Effect of heuristic post-processing on knowledge graph profile patterns: cross-domain study*

Gollam Rabby[1], Farhana Keya[1], Vojtěch Svátek[1], and Renzo Alva Principe[2]

[1] Prague University of Economics and Business, Prague, Czech Republic
{rabg00,keyf00,svatek}@vse.cz
[2] University of Milano-Bicocca, Milan, Italy
renzo.alvaprincipe@unimib.it

Abstract. Sets of frequent schema-level patterns characterizing a given knowledge graph (KG) represent a central output of profiling tools such as ABSTAT, as they could provide a quick overview of the coverage of the KG and its adequacy for various tasks. However, the number of patterns may be huge, and the most frequent ones might not be the most useful ones for semantically characterizing the KG, since they might feature generic (OWL, SKOS, etc.) classes and even XML data types. We hypothesize that the pattern profile suitability for a 'rapid skimming' scenario might be improved by applying a stop-list of namespaces or individual schema IRIs by which the original pattern set is pruned. We experimented with post-processing the patterns returned by ABSTAT with regard to reducing the quantity of patterns and re-ranking the patterns appearing in the first positions of the frequency-ordered results. We processed the sets of KGs from two different domains – COVID-19 and linguistics/lexicography.

Keywords: Profiling · knowledge graph · ABSTAT · pattern · linguistics · COVID-19

1 Introduction

In view of the high number and large size of knowledge graphs (KGs), which makes it difficult to rapidly identify the KG suitable for a certain application, KG *profiling* was recently introduced as a means of quantifying the structure and contents of KGs in order to judge their suitability for particular applications. Of the many kinds of quantitative and qualitative characteristics that can describe a KG, schema-level pattern of the form `<subjectType, pred, objectType>` as an abstract representation of the KG instances are particularly interesting from the point of view of knowledge engineering. Profiling tools based on schema patterns, such as ABSTAT [1] or Loupe [2], give the user certain insights to

* Supported by CHIST-ERA within the CIMPLE project (CHIST-ERA-19-XAI-003), and by Nexus Linguarum (COST Action CA18209).

frequent paths interconnecting entities at the instance level, while remaining relatively concise. The outcome is dependent on the ontology employed and on the the degree of explicit typing of entities.

The internals of these tools consist in sophisticated graph-theoretic algorithms, and some rely on massive parallelization of the computation. Yet, the results in their generic form may not always fit every kind of usage. The scenario we have particularly in mind is that of *rapid skimming through multiple KGs* in order to identify those having adequate coverage of some topic/s (contrasting to a scenario requiring detailed scrutiny of a dataset's schema). For this, the output of a state-of-the-art tool such as ABSTAT (even what is called a 'minimal', non-redundant set) still contains too many patterns that are 'boring' with respect to such skimming. In the short paper we demonstrate that a handful of post-processing heuristics can significantly prune such patterns. The study was carried out on KGs from two domains: COVID-19 and linguistics.

2 ABSTAT

ABSTAT is a scalable profiling tool that aims to support users in the exploration and understanding of large RDF KGs. Given a KG in the form of a dataset and an ontology (optional), ABSTAT computes a profile which consists of a summary about the dataset content and statistics. A summary is a set of data-driven ontology patterns in the form `<subjectType, pred, objectType>`, which represent the occurrence of the triples `<subj, pred, obj>` in the dataset. Minimalization is applied on types and properties that is, `subjectType` is a minimal type for `subj` (i.e., there is no type for `subj` that is in subsumption relation with `subjectType`), `objectType` is a minimal type of the `obj` and `subj` is linked to `obj` through `pred` or any other super-property of `pred`, hereby defining a clear distinction between patterns (a redundant pattern set) and minimal patterns. We will henceforth refer to minimal patterns as patterns. In addition, statistics such as the frequency of how many assertions in the dataset are represented by each pattern are also extracted. Alva Principe, Renzo Arturo, et al. described details regarding this KG profiling tool in [1].

The pruning effect of minimization becomes more effective when at the same time ontologies encode a rich type hierarchy and entities are mostly associated with many types (e.g., DBpedia). However, since ABSTAT is designed to summarize assertions in the KG while maintaining full coverage on them, it could be that a KG featuring many entities without a type and/or with a poor (absent) type hierarchy, fed to ABSTAT, leads to a summary with some pattern which may not be informative to the user because of its high generality.

3 Pattern stop-list

The motivation for heuristic post-processing was to eliminate the patterns that contain overly general namespaces or individual schema IRIs, so that ideally only patterns expressing ontological relationships properly characterizing the KG are

Class/property/datatype	L	C
http://www.w3.org/2002/07/owl#Thing	12	5
http://www.w3.org/2000/01/rdf-schema#label	7	4
http://www.w3.org/2001/XMLSchema#string	6	1
http://www.w3.org/2001/XMLSchema#integer	0	1
http://www.w3.org/2002/07/owl#Class	1	2
http://www.w3.org/2001/XMLSchema#float	0	1
http://www.w3.org/2008/05/skos-xl#Label	7	4
http://www.w3.org/2000/01/rdf-schema#literal	8	5
http://www.w3.org/2002/07/owl#namedindividual	1	0
http://www.w3.org/2002/07/owl#ontology	3	0
http://www.w3.org/2001/xmlschema#date	0	1
http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#String	3	0
http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#rfc5147string	2	0
http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#endindex	2	0
http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#beginindex	2	0
http://www.arg.dundee.ac.uk/aif#I-node	0	1
http://lemon-model.net/lemon#form	3	0

Table 1: Counts of KGs (L for linguistic and C for COVID-19 ones) for which at least one pattern was removed from the ABSTAT results because of a particular stop-listed class/property/datatype.

left (thus also reducing the overall size of the pattern set), or at least prioritized in the list. For finding the stop-list, we first analyzed the counts of KGs in whose ABSTAT profiles a particular concept appeared. The stop-list of schema IRIs with the number of KGs in whose profiles they appeared is in Table 1; we see that these are mostly datatype IRIs, or RDFS/OWL namespaces.We however also included some generic item that only appeared in few or even one KGs but were frequent there and thus had a strong pruning effect, such as, for example, http://www.arg.dundee.ac.uk/aif#I-node.

4 Experiments

As a first step, we processed all the KGs by ABSTAT; since we worked with the public web application, which has a maximum KG upload limit of 10 GB, this led to reduction of the number of KGs. More precisely, the KGs used in this experiment come from the linguistic and the COVID-19 domain and are listed in Table 2. We observe that KGs are very heterogeneous, for instance there are KGs that barely or not at all provide types for entities (the ones with low %TA), KGs with almost only assertions linking entities with literals (the ones with very high %DRA), KGs with only assertions linking entities (high %ORA) and also more balanced KGs.

Once profiles are computed, ABSTAT returns a set of patterns. Then we applied customizable heuristic post-processing relying on the stop-list (Table 1).

KG name	Before	After	TA	DRA	ORA	Type
basque-eurowordnet-lemon-lexicon-3.0	74	47	14%	25%	62%	L
catalan-eurowordnet-lemon-lexicon-3.0	78	47	14%	26%	60%	L
dbpedia-spotlight-nif-ner-corpu	52	37	35%	35%	31%	L
galician-eurowordnet-lemon-lexicon-3.0	74	47	15%	28%	57%	L
apertium-rdf-ca-it	15	2	0%	0%	100%	L
wordnet	39	36	27%	27%	66%	L
wn-wiki-instances	4	0	0%	0%	100%	L
asit-data	67	52	17%	33%	50%	L
Reuters-128	21	15	18%	53%	29%	L
clld-sails-sources	9	4	0%	67%	33%	L
lemonwiktionary	19	0	7%	23%	70%	L
apertium-rdf-en-es	8	0	0%	0%	100%	L
rss-500-nif-ner-corpus	18	1	25%	50%	25%	L
linked-hypernyms	0	0	100%	0%	0%	L
apertium-rdf-fr-ca	16	0	0%	0%	100%	L
galician-eurowordnet-lemon-lexicon-3.0	74	32	15%	28%	57%	L
SimpleEntries	4752	4445	36%	36%	27%	L
news-100-nif-ner-corpus	21	15	16%	55%	29%	L
drugbank	1408	13	6%	85%	9%	C
pro-sars2	12	0	7%	85%	8%	C
COKG-19	8	0	0%	98%	2%	C
COKG-19-Schema	7	0	0%	58%	42%	C
cord19-akg	108	55	39%	10%	51%	C

Table 2: Patterns before and after post-processing for Linguistic and COVID-19 KGs. For each raw KG, the percentage of typing assertions (%TA), datatype relational assertions (%DRA) and object relational assertions (%ORA) w.r.t. the full number of assertions are also provided.

The post-processing tool provides, for each element of the stop-list, the options “None”, “Put to the bottom” and “Remove”. In Table 2, we present the pattern frequency difference for the linguistic and COVID-19 KGs, upon application of the “Remove” option. For some KGs such as WordNet, asit-data, etc the difference is tiny, while for most others it is quite significant, outliers being apertium-rdf-ca-it or DrugBank having a reduction by two orders of magnitude. So, from Table 2, we can say that, for most of the KGs, ABSTAT pattern post-processing has a huge impact. The top 10 patterns before and after post-processing are available from an auxiliary page [3].

The reduction seems to be even more significant for the COVID-19 KGs than for the linguistic KGs, although the numbers of KGs are too small to make ultimate conclusions. A possible explanation is that since the COVID-19 KGs covered mainly contain textual annotations of scientific literature (rather than true conceptual relationships from the medical domain), their structure is mostly shallow, with dominance of data properties. Moreover, we can observe

[3] https://github.com/corei5/ABSTAT-patterns-post-processing

from Table 2 that the stop-list heuristic has a higher effect on (1) KGs with a very low percentage of typing assertions (e.g. apertium and lemonwiktionary KGs) as ABSTAT by default assigns `owl:Thing` as type for un-typed entities and (2) KGs with a majority of data type relational assertions (e.g., COVID-19 KGs except for CORD19-AKG) as many of the elements in the stop-list are referred to datatypes. On the other hand, the least affected KGs are those that the most balanced ones such as Wordnet and SimpleEntries.

While in this study we primarily aimed at reduction of the total count of patterns, the option "Put to the bottom" is also offered by the post-processing tool, since even patterns containing generic concepts and datatypes can actually be interesting, in particular for the subsequent detailed scrutiny of a chosen dataset. For example, KGs related to people or companies may contain informative datatype properties (containing information such as birth date, name, alias, web page, address, revenue, foundation date, number of employees, etc). After the familiarization with the essential nature of a KG, through the 'prioritized' patterns, the user may wish to study even such 'de-prioritized' patterns in the bottom of the list.

5 Conclusions and future work

The experiment suggests that simple heuristics leading to suppression of patterns containing generic concepts or datatypes might improve the output of state-of-art profiling tools in the context of rapid skimming of multiple KGs.

While the experiment was carried out via a separate post-processing tool, we will explore how a similar functionality could be achieved within ABSTAT itself, without compromising its current user experience or risking inadequate information loss. Additionally, while the present method of stop-list design was purely manual, we could also consider automated methods leveraging on a large number of previously created KG profiles. In particular, overly generic concepts that occur in a large proportion of KGs could be eliminated by applying a threshold value on the *inverse KG frequency* (analogous to the common IDF metric). One more potential feature of ABSTAT, which could provide a basis for pattern filtering, could be, for each pattern, the proportion of underlying triples having both the subject and object *within the given KG* (intra-KG links) and of triples the object of which is from an *external KG* (inter-KG links). In some profiling scenarios both can be equally important, while in some other focus would be given on patterns derived from either intra-KG or inter-KG links.

References

1. Alva Principe, R.A., Maurino, A., Palmonari, M. et al. ABSTAT-HD: a scalable tool for profiling very large knowledge graphs. The VLDB Journal (2021). https://doi.org/10.1007/s00778-021-00704-2
2. Mihindukulasooriya, N., Poveda-Villalón, M., García-Castro, R.,Gómez-Pérez, A.: Loupe-an online tool for inspecting datasets in the linked data cloud. In: International Semantic Web Conference (Posters and Demos) (2015)