Semantic Relatedness as an Inter-facet Metric for Facet Selection over Knowledge Graphs

Leila Feddoul[1,2(✉)], Sirko Schindler[2], and Frank Löffler[1]

[1] Heinz Nixdorf Chair for Distributed Information Systems, Friedrich Schiller University Jena, Jena, Germany
{leila.feddoul,frank.loeffler}@uni-jena.de
[2] Institute of Data Science, German Aerospace Center DLR, Jena, Germany
{leila.feddoul,sirko.schindler}@dlr.de

Abstract. Faceted Browsing is a wide-spread approach for exploratory search. Without requiring an in-depth knowledge of the domain, users can narrow down a resource set until it fits their need. An increasing amount of data is published either directly as Linked Data or is at least annotated using concepts from the Linked Data Cloud. This allows identifying commonalities and differences among resources beyond the comparison of mere string representations of metadata.

As the size of data repositories increases, so does the range of covered domains and the number of properties that can provide the basis for a new facet. Manually predefining suitable facet collections becomes impractical. We present our initial work on automatically creating suitable facets for a semantically annotated set of resources. In particular, we address two problems arising with automatic facet generation: (1) Which facets are applicable to the current set of resources and (2) which reasonably sized subset provides the best support to users?

Keywords: Faceted Browsing · Knowledge graph · Exploratory search

1 Introduction

Semantic annotations can considerably improve information retrieval by enriching resources with additional information. The Linked Data Cloud is a valuable source of such annotations. Its continuous growth in recent years, in both quantity of information and range of covered domains, enables applications to exploit semantic connections between resources, to ease information access, and to discover unexpected links across domains.

Consequently, semantic methods that are adapted to efficiently access Linked Data are gaining importance. Those methods allow to easily explore semantic data without expertise in the underlying technologies. In particular, non-expert users should not need to create complex queries to access this information.

P. Hitzler et al. (Eds.): ESWC 2019 Satellite Events, LNCS 11762, pp. 47–51, 2019.
https://doi.org/10.1007/978-3-030-32327-1_10

Faceted Browsing is a widely used technique to partition resource sets based on different dimensions. It provides a general overview of the characteristics of individual elements and allows exploring unknown data schemata. Users can leverage faceted interfaces to apply various filters, called *facets*, to incrementally refine the description of their information need. Possible tasks include searching for items with characteristic properties or exploring the whole set initially without specific goal in mind.

Considering a continuously changing and heterogeneous set of resources, manually predefining facets is often impractical. Furthermore, using concepts from *heterogeneous large scale* knowledge graph KGs, e.g., the Linked Data Cloud, for the semantic annotation of resources induces a large number of possible facets. Displaying all of them will negatively impact navigation efficiency. Hence, we require an automated method to select the most useful subset from any list of candidates matching a given collection of resources. For this purpose, we need to define metrics for measuring the "usefulness" of facets.

We focus on one component of this process that so far has not been extensively studied in the context of KGs: *inter-facet* metrics. In particular, we explore *semantic relatedness* as a measure to reduce the redundancy between selected facets while still maintaining a wide range of aspects. We adapt various techniques to develop a holistic system workflow for automatic facet generation over large scale KGs using the example of Wikidata [1].

2 Related Work

In the following, we give a short overview over selected publications on facet generation. Faceted Browsing over various data sources has been addressed by [2, 3]. In the context of Resource Description Framework (RDF) data, notable earlier research efforts include *BrowseRDF* [4], *mSpace* [5], and *Parallax* [6], followed by *gFacet* [7], and *Facete* [8]. Examples of recent works include *SemFacet* [9], and *GraFa* [10]. They focus on different aspects: *facet ranking* [4], *entity type pivoting* [6,9] *visualization* [7,8], or *indirect facet generation* [7,8].

Theoretical foundations of faceted search are defined in [9], while *performance issues* are described in [10]. Only few systems attempt to build facets over *large scale data* [6,10]. On the other hand, *domain heterogeneity* is disregarded by some approaches [4,5,8]. *Facetedpedia* [11] includes an inter-facet metric that relies on the category system of Wikipedia[1]. However, this does not provide the same generality as a generic *semantic relatedness* approach.

3 Automatic Facet Generation and Selection

Motivated by the idea that a faceted interface should not contain semantically overlapping facets, we consider the *semantic relatedness* as an inter-facet metric that can be used as an exclusion criterion. In effect, we want to prevent semantically close facets to appear together in the final result. For example, facets are

[1] https://www.wikipedia.org/.

semantically close when they are connected via an "is-a" relationship. The main reason behind this decision is to provide facets that partition the result space based on *different* aspects and thus helps avoid facets that generate the same subset of results.

In order not to overwhelm users, the number of shown facets should be limited. For this purpose, we need a ranking of candidate facets and, hence, a set of metrics to determine the degree of "usefulness". We categorize our metrics into two types: (i) *intra-facet metrics* rate the facets individually, and (ii) *inter-facet metrics* judge the relevance of a facet as part of a facet collection. Intra-facet metrics will be combined using a scoring function, mapping each facet to a score and providing an overall ranking. Inter-facet metrics will be used to decide which facets should not co-occur in the generated facet collection.

Figure 1 provides an overview of our proposed workflow for automatic facet generation and selection. It takes as input a *list of Internationalized Resource Identifiers*, e.g., the result of a keyword search, from which we generate a list of candidates using their direct properties and the first indirect ones. For example, considering an input list containing *universities*, both *location* and *location's country* are candidates, where *country* is linked to *location* and not to *university* itself.

Now, an initial filtering is performed to reduce the number of *candidates* and thereby the cost of subsequent ranking steps. For example, we remove candidates that apply only to a small subset of input IRIs. Details on this process are considered out of scope at this point.

To avoid co-occurrence of semantically related facets, we filter facets sharing a direct property. For this *selection of better categorization*, we group the facets by their direct property and only select the best-ranked candidate for further evaluation based on the already calculated intra-facet metrics. Considering the previous example, we select either *location* or *location's country*.

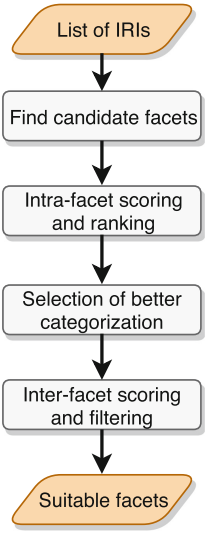

Fig. 1. Workflow for automatic facet generation and selection.

Out of the ranked list of candidates, a set of *facets that are semantically distant* from one another will be derived. For this, we consider solely the semantic relatedness of direct properties, even for facets based on indirect properties. We currently employ a structure-based relatedness measure [12] and use a selective approach: Let S be the final collection of suitable facets: (i) Initialize S with the best-ranked facet. (ii) Compare the next-best facet in terms of inter-facet metrics with the previous facets in S. (iii) Add it to S, iff it is not closely semantically related to previously chosen facets in S. (iv) Continue with Step (ii), until the desired number of facets is reached or there are no more candidates left.

After this process, S contains a collection of facets deemed suitable for the given set of resources, according to both our intra- and inter-facet metrics. These facets are now ready to be presented to users.

4 Conclusion

We propose a method for automatic facet generation including rankings based on intra- and inter-facet metrics. Our method also exploits indirect properties to find better categorizations and to create more useful facets. In particular, we focus on *semantic relatedness* as an inter-facet metric. This prevents facets in the final result that are too similar to one another and hence provide little additional assistance. Our goal is to automatically generate facets that are suitable for user navigation and consequently making a contribution in the improvement of accessibility to the semantic web for non-expert users.

In the future we aim to implement the proposed workflow, in a system that also scales with the size of big knowledge graphs like Wikidata. Furthermore, we plan an evaluation to test if our ranking approach provides facets that match user expectations and support them while browsing knowledge graphs.

References

1. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014). https://doi.org/10.1145/2629489
2. 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)
3. Tzitzikas, Y., Manolis, N., Papadakos, P.: Faceted exploration of RDF/S datasets: a survey. J. Intell. Inf. Syst. **48**(2), 329–364 (2016). https://doi.org/10.1007/s10844-016-0413-8
4. Oren, E., Delbru, R., Decker, S.: Extending faceted navigation for RDF data. In: Cruz, I., et al. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 559–572. Springer, Heidelberg (2006). https://doi.org/10.1007/11926078_40
5. Schraefel, M.C., Smith, D.A., Owens, A., Russell, A., Harris, C., Wilson, M.: The evolving mSpace platform: leveraging the semantic web on the trail of the Memex. In: Proceedings of the Sixteenth ACM Conference on Hypertext and Hypermedia, HYPERTEXT 2005, pp. 174–183. ACM, New York (2005). https://doi.org/10.1145/1083356.1083391
6. Huynh, D., Karger, D.: Parallax and companion: set-based browsing for the data web. Technical report, Metaweb Technologies Inc. (2009)
7. Heim, P., Ziegler, J., Lohmann, S.: gFacet: a browser for the web of data. In: Proceedings of the International Workshop on Interacting with Multimedia Content in the Social Semantic Web (IMC-SSW 2008), Aachen (2008)
8. Stadler, C., Martin, M., Auer, S.: Exploring the web of spatial data with Facete. In: Proceedings of the 23rd International Conference on World Wide Web, WWW 2014 Companion, pp. 175–178. ACM, New York (2014). https://doi.org/10.1145/2567948.2577022
9. Arenas, M., Grau, B.C., Kharlamov, E., Marciuska, S., Zheleznyakov, D.: Faceted search over RDF-based knowledge graphs. Web Semant. Sci. Serv. Agents World Wide Web **37**, (2016). https://doi.org/10.2139/ssrn.3199228
10. Moreno-Vega, J., Hogan, A.: GraFa: scalable faceted browsing for RDf graphs. In: Vrandečić, D., et al. (eds.) ISWC 2018. LNCS, vol. 11136, pp. 301–317. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00671-6_18

11. Li, C., Yan, N., Roy, S.B., Lisham, L., Das, G.: Facetedpedia: dynamic generation of query-dependent faceted interfaces for wikipedia. In: Proceedings of the 19th International Conference on World Wide Web, WWW 2010, pp. 651–660. ACM, New York (2010). https://doi.org/10.1145/1772690.1772757
12. Li, Y., Bandar, Z.A., Mclean, D.: An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans. Knowl. Data Eng. **15**(4), 871–882 (2003). https://doi.org/10.1109/TKDE.2003.1209005