Web Semantics: Science, Services and Agents on the World Wide Web 15 (2012) 71–72

Contents lists available at SciVerse ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

Letter

Reply to comment on "WebPIE: A Web-scale parallel inference engine using MapReduce"

Dear Sirs,

Thank you for having shared with us this letter. We are always keen to discuss any unclear points and any limitations of our work, and we thank Dr. Patel-Schneider for his letter.

The letter that we received can be divided in two parts. The first part addresses points which, as the letter states, "can be overcome without changing the thrust of the paper". This leads us to a *corrigendum* with clarifications and bug fixes in the original paper. Although these are valuable improvements, none of these invalidate our approach or our evaluation.

The second part of the letter addresses "the deep question whether the general approach in the paper can actually perform scalable RDFS materialization". By addressing the four points below, we conclude that fortunately the answer to this question is yes.

1. "There is a significant bug in Algorithm 6, which uses only two reducers…"

The algorithm does not create "only two reducers". The number of reducers is set a priori, depending on the number of machines used. The keys ("0 + type" and "1 + subclass") are used to calculate a hashcode that will determine which reducer will be assigned to that pair.

2. "The scalability is not properly analyzed. With two million triples in the input the transitive closure may include four trillion triples"

It is indeed possible to construct worst-case scenarios. In Section 8.2 (p. 71), the paper states: *"The computational worst-case complexity of even the RDFS/OWL Horst fragment precludes a solution that is efficient on all inputs. Any approach to efficient reasoning must make assumptions about the properties of realistic datasets, and optimize for those realistic cases"*. The paper then enumerates the key assumptions behind our algorithms.

3. "The authors say that they are performing RDFS materialization. On the other hand, the authors say that "they do not consider" certain aspects of the problem and that they "ignore the first case of rule 5".

In Section 3.5 (p. 62), the paper states: *"We ignore the first case of rule 5, following the advice against "ontology hijacking" from* [15] *that suggests that users may not redefine the semantics of standard RDFS constructs"*. One can differ in opinion about this design choice, but one can hardly argue that we were not explicit about it.

4. "A second basic problem is that it is possible to infer schema triples from combinations of schema triples and non-schema triples, which then participate in other inferences".

Our paper describes this case in detail at the end of Section 3.5: *"The only possible loop we can encounter is when we extend the schema by introducing subproperties of rdfs:subproperty. In this case rule 7 could fire rule 5, generating a loop. […] The loop generated by schema does not occur in our data. […] However, in case it happens, all the rules must be re-executed until fix-closure"*. The purpose of this re-execution loop is to cover also the case in which schema information is derived from non-schema information. This loop is not necessary if the data does not contain such special cases. In any case, WebPIE is configured by default to repeat the execution of the RDFS rules until fix point. This also correctly treats the alleged counterexample.

Conclusion

The corrigendum summarizes our improvements to the paper as prompted by the letter. None of these invalidate our approach or our evaluation. We welcome all further comments on our work, and all reported bugs in the code (which has always been publically available; see Ref. [8] in our paper) will be corrected and published in future releases.

We would like to end with a more general comment. We think that, underlying the points above, there is a difference of opinion about the value of incomplete inference on real world datasets versus complete inference under a worst-case analysis. It is certainly the case that WebPIE performs incomplete RDFS reasoning, and the paper makes no secret of this. Our impression is that Dr. Patel-Schneider views this as a bug, because it does not comply with the formal definition of a full RDFS closure. We, on the other hand, regard it as a feature, because it aims to ignore trivial or undesirable inferences, and it allows WebPIE to scale to very large datasets. Reconciling these two perspectives will certainly be the basis for fruitful future work.

Sincerely,

Jacopo Urbani *
Department of Computer Science, Vrije Universiteit Amsterdam, 1081 HV Amsterdam, Netherlands
E-mail address: jacopo@cs.vu.nl.

Spyros Kotoulas
IBM Research Ireland, IBM Technology Campus, Damastown Industrial Estate, Dublin 15, Ireland
E-mail address: Spyros.Kotoulas@ie.ibm.com.

1570-8268/$ – see front matter
doi:10.1016/j.websem.2012.09.002

Jason Maassen
Frank van Harmelen
Henri Bal
Department of Computer Science,
Vrije Universiteit Amsterdam,
1081 HV Amsterdam, Netherlands

E-mail addresses: jason@cs.vu.nl (J. Maassen), Frank.van.Harmelen@cs.vu.nl (F. van Harmelen), bal@cs.vu.nl (H. Bal).

Available online 14 September 2012

* Corresponding author.