On the Road to the Evaluation of RDF Stream Compression Techniques

Jesús Arias[1], Óscar Corcho[2], Javier D. Fernández[3], Norberto Fernández[4], Alejandro Llaves[2], and Luis Sánchez[1]

[1] Dpto. Ing. Telemática, Universidad Carlos III de Madrid (Spain)
{jaf, luiss}@it.uc3m.es
[2] Ontology Engineering Group (OEG), Univ. Politécnica de Madrid (Spain)
{ocorcho,allaves}@fi.upm.es
[3] Vienna University of Economics and Business (Austria)
javier.fernandez@wu.ac.at
[4] Centro Universitario de la Defensa, Escuela Naval Militar (Spain)
norberto@cud.uvigo.es

The popularization of data streaming applications, such as those related to social networks and the Internet of Things, has fostered the interest of the Semantic Web community for this kind of data. As a result of this interest, the W3C RDF Stream Processing (RSP) community group[5] has recently been started with the goal of defining a common model "*for producing, transmitting and continuously querying RDF Streams*".

In this EOI we focus on the transmission model. As pointed out by recent research efforts (e.g. Ztreamy [4] and CQELS Cloud [6]), the efficient transmission of RDF streams is a necessary step to ensure higher throughput in RDF stream processors.

Our previous work has contributed to this area with several research initiatives related to RDF stream processing and compression:

- *HDT* [2], a compressed binary format for RDF. Besides addressing an efficient transmission, built-in indexes allow RDF triples to be randomly retrieved in compressed space, i.e. without prior decompression. Building such compressed, ready-to-consume serialization requires non-negligible processing time. Therefore, its use for streaming is challenging.
- *RDSZ* [3], an algorithm for lossless RDF stream compression, which combines a differential encoding mechanism with the general purpose stream compressor Zlib.
- *ERI* [1], an efficient interchange format for RDF streams, which adapts the encoding mechanism of the Efficient XML Interchange (EXI) format [7]: acknowledging that the described entities in an RDF stream often follow a common schema, ERI multiplexes the information into structural (schema) and value (concrete data) channels. A standard compressor such as Zlib is then used in each channel, resulting in high compression ratios and competitive processing time.

[5] http://www.w3.org/community/rsp/

– *Ztreamy* [4], a scalable middleware which allows to publish and consume RDF streams through HTTP. Using adequate buffering policies, RDF stream compression and single-threaded non-blocking input/output, Ztreamy is able to publish a real-time RDF stream to tens of thousands of simultaneous clients with delays up to a few seconds.

We are currently interested in evaluating the foundational data compression techniques underlying to current RDF compression proposals and efficient serializations of RDF streams. Our research roadmap can be depicted as follows:

– Collecting information on state-of-the-art techniques for data stream compression, such as dictionary-based compression, differential coding, Huffman coding [5] or using different compression channels for different infosets.
– Analysing the applicability of these techniques in the context of RDF stream compression. This would include, but not limited to, evaluating aspects like the scalability of the technique, whether it allows for direct access to a particular datum in the compressed streams and its complexity (for instance, whether or not it can be run on a limited device).
– Collecting suitable real-world datasets that can be used to define a corpora for RDF stream compression evaluation.
– Evaluating the performance that can be expected from the different approaches, with regard to typical parameters for data compression evaluation, i.e. compression rate, processing time and bandwidth. Additionally, reporting the compromise between the cost of data processing and the achieved compression figures and provided functionality.

We currently pursue this research as active participants of the RSP Serialization Group[6]. Given that one of its main goals is to analyse different RDF serialization and transmission approaches, the outcomes of our evaluation will probably provide relevant feedback for this group and the whole RSP community.

Acknowledgments

This work is partially funded by Ministerio de Economía y Competitividad (Spain) under the projects "HERMES-SMARTDRIVER" (TIN2013-46801-C4-2-R) and "4V: Volumen, Velocidad, Variedad y Validez en la Gestión Innovadora de Datos" (TIN2013-46238-C4-2-R), and Austrian Science Fund (FWF): M1720-G11.

References

1. Fernández, J., Llaves, A., Corcho, O.: Efficient RDF Interchange (ERI) Format for RDF Data Streams. In: The Semantic Web - ISWC 2014, LNCS, vol. 8797, pp. 244–259. Springer (2014)

[6] http://www.w3.org/community/rsp/wiki/RSP_Serialization_Group

2. Fernández, J.D., Martínez-Prieto, M.A., Gutiérrez, C., Polleres, A., Arias, M.: Binary RDF representation for publication and exchange (HDT). Web Semantics: Science, Services and Agents on the World Wide Web 19(0), 22 – 41 (2013)
3. Fernández, N., Arias, J., Sánchez, L., Fuentes-Lorenzo, D., Corcho, O.: RDSZ: An Approach for Lossless RDF Stream Compression. In: The Semantic Web: Trends and Challenges, LNCS, vol. 8465, pp. 52–67. Springer (2014)
4. Fisteus, J.A., Fernández, N., Sánchez, L., Fuentes-Lorenzo, D.: Ztreamy: A middleware for publishing semantic streams on the Web. Web Semantics: Science, Services and Agents on the World Wide Web 25(0) (2014)
5. Huffman, D.A.: A method for the construction of minimum-redundacy codes. In: Proceedings of the I.R.E. pp. 1098–1101 (September 1952)
6. Le-Phuoc, D., Nguyen Mau Quoc, H., Le Van, C., Hauswirth, M.: Elastic and Scalable Processing of Linked Stream Data in the Cloud. In: The Semantic Web ISWC 2013, LNCS, vol. 8218, pp. 280–297 (2013)
7. Schneider, J., Kamiya, T., Peintner, D., Kyusakov, R.: Efficient XML Interchange (EXI) Format 1.0 (Second Edition). W3C Recommendation (11 February 2014)