OWLIM Reasoning over FactForge

Barry Bishop, Atanas Kiryakov, Zdravko Tashev, Mariana Damova, Kiril Simov

Ontotext AD, 135 Tsarigradsko Chaussee, Sofia 1784, Bulgaria

Abstract. In this paper we present the reasoning mechanism in the OWLIM family of semantic repositories, which is based on materialization. This mechanism is evaluated using a combination of datasets from the Linked Open Data cloud in a public service called FactForge, where the benefits of materialization are manifested in improved SPARQL query performance.

Keywords: LOD, materialization, OWLIM, RDF, semantic repository

1 Introduction

In this paper we present the reasoning mechanism employed in the OWLIM family of semantic repositories. These native RDF databases are implemented in Java and comprise storage components, inference-engine and query-answering engine. They are available in three editions: OWLIM-Lite, an in-memory and very fast RDF database that can load data at over 50,000 statements per second on a 1,000 USD machine using non-trivial inference; OWLIM-SE, that uses file-based, paged indices and data structures to be able to process tens of billions of RDF statements on standard desktop hardware; and OWLIM-Enterprise, a replication cluster based on OWLIM-SE that provides resilience and linearly scalable parallel query performance. OWLIM-Lite is free-for-use, whereas OWLIM-SE and OWLIM-Enterprise are the commercial editions licensed per CPU core. OWLIM-SE and OWLIM-Enterprise use a number of storage and query optimizations that allow it to sustain outstanding insert and delete performance even when managing tens of billions of statements of linked open data.

The experiments conducted were performed using datasets from the Linked Open Data cloud, see section 3, that constitute a reason-able view [2] named FactForge[1]. Entities described in more than one dataset are unified via `owl:SameAs` statements and a common ontology PROTON[2], called a unification ontology [1] for FactForge used for querying and data integration. PROTON is mapped to DBPedia[3], FreeBase[4] and Geonames[5]. Query performance with such dataset sizes is like-wise good, with sub-second response times for all the example queries found on the FactForge site.

---

1 http://factforge.net/

2 http://www.ontotext.com/proton-ontology

3 http://dbpedia.org/About

4 http://www.freebase.com/

5 http://www.geonames.org/

2 Reasoning in OWLIM

The inferencing strategy in OWLIM is one of materialization based on R-Entailment as defined by ter Horst [3], where Datalog like rules with inequality constraints operate directly on a single ternary relation that represents all triples. In addition, free variables in rule heads are treated as blank nodes. Materialization involves computing all the entailed statements at load time. While this introduces additional reasoning cost when loading statements into a repository, the desirable consequence is that query evaluation can proceed extremely quickly. Several standard rule sets are included in all editions of OWLIM and these include:

empty – no inference;

rdfs[6] – RDFS semantics using rule entailment, but without data-type reasoning, i.e. without the literal generalization and related rules;

owl-horst – equivalent to pD*, again without data-type reasoning;

owl-max – RDFS and OWL-Lite (that can be captured in rules);

owl2-ql – a fragment of OWL2 Full based on DL-Lite$_R$, a variant of DL-Lite that does not require the unique name assumption;

owl2-rl[7] – the OWL2 RL profile, a fragment of OWL2 Full amenable to implementation on rule-engines, but without data-type reasoning.

In addition to the standard semantics, user-defined rule-sets can be used. In this case the user provides the full pathname to a custom rule file that contains definitions of axiomatic triples, rules and consistency checks. For ease of use, the rule files for the standard rule-sets are included in the distribution and users can modify or extend these for their specific purposes.

Consistency checks are used to ensure that the data model is in a consistent state and are applied whenever an update transaction is committed, for example to ensure that `owl:Nothing` has no members or that no pair of individuals have both `owl:sameAs` and `owl:differentFrom` relationships.

During loading, all inferred statements are materialized, except those generated as a result of the semantics of `owl:sameAs`. OWLIM-SE uses special data structures to maintain equivalence classes and uses the URI of the first asserted resource in each equivalence class in the statement indices. This allows for the correct expansion of results during query-answering while keeping the index sizes manageable. This technique has the further advantage that it can be switched off during query answering in order to limit the number of 'duplicate' results.

3 FactForge - a Reason-able View on LOD

FactForge is a reason-able view [] to the Web of Linked Data, made up of 11 of the central LOD datasets, which have been selected and refined in order to serve as a useful index and entry point to the LOD cloud and to present a good use-case for large-scale reasoning and data integration. The compound dataset of FactForge is the largest

[6] http://www.w3.org/TR/rdf-schema/

[7] http://www.w3.org/TR/owl2-profiles/

body of heterogeneous general knowledge on which inference has been performed. It counts 1.7 billion explicit statements; 15 billion retrievable statements available after inference and owl:sameAs expansion (cf. section 2); including 1.4 billion inferred statements. The datasets combined in FactForge are:

- DBPedia - an RDF dataset derived from Wikipedia, designed to provide as full as possible coverage of the factual knowledge that can be extracted from the InfoBoxes of Wikipedia with a high level of precision;
- Freebase - a dataset containing information about 11 million things, including movies, books, locations, companies and more, with underlying schema based on properties, and not ontologies, which exploits user generated categories;
- Geonames - a geographic database that covers 6 million of the most significant geographical features on Earth, characterised by coordinates and relations to other features (e.g. 'parent feature' in which the feature is nested);
- CIA World Factbook[8] - a collection of structured data, including statistical, geographic, political, and other information about all countries;
- Lingvoj[9] - providing descriptions of the most popular human languages; currently it contains information about more than 500 languages;
- MusicBrainz[10] (RDF from Zitgist) – a comprehensive music information suitable for browsing or useful for tagging;
- WordNet[11] - a lexical database of English. Nouns, verbs, adjectives and adverbs that are grouped into sets of cognitive synonyms (synsets).

The interlinking of datasets is facilitated by DBPedia, which provides link-sets of owl:sameAs links of DBpedia with GeoNames, Lingvoj, Freebase, MusicBrainz, UMBEL and Wordnet. These link-sets are also loaded into FactForge along with the following ontologies and schemata:

- DCMI Metadata Terms[12] (Dublin Core - DC) - a relatively small, but very popular metadata schema. It defines attributes that can be used to describe information resources;
- SKOS[13] (Simple Knowledge Organization System) - a relatively simple RDF schema for describing taxonomies of concepts linked to each other by any sort of subsumption hierarchy;
- RSS - an RDF schema designed to enable syndication of machine-readable information about updates from Web sites;
- FOAF - an ontology for defining and linking personal profiles on the Web.

FactForge provides several methods to explore the combined dataset that exploits some of the advanced features of OWLIM-SE. Firstly, 'RDF Search and Explore' allows entities to be searched by keyword with a real-time auto-suggest feature ordered by 'RDF Rank' (similar to Google's Page Rank). The results page shows all triples where the selected node appears as the subject, predicate or object, together with the

8 http://www4.wiwiss.fu-berlin.de/factbook/
9 http://lingvoj.org/
10 http://musicbrainz.org/
11 http://wordnet.princeton.edu/
12 http://dublincore.org/
13 http://www.w3.org/2004/02/skos/

preferred label, RDF Rank indicator, etc. Secondly, a SPARQL page allows users to write their own queries with clickable options to add each of the known namespaces. The results are presented in a conveniently formatted table with the option to download results in various formats (SPARQL/XML, JSON, etc). Lastly, a graphical search facility called 'RelFinder' [4] that discovers paths between selected nodes. This is a computationally intensive activity and the results are displayed and updated dynamically during each iteration. The resulting graph can be reshaped by the user with simple click and drag operations. Entities within the emerging graph can be selected and a properties box provides links to the sources of information.

4 PROTON - Unification Ontology for FactForge

In addition to the above, FactForge also uses an ontology called PROTON (developed by Ontotext) to unify concepts in the main datasets. The PROTON ontology is a lightweight, upper-level ontology serving as a modelling basis for a number of tasks in different domains. PROTON is meant to serve as a seed for ontology generation, i.e. new ontologies constructed by extending PROTON. It can also be used for automatic entity recognition and more generally Information Extraction (IE) from text for the purpose of semantic annotation (metadata generation). The PROTON ontology contains about 500 classes and 150 properties, providing coverage of the general concepts necessary for a wide range of tasks. The design principles can be summarized as follows: (1) domain-independence; (2) light-weight logical definitions; (3) alignment with popular metadata standards; (4) good coverage of named entity types and concrete domains, e.g. people, organizations, locations, numbers, etc.; and (5) good coverage of instance data in Linked Open Data Reason-able views.

The ontology is encoded in a fragment of OWL Lite and split into two modules: Top and Extent. Top module is an upper ontology covering some basic philosophical distinctions between entity types, such as: `Object` – existing entities (agents, locations, vehicles); `Happening` – events and situations; `Abstract` – abstractions that are neither objects nor happenings. The Top module also contains the main classes for each of these types. The Extent module contains more domain and application oriented classes. In FactForge PROTON is used to join the ontological classes and properties of the main datasets. The mapping between PROTON and a given dataset ontology is done in three different ways: (1) using `rdfs:subClassOf` statements between classes in both ontologies and `rdfs:subPropertyOf` for properties; (2) using OWL expressions in the mappings where there is a difference in the conceptualization in both ontologies; and (3) using inference rules in cases where additional individuals are necessary in the repository in order to support the mapping. Only the PROTON ontology is loaded in FactForge. In this way the conceptual structure implied by the particular dataset ontologies is ignored and only the PROTON definitions are presented.

5 Evaluation

Table 1 shows the loading statistics for FactForge datasets. The figures are given in thousands, note ('000) in the header of the columns. The first column lists the datasets loaded. The column "Explicit Indexed Triples" shows the number of explicit facts loaded. The column "Inferred Indexed Triples" presents the number of triples that were generated as a result of the materialization during loading. The column "Total # of Indexed Triples" gives the sum of explicit and implicit triples loaded in OWLIM. The column "Entities" outlines the number of nodes in the graph generated for each dataset, and column "Inferred closure ratio" indicates the number of inferred triples per number of explicit triples loaded.

Dataset	Dataset Statistics				
	Explicit Indexed Triples ('000)	Inferred Indexed Triples ('000)	Total # of Indexed Triples ('000)	Entities ('000 of nodes in the graph)	Inferred closure ratio
Sechmata (RSS, DC) ontologies (geonames)	5	5	10	3	0,9
DBpedia (sameAs)	15 706	0	15 706	24 778	0
NY times	346	550	896	196	1,6
MusicBrainz	198 418	103 757	302 175	54 834	0,5
Lingvoj 2012 + ontology	22	27	49	20	1,2
Lexvo	693	542	1 235	584	0,8
CIA Factbook	40	39	79	24	1
Wordnet	2 724	13 234	15 959	1 081	4,9
Geonames 2.2.1	107 832	194 040	301 872	42 758	1,8
DBpedia core 3.7	659 738	205 602	865 341	155 209	0,3
Freebase	705 161	233 026	938 187	196 947	0,3
Proton	6	637 942	637 948	4	115 297,70
Total	**1 690 691**	**1 388 764**	**3 079 456**	**476 437**	**0,8**

Table 1 Statistics over statements loaded in FactForge

The effects of the materialization and the `owl:sameAs` optimization described in section 2 above result in 79% index compression, which means that close to 12 billion triples that are not indexed are available for querying, making the total of retrievable triples of FactForge close to 15 billion. The loading speed amounts to 8 032 explicit indexed statements per second, and 14 630 indexed statements per second on a CPU - 2 x Intel Xeon X5690, 3.46GHz, 12MB cache, 6 Core, RAM - 144 GB machine.

The utility of reasoning becomes apparent during the evaluation of SPARQL queries. For example the following SPARQL query about Mass media companies in Europe which uses PROTON predicates only:

```
PREFIX ptop: <http://proton.semanticweb.org/protontop#>
PREFIX pext: <http://proton.semanticweb.org/protonext#>
PREFIX dbpedia: <http://dbpedia.org/resource/>
```

```
SELECT * WHERE
       {
        ?Company ptop:locatedIn ?Place ;
                 pext:industryOf dbpedia:Mass_media .
        ?Place   ptop:subRegionOf dbpedia:Europe.
       }
```

returns answers indicating that "Associated Newspapers" is a media company located not only in the United Kingdom, but also in England and in London based on the materialization of the transitive relation `ptop:subRegionOf`.

Furthermore, the queries with formulated with PROTON only return results much faster than queries combining predicates and concepts from different LOD datasets in FactForge, which is due to optimization of the joins traversed.

6 Conclusion

In this paper we presented the inference mechanisms implemented in the OWLIM semantic repositories and their application to a dataset formed by several LOD datasets. The materialization of statements in the closure of the inference rules provides a sound basis for extracting inferred information at query time.

Acknowledgments

This work is partially supported by RENDER FP7-ICT-2009-5, Contract no.: 257790.

References

1. Damova, M., Kiryakov, A., Grinberg, M., Bergman, M., Giasson, F., Simov, K.. Creation and Integration of Reference Ontologies for Efficient LOD Management. In: Semi-Automatic Ontology Development: Processes and Resources, IGI Global, USA, Armando Stellato and Maria Teresa Pazienza (Eds.) 2012.
2. Kiryakov, A; Ognyanov, D; Velkov, R; Tashev, Z; Peikov, I; LDSR: a Reason-able View to the Web of Linked Data, in: SW Challenge (ISWC2009), 2009.
3. ter Horst, H. J. Combining RDF and Part of OWL with Rules: Semantics, Decidability, Complexity . In Proceedings of The Semantic Web ISWC 2005, LNCS volume 3729 pp. 668–684. Springer Berlin / Heidelberg, 2005.
4. Heim, P; Hellmann, S; Lehmann, J; Lohmann, S; Stegemann, T; (2009) RelFinder: Revealing Relationships in RDF Knowledge Bases. In Semantic Multimedia, volume 5887 of LNCS, pp. 182–187. Springer Berlin/Heidelberg, 2009.