OReCaP – Towards Ontology Reuse via Focused Categorization Power

Viet Bach Nguyen , Vojtěch Svátek ,
Gollam Rabby, and Ondřej Zamazal

Department of Information and Knowledge Engineering
University of Economics, Prague, Czech Republic
{nguv03,svatek,rabg00,ondrej.zamazal}@vse.cz

Abstract. Focused categorization power (FCP) has been recently introduced as a way of measuring the utility of an ontology by the count of concept expressions expressible using the ontology and subsumed by the given (focus) class/es. OReCaP is an ontology search interface with an integrated ontology ranking method based on the FCP value. The choice of ontologies for reuse is supported by the listing of different types of categories provided by the ontology for a particular focus class.

Keywords: focused categorization power · ontology ranking · ontology reuse · ontology search

1 Motivation

When reusing existing OWL ontologies for publishing a dataset in RDF or developing a new ontology, preference may be given to those providing extensive subcategorization for the classes deemed important in the new dataset schema or ontology (focus classes). The reused set of categories may not only consist of *named classes* but also of some *compound concept expressions* viewed as meaningful categories by the knowledge engineer and possibly later transformed to a named class, too, in a local setting. In our previous work [6] we defined the general notion of *focused categorization power* (FCP) of a given ontology, calculated with respect to a focus class and a particular *concept expression language*, as the (estimated) weighted count of the categories that can be built from the ontology's signature, conform to the language, and are subsumed by the focus class. For the sake of tractable experiments we then formulated and empirically justified a restricted concept expression language based on existential restrictions.

As an example, let us consider the case of ontology reuse for describing the dataset of a used car retailer. For the focus class `Vehicle` in a particular ontology, we can consider its named subclasses such as `Motorcycle`, but also anonymous concept expressions[1] such as `Vehicle and hadAccident some Thing` (vehicle that underwent an accident), `Vehicle and hasSeller some Company` (vehicle sold by a company, not by a person), or `Vehicle and hasFuel value CNG` (vehicle that uses CNG as fuel). These three examples belong each to a different concept expression type; however, all are a part of the 'existential restriction family'. Their Tbox templates are, in turn: $\exists P.\top$, $\exists P.C$, and $\exists P.\{i\}$. Exhaustively enumerating all such expressions that can be constructed from the signature of the given ontology would of course have little relevance. However, we assume that the expressions can be filtered using syntactic patterns over the Tbox axioms; most prominent role is played by the domain/range axioms, in this respect. For example, $\exists P.\top$ is more likely to be a meaningful subcategory for a focus class FC if there is an axiom in the form

$$P \ \texttt{rdfs:domain} \ FC$$

in the (inferential closure of the) ontology. The heuristic patterns for other axioms types [6] are a bit more complex, but still easy to detect in the ontology Tbox.

In this demo paper we present the first operationalization of the notion of FCP in its main target context: *ontology recommendation for dataset description*. Recommendation of ontologies and of individual terms from them have recently been an active field of research. The mainstream approach consists in various kinds of term/ontology popularity computation. For example, Atemezing & Troncy [1] used an information-theoretic approach, and Butt [2] employed a hub-authority graph analysis approach. Stavrakantonakis et al. [5] then combined popularity metrics with the credibility of the vocabulary designers (based on the previously developed ontologies) as an orthogonal feature. Kolbe et al. [3], analogously, measured the academic publication performance of the designers.

We believe that the FCP is yet another relatively orthogonal feature to be considered: while some existing approaches include a similar notion of class 'importance' within the ontology [2], they do not consider compound concepts as 'latent' entities influencing this importance.

The survey on ontology reuse strategies by Schaible et al. [4] indicates that reusing multiple entities from the same vocabulary (even if some of them are by themselves less popular than analogous entities from other vocabularies) is often preferred. This corroborates the relevance of measuring the FCP of ontologies: ontologies providing ample sub-categorization for the 'pillar' concepts of the to-be-published dataset deserve to be adopted in bulk.

2 Tool Description

OReCaP is a web application[2] that aims to demonstrate the calculation of FCP scores for ontologies in the context of an ontology search (for reuse) scenario.

[1] Here written in the human-readable Manchester syntax, see https://www.w3.org/TR/owl2-manchester-syntax/.

[2] Available as demo at https://fcp.vse.cz/orecap.

The interaction starts with a keyword-based search where the input consists of at least one *focused class keyword* and of optional *additional keywords*. The intuition is that the focused class keyword/s denotes the high- or medium-level type/s of entities whose instances are to be further sub-categorized using concepts from the ontology; the additional keywords, on the other hand, correspond to whatever domain terms. Imagine, for example, that the data is currently stored in a relational database. The focused class keyword might then often be the name of the top-level table (which can be, e.g., 'Client', 'Patient', 'Vehicle', 'Account', or the like); the additional keywords can be taken, e.g., from the names of subordinate tables, table columns, or predefined values for the fields.

The search returns a sorted list of ontologies whose classes match one or more of the provided keywords by their IRI, name or description; classes with a match of focused class keyword are listed first. The matched classes are listed for each ontology. Classes that match the *focused class keywords* are preselected (i.e., checked) by default; classes that matches the *additional keywords* are not preselected but can be selected (checked) manually by the user.

The next step is to execute the *FCP calculation* for a chosen ontology, given the selected classes as focus classes, by clicking on the 'Calculate FCP' button. In a pop-up window, metadata about the ontology including its URI and namespace is displayed, along with the total FCP score, which is calculated based on the FCP weight values and the categorizations listed at the bottom. This score is the sum of all partial scores for each focus class. The weight values can be adjusted for each calculated ontology according to the user's assessment of each category type, and the resulting FCP score will change accordingly. The global FCP weights can be changed in the settings section, so that every new FCP calculation would use them as the default weight values. The calculated FCP score is then saved to a comparison list, which shows the FCP-based ranking of the ontologies. Furthermore, the details of the calculations and categorizations can be inspected, where for each focus class, its categories are displayed. There are 4 types of categories considered, conforming to the earlier formulated [6] concept expression language (the FC symbol denotes the focus class):

- *t1*: named classes; specifically, we consider the subclasses of the focus class (C; $C \sqsubseteq FC$)
- *t2*: existential restriction to the top concept ($FC \sqcap \exists P.\top$)
- *t3*: existential restrictions to a named class ($FC \sqcap \exists P.C$)
- *t4*: existential restrictions to a particular individual ($FC \sqcap \exists P.\{i\}$).

OReCaP makes use of the Linked Open Vocabulary API[3] for the keyword-based search and for retrieving the ontology metadata. The FCP calculation itself [6] is implemented on top of OWL API, in combination with the jFact reasoner.[4] OWL API is used to load and parse the ontology source codes, and jFact is used to infer class expressions. Our implementations for this demo are open-source and available on GitHub[5] under the MIT license.

[3] https://lov.linkeddata.es/dataset/lov/api

[4] https://github.com/owlcs/owlapi, https://github.com/owlcs/jfact

[5] https://github.com/nvbach91/orecap, https://github.com/nvbach91/fcp-api

3 Usage Scenario

This scenario addresses sports event data publishing. For the focus class keywords *competition*, *round*, and *match*, and additional keywords *game*, *medal*, *player*, *team*, and *sport*, OReCaP lists the *BBC Sport Ontology* as one of the top matches. The FCP calculation for this particular ontology with 3 selected classes, *sport:Competition*, *sport:Match*, and *sport:Round*, yields, with previously empirically estimated default category type weights [6], a score of 162.00 (of which *sport:Competition* alone assures over 130). The detailed calculation is shown in Table 1. Among the meaningful categories usable for sub-categorizing instances of *sport:Competition* using the ontology, OReCaP lists, e.g., the following ones:

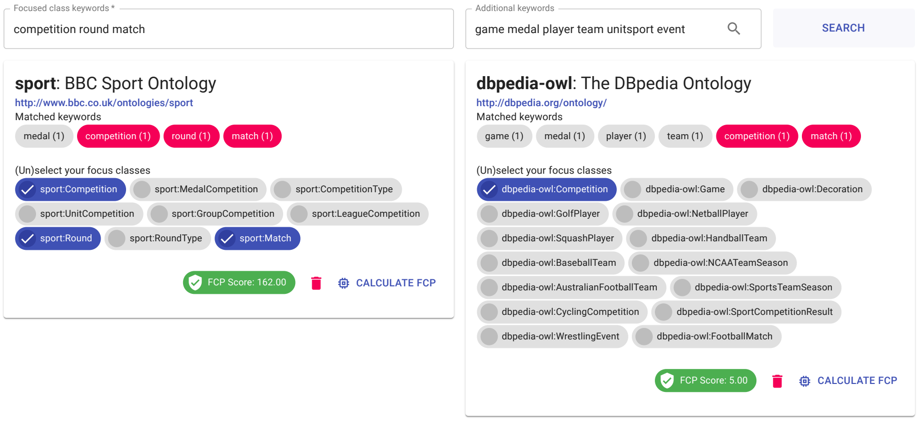
- *sport:GroupCompetition* (*t1*);
- ∃ *sport:promotesTo.owl:Thing* (*t2*) – competitions that lead to a promotion;
- ∃ *sport:lastStage.sport:KnockoutCompetition* (*t3*) – competitions that have a knock-out competition as their the last stage;
- ∃ *sport:eventGender*.{*http://www.bbc.co.uk/things/event-gender/mixed*} (*t4*) – competitions where the gender of competitors is mixed.

Focus class	**Category type**	**Categories**	**Weight**	**Score**
sport:Competition	t1	12	1	12.00
sport:Match	t2	3	0.3	0.90
sport:Competition	t2	39	0.3	11.70
sport:Round	t2	9	0.3	2.70
sport:Match	t3	20	0.5	10.00
sport:Competition	t3	158	0.5	79.00
sport:Round	t3	6	0.5	3.00
sport:Match	t4	2	0.7	1.40
sport:Competition	t4	43	0.7	30.10
sport:Round	t4	16	0.7	11.20
Total FCP score				**162.00**

Table 1: Detailed FCP calculation for BBC Sport Ontology

Of course, not all categories are equally meaningful. To reflect that, the user can adjust the weight values as described in Section 2. At the moment, OReCaP only allows to change the weight values at the level of the whole category types (we also plan to add the option of altering the weight of the individual categories).

Another result retrieved for the same keyword setting is the *The DBpedia Ontology*. Even if it has more (additional) keyword matches, it only matches a single focus class keyword (with *dbpedia-owl:Competition*), and its (default) FCP score is only 5.00. A partial screenshot with these two results is in Fig. 1.



Focused class keywords *: competition round match

Additional keywords: game medal player team unitsport event

sport: BBC Sport Ontology
http://www.bbc.co.uk/ontologies/sport
Matched keywords
medal (1) competition (1) round (1) match (1)

(Un)select your focus classes
☑ sport:Competition ○ sport:MedalCompetition ○ sport:CompetitionType
○ sport:UnitCompetition ○ sport:GroupCompetition ○ sport:LeagueCompetition
☑ sport:Round ○ sport:RoundType ☑ sport:Match

FCP Score: 162.00

dbpedia-owl: The DBpedia Ontology
http://dbpedia.org/ontology/
Matched keywords
game (1) medal (1) player (1) team (1) competition (1) match (1)

(Un)select your focus classes
☑ dbpedia-owl:Competition ○ dbpedia-owl:Game ○ dbpedia-owl:Decoration
○ dbpedia-owl:GolfPlayer ○ dbpedia-owl:NetballPlayer
○ dbpedia-owl:SquashPlayer ○ dbpedia-owl:HandballTeam
○ dbpedia-owl:BaseballTeam ○ dbpedia-owl:NCAATeamSeason
○ dbpedia-owl:AustralianFootballTeam ○ dbpedia-owl:SportsTeamSeason
○ dbpedia-owl:CyclingCompetition ○ dbpedia-owl:SportCompetitionResult
○ dbpedia-owl:WrestlingEvent ○ dbpedia-owl:FootballMatch

FCP Score: 5.00

Fig. 1. Two ontologies found via keyword search, with focus classes and FCP scores

4 Conclusions and Future Work

The notion of FCP is fundamentally novel within the family of content-based ontology recommendation approaches. The current demo is meant to demonstrate its contribution. In the future we plan to compare the results obtained through this measure with those obtained by popularity, credibility, and other existing measures, to see how they could support the dataset publisher in a complementary way, within a coherent methodology. We would also like to perform experiments with the tool in various domains, in order to devise novel heuristics for setting parameters on the onset of the ontology search and reuse sessions.

Acknowledgment

This research is being supported by project no. 18-23964S of the Czech Science Foundation, “*Focused categorization power of web ontologies*”.

References

1. Atemezing, G. A., Troncy, R.: Information Content based Ranking Metric for Linked Open Vocabularies. In: 10th Int. Conf. Semantic Systems, 2014, ACM.
2. Butt, A. S.: Ontology search: Finding the right ontologies on the web. In: WWW 2015, Companion volume.
3. Kolbe, N., Kubler, S., Le Traon, Y.: Popularity-Driven Ontology Ranking Using Qualitative Features. In: ISWC 2019. Springer LNCS 11778.
4. Schaible, J., Gottron, T., Scherp, A.: Survey on Common Strategies of Vocabulary Reuse in Linked Open Data Modeling. In: ESWC 2014, Springer, LNCS 8465.
5. Stavrakantonakis, I., Fensel, A., Fensel, D.: Linked Open Vocabulary Ranking and Terms Discovery. In: SEMANTiCS 2016, ACM.
6. Svátek V., Zamazal O., Vacura M.: Categorization Power of Ontologies with Respect to Focus Classes. In: EKAW 2016, Springer, LNCS 10024.