ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-1551

Towards SPARQL-Based Induction for Large-Scale RDF Data Sets

Simon Bin[1] and **Lorenz Bühmann**[1] and **Jens Lehmann**[2] and **Axel-Cyrille Ngonga Ngomo**[1]

Abstract. We show how to convert OWL Class Expressions to SPARQL queries where the instances of that concept are with a specific ABox equal to the SPARQL query result. Furthermore, we implement and integrate our converter into the CELOE algorithm (Class Expression Learning for Ontology Engineering), where it replaces the position of a traditional OWL reasoner. This will foster the application of structured machine learning to the Semantic Web, since most data is readily available in triple stores. We provide experimental evidence for the usefulness of the bridge. In particular, we show that we can improve the run time of machine learning approaches by several orders of magnitude.

1 INTRODUCTION AND MOTIVATION

A growing amount of data from diverse domains is being converted into RDF[3] as demonstrated by the growth of the Linking Open Data Cloud.[4] With this conversion come a significant number of complex applications which rely on large amounts of data in RDF and OWL[5] to perform demanding tasks, such as detecting patients with particular diseases [11]. While OWL reasoners can provide the required information for structured machine learning, they do not scale to large data sets. On the other hand, the SPARQL query language was developed specifically to query large amounts of data. By creating a bridge between SPARQL and OWL, we are able to answer OWL queries on large amounts of data.

Description Logics is the name of a family of knowledge representation (KR) formalisms. They emerged from earlier KR formalisms like semantic networks and frames. Their origin lies in the work of Brachman on structured inheritance networks [3]. $\mathcal{SROIQ}$ is a well-known description language, as it is the basis for OWL 2. $\mathcal{SROQ}$ is the subset of that language lacking inverse properties. We refer to [6] for details. For a complete definition of the SPARQL syntax and semantics, we refer to [1, 10] and the official W3C recommendation.[6]

2 OWL CLASS EXPRESSION REWRITING ALGORITHM

Proposition. Given an ABox $\mathcal{A}$, which contains class assertions to named classes and role assertions, we define $\mathcal{I}$ as the canonical interpretation [2]. Then we can show that executing the query converted from a concept C using τ as given in Tables 1 and 2 over $\mathcal{A}$ is the same as the canonical interpretation of C. Due to space constraints, please refer to our technical report for the proof.[7]

Table 1. Conversion of class expressions into a SPARQL graph pattern.

Class Expression C_i	Graph Pattern $\mathfrak{p} = \tau(C_i,$?var)
A	{?var rdf:type A.}
$\neg C$	{?var ?p ?o . FILTER NOT EXISTS { $\tau(C$, ?var)}}
$\{a_1, \ldots, a_n\}$	{?var ?p ?o . FILTER (?var IN ($a_1, \ldots, a_n$))}
$C_1 \sqcap \ldots \sqcap C_n$	{$\tau(C_1$, ?var) $\cup \ldots \cup$ $\tau(C_n$, ?var)}
$C_1 \sqcup \ldots \sqcup C_n$	{$\tau(C_1$, ?var)} UNION ... UNION {$\tau(C_n$, ?var)}
$\exists\, r.C$	{?var r ?s.} $\cup$ $\tau(C$, ?s)
$\exists\, r.\{a\}$	{?var r a.}
$\exists\, r.SELF$	{?var r ?var.}

$\forall\, r.C$	{ ?var r ?s0.
	{ SELECT ?var
	(count(?s1) AS ?cnt1)
	WHERE { ?var r ?s1 . $\tau(C,?s1)$ }
	GROUP BY ?var }
	{ SELECT ?var
	(count(?s2) AS ?cnt2)
	WHERE { ?var r ?s2 }
	GROUP BY ?var }
	FILTER (?cnt1 = ?cnt2) }

$\Theta n\, r.C$
$\Theta \in \{\leq, \geq, =\}$
	{ ?var r ?s0.
	{ SELECT ?var
	WHERE { ?var r ?s . $\tau(C,?s)$ }
	GROUP BY ?var
	HAVING (count(?s) $\Theta\, n$) } }

Table 2. Conversion of property expressions into a SPARQL g.p.

Property Expression p_i	Graph Pattern $\mathfrak{p} = \tau(p_i,$?var)
p	{?var p ?o.}
p^{-1}	{?s p ?var .}
$p_1 \circ \cdots \circ p_n$	{?var p_1 ?o1. $\vdots$ $\vdots$ $\vdots$?o_n-1 p_n ?o_n.}

3 LEARNING PROBLEM AND ALGORITHM

We consider supervised machine learning from positive and negative examples. All our experiments are binary classification tasks. The *CELOE* algorithm (Class Expression Learning for Ontology Engineering) iteratively generates class expressions and evaluates their performance on the positive and negative examples [8]. CELOE relies

[1] Universität Leipzig, Germany, email: {sbin,buehmann,ngonga}@informatik.uni-leipzig.de

[2] University of Bonn and Fraunhofer IAIS, Germany, email: jens.lehmann@cs.uni-bonn.de and jens.lehmann@iais.fraunhofer.de

[3] http://www.w3.org/TR/rdf11-concepts/

[4] http://lod-cloud.net/

[5] http://www.w3.org/TR/owl2-overview/

[6] http://www.w3.org/TR/sparql11-query/

[7] http://svn.aksw.org/papers/2016/ECAI_SPARQL_Learner/tr_public.pdf

Table 3. Data set characteristics.

number of	triples	classes	data/object properties		expressivity
carcinogenesis	74,567	142	15	4	$\mathcal{ALC}$(D)
mutagenesis	62,067	86	6	5	$\mathcal{AL}$(D)
mammograms	6,809	19	2	3	$\mathcal{AL}$(D)
TCGA-A	35,329,868	24	113	48	$\mathcal{AL}$(D)

on a top-down algorithm based on refinement operators. We use the refinement operator defined in [9].

4 EVALUATION

We implemented the SPARQL querying method according to Section 2 as an extension to the DL-Learner [7] framework in the place of an OWL API reasoner.[8] We use four data sets to compare the SPARQL induction to OWL reasoners. The *Carcinogenesis* and *Mutagenesis* data sets are moderately-sized data sets converted from data provided by the Oxford University Machine Learning group.[9] The mutagenesis data set is based on the results of [4]. The *Mammographic Mass data set* (Mammograms) was published in [5].[10] Additionally, we evaluated our approach on an excerpt from the cancer patient data in LinkedTCGA[11] (35 million triples). This data set has not previously been possible to use with DL-Learner due to its size.

In Table 3, the main characteristics of the data sets are described. The number of RDF triples describes the total size of the data set. Furthermore, the total number of classes and data as well as object properties (across all classes) is indicated. The expressivity refers to the description logic language features that are used in the data set as customary in description logics.

The approaches are evaluated using three different access options inside the DL-Learner framework. We tested the popular HermiT,[12] Pellet[12] and FaCT[13] OWL reasoners. For SPARQL, the data was loaded into an in-memory Jena[13] model with OWL/Lite inference rules[14] as well as a pre-computed model, in which case the SPARQL back-end acts as a pure graph database. In our evaluation setup, 22 seconds (mutagenesis), 36 seconds (carcinogenesis), 7 seconds (mammograms) were spent on pre-computing all inferences externally beforehand. We loaded the LinkedTCGA data set[15] into a SPARQL endpoint running OpenLink Virtuoso[16] 7.1. All experiments were run on an AMD Opteron 6376 @ 2.3GHz with 256 GB system memory, of which the DL-Learner framework used 32 GB. The algorithm itself is single-threaded.

The configuration files to reproduce our experiment can be found in the DL-Learner repository.[17]

5 CONCLUSION

The SPARQL approaches were nearly two orders of magnitude faster as can be seen in Tables 4 and 5. However, for the moment we have excluded the pressing lack of inference support for larger data sets.

Table 4. Number of concept tests within 4000 seconds (mean average in ten runs).

	SPARQL Precomp.	SPARQL Micro-rule	HermiT	Pellet
carcinogenesis	162,430	60,527	87	93
mutagenesis	11,713	4,552	timeout	176
mammograms	26,036	12,260	28	173

Table 5. Run time (seconds) until top accuracy.

	Precomp.	Micro-rule	HermiT	Pellet	FaCT
carcinogenesis	95	1,380	timeout	timeout	713
mutagenesis	1	3	timeout	159	-
mammograms	286	647	494	365	-
TCGA-A	1,243	timeout	timeout	timeout	-

ACKNOWLEDGEMENTS

This work was supported in part by a research grant from the German Ministry for Finances and Energy under the SAKE project (Grant No. 01MD15006E) and by a grants from the EU's 7th Framework Programme provided for the projects GeoKnow (GA no. 318159), Big Data Europe (GA no. 644564) and HOBBIT (GA no. 688227).

REFERENCES

[1] M. Arenas, C. Gutierrez, and J. Pérez, 'On the semantics of SPARQL', in *Semantic Web Information Management*, 281–307, Springer, (2010).

[2] F. Baader, D. Calvanese, D.L. McGuinness, D. Nardi, and P.F. Patel-Schneider, eds. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.

[3] R.J. Brachman, 'A structural paradigm for representing knowledge', Technical Report BBN Report 3605, Bolt, Beraneck and Newman, Inc., Cambridge, MA, (1978).

[4] A.K. Debnath, R.L. Lopez de Compadre, G. Debnath, A.J. Shusterman, and C. Hansch, 'Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity', *Journal of medicinal chemistry*, **34**(2), 786–797, (1991).

[5] M. Elter, R. Schulz-Wendtland, and T. Wittenberg, 'The prediction of breast cancer biopsy outcomes using two cad approaches that both emphasize an intelligible decision process', *Medical Physics*, **34**(11), 4164–4172, (2007).

[6] I. Horrocks, O. Kutz, and U. Sattler, 'The even more irresistible SROIQ', in *Proceedings, Tenth International Conference on Principles of Knowledge Representation and Reasoning, Lake District of the United Kingdom, June 2-5, 2006*, eds., P. Doherty, J. Mylopoulos, and C.A. Welty, pp. 57–67. AAAI Press, (2006).

[7] J. Lehmann, 'DL-Learner: learning concepts in description logics', *Journal of Machine Learning Research (JMLR)*, **10**, 2639–2642, (2009).

[8] J. Lehmann, S. Auer, L. Bühmann, and S. Tramp, 'Class expression learning for ontology engineering', *Journal of Web Semantics*, **9**, 71 – 81, (2011).

[9] J. Lehmann and P. Hitzler, 'Concept learning in description logics using refinement operators', *Machine Learning journal*, **78**(1-2), 203–250, (2010).

[10] J. Pérez, M. Arenas, and C. Gutierrez, 'Semantics and complexity of SPARQL', *ACM Transactions on Database Systems (TODS)*, **34**(3), 16, (2009).

[11] M. Saleem, S.S. Padmanabhuni, A. Ngonga Ngomo, J.S. Almeida, S. Decker, and H.F. Deus, 'Linked cancer genome atlas database', in *Proceedings of I-Semantics2013*, (2013).

[12] E. Sirin, B. Parsia, B. Cuenca Grau, A. Kalyanpur, and Y. Katz. Abstract Pellet: A practical OWL-DL reasoner, 2008.

[13] D. Tsarkov and I. Horrocks, 'FaCT++ description logic reasoner: System description', in *Proc. of the Int. Joint Conf. on Automated Reasoning (IJCAR 2006)*, volume 4130 of *Lecture Notes in Artificial Intelligence*, pp. 292–297. Springer, (2006).

[8] http://owlapi.sourceforge.net/reasoners.html

[9] http://www.cs.ox.ac.uk/activities/machlearn/applications.html

[10] http://archive.ics.uci.edu/ml/datasets/Mammographic+Mass

[11] http://aksw.org/Projects/LinkedTCGA

[12] http://www.hermit-reasoner.com/

[13] https://jena.apache.org/

[14] https://jena.apache.org/documentation/inference/

[15] https://code.google.com/p/bigrdfbench/

[16] http://virtuoso.openlinksw.com/

[17] https://github.com/AKSW/DL-Learner/tree/sparql-comparison/test/sparql-comparison