6-2019

MementoMap: An Archive Profile Dissemination Framework

Sawood Alam

Michele C. Weigle

Michael L. Nelson

MementoMap: An Archive Profile Dissemination Framework

Sawood Alam, Michele C. Weigle, and Michael L. Nelson
Department of Computer Science Old Dominion University, Norfolk, Virginia – 23529 (USA)
{salam,mweigle,mln}@cs.odu.edu

ABSTRACT

We introduce *MementoMap*, a framework to express and disseminate holdings of web archives (archive profiles) by themselves or third parties. The framework allows arbitrary, flexible, and dynamic levels of details in its entries that fit the needs of archives of different scales. This enables Memento aggregators to significantly reduce wasted traffic to web archives.

1 INTRODUCTION AND BACKGROUND

The Memento framework [12] introduced a uniform means for various web archives to interoperate when it comes to accessing archived representations of a given Uniform Resource Identifier (URI). This enabled easy aggregation of mementos from many web archives. However, there is no standard way to access holdings of web archives without providing a full or partial lookup URI. Without the prior knowledge of holdings of each web archive, a naive Memento aggregator might flood all aggregated archives with unnecessary broadcast lookup requests for which they might not have any good results to return. The log of a Memento aggregator service[1] running at ODU using MemGator [3] shows that in over three years of its service it performed about 62M lookup requests to 14 different web archives of which only 5.44% requests returned any mementos. By knowing the holdings of these web archives (i.e., archive profiles), remaining over 94% wasted requests could have been avoided, which would have resulted in quicker response to users and reduced overhead to archives.

Previous web archive profiling works were focused on identifying different means to discover holdings of an archive and create archive profiles that maximize efficiency of aggregator lookup routing while minimizing various associated costs [4, 5, 7–10]. In our previous work we proposed CDXJ [2] as a potential format to represent archive profiles. In this work we are focusing on standardization of a format called *MementoMap*, which allows an efficient and flexible way of representing and disseminating archive profiles. This is an improvement over the previously proposed CDXJ format as it allows simpler and more storage efficient syntax and arbitrary depth in keys of the same file by using wildcards. This format is evaluated against the complete index of Portuguese Web Archive called Arquivo.pt (PWA) with various levels of details to assess associated costs and accuracy [6].

```
!context  ["https://git.io/mementomap"]
!id       {uri: "https://archive.example.org/"}
!fields   {keys: ["surt"], values: ["frequency"]}
!meta     {name: "Example Archive", year: 1996}
!meta     {type: "MementoMap"}
!meta     {updated_at: "2018-09-03T13:27:52Z"}
*                     54321/20000
com,*                 10000+
org,arxiv)/           100
org,arxiv)/*          2500~/900
org,arxiv)/pdf/*      0
uk,co,bbc)/images/* 300+/20-
```

Figure 1: A Sample MementoMap in UKVS Format

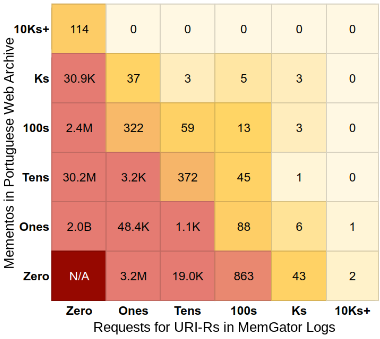
2 MEMENTOMAP USING UKVS

MementoMap is a framework for profiling web archives and expressing their holdings in an adaptive and flexible way to easily scale by utilizing Unified Key Value Store (UKVS) [1]. It is inspired by the simplicity of widely used `robots.txt` and `sitemap.xml` formats, but for a purpose other than search engine optimization. Figure 1 illustrates a sample *MementoMap* file that starts with some metadata headers. Header lines are prefixed with "`!`" sign to ensure they are separated from data lines and surfaced on top when the file is sorted. The "`!fields`" header tells that the first column is a *SURT* (i.e., Sort-friendly URI Reordering Transform) [11] and is used as a lookup key (there can be more than one key column such as *Datetime* or *Language*) that is followed by a value column which holds "`frequency`" information. Unlike the standard SURT, *MementoMap* allows wildcard-based partial *URI Keys* to enable flexibility in how detailed or concise one wants it to be depending on use cases, full or partial knowledge about the archive's holdings, and available resources. Each data line can optionally also contain a single-line JSON block, which is not illustrated here for simplicity sake. The `frequency` column is formatted as "`[URI-M Count]/[URI-R Count]`" where both counts are optional and the separator is also optional if only the `URI-M Count` is present. Moreover, these counts can have an optional suffix character `+`, `-`, or `~` to express that the numbers are not exact and represent a lower bound, an upper bound, and a rough estimate respectively. The first data line in the example means there are a total of exactly 54,321 mementos (*URI-Ms*) of exactly 20,000 *URI-Rs* in the archive and the next line suggests that there are at least 10,000 mementos from the "`.com`" *TLD* (i.e., Top Level Domain). The next two lines suggest that there are 100 mementos of the `arxiv.org` homepage and many more captures of pages with deeper paths. However, the next line illustrates an exclusion of a subtree by being more specific under `/pdf/*` that has zero mementos. For a detailed description of the format, more variations of representations of archive profiles, and some other use case refer to the UKVS.

A *MementoMap* can either be generated by the archives themselves or by third parties based on their external observations. We propose the "`mementomap`" link relation for its dissemination and discovery. We implemented a tool to generate *MementoMap* efficiently from the index of an archive (or other means of listing archival holdings) and open-sourced it under MIT license[2]. The tool allows configuration options to identify criteria of rolling multiple occurrences of a *URI Key* prefix into a shorter key with wildcard recursively. The tool also implements a binary search mechanism in *MementoMap* files.

[1] https://memgator.cs.odu.edu/api.html

[2] https://github.com/oduwsdl/MementoMap



Mementos in Portuguese Web Archive	Zero	Ones	Tens	100s	Ks	10Ks+
10Ks+	114	0	0	0	0	0
Ks	30.9K	37	3	5	3	0
100s	2.4M	322	59	13	3	0
Tens	30.2M	3.2K	372	45	1	0
Ones	2.0B	48.4K	1.1K	88	6	1
Zero	N/A	3.2M	19.0K	863	43	2

Figure 2: Overlap Between Archived and Accessed Resources

3 EVALUATION

For evaluation, we used the complete index of Arquivo.pt containing about 5B mementos of over 2B unique URI-Rs (i.e., original URIs) and 3.3M unique URIs in ODU's MemGator logs looked up over 5.2M times over a period of more than three years. Figure 2 shows a breakdown of what people are looking for in archives and what web archives hold. The 1.1K entry in the "Ones" row and "Tens" column shows that there are over a thousand *URI-Rs* that were requested 10–99 times in *MemGator* and each has 1–9 mementos in PWA. Large numbers in the "Zero" column show there are a lot of mementos that are never requested from *MemGator* (a usage-based profile might miss this data). Similarly, the "Zero" row shows there are a lot of requests that have zero mementos in PWA (a content-based profile might miss this data). The (Zero, Zero) corner suggests there are undetermined number of *URI-Rs* that were never archived or accessed. Irrespective of how the profile information was generated, *MementoMap* framework allows efficient means to express both what an archive holds and what it does not.

To evaluate the accuracy of *MementoMap*-based profiles, we first sampled all the unique *HTML* URI-Rs from the index that return `200` status code, which resulted in a dataset of about 1B unique entries (almost half of the original index). We then generated many *MementoMap* files from this using different configuration options to optimize the holdings representation and measured the accuracy of these when evaluated against MemGator logs. Figure 3 shows the relationship of relative cost (i.e., the ratio of number of *URI Keys* in the *MementoMap* and total number of unique *URI-Rs* in the archive) vs. routing accuracy (i.e., URIs who's presence or absence in the archive was correctly identified). The highlighted data point in the figure shows that there is a configuration that produces a *MementoMap* with only about 1.5% of the unique entries in the index and still correctly routes lookup requests with 60% accuracy with 100% recall. The accuracy can further be improved by 1) exploring other optimal configurations for subtree pruning, 2) generating profiles with the full index, not just a sample, and 3) including entries for absent resources from the "Zero" row of the Figure 2.

4 CONCLUSIONS AND FUTURE WORK

We introduced *MementoMap*, a serialization and dissemination format for archive profiles based on *UKVS*. The format allows arbitrary level of details for individual records and gives ability to dynamically roll well-populated subtrees up. We evaluated it against a large index of Arquivo.pt with billions of mementos and three years of a Memento aggregator log. The format is suitable for both small and large web archives and scales well. We open-sourced our implementation of *MementoMap* generation and binary search.

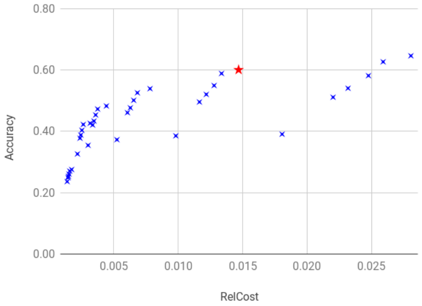


Figure 3: Relative Cost vs. Lookup Routing Accuracy

UK Web Archive, Arquivo.pt, National Records of Scotland, and National Library of Australia have recently shown interest in being able to express the summary of their holdings. As a pilot project these archives can be invited to generate *MementoMap* from their collections using our tool and advertise it using the "`mementomap`" link relation. Memento Aggregator services such as LANL's Time-Travel and ODU's MemGator can then leverage this information to better route lookup requests to these archives.

5 ACKNOWLEDGEMENTS

We thank Fernando Melo and Daniel Gomes from Arquivo.pt for the CDX dataset. Supported in part by NSF grant IIS-1526700.

REFERENCES

[1] Sawood Alam. 2019. Unified Key Value Store (UKVS). https://github.com/oduwsdl/ORS/blob/master/ukvs.md.

[2] Sawood Alam, Ilya Kreymer, and Michael L. Nelson. 2015. Object Resource Stream (ORS) and CDX-JSON (CDXJ). https://github.com/oduwsdl/ORS.

[3] Sawood Alam and Michael L. Nelson. 2016. MemGator - A Portable Concurrent Memento Aggregator. In *Proceedings of the 16th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '16)*.

[4] Sawood Alam, Michael L. Nelson, Herbert Van de Sompel, Lyudmila L. Balakireva, Harihar Shankar, and David S. H. Rosenthal. 2016. Web Archive Profiling Through CDX Summarization. *International Journal on Digital Libraries* 17, 3 (2016), 223–238. https://doi.org/10.1007/s00799-016-0184-4

[5] Sawood Alam, Michael L. Nelson, Herbert Van de Sompel, and David S. H. Rosenthal. 2016. Web Archive Profiling Through Fulltext Search. In *Proceedings of 20th International Conference on Theory and Practice of Digital Libraries, TPDL 2016*. 121–132.

[6] Sawood Alam, Michele C. Weigle, Michael L. Nelson, Fernando Melo, Daniel Bicho, and Daniel Gomes. 2019. MementoMap Framework for Flexible and Adaptive Web Archive Profiling. In *Proceedings of the 19th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '19)*.

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

[8] Nicolas Bornand, Lyudmila Balakireva, and Herbert Van de Sompel. 2016. Routing Memento Requests Using Binary Classifiers. In *Proceedings of the 16th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '16)*. 63–72.

[9] Robert Sanderson. 2012. Global Web Archive Integration with Memento. In *Proceedings of the 12th ACM/IEEE Joint Conference on Digital Libraries*. 379–380.

[10] Robert Sanderson, Herbert Van de Sompel, and Michael L. Nelson. 2012. IIPC Memento Aggregator Experiment. http://www.netpreserve.org/sites/default/files/resources/Sanderson.pdf.

[11] Kristinn Sigurðsson, Michael Stack, and Igor Ranitovic. 2006. Heritrix User Manual: Sort-friendly URI Reordering Transform. http://crawler.archive.org/articles/user_manual/glossary.html#surt.

[12] Herbert Van de Sompel, Michael L. Nelson, and Robert Sanderson. 2013. HTTP Framework for Time-Based Access to Resource States – Memento, Internet RFC 7089. https://tools.ietf.org/html/rfc7089.