ skos: Concept\vert$

Table 1. Detection formulas of COs. *V.* refers to variant and *P.* refers to pattern.

measure at the level of the whole ontology, denoted as fcp_1 to fcp_5, Table 1. Note that these FCP are only (often quite imprecise) approximations of 'ground truth' FCP that would only take into account 'true' OCs approved by some consensus of human ontologists. We also *prune* candidate COs whose ineligibility as OC follows from the ontology structure, e.g., for pattern 2 we prune those properties P that appear in an existential restriction $FC \sqsubseteq \exists P.C$; for such properties the CO would contain *all* instances of the FC.[6] For detection we always use the inferential closure of the ontology, e.g., in order to get inferred domain or inferred subclasses, however, with the exception of the range axiom, which is only considered as asserted (therefore the 'a' subscript) – otherwise not only subclasses of D but also its superclasses would be inferred as range of P. Because we use inference we employed OWLAPI framework instead of using SPARQL.

3 Implementation in OOSP

Currently, ontology search based on focus classes is available for three ontology collections, referred to as 'pools' in OOSP. *LOV*[7] is a well-curated collection of linked open vocabularies used in the Linked Data Cloud. Out of the 529 ontologies (Jan. 2016 snapshot) 1 was not parseable by OWL-API and 19 ontologies were not processable due to unavailable imports. In all, our Jan. 2016 snapshot contains 509 LOV ontologies (96%). *NanJing Vocabulary Repository*[8] (NJVR) is a vocabulary repository extracted from the index of the Falcons search engine.[9] The latest release is from June 2015 and it consists of 1763 vocabularies from which 135 were not parseable by OWL-API or were not processable due to unavailable imports. In all, our Jan. 2016 snapshot contains 1628 vocabularies divided into *Nanjing* vocabularies extracted from single files (1403) and *Nanjing merged* vocabularies extracted from more than one RDF file (225).[10] Finally,

[6] We plan to present a rigorous and evaluated framework for FCP in the future, which will benefit from the usage of this application. For more explanation about FCP and its computation based on variants 1-5 see `http://owl.vse.cz:8080/SumPre2016-FCP.pdf`.

[7] `http://lov.okfn.org/dataset/lov/`

[8] `http://ws.nju.edu.cn/njvr/`

[9] `http://ws.nju.edu.cn/falcons/objectsearch/`

[10] The latter are experimental ontologies which were created as a merge of their definitions spreading over RDF files.

we included the *OntoFarm* ontology collection, which includes 16 small but relatively rich ontologies from the conference organization domain. The collection has previously been used for experiments in Ontology Matching and elsewhere.[11]

Sc1, *Ontology Search Based on Focus Classes*,[12] is supported in a four-step workflow. First, an initial *ontology pool* is selected and a keyword for FC search is provided by the user; the keyword can be searched as a whole localname or as a part of it. Second, the user obtains ontologies divided into two tables, the first providing the ontologies containing FCs according to the input keyword, and the second providing ontologies with classes (non-FCs) matching the keyword but having no COs. For both tables OOSP also provides ontology metrics values. Showing not only FCs but also non-FCs can be of interest especially in single-domain collections. For example, although 12 ontologies from OntoFarm contain the *review* concept, only 5 allow to categorize reviews according to one of the patterns. (The highest number of COs is attributed to the ekaw ontology where there are 23 COs out of which 18 COs is related to variant/pattern 3.) Third, for each ontology for which some FC was found the user can check which FCs were discovered and with which COs. For example, in the *ekaw* ontology reviews can be categorized, in variant 3, either according to the type of paper being reviewed (regular paper, workshop paper etc.) or by the role of person who wrote the review (workshop chair, session chair etc.). For large ontologies with many relevant COs, the user can ask for a random sample of size between 10 and 100 COs. The fourth, optional step consists in providing feedback to the system on which COs are not proper OCs: it is going to be used for training a classifier.

An assumption behind **Sc1** is that if an important term of a domain is in the role of FC in an ontology then this ontology covers an important part of this domain overall. For example, searching in LOV ontology pool for FCs based on the 'university' keyword retrieves three ontologies with FCs and five ontologies with non-FCs. While in the former group all are from the university/education domain, the latter group only contains 60% of ontologies from this domain.

Sc2, *Ontology Search Based on Overall Categorization Power*,[13] is supported in a four-step workflow originally introduced for OOSP in [5]. After the ontology pool selection the user can filter the ontologies not only according to axiom-based, taxonomy-based, and similar ontology metrics, but also according to the number of COs the ontology provides and/or FCs an ontology contains. For example, let us consider the user wants to find ontologies providing a relatively large ABox and at the same time a large proportion of their classes can be categorized. The user sets the 'number of instances' metric to 'at least 100' and the ratio of (any type of) FCs to all classes to 'at least 0.8'. In LOV (01-2016) there are 10 ontologies having *both* these characteristics, while there are 55 ontologies having more than 100 instances and 176 ontologies with FC ratio higher than 0.8.

[11] http://owl.vse.cz:8080/ontofarm/

[12] In OOSP, it is available after 'Go to Ontology Search Based on Focus Classes' option.

[13] In OOSP, after 'Go to Metrics Selection' and an ontology pool selection, it is available as 'Metrics Based On Focus Classes' option on 'Entity' metrics page.

4 Related Work

The categorization power of ontologies has not been, to our knowledge, studied with the flavor presented here. However, our approach can be compared to existing ontology search tools. Users of Watson can search by keywords using a number of parameters: entity types, match level and scope of keywords [1]. Extensive search is provided by the *LOV portal* [4] in terms of metadata, ontology and terms. Besides full-text search it is possible to narrow the search by filtering on term type (class, property, datatype or instance), language, etc. Other approaches focus mainly on term search, e.g. *vocab.cc*[14] [2], LOVR Framework [3]. Since OOSP is primarily ontology-oriented, we do not foresee term-level search, however we plan to provide ontology full-text search similar as in Watson.

5 Conclusions and Future Work

The demo introduces ontology search based on categorization power as an extension to the OOSP tool. The presented scenarios are based on the assumption that discovery of FCs and their categorization power can support ontology search since categorization is an often (implicitly) required feature. Currently we only discover FCs and their categorization power based on the ontology TBox, however, when available, we also plan to process the *ABox*. Property value pairs assigned to different instances can be used to automatic discovery of COs. We will also consider more variants of CO construction in future. While we currently offer ontology search based on *a priori* FCs, we also plan to deal with *a posteriori* FCs, in the sense of clustering the selected ontologies according to their *shared* FCs. Sharing FCs may indicate that the ontologies are close to each other.

Acknowledgement Ondřej Zamazal has been supported by the CSF grant no. 14-14076P.

References

1. d'Aquin M., Sabou M., Motta E., Angeletou S., Gridinoc L., Lopez V., Zablith F. What can be done with the Semantic Web? An Overview of Watson-based Applications. In: 5th Workshop on Semantic Web Applications and Perspectives, SWAP 2008, Rome, Italy.
2. Stadtmüller S. Harth A. Grobelnik M. Accessing information about linked data vocabularies with vocab.cc. In: *Semantic Web and Web Science*. 2013. Springer.
3. Stavrakantonakis I., Fensel A., Fensel D. Linked Open Vocabulary Recommendation Based on Ranking and Linked Open Data. In: Joint International Semantic Technology Conference. 2015. Springer.
4. Vandenbussche P. Y., Atemezing G. A., Poveda-Villalón M., Vatant B. Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the Web. *Semantic Web Journal*. 2015.
5. Zamazal O., Svátek V.: OOSP: Ontological Benchmarks Made on the Fly. In: Workshop SumPre'15 at ESWC 2015.

[14] `http://vocab.cc/`