Log File Analysis for Web of Data Endpoints

Markus Luczak-Rösch and Hannes Mühleisen

Freie Universität Berlin, Institute of Computer Science,
Networked Information Systems,
Berlin D-14195, Germany,
luczak|muehleis@inf.fu-berlin.de,
http://www.ag-nbi.de

1 Introduction

Since the usage of Linked Data has evolved that far, that linked datasets can be seen as one of the most actively used application areas of ontologies on the Web, usage analysis in the context of this Web of data is an interesting and emerging research topic. Eventhough it is not a requirement of a Linked Data endpoint to offer a SPARQL endpoint lots of dataset providers on the so called *Web of data* do so. Especially the SPARQL requests against these endpoints carry an additional and very often extrinsic view to served data, due to the structure of basic graph patterns in queries. Thus it is promising to evaluate them in a special way and to use the retrieved information for the maintenance of datasets. To date we recognized one other approach that also focuses usage analysis and statistics in the context of the Web of Data. Möller et al.[1] analyze user clients, requested content and in a preliminary form the structure of SPARQL queries. In this poster proposal we introduce our preprocessing method for an in-depth analysis of log files of Web of Data endpoints.

2 Log File Preprocessing for Web of Data Usage Analysis

Our approach is intended to run on server log files following the extended common log format[1]. These logs contain information about the direct access to RDF resources and about SPARQL queries. The first step of our preprocessing is to clean the log from all entries that contain 40x and 50x response codes. Afterwards we transform all request for primitives resources into SPARQL DESCRIBE queries and retrieve a normalized view to the usage of the dataset on the level of SPARQL queries. It is our goal to retrieve information about (1) which graph and triple patterns cause non empty/empty result sets, (2) to which classes do requested resources belong to very often/seldom, (3) which predicates are requested very often/seldom (in combination with resources of certain types) and (4) which filters are used and how they effect on the result sets. So we generate a database that reflects information about the success of the data access of the original request in the log entry as shown in Figure 1. We perform this step

[1] http://www.w3.org/TR/WD-logfile.html

for all (1) basic graph patterns, (2) triple patterns and (2) filters of each single query. That means that each of these atomic parts and also the original query itself is re-ran against a mirror of the dataset of which the usage data is from, to check if this yields a non-empty result set and which resource or predicate in the query is actually existing in the data and the underlying ontology.

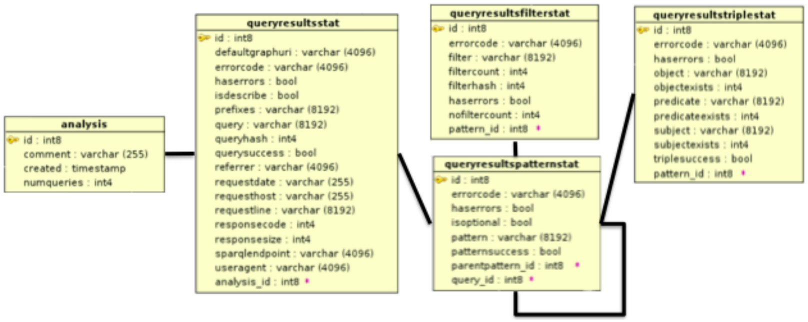

Fig. 1. The resulting database schema of the preprocessing

3 Potential for Web of Data Usage Mining

We preprocessed logfiles from the DBpedia[2] dataset, the Semantic Web Dog Food server[3] and the Linked Open Geo Data endpoint[4]. We detected that the results are very promising for various areas such as index selection for triple stores based on the analysis of the used SPARQL patterns so we are currently working on studies on this. We also visualized several ontology heat maps using a network analysis tool based on statistics on the collected data. Our ongoing work in this field is to perform further statistical analysis as well as data mining methods on our data, such as sequencial pattern analysis and clustering, with the goal to recommend ontology maintenance efforts to be performed by dataset hosts.

References

1. Möller, K., Hausenblas, M., Cyganiak, R., Grimnes, G.A.: Learning from linked open data usage: Patterns metrics. In: Proceedings of the WebSci10: Extending the Frontiers of Society On-Line. (2010)

[2] http://dbpedia.org
[3] http://data.semanticweb.org
[4] http://linkedgeodata.org