rdf:SynopsViz – A Framework for Hierarchical Linked Data Visual Exploration and Analysis

Nikos Bikakis[1,2] Melina Skourla[1] George Papastefanatos[2]

[1]National Technical University of Athens, Greece
[2]IMIS, ATHENA Research Center, Greece

Abstract. The purpose of data visualization is to offer intuitive ways for information perception and manipulation, especially for non-expert users. The Web of Data has realized the availability of a huge amount of datasets. However, the volume and heterogeneity of available information make it difficult for humans to manually explore and analyse large datasets. In this paper, we present rdf:SynopsViz, a tool for hierarchical charting and visual exploration of Linked Open Data (LOD). Hierarchical LOD exploration is based on the creation of multiple levels of hierarchically related groups of resources based on the values of one or more properties. The adopted hierarchical model provides effective information abstraction and summarization. Also, it allows efficient -on the fly- statistic computations, using aggregations over the hierarchy levels.

Keywords: Visual analytics, Semantic Web, LOD, RDF visualization, Data exploration, RDF Statistics, RDF Charts, Faceted search, RDF Facets.

1 Introduction

The purpose of data visualization is to offer intuitive ways for information perception and manipulation that essentially amplify, especially for non-expert users, the overall cognitive performance of information processing. This is of great importance in the Web of Data, where the volume and heterogeneity of available information make difficult for humans to manually explore and analyse large datasets. An important challenge is that visualization techniques must offer scalability and efficient processing for on the fly visualization of large datasets. They must also employ appropriate data abstractions and aggregations for avoiding information overloading due to the size and diversity of the data presented to the user. Finally, they must be generic and provide uniform and intuitive visualization results across multiple domains.

In this work, we present rdf:SynopsViz, a framework for hierarchical charting and exploration of Linked Open Data (LOD). Hierarchical LOD exploration realized through the creation of multiple levels of hierarchically related groups of resources based on the values of one or more properties. For example, a numerical group, characterized by a numerical range, comprises all resources with a property value within the range of this group. Hierarchical browsing can address the problem of information overloading as it provides information abstraction and summarization [1]. It can also offer rich insights on the underlying data when combined with rich statistical information on the groups and their contents.

The key features of rdf:SynopsViz framework are summarized as follows: (1) It adopts a *hierarchical model* for RDF data visualization, browsing and analysis. (2) It offers *automatic* on-the-fly hierarchy construction based on data distribution, as well as *user-defined* hierarchy construction based on user's preferences. (3) Provides *faceted* browsing and filtering over classes and properties. (4) Integrates *statistics with visualization*; visualizations have been enriched with useful statistics and data information. (5) Offers several visualizations techniques (e.g., timeline, chart, treemap). (6) Provides a large number of dataset's *statistics* regarding the: data-level (e.g., number of sameAs triples), schema-level (e.g., most common classes/properties), and structure level (e.g., entities with the larger in-degree). (7) Provides numerous *metadata* related to the dataset: licensing, provenance, linking, availability, undesirability, etc. The latter are useful for assessing data quality [13].

2 Framework Overview

The architecture of rdf:SynopsViz is presented in Figure 1. Our scenario involves three main parts: the Client GUI, the rdf:SynopsViz framework, and the input data. The *Client* part, corresponds to the framework's front-end offering several functionalities to the end-users (e.g., statistical analysis, facet search, etc.). rdf:SynopsViz consumes RDF data as *Input data*; optionally, OWL-RDF/S vocabularies/ontologies describing the input data can be loaded. Next, we describe the basic components of the rdf:SynopsViz framework.

In the preprocessing phase, the *Data and Schema Handler* parses the input data and inferes schema information (e.g., properties domain(s)/range(s), class/property hierarchy, type of instances, type of properties, etc.). *Facets Generator* generates class and property facets over input data. *Statistics Generator* computes several statistics regarding the schema, instances and graph structure of the input dataset, such as the number of different types of classes and properties, or the number of sameAs triples, or finally the average in/out degree of the RDF graph, respectively. *Metadata Extractor* collects dataset metadata which can be used for data quality assessment. *Hierarchical Model Module* adopts our

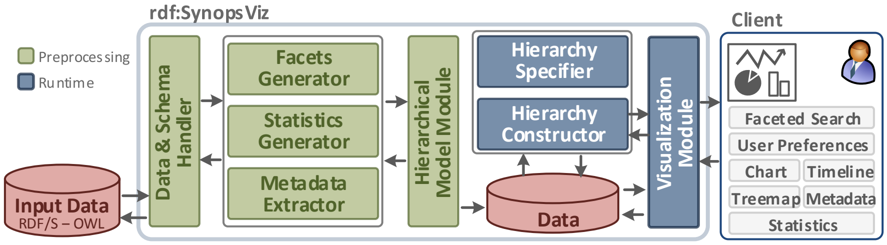

Fig. 1: System Architecture

hierarchy model and stores the initial data enriched with the information computed during the preprocessing phase.

During runtime the following components are involved. *Hierarchy Specifier* is responsible for managing the configuration parameters of our hierarchy model, e.g., the number of hierarchy levels, the number of nodes per level, and providing this information to the Hierarchy Constructor. *Hierarchy Constructor* implements the hierarchy model. Based on the selected facets, and the hierarchy configuration: it determines the hierarchy of groups and the contained triples, and computes the statistics about their contents (e.g., range, variance, mean, number of triples contained, etc.). *Visualization Module* allows the interaction between the user and the framework, allowing several operations (e.g, navigation, filtering, hierarchy specification) over the visualized data.

3 Implementation & Demonstration Outline

Implementation. rdf:SynopsViz is implemented on top of several open source tools and libraries. Regarding visualization libraries, we use Highcharts[1], for the area and timeline charts. and Google Charts[2] for treemap and pie charts. Additionally, it uses Jena framework[3] for RDF data handing and Jena TDB for RDF storing.

The web-based prototype of rdf:SynopsViz is available at `83.212.125.131:8084/synopsViz`. Also a video demonstrating the scenario presented below is available at `http://youtu.be/8v-He1U4oxs`.

Demonstration scenario. First, the attenders will be able to select a dataset from a number of offered real-word datasets (e.g., dbpedia, Eurostat, World Bank, U.S. Census, etc.) or upload their own. Then, for the selected dataset, the attendees are able to examine several of the dataset's *metadata*, and explore several datasets's *statistics*.

Using the facets panel, the attenders are able to navigate and filter data based on classes, numeric and date properties. In addition, through facets navigation several information about the classes and properties (e.g., number of instances, domain(s), range(s), IRI, etc.) are provided to the users through the UI.

The attenders are able to navigate over data by considering properties' values. Particularly, area charts and timeline-area charts are used to visualize the resources considering the user's selected properties. Classes' facets can also be used to filter the visualized data. Initially, the top level of the hierarchy is presented providing an overview of the data, organized into top-level groups; the user can interactively zoom in and out the group of interest, up to the actual values of the raw input data. At the same time, statistical information concerning the hierarchy groups as well as their contents (e.g., mean value, variance, sample data, etc.) are presented.

[1] www.highcharts.com

[2] developers.google.com/chart

[3] jena.apache.org

In addition, the attenders are able to navigate over data, through class hierarchy. Selecting one or more classes, the attenders can interactively navigate over the class hierarchy using treemaps. In rdf:SynopsViz the treemap visualization has been enriched with schema and statistical information. For each class, schema metadata (e.g., number of instances, subclasses, datatype/object properties) and statistical information (e.g., the cardinality of each property, min, max value for datatype properties' ranges, etc.) are provided.

Finally, the attenders can interactively modify the hierarchy specifications. Particularly, they are able to increase or decrease the level of abstraction/detail presented, by modifying modifying both the number of hierarchy levels, and number of nodes per level.

4 Related Work

A large number of works studying issues related to RDF or LOD visualization and analysis have been proposed in the literature [2,3,4,5]. Additionally, numerous tools offering RDF or Linked Open Data visualization have been developed, e.g., *Sgvizler* [6], *LODWheel* [7], *Payola* [8], *CubeViz* [9], *KC-Viz* [10], *RelFinde*[4], *Welkin*[5], *IsaViz*[6], *RDF-Gravity*[7], etc.

In the context of RDF and Linked Open Data statistics, *RDFStats* [14] calculates statistical information about RDF datasets. *LODstats* [11] is an extensible framework, offering scalable statistical analysis of Linked Open Data datasets.

Regarding the quality assessment issues, [13] studies the criteria which can be used in Linked Data quality assessment. [14] review millions of RDF documents to analyse Linked Data conformance. Finally, several frameworks for the quality assessment in the Web of Data, have been proposed *LINK-QA* [15], *Sieve* [16], *WIQA* [17]. In contrast to existing approaches, we provide hierarchical RDF data visualization enriched with data statistics. The hierarchical model solves the visualization overload issues, offering efficient, on the fly statistical computations over hierarchy levels. Finally, due to hierarchical model our tool can efficiently handle and analyse very large datasets.

5 Conclusions

In this paper we have presented rdf:SynopsViz, a framework for hierarchical charting and exploration of Linked Open Data. The hierarchical model adopted by our framework can address the problem of information overloading, offering an effective mechanism for information abstraction and summarization. Additionally, the adopted model allows the efficient statistic computations, using aggregations over the hierarchy levels.

[4] www.visualdataweb.org/relfinder.php

[5] simile.mit.edu/welkin

[6] www.w3.org/2001/11/IsaViz

[7] semweb.salzburgresearch.at/apps/rdf-gravity

Some future extensions of our tool include the application of more sophisticated filtering techniques (e.g., SPARQL-enabled browsing over the data), as well as the addition of more visual techniques and libraries.

Acknowledgement. This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: THALIS and KRIPIS - Investing in knowledge society through the European Social Fund.

References

1. Elmqvist N., Fekete J-D., "Hierarchical Aggregation for Information Visualization: Overview, Techniques, and Design Guidelines.", IEEE Trans. Vis. Comput. Graph. 16(3) 2010
2. Dadzie A., Rowe M.,"Approaches to visualising Linked Data: A survey", Semantic Web 2(2), 2011
3. Brunetti J., Auer S., Garcia R., "The Linked Data Visualization Model", ISWC 2012
4. Dadzie A., Rowe M., Petrelli D."Hide the Stack: Toward Usable Linked Data", ESWC 2011
5. Alonen M., Kauppinen T., Suominen O., Hyvonen E. "Exploring the Linked University Data With Visualization Tools", ESWC 2013
6. Skjveland M.: "Sgvizler: A JavaScript Wrapper for Easy Visualization of SPARQL Result Sets", ESWC 2012
7. Stuhr M., Dumitru R., Norheim D.,"LODWheel - JavaScript-based Visualization of RDF Data", Workshop on Consuming Linked Data 2011
8. Klímek J., Helmich J., Necaský M., "Payola: Collaborative Linked Data Analysis and Visualization Framework", ESWC 2013
9. Salas P., Mota F., Breitman K., Casanova M., Martin M., Auer S., "Publishing Statistical Data on the Web", IEEE Semantic Computing 2012
10. Motta E., Mulholland P., Peroni S., d'Aquin M., Gomez-Perez J., Mendez V., Zablith F., "A Novel Approach to Visualizing and Navigating Ontologies", ISWC 2011
11. Auer S., Demter J., Martin M., Lehmann J.: "LODStats - An Extensible Framework for High-Performance Dataset Analytics", Knowledge Engineering and Knowledge Management 2012
12. Langegger A., Wöß W.: "RDFStats - An Extensible RDF Statistics Generator and Library" Workshop on Web Semantics 2009
13. Zaveri A., Rula A., Maurino A., Pietrobon R., LehmannJ., Auer S.: "Quality assessment methodologies for linked open data", Under review, available at Semantic Web Journal site.
14. Hogan A., Umbrich J., Harth A., Cyganiak R., Polleres A., Decker S.: "An empirical survey of Linked Data conformance.", J. Web Sem. 14, 2012
15. Guéret C., Groth P. T., Stadler C., LehmannJ. "Assessing linked data mappings using network measures", ESWC 2012.
16. Mendes P., Mühleisen H., Bizer C.: "Sieve: linked data quality assessment and fusion", Workshop on Linked Web Data Management 2012
17. Bizer C., Cyganiak R. "Quality-driven information filtering using the WIQA policy framework", J. Web Sem. 7(1) 2009