The Schema-Agnostic Queries (SAQ-2015) Semantic Web Challenge: Task Description

André Freitas[1(✉)] and Christina Unger[2]

[1] Department of Computer Science and Mathematics, University of Passau, Passau, Germany
andre.freitas@uni-passau.de
[2] Semantic Computing Group, Cognitive Interaction Technology, Center of Excellence (CITEC), Bielefeld University, Bielefeld, Germany

Abstract. As datasets grow in schema-size and heterogeneity, the development of infrastructures which can support users querying and exploring the data, without the need to fully understand the conceptual model behind it, becomes a fundamental functionality for contemporary data management. The first edition of the Schema-agnostic Queries Semantic Web Challenge (SAQ-2015) aims at creating a test collection to evaluate *schema-agnostic/schema-free* query mechanisms, i.e. mechanisms which are able to semantically match user queries expressed in their own vocabulary to dataset elements, allowing users to be partially or fully abstracted from the representation of the data.

1 Introduction

The evolution of data environments towards the consumption of data from multiple data sources and the growth in the schema size, complexity, dynamicity and decentralisation (SCoDD) of data [4,7] increases the complexity of contemporary data management. The SCoDD trend emerges as a central data management concern in Big Data scenarios, where users and applications have a demand for more complete data, produced by independent data sources, under different semantic assumptions and contexts of use, which is the typical scenario for Semantic Web/Linked Data applications.

The evolution of databases in the direction of heterogeneous data environments strongly impacts the usability, semiotic and semantic assumptions behind existing data accessibility methods such as structured queries, keyword-based search and visual query systems. With schema-less databases containing potentially millions of dynamically changing attributes, it becomes unfeasible for some users to become aware of the 'schema' or vocabulary in order to query the database. At this scale, the effort in understanding the schema in order to build a structured query can become prohibitive.

This Semantic Web Challenge focuses on catalyzing the development and evaluation of methods and tools which can help data consumers to query structured data without the understanding of the representation behind the data.

F. Gandon et al. (Eds.): SemWebEval 2015, CCIS 548, pp. 191–198, 2015.
DOI: 10.1007/978-3-319-25518-7_16

At the center of this discussion is the semantic gap between users and databases, which becomes more central as the scale and complexity of the data grows. Addressing this gap is a fundamental part of the Semantic Web vision.

Schema-agnostic query mechanisms aim at allowing users to be abstracted from the representation of the data, supporting the automatic matching between queries and databases [1,2,5]. This challenge aims at emphasizing the role of schema-agnosticism as a key requirement for contemporary database management, by providing a test collection for evaluating flexible query and search systems over structured data in terms of their level of *schema-agnosticism* (i.e. their ability to map a query issued with the users' terminology and structure, mapping it to the dataset vocabulary). The challenge is instantiated in the context of Semantic Web datasets.

2 Schema-Agnostic Queries

Schema-agnostic queries can be defined as query approaches over structured databases which allow users satisfying complex information needs without the understanding of the representation (schema) of the database. Similarly, [5] defines it as "search approaches, which do not require users to know the schema underlying the data". Approaches such as keyword-based search over databases allow users to query databases without employing structured queries. However, as discussed by [5]: "From these points, users however have to do further navigation and exploration to address complex information needs. Unlike keyword search used on the Web, which focuses on simple needs, the keyword search elaborated here is used to obtain more complex results. Instead of a single set of resources, the goal is to compute complex sets of resources and their relations".

The development of approaches to support natural language interfaces (NLI) over databases have aimed towards the goal of schema-agnostic queries. Complementarily, some approaches based on keyword search have targeted keyword-based queries which express more complex information needs. Other approaches have explored the construction of structured queries over databases where schema constraints can be relaxed. All these approaches (natural language, keyword-based search and structured queries) have targeted different degrees of sophistication in addressing the problem of supporting a flexible semantic matching between queries and data, which vary from the completely absence of the semantic concern to more principled semantic models.

While the demand for schema-agnosticism has been an implicit requirement across semantic search and natural language query systems over structured data, it is not sufficiently individuated as a concept and as a necessary requirement for contemporary database management systems. Recent works have started to define and model the semantic aspects involved on schema-agnostic queries [1,2,5].

3 Challenge Description

The challenge aims at providing an evaluation test collection for schema-agnostic query mechanisms, focusing on Semantic Web scenarios. The large-schema and

semantically heterogeneous nature of Semantic Web datasets brings schema-agnosticism as a fundamental data management concern for this community.

The test collection supports the quantitative and qualitative evaluation of degree of schema-agnosticism of different approaches. Since addressing schema-agnostic queries is dependent on semantic approaches which need to cope with different types of semantic matching between query and dataset, the test collection explores different categories of semantic phenomena involved in the challenge of matching schema-agnostic queries. Each query is categorized according to the semantic mapping types. This categorization supports a fine-grained qualitative and quantitative interpretation of the evaluation results.

4 Evaluation Description

The challenge provides a gold standard with the correct answers for each *schema-agnostic query*. Queries are issued over DBpedia 3.10. A training dataset consisting of 30 queries is be made available for the participants. In order to participate in the challenge, each system submitted the results in the format proposed by the challenge. The organizers then automatically calculated *precision*, *recall*, *mean reciprocal rank* for each query and the associated averages. Participants are recommended to submit their *query execution time*, *dataset semantic enrichment time*, and *user-interaction and disambiguation effort*.

The challenge consists of addressing a set of 103 schema-agnostic queries over DBpedia 2014[1] and associated YAGO classes[2]. The training and test sets are available at[3].

The schema-agnostic queries were derived from the natural languages present at the Question Answering over Linked Data (QALD-4) test collection [6]. These natural language questions were manually converted to schema-agnostic queries, preserving its vocabulary and using a consistent set of conversion guidelines.

Two categories of schema-agnostic queries (tasks) are available: *schema-agnostic SPARQL query* and *schema-agnostic keyword query*. Evaluation systems can compete in one or in both categories.

4.1 Schema-Agnostic SPARQL Query

Consists of schema-agnostic queries following the syntax of the SPARQL standard without namespace prefixes. The syntax and semantics of operators are maintained, while different terminologies are used.

Example I:

```
SELECT ?y {
  BillClinton hasDaughter ?x .
  ?x marriedTo ?y .
}
```

[1] http://wiki.dbpedia.org/Downloads2014.

[2] http://data.dws.informatik.uni-mannheim.de/dbpedia/2014/links/yago_types.nt.bz2.

[3] https://sites.google.com/site/eswcsaq2015/resources.

which maps to the following SPARQL query in the dataset vocabulary:

```
PREFIX : <http://dbpedia.org/resource/>
PREFIX dbpedia2: <http://dbpedia.org/property/>
PREFIX dbpedia: <http://dbpedia.org/ontology/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX dbo: <http://dbpedia.org/ontology/>

SELECT   ?y  {
 :Bill_Clinton dbpedia:child ?x .
 ?x dbpedia2:spouse ?y .
 }
```

Example II:

```
SELECT   ?x {
         ?x isA book .
         ?x by William_Goldman .
         ?x has_pages ?p .
         FILTER (?p > 300) .
  }
```

which maps to the following SPARQL query in the dataset vocabulary:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX : <http://dbpedia.org/resource/>
PREFIX dbpedia2: <http://dbpedia.org/property/>
PREFIX dbpedia: <http://dbpedia.org/ontology/>
SELECT ?x {
         ?x rdf:type dbpedia:Book .
         ?x dbpedia2:author :William_Goldman .
         ?x dbpedia:numberOfPages ?p .
                 FILTER(?p > 300) .
}
```

4.2 Schema-Agnostic Keyword Query

Consists of schema-agnostic queries using keyword queries. In this case the syntax and semantics of operators are different from the SPARQL syntax.

Example I: “Bill Clinton daughter married to”

Example II: “Books by William Goldman with more than 300 pages”

4.3 Returned Result

In order to participate in the challenge, systems submitted the results in the format proposed by the challenge. For queries which return a list of URIs (uri1, uri2) or values:

```
<dataset id="saq-2015_test">
<query id="1">
<answers>
<answer> uri1 </answer>
<answer> uri2 </answer>
</answers>
</query>

<query id="2">
<answers>
<answer> value </answer>
</answers>
</query>

</dataset>
```

For queries of the type YES/NO:

```
<dataset id="saq-2015_test">
<query id="1">
<answers>
<answer> true </answer>
</answers>
</query>
</dataset>
```

Teams had 24 h after receiving the test query set to return their results.

5 Schema-Agnostic Mappings

In the test set, each schema-agnostic query contains a classification of the query-data alignments. For example:

```
<query id="14">
<keyword_query lang="en">
<![CDATA[ships called after Benjamin Franklin]]>
</keyword_query>
<schema_agnostic_query>
<![CDATA[
SELECT DISTINCT ?uri
WHERE {
        ?uri type Ship .
        ?uri calledAfter Benjamin_Franklin .
}
]]>
```

```
</schema_agnostic_query>
<resolved_query><![CDATA[
PREFIX res: <http://dbpedia.org/resource/>
PREFIX dbp: <http://dbpedia.org/property/>
SELECT DISTINCT ?uri
WHERE {
        ?uri dbp:shipNamesake res:Benjamin_Franklin.
}
]]>
</resolved_query>
<alignments>
<alignment> Ship (c o) -> shipNamesake (p) | substring </alignment>
<alignment> Benjamin_Franklin (i o) -> Benjamin_Franklin (i o) | substring </alignment>
<alignment> calledAfter (p) -> shipNamesake (p) | related </alignment>
<op> select -> select </op>
</alignments>
<answers>
<answer>http://dbpedia.org/resource/HMS_Canopus_(1798)</answer>
<answer>http://dbpedia.org/resource/USS_Franklin_(1815)</answer>
<answer>http://dbpedia.org/resource/USS_Franklin_(1795)</answer>
<answer>http://dbpedia.org/resource/Ben_Franklin_(PX-15)</answer>
</answers>
</query>
```

In the alignment below, the schema-agnostic query term 'calledAfter' is associated with a predicate '(p)' data type, mapping to the predicate 'shipNamesake' in the dataset, and that the type of relationship between two terms are described as *semantically related*.

```
<alignment> calledAfter (p) -> shipNamesake (p) | related </alignment>
```

Alignments are categorized according to 6 categories:

- **semantically related:** If a query term and its associated database entity are *semantically related*. Example: *languageOf* in the query maps to *spokenIn* in the dataset.
- **semantically similar:** If a query term and its associated database entity are *semantically similar*, i.e. it follows a taxonomic relation. Example: *wifeOf* in the query maps to *spouseOf* in the dataset.
- **synonym:** If a query term and its associated database entity are *synonyms*. Example: *startDate* in the query maps to *beginDate* in the dataset.
- **string similar:** If a query term has a *string similarity* relationship to its associated database entity. Example: *startDate* in the query maps to *beginDate* in the dataset.
- **substring:** If a query term is a *substring* of its associated database entity or vice-versa. Example: *wifeOf* in the query maps to *wife* in the dataset.
- **functional content:** Consists on the mapping of *function words* (e.g. prepositions) in the query to *other function words* or *content words* in the dataset entity. Example: *in* in the query maps to *location* in the dataset.
- **abbreviation:** If a query term is an *abbreviation* of its associated database entity or vice-versa. Example: *extinct* in the query maps to *'EX'* in the dataset.

Other examples of alignments (including compositions of different categories) include:

```
<alignment> languageOf (p) -> spokenIn (p) | related </alignment>
<alignment> writtenBy (p) -> author (p) | substring, related </alignment>
<alignment> in (p) -> location (p) | functional_content </alignment>
<alignment> in (p) -> isPartOf (p) | functional_content </alignment>
<alignment> FemaleFirstName (c o) -> gender (p) | substring, related </alignment>
<alignment> state (p) -> locatedInArea (p) | related </alignment>
<alignment> extinct (p) -> conservationStatus (p) | related </alignment>
<alignment> extinct (p) -> 'EX' (v o) | substring, abbreviation </alignment>
<alignment> startAt (p) -> sourceCountry (p) | substring, synonym </alignment>
<alignment> U.S._State (c o) -> StatesOfTheUnitedStates (c o) | string_similar </alignment>
<alignment> calledAfter (p) -> shipNamesake (p) | related </alignment>
<alignment> wifeOf (p) -> spouse (p) | substring, similar </alignment>
<alignment> constructionDate (p) -> beginningDate (p) | substring, related </alignment>
```

Alignment terms are classified according to their data model types, with regard to the position within the triple (*subject* (s), *predicate* (p), *object* (o)) and entity type (*instance* (i), *class* (c), *property* (p), *value* (v)).

The alignment classifications are a simplification of the schema-agnostic alignments described in [1].

6 Results

Just one system competed officially in the SAQ-2015 Semantic Web Challenge: the UMBC_Ebiquity-SFQ system from the University of Maryland Baltimore County (Syed et al. [3]).

The results are described in Table 1:

Table 1. Evaluation of the participating system for the SAQ-2015 challenge.

System	Avg. precision	Avg. recall	Avg. f1-measure	% of answered queries
UMBC_Equity-SFQ	0.33	0.36	0.31	0.44

7 Summary

The ability to abstract users from the specifics of the representation of the data, including its vocabulary and structural relations is a fundamental functionality for large-scale and heterogeneous data. The Schema-agnostic Queries Semantic Web Challenge (SAQ-2015) aims at providing a test collection for supporting the development of schema-agnostic query mechanisms, i.e. query approaches which supports automatically crossing the semantic gap between users and the data. The test collection provides a categorized set of schema-agnostic queries, covering a range of different alignments from string variations to different types

of semantic relations. The performance of the participating system indicates that state-of-the-art systems are able to provide an initial solution for the problem. However, the initial results show that schema-agnostic queries are still a challenging problem and that there is space for major improvements.

References

1. Freitas, A., Da Silva, J.C.P., Curry, E.: On the semantic mapping of schema-agnostic queries: a preliminary study. In: 13th International Semantic Web Conference (ISWC) Workshop of the Natural Language Interfaces for the Web of Data (NLIWoD), Rival del Garda (2014)
2. Bischof, S., Krötzsch, M., Polleres, A., Rudolph, S.: Schema-agnostic query rewriting in SPARQL 1.1. In: Mika, P., et al. (eds.) ISWC 2014, Part I. LNCS, vol. 8796, pp. 584–600. Springer, Heidelberg (2014)
3. Syed, Z.: UMBC_Ebiquty-SFQ: schema free querying system. In: 12th Extended Semantic Web Conference on SAQ-2015 Semantic Web Challenge (ESWC) (2015)
4. Helland, P.: If you have too much data, then 'good enough' is good enough. Commun. ACM **54**(6), 40–47 (2011)
5. Tran, T., Mathäß, T., Haase, P.: Usability of keyword-driven schema-agnostic search. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) ESWC 2010, Part II. LNCS, vol. 6089, pp. 349–364. Springer, Heidelberg (2010)
6. Unger, C., et al.: Question answering over linked data (QALD-4). In: Proceedings of CLEF (2014)
7. Brodie, M.L., Liu, J.T.: The power and limits of relational technology in the age of information ecosystems. In: Keynote, On The Move Federated Conferences, Heraklion, Greece, 25–29 October 2010