2013

Structure inference for linked data sources using clustering

Christodoulou, Klitos

ACM Digital Library

Title:	**STRUCTURE INFERENCE FOR LINKED DATA SOURCES USING CLUSTERING**
Year:	2013
Author:	Christodoulou, Klitos ; Paton, Norman W. ; Fernandes, Alvaro A.A.
Abstract:	Linked Data (LD) is supplementing the World Wide Web of documents with a Web of data. This is becoming apparent from the number of LD repositories available as part of the Linked Open Data (LOD) cloud. At the instance-level, LD sources use a combination of terms from various vocabularies, expressed as RDFS/OWL, to describe their data and publish them to the Web. However, LD sources do not organise their data under a specific structure analogous to a relational schema; instead data can adhere to multiple vocabularies. Expressing SPARQL queries over LD sources -- usually over a SPARQL endpoint that is presented to the user -- requires a knowledge of the predicates used, to allow queries to express user requirements as graph patterns. Although LD provides low barriers to data publication using a homogeneous language (i.e., RDF), sources organise their data with different structures and terminologies. We would like to have a synopsis of how such data are organised in LD sources to inform the expressing of queries over such sources. With this paper we make the case that structural summaries over LD sources can inform query formulation and provide support for data integration and query processing over multiple LD sources. To fulfil this aim we propose an approach, that builds on a hierarchical clustering algorithm, for inferring structural summaries over LD sources. We have conducted an experimental evaluation using various LD sources to ascertain the extent to which our technique can successfully infer structural summaries from LD sources.