Connecting the Dots in Million-Nodes Knowledge Graphs with SemSpect

Thorsten Liebig, Vincent Vialard, and Michael Opitz

derivo GmbH, Ulm, Germany

Abstract. Knowledge Graphs are becoming more common within industrial and scientific applications. Field experience with our customers have revealed that existing graph visualization and querying tools fail to adequately support users in understanding and querying real-world datasets. We present SemSpect, a solution that enables users to visualize, interactively explore and easily retrieve answers to sophisticated request from large RDF/OWL knowledge graphs without being fluent in a query language. The demo will showcase its usage with various datasets such as SpringerNature's SciGraph[1], the Panama Papers[2] or legislative open data from GovTrack.us[3] and invites attendees to gain insight into the data by hands on experience.

1 Exploring Large Knowledge Graphs

Analyzing and utilizing the content of large Knowledge Graphs is challenging. A proper SPARQL query can deliver some insight when the basic structure of the data and the request idea is known to the questioner. But how to proceed when facing complex, cross-linked data pools or when the insightful queries are not transparent because of a lack of the big picture? For instance, how to identify the real beneficial owners of an offshore entity within the Panama Papers whose shares ownership are intentionally blurred without knowing the obfuscation patterns?

Typical visualization approaches and query interfaces fall short to provide real support for exploring large unknown datasets. The usual graph visualization tools are cumbersome as soon as the number of vertices and their connection degree rises. Textual, form-based or graphical query interfaces are not well suited for the trial and error process inherent to exploratory tasks. Faceted search tools like Sparklis [1] or SemFacet [2] avoid requests that return no results by offering selections supported by the data, but the view on the data is limited on the current focus and bears the risk for users to lose the plot. Furthermore, the linear or tabular presentation of the results make their interpretation laborious. Another approach is followed by VisualRDF [3] which proposes a graphical aggregated

[1] Springer Nature SciGraph, 2017. `http://www.springernature.com/scigraph`
[2] ICIJ Offshore Leaks Database, 2016. `https://offshoreleaks.icij.org`
[3] United States legislation data, GovTrack, 2016. `https://www.govtrack.us/`

overview with some drill down possibilities, but it is restricted to a meta level summary and does not provide access the actual objects.

SemSpect is a tool that enables even domain or query novices to carry out sophisticated research by interacting with a visual representation of the data. One unique feature is the aggregated tree-like overview of the relations between groups of objects built step by step by the user. This interaction, guided by the data as well as by the data model, makes the relations between the different categories of objects within the graph tangible. At the same time the quantity of displayed information is reduced. Details of particular objects, group of objects or their relations are displayed only on user demand.

2 Overview First – Details on Demand

Let us consider the SpringerNature SciGraph dataset[4] that contains data about scientific articles, journals, authors, institutions and grants as well as their links. For the year 2016 the dataset comprises some three million data objects (53 million triples after inferencing). What is the basic structure of the graph? How are the data objects connected with objects of other categories? In SemSpect the 351k articles of 2016 are visually depicted as a group that lists its related categories and their relationships on double click.

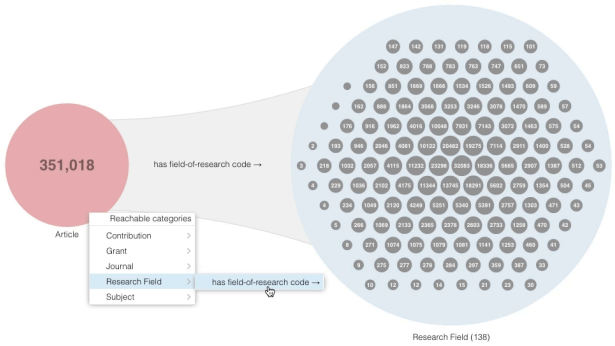

Fig. 1. Exploring relations between articles and research fields in SN SciGraph

When a relationship is selected from the exploration menu, a branch is added to the exploration tree showing the related objects. Note that in contrast to common graph renderings, SemSpect aggregates objects and relations. Individual data objects – here the research fields – are shown as dots within a group only when their number is below an adjustable threshold. A number in a dot of an object indicates the number of related objects in the preceding group.

[4] The SN SciGraph schema and data is available in N-Triples format and can be loaded into SemSpect right away. As key partner of SN's SciGraph initiative, we enriched the model with facets for the core categories.

Pointing at an object highlights the objects from the other groups that are related to it. Selecting an object displays its attributes in a dedicated side panel on the right. Selecting a group displays statistical information about the class membership of its objects within this side panel (this information can be structured in independent facets by annotating the model accordingly). A tabular view is available on demand for listing the values of the attributes of the objects of a group or for listing all the object pairs of two connected groups and their relations.

In order to focus on objects with specific properties, each object group of the exploration graph can be filtered using the predefined classification as well as using the values of the attributes of its objects. For instance, to find the top journals and grants related to articles about cancer in SciGraph, start with the subjects, restrict them to those whose description contain 'cancer' in the tabular view and expand this group to journals and grant via articles (Fig. 2).

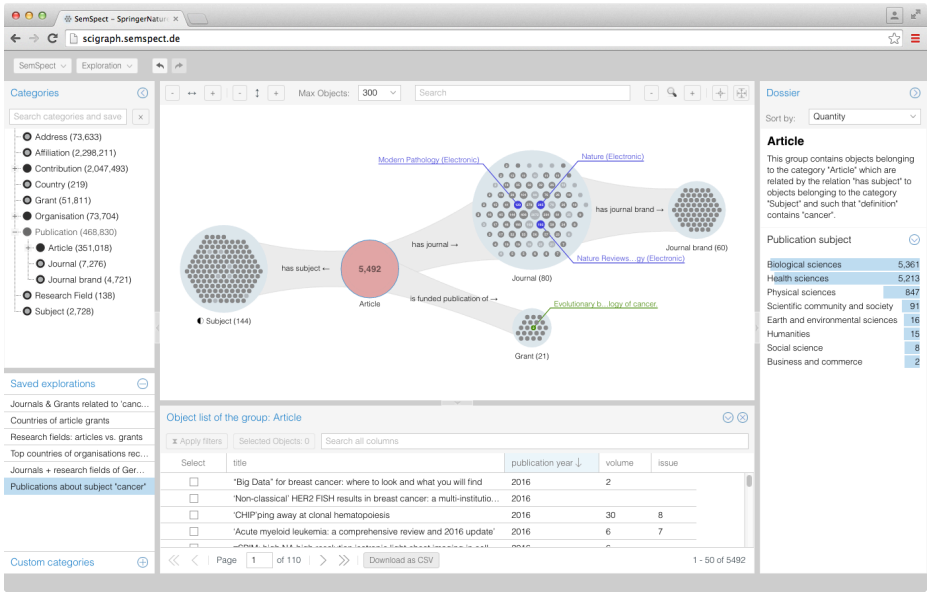

Fig. 2. Exploring the objects related to articles about cancer in SN SciGraph

These functionalities coupled with the possibility to store an exploration for later usage enables the creation of an equivalent to complex parameterized queries. Moreover, each object group obtained during an exploration can be used to define a new category (defined by the implicit request used to obtain the group). The user can therefore step by step build his personalized exploration toolbox.

3 Technical Details of SemSpect

SemSpect is a client-server application with a HTML5/JavaScript UI[5] and a Java REST backend. It is able to process knowledge graphs formalized according to RDF or OWL – W3C's ontology languages. The reasoning and request processing engine is our GraphScale [4] system with a Neo4j[6] graph database as storage component in the present case.

The apparent simplicity of the aggregated exploration graph used by SemSpect comes with high requirements on the request processing engine. Indeed, the query for retrieving the objects of a group gets more complex the further the group is from the root and many informations such as predecessor count and category membership have to be aggregated. Moreover, in order to build the exploration menu of each group, the list of all reachable object categories with their specific relation types must be collected. As each group can be filtered at any time, and as this filtering propagates down the tree, this retrieval has to be processed at run time for each updated group.

In order to achieve an acceptable performance at the million objects scale, SemSpect relies on the patented technology used in GraphScale [4] based on a condensed version of the original data called abstraction that is small enough to fit in memory. The abstraction contains only partial information but it allows for a significant reduction of the lookups at the actual data in the store.

4 Demo Description

This demo will showcase the visualization and exploration features of SemSpect with the help of variety of hand-on sample data sets such as SpringerNatures SciGraph (`http://scigraph.semspect.de`), the Panama Papers (`http://panama.semspect.de`) or GovTrack (`http://govtrack.semspect.de`). Futhermore, we will provide information about the system architecture, underlying technology and the type of application domains where SemSpect is already deployed or can be beneficial.

References

1. Sébastien Ferré. SPARKLIS: a SPARQL Endpoint Explorer for Expressive Question Answering. Proceedings of the ISWC 2014 Posters & Demonstrations Track.
2. B. Cuenca Grau, E. Kharlamov, S. Marciuska, D. Zheleznyakov, M. Arenas. SemFacet: Faceted Search over Ontology Enhanced Knowledge Graphs. Proceedings of the ISWC 2016 Posters & Demonstrations Track.
3. F. Florenzano, D. Parra, J. L. Reutter, F. Venegas. An Interactive Visualisation for RDF Data. Proceedings of the ISWC 2016 Posters & Demonstrations Track.
4. T. Liebig, V. Vialard, M. Opitz, S. Metzl. GraphScale: Adding Expressive Reasoning to Semantic Data Stores. Proceedings of the ISWC 2015 Posters & Demonstrations Track.

[5] works best with current Chrome/Firefox or IE11+
[6] https://neo4j.com/