Fokusovaná kategorizační síla webových ontologií

Vojtěch Svátek[1], Ondřej Zamazal[1], Miroslav Vacura[2]

[1]Katedra informačního a znalostního inženýrství, Vysoká škola ekonomická v Praze,
nám. W. Churchilla 1938/4, 130 67 Praha 3
[2]Katedra filozofie, Vysoká škola ekonomická v Praze,
nám. W. Churchilla 1938/4, 130 67 Praha 3

{svatek,ondrej.zamazal}@vse.cz
vacuram@vse.cz

Abstrakt: Přepoužívání ontologií je jedním z předpokladů efektivního využívání propojených dat na sémantickém webu. Výběr ontologie pro přepoužití bývá v současnosti založen na textovém vyhledávání a metrikách popularity ontologie. Kritériem přepoužitelnosti ontologie však může být i její schopnost vyjádřit podkategorie klíčových tříd objektů z aktuální datové sady, tzv. fokusových tříd. Tuto schopnost navrhujeme kvantifikovat pomocí fokusované kategorizační síly ontologie: míry odvozené z velikosti množiny binárních možností, které model nabízí pro kategorizaci objektů již přiřazených k obecnější fokusové třídě. Přepoužitelnými podkategoriemi mohou být i složené konceptové výrazy, a to s různou mírou jistoty, která se promítá do váhových koeficientů používaných ve výpočtu kategorizační síly. Pravděpodobnost, s jakou je výraz přepoužitelnou kategorií, můžeme odhadovat mj. na základě hodnocení lidskými respondenty.

Klíčová slova: ontologie, sémantický web, propojená data, kategorizace.

1 Úvod

Webové ontologie lze zhruba charakterizovat jako soubory konceptů (neboli tříd) a typů vztahů (vlastností), označených globálními identifikátory IRI a propojených množinou logických tvrzení v ontologickém jazyce OWL[1] odpovídajícímu relativně expresivní variantě deskripční logiky [1]. Ontologie jsou stále více využívány pro přiřazování strojově čitelné sémantiky strukturovaným datům vystavovaným na webu. Motivací je zejména možnost souběžného automatického zpracování nezávisle vzniklých datových zdrojů. Například, pokud jsou výrobky stejného druhu nabízené různými e-shopy sémanticky popsány pomocí stejných tříd a vlastností (souhrnně, entit), lze relativně snadno implementovat jejich automatické porovnávání a doporučování. To ovšem vyžaduje, aby pojmy z různých ontologií byly *přepoužívány*. Vystavovatel datové sady pak musí při tvorbě jejího schématu, namísto izolovaného návrhu nových

[1] https://www.w3.org/TR/owl2-primer/

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 146-150.

entit, vložit úsilí do nalezení relevantních existujících ontologií a jejich entity do schématu zaintegrovat – ať už přímo, nebo pomocí subsumpčního či ekvivalenčního mapování.

Výběr ontologie pro přepoužití je ovšem netriviální úlohou, pro kterou teprve v posledních letech vznikají exaktní metody. Ty vesměs, vedle *textového vyhledávání*, spoléhají na *metriky popularity* (ontologie nebo jednotlivých entit), např. kolik instancí v kolika různých datových sadách se na ně odkazuje, případně na míry apriorní *důvěryhodnosti* ontologie [2], např. zda jsou ontologie zachyceny v autoritativním katalogu, jako je LOV.[2] Pilotní studie zaměřená na strategie přepoužívání webových ontologií [2] naznačila, že vydavatelé datových sad preferují přepoužít větší počet entit z nižšího počtu ontologií i za cenu nižší průměrné míry jejich popularity. Míry založené na popularitě navíc trpí problémem studeného startu: mnoho kvalitních ontologií je nových a jejich potenciál proto nelze odkazovou popularitou spolehlivě hodnotit.

Navrhovaný přístup ke zlepšení procesu přepoužívání ontologií je postaven na následující intuici:

1. Využití ontologií na webu má často charakter přiřazení objektů k určitým kategoriím, s tím, že již před tímto přiřazením je o objektech známo, že jsou instancemi určité (obecnější) třídy, kterou označíme jako *fokusovou třídu*.
2. Přiřazované kategorie nemusí být nutně v ontologii uvedeny jako pojmenované třídy, ale může se jednat o *složené konceptové výrazy* zkonstruované z pojmenovaných entit pomocí operátorů příslušné deskripční logiky.
3. Počet a „kvalita“ kategorií, které ontologie nabízí pro určitou fokusovou třídu, je *indikátorem přepoužitelnosti* ontologie pro datovou sadu obsahující objekty patřící do této třídy.

Téma bylo in extenso zpracováno v příspěvku na evropské konferenci EKAW 2016 [3] (zařazen v nominaci na *Best Paper Award*). V tomto referativním příspěvku pouze nastíníme navrženou metodu a shrneme hlavní dosažené výsledky.

2 Schéma výpočtu kategorizační síly

Uvažujme množinu n typů konceptových výrazů nepřímo vymezenou formálním jazykem $\mathcal{L}$ (nad konstrukty deskripční logiky) a množinu *vzorů* používaných pro jejich detekci v ontologiích, $P = \{p_1, \ldots, p_n\}$. Odhad fokusované kategorizační síly ontologie O vzhledem k fokusové třídě FC pak můžeme vyjádřit jako

$$\widehat{FOCP}(FC,\mathcal{L},O) = Occ(p_1, FC,O) \,.\, w_1 + \ldots + Occ(p_n, FC,O) \,.\, w_n$$

kde $Occ(p_i, FC,O)$ je funkce vracející počet výskytů vzoru p_i v O, a w_i jsou váhové koeficienty z intervalu (0, 1]. Jedním z typů konceptových výrazů je i „pojmenovaná třída“, reprezentující tranzitivní podtřídy třídy FC. Ta je detekována vzorem vyjádřeným ve formě subsumpčního tvrzení C `rdfs:subClassOf` FC a vyhledávaným nad

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

tranzitivním uzávěrem ontologie; *C* je zde proměnná, za kterou se podtřídy dosazují. V případě tohoto vzoru budeme předpokládat hodnotu w_i rovnou 1.

Pro stanovení adekvátních *váhových koeficientů* složených konceptových výrazů se nabízejí dva hlavní zdroje: zpětná vazba od uživatelů, a automatická empirická analýza – jednak samotných ontologií, jednak datových sad, které se na ně odvolávají.

3 Provedené experimenty a jejich výsledky

V první fázi výzkumu byla využívána zejména zpětná vazba od uživatelů k jednotlivým konceptovým výrazům různých typů. Rozlišovány byly tři typy složených výrazů (konkrétní příklady jsou níže), které lze vyjádřit variantami existenční restrikce nad vlastností *R*: $FC \sqcap \exists R.C$ (s hodnotou vlastnosti vymezenou pomocí třídy *C*), $FC \sqcap \exists R.\{i\}$ (tzv. "value restriction" s hodnotou vyjádřenou konkrétní instancí *i*), a $FC \sqcap \exists R.\top$ (s hodnotou vlastnosti "vymezenou" univerzálním "super-konceptem" $\top$, tedy nijak neomezenou). Připomeňme si ovšem, že v ontologiích se nevyskytují samotné konceptové výrazy, ale logická tvrzení. Na ně je nutno aplikovat ontologické vzory z *P*, abychom získali hodnoty jednotlivých $Occ(p_i, FC, O)$. Mírně zjednodušeným příkladem takového "konstrukčního" vzoru pro výraz $FC \sqcap \exists R.C$ je

$\exists D$ (R `rdfs:domain` $FC \wedge R$ `rdfs:range` $D \wedge C$ `rdfs:subClassOf` D)

tj. výraz zkonstruujeme tehdy, jestliže má vlastnost R jako svůj definiční obor (ať už přímo, nebo s využitím dědičnosti) fokusovou třídu *FC*, a zároveň má jako svůj obor hodnot určitou třídu *D* takovou, že třída *C* je její podtřídou.

Uživatelé měli za úkol rozhodnout, zda daný konceptový výraz považují za přepoužitelnou[3] kategorii vzhledem k určité jeho logické nadtřídě *FC*, nebo ne. Příklady konceptových výrazů (tří výše uvedených typů), které byly uživateli, konkrétně, 27 studenty dvou předmětů zaměřených na ontologické inženýrství, resp. propojená data, relativně často vnímány jako přepoužitelné kategorie, jsou:[4]

- $Place \sqcap \exists\ isEquippedBy.AudiovisualEquipment$
- $FridgeFreezer \sqcap \exists\ styleOfUnit.\{SingleDoor\}$
- $ProgramCommitteeMember \sqcap \exists\ writeReview.\top$

V prvním případě je kategorie "místa" upřesněna třídou *AudiovisualEquipment* omezující hodnoty vlastnosti *isEquippedBy*. Ve druhém případě je *SingleDoor* formálně individuem, avšak v realitě jde opět o vyjádření obecné kategorie (stylu chladničky). Ve třetím případě je vymezení kategorie dáno pouze vlastností *writeReview*; omezující kategorie hodnoty vlastnosti ("recenze") je zde přítomna implicitně v jejím názvu, proto kategorie dává smysl i bez explicitního upřesnění hodnoty ve struktuře výrazu.

[3] Za přepoužitelnou kategorii měl uživatel považovat takovou, u které by ho "nepřekvapilo", kdyby byla vyjádřena i jako pojmenovaná třída. Kategorie, které takto vnímá většina uživatelů ("ontologistů") pak označíme jako *ontologistické kategorie*.

[4] Prvním členem konjunkce je vždy fokusová třída, kterou výraz specializuje. Jmenné prostory ontologií používané na webu pro stručnost neuvádíme.

Z četnosti odpovědí uživatelů (na Likertově škále) byla pracovně odvozena empirická pravděpodobnost, že náhodně vybraný konceptový výraz daného typu bude "průměrným uživatelem" chápán jako přepoužitelná kategorie; tato pravděpodobnost může být použita jako váhový koeficient ve výše uvedeném vzorci pro $\widehat{FOCP}$. Nejvyšší hodnoty (cca 0,7) dosáhl vzor $FC \sqcap \exists R.\{i\}$, následován $FC \sqcap \exists R.C$ (cca 0,5); nejnižší, ale stále nezanedbatelná pravděpodobnost (cca 0,3) pak byla zjištěna u vzoru $FC \sqcap \exists R.\top$.

Komplementární analýzou k výše uvedenému ručnímu hodnocení malého vzorku byl dále automatický průzkum výskytu relevantních vzorů nad rozsáhlými kolekcemi (více než 500) ontologií. Tento průzkum prokázal nejvyšší zastoupení vzoru odpovídajícího $FC \sqcap \exists R.\top$, který se ve většině ontologií uplatňuje pro relativně vysoký počet různých fokusových tříd. Další vzory se uplatní jen pro omezený počet fokusových tříd – zřejmě těch, které jsou v dané ontologii skutečně „stěžejní".

4 Závěr

Metoda výpočtu fokusované kategorizační síly představuje zcela nový[5] příspěvek k řešení problému přepoužívání ontologií, přičemž samotný pojem fokusované kategorizace dosud nebyl, přinejmenším v kontextu ontologického inženýrství, explicitně formulován. Navazující výzkum se věnuje mj. návrhu automatické procedury využívající jednoduché techniky analýzy přirozeného jazyka a umožňující na základě jmen prvků složených konceptových výrazů predikovaných jako „ontologistické" automaticky navrhnout jména pro odpovídající pojmenované třídy, např. *PlaceEquippedByAudiovisualEquipment* nebo *SingleDoorFridgeFreezer* (zde by zafungovala heuristika, že pokud vlastnost obsahuje slovo typu „style", „type", „model", npod., není ji třeba do nového názvu zahrnout a stačí název fokusové třídy zleva rozšířit o hodnotu této vlastnosti).

Literatura

1. Baader, F. et al.: The description logic handbook: theory, implementation, and applications. Cambridge University Press New York, NY, USA, 2003.
2. Schaible, J., Gottron, T., Scherp, A.: Survey on Common Strategies of Vocabulary Reuse in Linked Open Data Modeling. In ESWC 2014: 457-472.
3. Stavrakantonakis, I., Fensel, A., Fensel, D.: Linked Open Vocabulary Ranking and Terms Discovery. In: SEMANTICS 2016: 1-8.
4. Svátek, V., Zamazal, O., Vacura, M.: Categorization Power of Ontologies with Respect to Focus Classes. In: EKAW 2016, LNCS 10024, Springer, 2016: 636–650.

[5] Vzhledem k novosti problému i jeho řešení v tomto stručném příspěvku neuvádíme „srovnání s existujícím výzkumem" – dříve řešené projekty s popisovaným souvisí pouze nepřímo.

Annotation:

Focused Categorization Power of Web Ontologies

Ontology reuse is a pre-requisite of effective use of linked data on the semantic web. The selection of an ontology for reuse currently relies on text search and entity popularity metrics. A further criterion may however be the capability of the ontology to express a subcategorization of a given focus class, where the subcategories may include compound concept expressions only implicitly present in the ontology. We propose a method of focused categorization power calculation in which the different compound concept expressions partially contribute via weight coefficients. The weights can be derived, among other, from the assessment by human users – ontologists.