2015-02-24

Structure Inference for Linked Data Sources Using Clustering

Christodoulou, Klitos

Springer Link

Title:	**STRUCTURE INFERENCE FOR LINKED DATA SOURCES USING CLUSTERING**
Year:	2015-11
Author:	Christodoulou, Klitos ; Paton, Norman W. ; Fernandes, Alvaro A.A.
Abstract:	Linked Data (LD) overlays the World Wide Web of documents with a Web of Data. This is becoming significant as shown in the growth 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 data and publish it to the Web. However, LD sources do not organise data to conform to 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 knowledge of the predicates used so as to allow queries to express user requirements as graph patterns. Although LD provides low barriers to data publication using a single language (i.e., RDF), sources organise data with different structures and terminologies. This paper describes an approach to automatically derive structural summaries over instance-level data expressed as RDF triples. The technique builds on a hierarchical clustering algorithm that organises RDF instance-level data into groups that are then utilised to infer a structural summary over a LD source. The resulting structural summaries are expressed in the form of classes, properties and, relationships. Our experimental evaluation shows good results when applied to different types of LD sources.