Profiling Web Archives

For Efficient Memento Query Routing

Sawood Alam
Old Dominion University
Norfolk, VA (USA)
salam@cs.odu.edu

Michael L. Nelson
Old Dominion University
Norfolk, VA (USA)
mln@cs.odu.edu

Herbert Van de Sompel
Los Alamos National Lab
Los Alamos, NM (USA)
herbertv@lanl.gov

Lyudmila L. Balakireva
Los Alamos National Lab
Los Alamos, NM (USA)
ludab@lanl.gov

Harihar Shankar
Los Alamos National Lab
Los Alamos, NM (USA)
harihar@lanl.gov

David S. H. Rosenthal
Stanford University Libraries
Stanford, CA (USA)
dshr@stanford.edu

We present the results to date of the IIPC-funded project for profiling the contents of web archives [2]. We review the proposed, light-weight JSON-LD format for describing the contents of a web archive. Based on our findings, we propose a more suitable serialization method for profiles than JSON. We review one of the primary methods for generating profiles when CDX files [4] are available; we analyze the CDX files and populate the profiles. We review the cost (growth) of profiles in terms of storage size, number of keys, and generation time with various profiling policies as the size of the archive grows. We present the effectiveness of various profiles in predicting the presence or absence of lookup URIs in an archive. We also review the ways to merge partial profiles generated from new data into existing profile of an archive to facilitate incremental profile updates. We briefly describe our ongoing approach to generate archive profiles when CDX files are not available. We review our methods of sampling the archive with URIs or search terms if full-text search is available.

The resulting profiles are written to well-known locations in a GitHub repository where they can be accessed by any party and are subject to version control. We briefly show how the resulting profiles can be used to inform URI routing decisions by the Memento Aggregator [5] as well as visualizing archive contents.

This is an oral presentation of [1]. Previous work in profiling ranged from using full URIs (no false positives, but with large profiles) [6] to using only top-level domains (TLDs) (smaller profiles, but with many false positives) [3]. This work explores strategies in between these two extremes. In our experiments, we gained up to 22% routing precision with less than 5% relative cost as compared to the complete knowledge profile without any false negatives. With respect to the TLD-only profile, the registered domain profile doubled the routing precision, while complete hostname and one path segment gave a five fold increase in routing precision.

We made our profiler code public[1] that includes implementation of various profiling strategies, serialization, and benchmarking.

[1] `https://github.com/oduwsdl/archive_profiler`

Acknowledgments

This work is supported in part by the International Internet Preservation Consortium (IIPC). Andy Jackson (BL) helped us with the UKWA datasets. Kris Carpenter (IA) and Joseph E. Ruettgers (ODU) helped us with the Archive-It data sets. Ilya Kreymer contributed to the discussion about CDXJ profile serialization format.

REFERENCES

[1] S. Alam, M. L. Nelson, H. Van de Sompel, L. L. Balakireva, H. Shankar, and D. S. H. Rosenthal. Web Archive Profiling Through CDX Summarization. In *Proceedings of the International Conference on Theory and Practice of Digital Libraries, TPDL 2015*, 2015.

[2] S. Alam, M. L. Nelson, H. Van de Sompel, and D. S. H. Rosenthal. Web Archive Profiling Via Sampling. `http://netpreserve.org/projects/web-archive-profiling-sampling`, 2015.

[3] A. AlSum, M. C. Weigle, M. L. Nelson, and H. Van de Sompel. Profiling Web Archive Coverage for Top-Level Domain and Content Language. *International Journal on Digital Libraries*, 14(3-4):149–166, 2014.

[4] Internet Archive. CDX File Format. `http://archive.org/web/researcher/cdx_file_format.php`, 2003.

[5] R. Sanderson. Global Web Archive Integration with Memento. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 379–380. ACM, 2012.

[6] R. Sanderson, H. Van de Sompel, and M. L. Nelson. IIPC Memento Aggregator Experiment. `http://www.netpreserve.org/sites/default/files/resources/Sanderson.pdf`, 2012.