Scalable Integration and Processing of Linked Data[*]

Andreas Harth
Institute AIFB, Karlsruhe
Institute of Technology
76128 Karlsruhe, Germany
harth@kit.edu

Aidan Hogan
Digital Enterprise Research
Institute (DERI), National
University of Ireland, Galway
Galway, Ireland
aidan.hogan@deri.org

Spyros Kotoulas
Dept. of Computer Science,
Vrije Universiteit Amsterdam
1081 HV, Amsterdam, The
Netherlands
kot@few.vu.nl

Jacopo Urbani
Dept. of Computer Science,
Vrije Universiteit Amsterdam
1081 HV, Amsterdam, The
Netherlands
jacopo@cs.vu.nl

ABSTRACT

The goal of this tutorial is to introduce, motivate and detail techniques for integrating heterogeneous structured data from across the Web. Inspired by the growth in Linked Data publishing, our tutorial aims at educating Web researchers and practitioners about this new publishing paradigm. The tutorial will show how Linked Data enables uniform access, parsing and interpretation of data, and how this novel wealth of structured data can potentially be exploited for creating new applications or enhancing existing ones.

As such, the tutorial will focus on Linked Data publishing and related Semantic Web technologies, introducing scalable techniques for crawling, indexing and automatically integrating structured heterogeneous Web data through reasoning.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: [On-line Information Services]

General Terms

Management

1. OVERVIEW

Linked Data is now a central topic in the Semantic Web; with recent adoption by governmental agencies (e.g., `data.gov`, `data.gov.uk`) and corporate entities (e.g., BestBuy, BBC, New York Times, Facebook, Metaweb) complimenting rich community-driven exports (e.g., DBpedia, Geonames, FOAF and SIOC), its importance is steadily growing. Linked Data provides a rich source of openly-available structured data which current Web applications and research can exploit. As more and more Linked Data is published, such growth challenges the community to develop technologies which can handle data at Web scale. Taking these cues, our tutorial will focus on (i) pragmatic and (ii) scalable techniques for consuming Linked Data which are (iii) applicable to heterogeneous datasets as now available on the Web. These techniques include crawling, querying, reasoning and composing pipelined workflows for processing Linked Data. Semantic Web technologies are often presented in a more theoretical or research-focused scope, whereas we will provide a more tangible introduction and hands-on approach to the topic, thus targeting the wider Web community—as such, our tutorial is tailored to be of interest to a wide catchment of WWW 2011 attendees.

2. CONTENT

2.1 Introduction to RDF and Linked Data

The first session gives an overview of RDF and Linked Data publishing. We will discuss the RDF data model and Linked Data principles for publishing RDF data on the Web. In particular, this session will cover:

- RDF rationale and basics
- Linked Data principles and introduction
- Current adoption and trends in Linked Data

Example materials include slides discussing end-user benefits of Governmental Linked Data, presented at the ICT 2010 event organised by the European Commission[1] and a paper discussing Linked Data publishing (see [2]).

2.2 Scalable Linked Data Crawling

This session gives an overview of the state of the art in efficient data-retrieval techniques, including novel challenges and techniques for crawling Linked Data from the Web. We will present the architecture of a crawler for small to medium-sized datasets in the range to several hundred million triples. In particular, this session will cover:

- Linked Data location, access and crawling

[*] Supported by the European Commission through projects PlanetData (FP7-257641) and LarKC (FP7-215535)

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

[1] `http://www.w3.org/2010/09/egov-session-ict2010/open-data-applications.pdf`

- Scalable crawling techniques and algorithms
- Description of the open-source LDSpider Linked Data crawler

Example materials include open-source code for crawling Linked Data[2], papers discussing scalable RDF crawling (see [3]); and presentation on the MultiCrawler architecture for distributed crawling of structured data[3].

2.3 Scalable RDF Indexing Techniques

This session presents scalable techniques for indexing and querying local repositories of Linked Data. We will discuss the standardised SPARQL query-language and thereafter discuss the state-of-the-art in RDF storage with respect to research, directions and applications. In particular, this session will cover:

- Overview and challenges
- Introduction to the SPARQL standard
- Scalable RDF indexing systems

Example materials include the paper [1] and presentation[4] on the YARS2 distributed RDF storage system.

2.4 Reasoning: Motivation and Overview

This session gives an introduction to the RDFS and OWL standards and to rule-based reasoning, with heavy emphasis on motivating reasoning for the Linked Data use-case and for integrating heterogeneous data from a large number of diverse sources. We also introduce algorithms which incorporate information about the provenance of data during reasoning to ensure robustness in the face of noisy or impudent remote data. In particular, this session will cover:

- Introduction to RDFS/OWL
- Introduction to rule-based reasoning
- Motivating reasoning with respect to Linked Data
- Web-reasoning approaches (web-scale/web-tolerant)

Example materials include papers relating to Linked Data reasoning (see [6, 4]); a tutorial introduction to Linked Data and OWL[5]; Talis Research Seminar on reasoning over Web Data[6].

2.5 Scalable Distributed Reasoning over MapReduce

This session presents scalable distributed reasoning using the MapReduce distribution framework, enabling high performance over a cluster of commodity hardware. This session details the MapReduce framework (employed by Google and Yahoo, among others) and the award-winning WebPIE system which integrates optimised execution strategies for rules supporting a (pragmatic) fragment of OWL semantics.

- MapReduce architecture
- Core optimisations and approach for distributed reasoning
- Implementing RDFS
- Extension to pD* (OWL-Horst)
- Hands-on: how to launch the reasoner on a cluster and on the Amazon cloud

Example materials include papers on reasoning over the MapReduce framework (see [6, 5]); Talis Research Seminar on scalable reasoning[7]; the official WebPIE page[8].

2.6 Hands-on: Implementing a LarKC Workflow

This session allows attendees to get hands-on with building scalable linked data applications. Some of the technologies presented in the previous sessions will be put together using a scalable workflow engine tailored for Linked Data: the Large Knowledge Collider (LarKC).

- Workflow overview and rationale
- LarKC platform overview
- Hands-on: Building a LarKC Workflow for crawling and reasoning over Linked Data

Example materials include a tutorial on the LarKC architecture[9].

3. REFERENCES

[1] A. Harth, J. Umbrich, A. Hogan, and S. Decker. YARS2: A federated repository for querying graph structured data from the web. In *6th International Semantic Web Conference, 2nd Asian Semantic Web Conference*, pages 211–224, 2007.

[2] A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres. Weaving the Pedantic Web. In *3rd International Workshop on Linked Data on the Web (LDOW2010)*, Apr. 2010.

[3] A. Hogan, A. Harth, J. Umbrich, S. Kinsella, A. Polleres, and S. Decker. Searching and Browsing Linked Data with SWSE: the Semantic Web Search Engine. Technical report, Digital Enterprise Research Institute, Galway, 2010. http://www.deri.ie/fileadmin/documents/DERI-TR-2010-07-23.pdf.

[4] A. Hogan, J. Z. Pan, A. Polleres, and S. Decker. SAOR: Template Rule Optimisations for Distributed Reasoning over 1 Billion Linked Data Triples. In *International Semantic Web Conference*, 2010.

[5] J. Urbani, S. Kotoulas, J. Maassen, F. van Harmelen, and H. E. Bal. OWL Reasoning with WebPIE: Calculating the Closure of 100 Billion Triples. In *ESWC (1)*, pages 213–227, 2010.

[6] J. Urbani, S. Kotoulas, E. Oren, and F. van Harmelen. Scalable Distributed Reasoning Using MapReduce. In *International Semantic Web Conference*, pages 634–649, 2009.

[2] http://code.google.com/p/ldspider/

[3] http://videolectures.net/iswc06_harth_ciswd/

[4] http://videolectures.net/iswc07_harth_frs/

[5] http://www.abdn.ac.uk/~csc280/tutorial/eswc2010/

[6] http://vimeo.com/15255566

[7] http://vimeo.com/15256587

[8] http://cs.vu.nl/webpie

[9] http://videolectures.net/eswc2010_kotoulas_iotl/