Link Traversal Querying for a diverse Web of Data

Abstrakt

Václav Pospíšil

"Tradiční přístupy pro dotazování se nad Web of Data často zahrnují centralizované sklady, které replikují vzdálená data. Naopak principy v rámci Linked Data umožňují zodpovídat dotazy "živě" na webu pomocí dereferencování URI pro procházení vzdálených zdrojů dat."[1] Tento - trochu volnější - překlad je převzatý abstrakt autorů článku *Link Traversal Querying for a diverse Web of Data*. Autoři tím nastiňují výhodu "on-line" přístupu k dotazování nad Web of Data oproti přístupu tradičnímu, který je mnohdy časově, finančně či procesně náročný na získání aktuálních (a tedy relevantních) dat. Na základě návrhu O. Hartiga je k Web of Data přistupováno jako ke globální databázi, a při SPARQL dotazování se spoléhá na dereferencovatelná URI a RDF linky / zdroje, aby bylo zaručeno objevení všech relevantních zdrojů.

Nicméně i tento přístup - nazvaný "Linked traversal based query execution", či zkráceně LTBQE - má své neduhy. Dle autorů k nim patří převážně délku běhu dotazu a získání dostatku relevantních odpovědí. Tento přístup (údajně) zároveň nedokáže poskytnout výsledky, pokud v dotazu nejsou dereferencovatelná URI. Některé potenciální výsledky mohou být "přehlédnuty" v takových případech, kdy jsou dokumenty propojeny na základě literálu, blank-nodes či nedereferencovatelných URI.

Autoři navrhli určité kroky, resp. způsoby optimalizace, tedy řešení prezentovaných problémů. Mezi takové kroky patřilo zavedení RDF sémantiky, a užití *property* owl:sameAs. Optimalizace jako taková byla autory nazvána "LiDaQ", a měla obsahovat celkem pět komponent, které měly napomoci k získání co nejvíce možných výsledků, resp. odpovědí na

[1] Jürgen Umbrich, Aidan Hogan, Axel Polleres, Stefan Decker. *Link Traversal Querying for a diverse Web of Data*, **The Semantic Web** [online]. 2013-07-10 [cit. 2016-15-11]. Dostupné z: http://www.semantic-web-journal.net/content/link-traversal-querying-diverse-web-data-0

dotazy za co nejkratší čas (nebo - lépe řečeno - mít tento poměr “co nejlepší”), a k získání dodatečných - k danému dotazu relevantních - zdrojů.

V praktické části studie se autoři věnovali testování základního (resp. původního) přístupu k dotazování nad Web of Data, a porovnali ho s přístupem “obohaceným” o jimi navržené optimalizace. Testování probíhalo analýzou poměrně rozsáhlého vzorku dat z Web of Data; jeho výsledky inspirovaly autory k závěru, že standardní přístup LTBQE je vhodné užít u poměrně jednoduchých dotazů (myšleno takových dotazů, které uvažují / projdou nejvíce přibližně 100 zdrojů).

Optimalizace autorů dokázala poskytnout dodatečné odpovědi, nicméně s výšším vynaložením času běhu dotazu, a při dotazování se nad některými doménami (např. DBpedií) vedla k potížím.

Užití owl:sameAs vedlo ke generování nebo zpřístupnění signifikantně více dat (až 2,5 krát), nad kterými bylo možné se dotazovat. Užití RDFS reasoning vedlo často k navýšení rozsahu relevantních dat k dotazování, a tedy lepším výsledkům.

Zdroje

Jürgen Umbrich, Aidan Hogan, Axel Polleres, Stefan Decker. *Link Traversal Querying for a diverse Web of Data*, **The Semantic Web** [online]. 2013-07-10 [cit. 2016-15-11]. Dostupné z: http://www.semantic-web-journal.net/content/link-traversal-querying-diverse-web-data-0