Exploiting Semantic Web Datasets: A Graph Pattern Based Approach

Honghan Wu[1], Boris Villazon-Terrazas[2], Jeff Z. Pan[1], and Jose Manuel Gomez-Perez[2]

[1] Department of Computing Science, University of Aberdeen, UK
[2] iSOCO, Intelligent Software Components S.A., Spain

Abstract. In the last years, we have witnessed vast increase of Linked Data datasets not only in the volume, but also in number of various domains and across different sectors. However, due to the nature and techniques used within Linked Data, it is non-trivial work for normal users to quickly understand what is within the datasets, and even for tech-users to efficiently exploit the datasets. In this paper, we propose a graph pattern based approach to guide the exploitation of Linked Data.

1 Introduction

So far, Linked Data principles and practices are being adopted by an increasing number of data providers, getting as result a global data space on the Web containing hundreds of LOD datasets [1]. However the technical prerequisites of using Semantic Web dataset prevent efficient exploitations on these datasets. To tackle this problem, in this paper we present a summarisation based approach which can not only provide a quick understanding of the dataset in question, but also is able to guide users in exploiting it in various ways.

In this demo, we will introduce our graph pattern based exploitation system[3] and demonstrate three exploitation tasks of *(Quick Understanding)* big picture presenting and summary browsing, *(Guided Exploitation)* two query generation methods, and *(Dataset Enrichment)* atomic pattern based dataset linkage. This work is supported by K-Drive project[4].

The rest of paper is organized as follows: in section 2, we briefly discuss the related work. Then, in section 3 we introduce the details of the summarisation definition and generation. In section 4, we demonstrate several typical data exploitation scenarios based on the information and properties of our summarisation. Conclusions and future work are briefly given in the final section.

2 Related Work

There are existing work such as (1) *LODStats*[5] that provides the information related to a dataset, and (2) *make-void* [6] that computes statistics about RDF

[3] http://homepages.abdn.ac.uk/honghan.wu/pages/kd.wp3/
[4] EU FP7 IAPP / Grant agreement no.: 286348
[5] http://stats.lod2.eu/
[6] https://github.com/cygri/make-void

files. However, LODStats is thought for the whole set of LOD datasets registered in The Data Hub [7], and it is based on declarative descriptions of those datasets; and *make-void* is thought for RDF files but not for RDF datasets.

Moreover, there are some existing efforts such as A-Box Summary [2] for efficient consistency checking, Zhang et al. [5] for summarising ontologies based on RDF sentence graphs, and Li et al. [3] for user-driven ontology summarisation. However, both help the understanding rather than the exploitation, which is usually task oriented.

3 RDF Summarisation: The Entity Description Pattern

Given an RDF graph, the summarisation task is to generate a condensed description which can facilitate data exploitations. Different from existing ontology summarisation work, we put special emphasises on identifying a special type of basic graph patterns in RDF data, which is suitable for data exploitation. The assumption of this special focus is that there exist such building blocks for revealing the constitution of an RDF dataset in a way which can not only help the understanding of the data but also is capable to guide RDF data exploitation. The rationale behind the assumption is that RDF data exploitation are usually based on graph patterns, e.g., SPARQL queries are based on BGP: basic graph patterns.

Specifically, in this paper, we propose one definition of such building blocks, i.e., *Entity Description Patterns* (EDPs for short). Given a resource r in an RDF graph G, the description pattern of r is $EDP(r, G) = \{C, A, P, R\}$, in which C is the set of its classes, A is a set of its data valued properties, P is a set of its object properties, and R is a set of r's inverse properties.

While EDP is a way to summarise the descriptions of resources, the relations between resources can be characterised by connections between EDPs. Two EDP E and F is said to be linked by p if and only if there exits $< r_e, p, r_f > \in G$ where $EDP(r_e, G) = E$ and $EDP(r_f, G) = F$.

Furthermore the statistics of EDP and their relations are useful information for guiding the data exploitation, e.g., the most cited papers can be interesting entities to the user. Hence, we provide statistical analysis result on EDPs and their relations. For each EDP E, it is annotated with a number which is the number of solutions to $Q(x) \leftarrow C_E(x)$. For each EDP relation $p(C_E, C_F)$, there is a tuple of (n_1, n_2), whose elements are the numbers of solutions to $Q_1(x) \leftarrow C_E(x), p(x, y), C_F(y)$ and $Q_2(y) \leftarrow C_E(x), p(x, y), C_F(y)$ respectively.

4 Demos: The Summary Based Data Exploitations

To evaluate and demonstrate the effectiveness of our definition of data building blocks i.e., EDP, in data exploitation scenarios, we implemented an EDP based data exploitation system for three types of tasks i.e., gaining big picture and browsing, generating queries and enriching datasets.

The user interface is shown in Figure 1 which contains three panels. The upper part is the *Dataset Selection Panel*, which displays the list of datasets in

[7] http://thedatahub.com

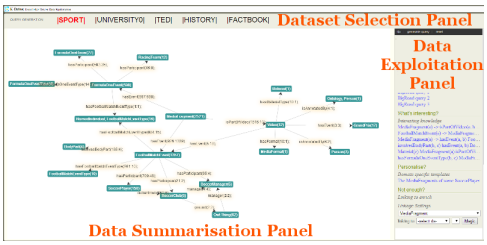

Fig. 1. Data Exploitation UI

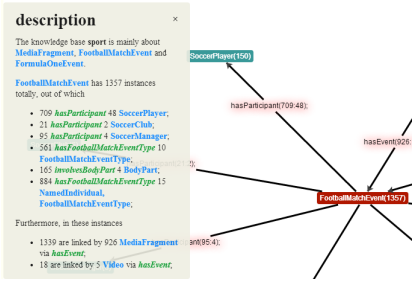

Fig. 2. Node Browsing

current demo system. To switch to another dataset, one can simply click on its name in this panel. The middle panel is the main interaction and visualisation panel, the *Data Summarisation Panel*. By default, it displays the summarisation of the selected dataset as an interactive graph i.e., the EDP graph. In other situations, relevant subgraphs of the EDP graph will be shown in the data exploitation process. The right panel is the *Data Exploitation Panel*, which shows a bunch of UI components supporting various data exploitation operations.

Given the UI, we now demonstrate a list of data exploitation scenarios to illustrate how the summarisation can help the data exploitation tasks.

The Big Picture and Browsing Operations When facing an unfamiliar dataset, users usually pursue a quick and rough *big picture* of it before (s)he can assess whether it is interesting or not, e.g, what are the data describing (concepts), how are the main concepts connected to each other (relations) and which are the important parts (clusters). To help the users gain answers to these questions quickly, as shown in the *Data Summarisation Panel* of Figure 1, the EDP graph is visualised by using force-directed graph drawing techniques [8]. Each node in the graph describes a concept. In addition to the concept name, a node is also attached with the number of instances it has in the dataset. Such statistics(c.f. Figure 2) helps to assess the importance of each concept in the dataset (in terms of data portions). The relations between (instances of) these concepts are rendered as edges, and such edges are used to calculate groups of closely related nodes, which are in turn rendered as clusters in the graph.

Two browsing operations are supported on the summary graph. The first is *node browsing*. By clicking on one node in the graph, users can gain detailed description about the concept (c.f. Figure 2) including the subgraph centralised on this node which is displayed in the middle panel and the natural language description of the node displayed in a pop-up panel on the left. The second browsing operation is *graph browsing*. After selecting a node, users can keep selecting/de-selecting interconnected nodes in current subgraph to grow or shrink it. This operation enables focused investigation on relations between interested nodes.

[8] Arbor Javascript Library (http://arborjs.org/introduction) is used for the EDP graph rendering.

Query Generation A typical usage on Semantic Web datasets is querying it. Query generation techniques [4] are helpful for either novice or advanced users because technical skills and dataset knowledge are prerequisites to write SPARQL queries. Based on the EDP summarisation, we implemented two types of query generation techniques. One is called guided query generation, which generates queries by utilising the EDP graph and statistics information attached in the graph. Such technique is good at generating queries for revealing main concepts and relations in the datasets. These two query types are called *Big City Queries* and *Big Road Queries* in the *Data Exploitation Panel* of the system. They are analogous to big cities and highways in a geography map. The other generation technique makes use of the links in the summarisation to do efficient association rule mining [4]. This method is good at revealing insightful knowledge in the data in the form of corresponding graph patterns. Such queries are called *interesting knowledge* in the system. Clicking on any of these generated queries will bring out an illustrating subgraph in the middle part of the UI.

Dataset Enrichment One of the promising features of Semantic Web techniques is the ability to link data silos to form a more valuable information space. Instead of instance-level linkage or ontology mapping, in our system, we introduce a new data linkage operation on EDPs. Such EDP-level linkage makes it possible to investigate what kinds of possibilities would be enabled after cross-dataset EDPs are linked, e.g., previously unanswerable queries might turn to be answerable by linking another dataset via EDP linkage. In the demo, we will demonstrate EDP-linkage between TED and Factbook datasets and show how such linkage can benefit a specific scenario of filtering tenders by country relations.

5 Conclusions and Future Work

We described an EDP centralised summarisation, which was shown to be useful for various data exploitation scenarios. The future work will focus on investigating the properties of the summary and in-depth studies in above scenarios.

References

1. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data-the story so far. *International journal on semantic web and information systems*, 5(3):1–22, 2009.
2. Achille Fokoue, Aaron Kershenbaum, Li Ma, Edith Schonberg, and Kavitha Srinivas. The summary abox: Cutting ontologies down to size. In *The Semantic Web-ISWC 2006*, pages 343–356. Springer, 2006.
3. Ning Li and Enrico Motta. Evaluations of User-Driven Ontology Summarization. In Philipp Cimiano and Helena Sofia Pinto, editors, *EKAW*, volume 6317 of *Lecture Notes in Computer Science*, pages 544–553. Springer, 2010.
4. Jeff Z Pan, Yuan Ren, Honghan Wu, and Man Zhu. Query generation for semantic datasets. In *Proceedings of the seventh international conference on Knowledge capture*, pages 113–116. ACM, 2013.
5. Xiang Zhang, Gong Cheng, and Yuzhong Qu. Ontology summarization based on rdf sentence graph. In Carey L. Williamson, Mary Ellen Zurko, Peter F. Patel-Schneider, and Prashant J. Shenoy, editors, *WWW*, pages 707–716. ACM, 2007.