GraFa: Faceted Search & Browsing for the Wikidata Knowledge Graph

José Moreno-Vega and Aidan Hogan

IMFD Chile & Department of Computer Science, University of Chile

Abstract. We present a demo of the GraFa faceted search and browsing interface over the Wikidata knowledge graph. We describe the key aspects of the interface, including the types of interactions that the system allows, the ranking schemes employed, and other features to aid usability. We also discuss future plans for improving the system.
Online Demo: http://grafa.dcc.uchile.cl/

1 Introduction

Faceted browsing [7] has become a popular paradigm for interacting with data on the Web, where a number of authors have proposed systems for faceted browsing interfaces over Semantic Web knowledge-bases (see [3,5] for surveys). However, many such systems are demonstrated for small, uniform datasets, where we could not find an available system that would work for a dataset as large (many triples) and diverse (many properties and classes) as Wikidata [6]. These large, diverse datasets are precisely those most in need of intuitive user interfaces.

To bridge this gap, we propose the GRaph FAcets (GraFa) system designed to offer faceted search and browsing over large, diverse RDF graphs. An important feature of the GraFa system is that it provides exact faceted views, meaning that the facets offered to restrict the current results set offer exact counts and are exactly those that will lead to non-empty results upon selection. However, result sets for very common types of (intermediate) queries – such as human or human AND gender male, etc. – can reach into the millions of entities. Computing the exact facets for such large results while maintaining interactive response times is technically challenging and does not appear to be well-supported by available faceted search tools. To improve scalability while maintaining efficiency, the GraFa system thus incorporates novel indexing schemes that, in an offline phase, pre-compute and store exact facets for large results sets.

In our paper accepted in the research track [4], we describe the GraFa system in detail, including the faceted browsing interactions it permits, the indexing scheme used to improve query performance, the implementation based on Lucene, performance experiments over Wikidata, as well as an initial user evaluation of the system. These results show that by pre-indexing the exact facets for 141 queries identified as generating more than 50,000 results, the worst-case response times for the system are under 3 seconds. In addition, initial user evaluation results provide feedback for further directions in which the system can be improved. We refer the reader to [4] for detailed results.

In the demo track, we propose to offer attendees of ISWC a live demo of the GraFa system for performing faceted search and browsing over Wikidata. We also wish to discuss possible features that could be added to the system in future, other datasets or use-cases to which Grafa could be applied, as well as to gain feedback and identify potential topics for collaboration. The demo we plan to provide is publicly available here: `http://grafa.dcc.uchile.cl/`. In this paper, we provide details on user interactions, details on the prototype, and current limitations; for more information on performance, back-end, indexing, usability, etc., we refer to our paper in the research track [4].

2 User Interactions

Figure 1 provides an overview of the user interactions that the GraFa system currently supports. The user is first presented with the option of searching by keyword (e.g., `"nick drake"`) or by selecting a type IRI (e.g., wd:Q5 (human)). Each result set is associated with a list of *facets* from which the user may iteratively select further restrictions of the current results. We will now discuss each of these interactions in further detail.

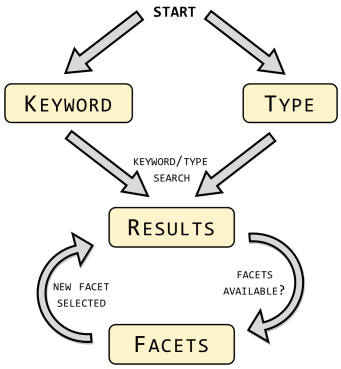

Fig. 1: Overview of user interactions supported by GraFa

For the initial keyword search, a ranked list of entities are returned that have matching keywords in their labels, aliases or descriptions; the properties corresponding to such values must be configured in the system (in the case of Wikidata, we use `rdfs:label`, `skos:altLabel` and `schema:description`). Furthermore, a set of supported languages must be configured, where the labels, aliases and descriptions of these languages will be indexed, as available; the demo is configured for both Spanish and English.

Rather than search by keyword, the user can instead opt to perform a type search; here, types are defined as values for a configured list of type properties (in the case of Wikidata, we use wdt:P31 (instance of)). Given that users will often not know the required IRI of a type, the GraFa interface offers auto-completion on the labels and aliases of types in the graph, where suggested types are ranked according to a PageRank score and are annotated with the number of results with that type. For example, if a user types in the partial label `hum`, GraFa will suggest `human (3595226 results)` as the first result; if selected, GraFa will search for entities of type wd:Q5; on the other hand, if the user types in `per`, the first suggestion will be `person (3595226 results)`, which if selected will also trigger a query for entities of type wd:Q5 (for which person is an alias).

Whether the user begins with a keyword or type search, GraFa will generate a list of resulting entities. For each entity in the results, its label, aliases, description and an image are displayed; the label is presented as a hyperlink

that will dereference the entity IRI. In the case of keyword search, results are ranked by a combined query-relevance and PageRank score. In the case of type search, results are ranked purely by PageRank score. In either case, a list of facets are computed for the current results set. Each facet is a property and a list of values that at least one entity in the current result set is associated with. The facet view displays all possible properties, with a count of the number of entities with some value for that property. Upon selecting a property, the user can perform an auto-complete search on the label of a particular value, or can browse possible values in a drop-down list. Once a value is selected, a new result set is generated for the active conjunction of restrictions, where these results are ordered by PageRank; a new facet view is also generated. The user may continue iteratively adding facets until they are satisfied with the results, or until they reach a single result (facets with zero results are never offered).

3 Demo

We have implemented a prototype of the GraFa system for experimental purposes and to gather initial feedback and expressions of interest. The system uses Apache Lucene as a back-end store, managing (1) full-text search indexing for keyword and auto-complete prefix searches, (2) structured indexes for type and facet selection, (3) indexes for cached queries generating more than 50,000 results. The front-end is implemented as a Java servlet, with interactive autocomplete features being based on Javascript libraries. The source code is available from the following repository: `https://github.com/joseignm/GraFa/`.

We will demonstrate an instance of the GraFa prototype indexing a dump of Wikidata. More specifically, the demo indexes the "truthy" dump of Wikidata from 2017/09/13, containing 1.77 billion triples and 74.1 million entities. On a machine with an Intel Xeon E5-2609 v3 CPU, 32 GB of RAM, and 2 × 2TB Seagate 7200 RPM 32MB Cache SATA, indexing the data takes approximately 5 days, with the majority of time (4.5 days) taken to compute and index all queries with more than 50,000 results [4]. In performance experiments on type and facet selections, the worst-case response times are around 3 seconds [4].

In Figure 2, we provide an example screenshot of the GraFa demo where the user has searched for lighthouses on the Adriatic Sea, and is now considering further filtering the 14 available results by country.[1]

4 Limitations and Future Work

We see the current GraFa system as offering a baseline system for future development, where indeed the current version has a number of limitations that we have yet to address, including (1) support for datatypes and range queries, (2) support for existential value queries, (3) support for class/property hierarchies and potentially other forms of inference, (4) incremental updates. We note that

[1] See `http://grafa.dcc.uchile.cl/search?instance=Q39715&properties=P206%23%23Q13924`

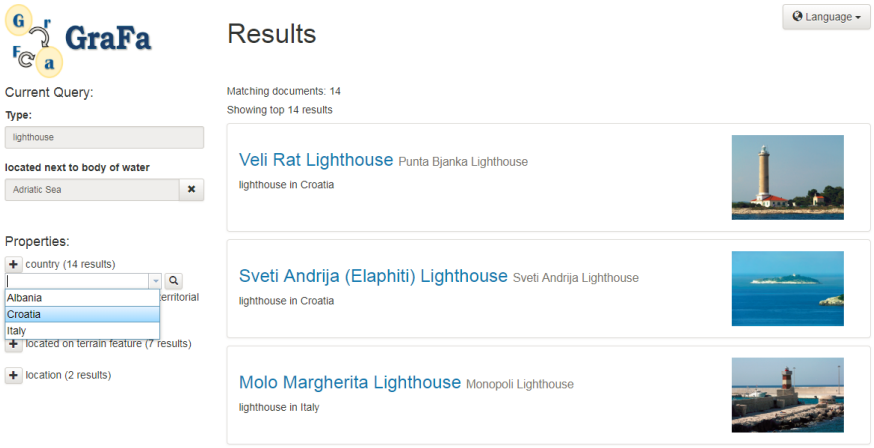

Fig. 2: Results for lighthouses located next to the Adriatic Sea, further showing possible values for countries by which the current results can be restricted

existing faceted browsing systems support some of these features, where, for example, Broccoli [2] offers range queries, while SemFacet [1] offers reasoning capabilities. It would thus be interesting to see if similar techniques could be combined into GraFa in the future, what performance cost such new features would imply for faceted browsing over a dataset such as Wikidata, and indeed, what sorts of benefits they could bring for users of GraFa.

Acknowledgements This work was supported by the Millennium Institute for Foundational Research on Data (IMFD) and by Fondecyt Grant No. 1181896.

References

1. Arenas, M., Grau, B.C., Kharlamov, E., Marciuska, S., Zheleznyakov, D.: Faceted search over RDF-based knowledge graphs. J. Web Sem. 37-38, 55–74 (2016)
2. Bast, H., Bäurle, F., Buchhold, B., Haußmann, E.: Easy access to the Freebase dataset. In: International World Wide Web Conference (WWW). pp. 95–98 (2014)
3. Dadzie, A., Rowe, M.: Approaches to visualising Linked Data: A survey. Semantic Web 2(2), 89–124 (2011)
4. Moreno-Vega, J., Hogan, A.: GraFa: Scalable Faceted Browsing for RDF Graphs. In: International Semantic Web Conference (ISWC) (2018), (to appear)
5. Tzitzikas, Y., Manolis, N., Papadakos, P.: Faceted exploration of RDF/S datasets: a survey. J. Intell. Inf. Syst. 48(2), 329–364 (2017)
6. Vrandecic, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM 57(10), 78–85 (2014)
7. Wei, B., Liu, J., Zheng, Q., Zhang, W., Fu, X., Feng, B.: A survey of faceted search. J. Web Eng. 12(1&2), 41–64 (2013)