Exploring Linked Data Graph Structures

Anja Jentzsch[1], Christian Dullweber[1], Pierpaolo Troiano[2], Felix Naumann[1]

[1] Hasso Plattner Institute, Potsdam, Germany {anja.jentzsch, felix.naumann}@hpi.de, christian.dullweber@student.hpi.de
[2] DII, University of Modena and Reggio Emilia, Modena, Italy
78242@studenti.unimore.it

Abstract. The true value of Linked Data becomes apparent when datasets are analyzed and understood already at the basic level of data types, constraints, value patterns etc. Such *data profiling* is especially challenging for RDF data, the underlying data model on the Web of Data. In particular, graph analysis can be used to gain more insight into the data, induce schemas, or build indices. We present ProLOD++, a tool for various profiling and mining tasks and in particular its recent extension GraphLOD, which offers RDF *graph analysis* features. ProLOD++ features many interactive profiling results specific for open data, such as schema discovery for user-generated attributes, association rule discovery to uncover synonymous predicates, and key discovery along ontology hierarchies. GraphLOD enhances it with subgraph pattern mining, node degree distribution, component visualization and analysis, and more.

1 RDF Data and Graph Exploration

In comparison to other data models, RDF lacks explicit schema information that precisely defines the types of entities and their attributes. Therefore, datasets can provide ontologies that categorize entities and define the semantics of properties. However, ontology information is often not available or incomplete, and even if present, datasets do not always adhere to them. Algorithms and tools are needed that *profile* the dataset to retrieve relevant and interesting metadata.

While there is a plethora of tools for profiling Linked Datasets and gathering comprehensive statistics [3, 7–10], most tools focus on a specific profiling task. Some approaches tackle the modeling and publication of profiling results [2, 11] to the Web of Data. Others focus on the visualization to explore RDF graph structures. For instance, LODlive [5] is a browser-based tool to browse and search in RDF datasets using a dynamic visual graph. LODeX [4] is a web tool to browse and visualize Linked Dataset schematas accompanied by various statistics.

Graph patterns are of interest to many communities, e.g., for protein structures, network traffic, crime detection, modeling object-oriented data, and querying RDF data. We leverage the graph pattern mining approaches gSpan [12] and GRAMI [6], to analyze Linked Datasets. To this end, we have significantly extended our prototype ProLOD++, which features many basic as well as specific profiling tasks for a given RDF dataset, such as schema discovery for user-

Profiling	Mining
• Graph feature analysis	• Unsupervised clustering & labeling
• Key analysis	• Association rules on S, P, and O
• Predicate & value distribution	• Inverse predicate discovery
• String pattern analysis	• Synonym predicate discovery
• Data type and link analysis	

Table 1. Functionalities of ProLOD++

generated attributes, association rule discovery to uncover synonymous predicates, and key discovery along ontology hierarchies [1]. ProLOD++ now is a Play application and allows easy extension by further techniques. It is available at `http://prolod.org`. We implemented and added the GraphLOD library, which provides the following new functionality:

- Basic graph statistics, such as the number of connected components and strongly connected components, their corresponding diameter, chromatic number, and node degree distribution.
- Connected components are visualized, and grouped if isomorphic.
- Three graph pattern mining algorithms.
- Visualization of mined patterns with class coloring.
- Interactive graph structure exploration in a faceted fashion.

2 Profiling and Mining Features

The features of ProLOD++ can be categorized into profiling and mining tasks, as illustrated in Table 1.

Basic Analysis. Imported data is clustered by hierarchical topic clustering if no underlying schema is available, otherwise it is grouped based on the underlying taxonomic hierarchy. The profiling and mining tasks are executed on import and results are stored in a relational database. These include statistics on frequencies and distributions of distinct subjects, properties, and objects. Pattern analysis provides the user with statistics on data types and value pattern distributions of particular properties. ProLOD++ discovers positive and negative association rules, e.g., to discover synonymous properties or inverse properties. To cope with the sparsity of property values on the Web of Data when discovering key candidates, ProLOD++ calculates the keyness measure for each property along the ontology class hierarchy. These features were already demonstrated in [1]; the main contributions of this demonstration are described next.

Graph Feature Analysis. ProLOD++ allows exploring the graphical structures of Linked Datasets by visualizing the connected components and the graph patterns mined from them. Given the underlying graph for a Linked Dataset, containing all entities as nodes and object properties between them as links, we detect graph patterns for its directed as well as undirected version. The latter allows for pattern mining on a more general level. Bigger graph components (> 1000 nodes) are mined for subgraph patterns using three different approaches:

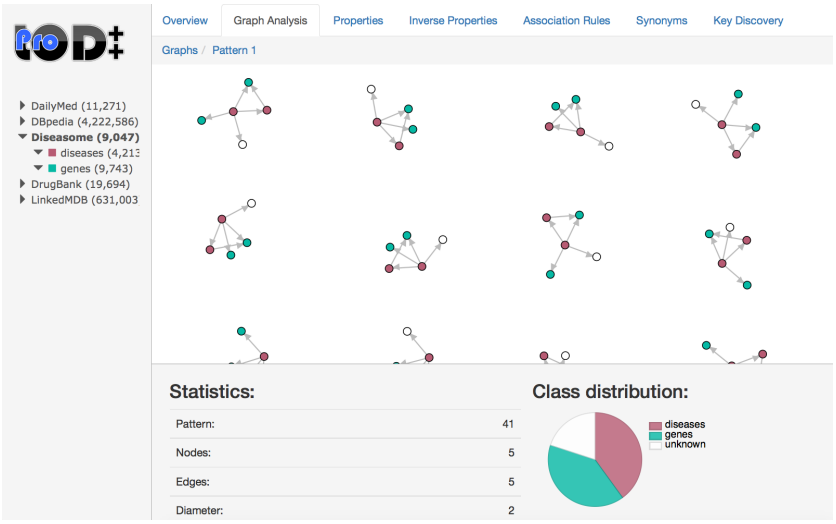

Fig. 1. Occurrences of a pattern in Diseasome visualized by ProLOD++

gSpan, GRAMI, and a new approach that mines for predefined patterns. Our goal is to define a set of graph patterns that can be considered the core of most Linked Datasets. We identify graph patterns such as paths, cycles, stars, siamese stars, antennas, caterpillars, and lobsters. Figure 1 is a screenshot of ProLOD++ showing all occurrences of a selected pattern and their class distribution along with some statistical information.

ProLOD++ allows faceted browsing through the graph patterns. Patterns are grouped when isomorphic, first based on their underlying structure and then based on the class membership (color). This allows for finding not only common, re-occurring patterns but also patterns that are dominant for certain class-combinations. E.g., astronomers in DBpedia are often to be found in star patterns, surrounded by their discovered astronomical objects.

Based on the graph features provided by ProLOD++ and its underlying GraphLOD library, an overall model for Linked Datasets can be given: We observe that most of the Linked Datasets consist of a number of small satellite graphs and a giant component that contains more than 80% of the nodes and thus resemble scale-free networks as they occur in social networks.

When jointly profiling multiple datasets, ProLOD++ highlights the connectivity of connected components across them based on inter-dataset links. This, for instance, identifies the potential of dataset integration.

3 ProLOD++ Demonstration

ProLOD++ is a web-based tool to be either distributed for local execution or hosted as a service at `http://prolod.org`. Some of the described features are still under development, but at the time of submission ProLOD++ is already a useful tool to explore RDF datasets and their graph structure. During the demo,

users can bring along their own RDF dataset, import it into ProLOD++ and begin the analysis. A number of several interesting datasets from various domains have been already imported, including DBpedia, Diseasome, and LinkedMDB.

After the initial analysis phase, users can select datasets and clusters in a tree model and browse the profiling results across several tabs. The graph feature analysis shows graph statistics, such as number of nodes and edges, and the diameter for the connected and strongly connected components. A node degree distribution chart is displayed to analyze the underlying graph model. Besides statistical information, ProLOD++ allows faceted browsing through the graph patterns, from general patterns to class-colored patterns down to concrete pattern examples. The class distribution is visualized at each facet level.

References

1. Z. Abedjan, T. Grütze, A. Jentzsch, and F. Naumann. Mining and Profiling RDF Data with ProLOD++. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 1198–1201, 2014. Demo.
2. A. Assaf, R. Troncy, and A. Senart. Roomba: An extensible framework to validate and build dataset profiles. In *ESWC International Workshop on Dataset Profiling & Federated Search for Linked Data (PROFILES)*, 2015.
3. S. Auer, J. Demter, M. Martin, and J. Lehmann. LODStats – an extensible framework for high-performance dataset analytics. In *Proceedings of the International Conference on Knowledge Acquisition, Modeling and Management (EKAW)*, volume 7603, pages 353–362, 2012.
4. F. Benedetti, L. Po, and S. Bergamaschi. A visual summary for linked open data sources (demo). In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 173–176, 2014.
5. D. V. Camarda, S. Mazzini, and A. Antonuccio. LodLive, exploring the web of data. In *Proceedings of the International Conference on Semantic Systems*, I-SEMANTICS, pages 197–200, 2012.
6. M. Elseidy, E. Abdelhamid, S. Skiadopoulos, and P. Kalnis. GRAMI: frequent subgraph and pattern mining in a single large graph. *PVLDB*, 7(7):517–528, 2014.
7. T. Käfer, A. Abdelrahman, J. Umbrich, P. O'Byrne, and A. Hogan. Observing linked data dynamics. In *Proceedings of the Extended Semantic Web Conference (ESWC)*, volume 7882 of *LNCS*, pages 213–227. Springer, 2013.
8. S. Khatchadourian and M. P. Consens. ExpLOD: Summary-based exploration of interlinking and RDF usage in the linked open data cloud. In *Proceedings of the Extended Semantic Web Conference (ESWC)*, 2010.
9. A. Langegger and W. Wöß. RDFStats – an extensible RDF statistics generator and library. In *Proceedings of the International Workshop on Database and Expert Systems Applications (DEXA)*, pages 79–83, 2009.
10. H. Li. Data Profiling for Semantic Web Data. In *Proceedings of the International Conference on Web Information Systems and Mining (WISM)*, pages 472–479, 2012.
11. E. Mäkelä. Aether – generating and viewing extended VoID statistical descriptions of RDF datasets. In *ESWC (Satellite Events)*, pages 429–433, 2014.
12. X. Yan and J. Han. gSpan: Graph-based substructure pattern mining. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 721–724, 2002.