Web Archive Profiling for Efficient Memento Aggregation

Sawood Alam
Department of Computer Science
Old Dominion University
Norfolk, Virginia - 23529 (USA)
salam@cs.odu.edu

ABSTRACT

With the proliferation of public web archives, it is becoming more important to better summarize their contents, both to understand their immense holdings as well as support routing of requests in the Memento Aggregator (an archival aggregator service using Memento framework). To save resources, the Memento Aggregator should only poll the archives that are likely to have a copy of the requested Uniform Resource Identifier (URI). Using the Crawl Index (CDX) files produced after crawling, I can generate profiles of the archives that summarize their holdings and can be used to inform routing of the Memento Aggregator's URI requests. Previous work in profiling ranged from using full URIs (no false positives, but with large profiles) to using only top-level domains (TLDs) (smaller profiles, but with many false positives). This work explores strategies in between these two extremes and some other dimensions of profiles such as time and language. In my early experiments with various URI based profiling policies, I successfully identified about 78% of the URIs that were not present in the archive with less than 1% relative cost as compared to the complete knowledge profile and 94% URIs with less than 10% relative cost without any false negatives. In another experiment I found that we can correctly route 80% of the requests while maintaining about 0.9 recall by discovering only 10% of the archive holdings and generating a profile that costs less than 1% of the complete knowledge profile. I want to expand upon the foundation of my early work to include time and content-type profiles and analyze the trade-off between precision and recall of routing queries among archives while utilizing various types of profiles separately or together. I also want to generate profiles of archives with the help of URI samples and fulltext search. This profiling framework and analysis will allow building services that will predict and rank archives where desired Mementos of a requested URI are likely to be present.

Keywords

Web Archives, Profiling, CDX Files, Memento

1. MOTIVATION

The number of public web archives supporting the Memento protocol [19] natively or through proxies continues to grow. Figures 2(a) and 2(b) illustrate a naive implementation of the Memento Aggregator [15] where each request is broadcasted to all the known archives, but only a few archives return good results. The Memento Aggregator, the Time Travel Service[1], and other services, both research and production, need to know which archives to poll when a request for an archived version of a file is received. An efficient Memento routing in aggregators is desired from both aggregators' and archives' perspective. Aggregators can reduce the average response time, improve overall throughput, and save network bandwidth. Archives benefit by the reduced number of requests for which they have no holdings, hence saving computing resources and bandwidth.

In December 2015, soon after the surge of OldWeb.today[2] many archives struggled with the increased traffic. We found that fewer than 5% of the queried Uniform Resource Identifiers (URIs) are present in any individual archive other than the Internet Archive as illustrated in Table 1 and Figure 1. In this case, being able to identify a subset of archives that might return good results for a particular request becomes very important.

Figure 2(c) illustrates a setup where a Memento Aggregator queries an archive profile service and gets a list of archives sorted in the order of the probability of finding a copy of the queried file in them so that the aggregator can choose the top-K archives from the list to make the requests. Previous work proved that simple rules are insufficient to accurately model a web archive's holdings [7, 6]. For example, simply routing requests for `*.uk` URIs to the UK National Archives is insufficient: many other archives hold `*.uk` URIs, and the UK National Archives holds much more than just `*.uk` URIs. This is true for the many other national web archives as well.

In this work, I examine various strategies for producing *profiles* of web archives. The idea is that profiles are a lightweight description of an archive's holdings to support applications such as coordinated crawling between archives, visualization of the archive's holdings, or routing of requests to the Memento Aggregator. It is the latter application that is the focus of this work.

I envision following four approaches to discover the holdings of an archive based on the URIs:

1. *CDX Profiling* – This approach involves acquiring the CDX files[3], hence having complete knowledge of the archive holdings, and then generate profiles from it. Alternatively, the archives can run the profiling script on their collection and return the generated profiles.

[1] http://timetravel.mementoweb.org/
[2] http://oldweb.today/
[3] CDX (Capture/Crawl inDeX) files are created as an index of the WARC [11] files generated from the Heritrix web crawler; see [10] for a description of the CDX file format.

Table 1: Presence of the Sample Lookup URI-Rs in Each Archive.

Sample (1M URIs Each)	Archive-It	UKWA	Stanford	Union of {AIT, UKWA, SUA}
DMOZ	4.097%	3.594%	0.034%	7.575%
Memento Proxy Logs	4.182%	0.408%	0.046%	4.527%
IA Wayback Logs	3.716%	0.519%	0.039%	4.165%
UKWA Wayback Logs	0.108%	0.034%	0.002%	0.134%

2. *Fulltext Search Profiling* – In this approach I send random query terms to the fulltext search interface of the archive (if present) and from the search response we learn the URIs that it holds. These URIs are then utilized to build archive profiles.

3. *Sample URI Profiling* – This approach requires a sample set of URIs to query them against the archive and build the profile from the successful responses. This is quite wasteful, as few ($< 5\%$) of the sample URIs are found in any archive. This approach should only be used when the earlier two approaches are not possible.

4. *Response Cache Profiling* – This approach depends on the response data collected from the archive over a period of time as queries are made to the archives from an aggregator. Cached responses are analyzed and profiles are generated from them.

In my recent work [3] I established a generic archive profiling framework and explored the first approach, the *CDX Profiling*. I examined an extended set of 23 different policies to build archive profiles and measured their routing efficiency. In another work [5] I explored the second approach, the *Fulltext Search Profiling* within the framework established by my recent work. It is important to note that some web archives (including the Internet Archive) do not provide fulltext search, hence this approach is not applicable for them. The third approach was briefly explored by AlSum [6], but I need to explore it further within the framework I established. The fourth approach is different from the other three in a way that it is a usage-based profiling approach while the other three are content-based. The responses from various archives in the cache of the Memento aggregator over a period of time are analyzed to learn about their holdings. As a result the *Response Cache Profiling* is based on what people were looking for as opposed to what is in the archives. Bornand et al. implemented this approach for Memento routing by building binary classifiers from the aggregator cache data [8]. I need to explore this in terms of archive profiles rather than a classifier and evaluate the results against their work.

An archive profile has an inherent trade-off in its size vs. its ability to accurately describe the holdings of the archive. If a profile records each individual original URI (URI-R in Memento terminology [19]) the size of the profile can grow quite large and difficult to share, query, and update. On the other hand, an aggregator making routing decisions will have perfect knowledge about whether or not an archive holds archived copies of the page, or mementos (URI-Ms in Memento terminology). In contrast, if a profile keeps just the summaries of top-level domains (TLDs) of an archive the profile size will be small, but can result in many unnecessary queries being sent to the archive. For example, the presence of a single memento of bbc.co.uk will result in the profile advertising .uk holdings even though this may not be reflective of the archive's collection policy.

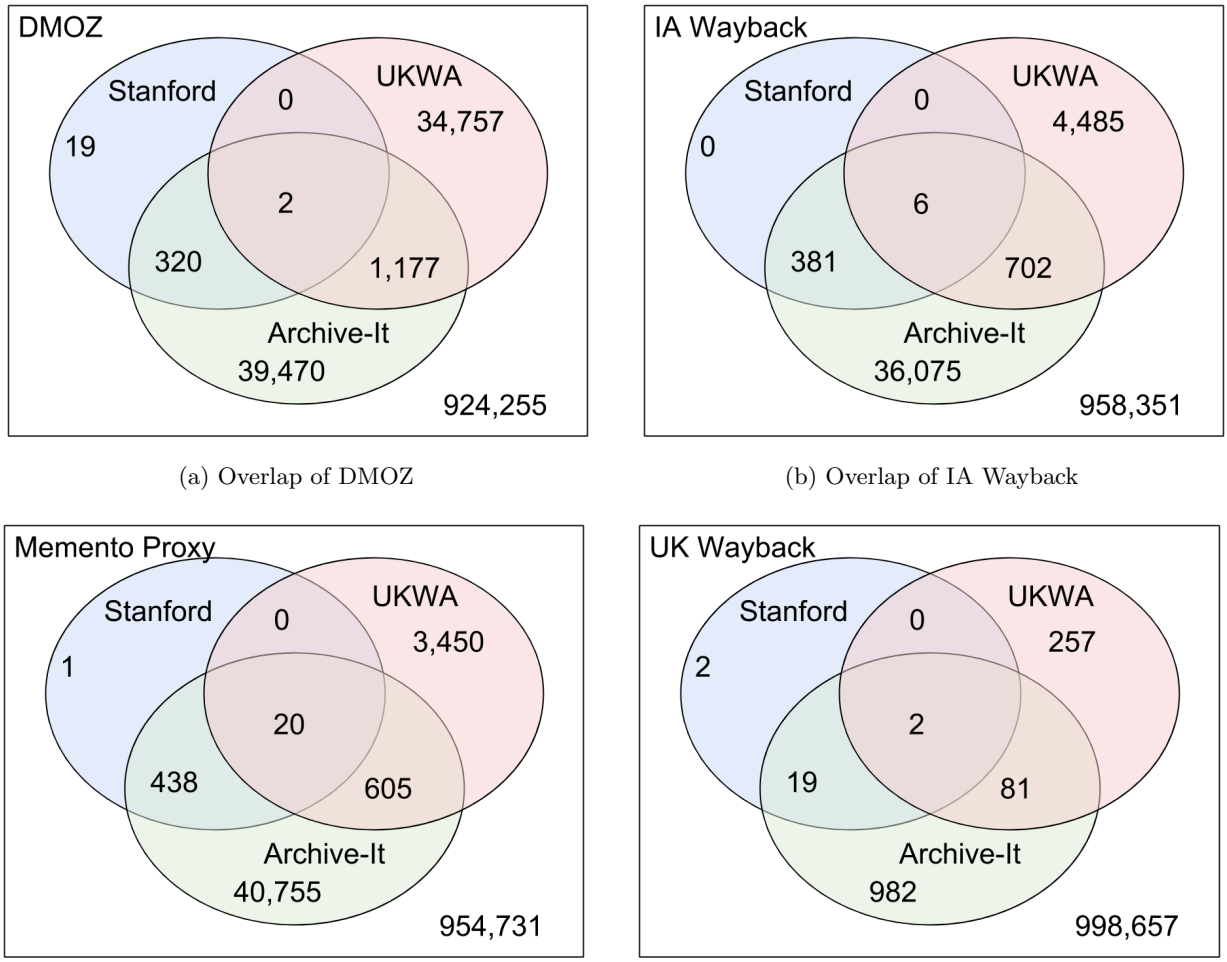
So far, I examined 23 different policies based on the URI-R for generating profiles, from the extremes of using the entire URI-R to just the TLD. Using the CDX files of the UK Web Archive (covering 10 years and 0.5 TB) and the ODU copy of the Archive-It (covering 14 years and 1.8 TB), I examined the trade-offs in profile size and routing efficiency for four million sample URIs. I further developed a *Random Searcher Model* to perform fulltext searching in archives to discover their holdings. I estimated the work required to discover certain percentage of the archives' holdings and the efficiency of profiles based on partial knowledge.

2. RELATED WORK

Query routing is a common practice in various fields including meta-searching and search aggregation [12, 17, 18]. Memento query routing was earlier explored in the two efforts described below, but they explored extreme cases of profiling. In an earlier study [3, 4] I found that an intermediate approach that gives flexibility with regards to balancing accuracy and effort can result in better and more effective Memento routing.

Sanderson et al. created exhaustive profiles [16] of various International Internet Preservation Consortium (IIPC) member archives by collecting their CDX files and extracting URI-Rs from them (I denote it as *URIR Profile*). This approach gave them complete knowledge of the holdings in each participating archive, hence they can route queries precisely to archives that have any mementos (URI-Ms) for the given URI-R. It is a resource and time intensive task to generate such profiles and some archives may be unwilling or unable to provide their CDX files. Such profiles are so big in size (typically, a few billion URI-R keys) that they require special infrastructure to support fast lookup. Acquiring fresh CDX files from various archives and updating these profiles regularly is not easy. As a result the profile fails to route requests for mementos that were added in the archive after the profile was generated until the profile is updated again with the fresh data.

Many web archives tend to limit their crawling and holdings to some specific TLDs, for example, the British Library Web Archive prefers sites with .uk TLD. AlSum et al. created profiles based on TLD [7, 6] in which they recorded *URI-R Count* and *URI-M Count* under each TLD for 12 public web archives. Their results show that they were able to retrieve the complete TimeMap [19] in 84% of the cases using only the top three archives and in 91% of the cases when using the top six archives. This simple approach can reduce the number of queries generated by a Memento Aggregator significantly with some loss in coverage. The issue of the *TLD-Only* profile is large number of false positives. For example, if an archive has copies of just a few `*.com`



(a) Overlap of DMOZ

(b) Overlap of IA Wayback

(c) Overlap of MementoProxy

(d) Overlap of UK Wayback

Figure 1: Sample Lookup URI Overlap in Various Archives

URIs, all the `*.com` queries will be routed there.

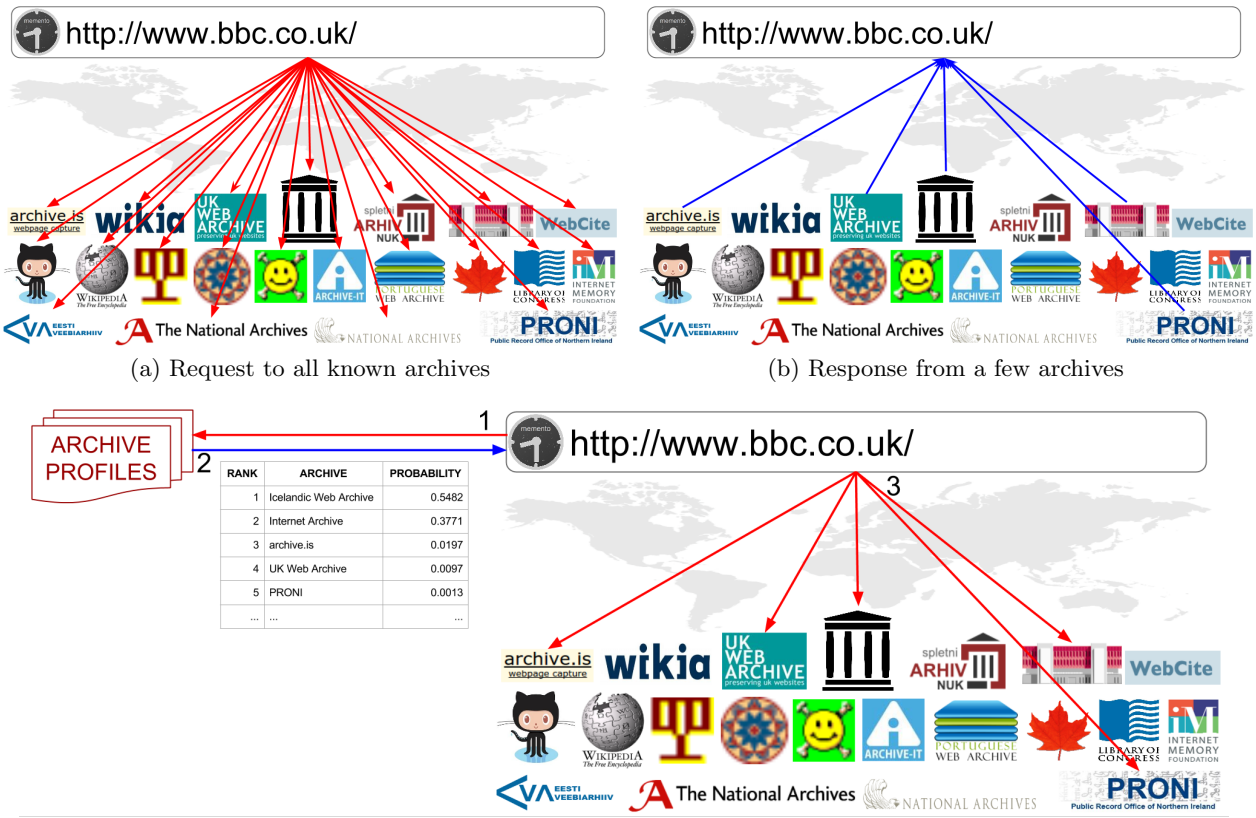
Recently, Bornand et al. implemented a different approach for Memento routing by building binary classifiers from the aggregator cache data [8]. They report a 77% reduction in the number of requests and a 42% reduction in response time while maintaining a 0.847 recall value.

There have been many efforts on crawling the hidden web that have no hyperlinks and are accessible only by filling out HTML forms [13, 20, 14]. My keyword search based archived content discovery is related to these efforts because archived contents span over a long period of time, which causes a disconnect between old and contemporary pages. As a result, hyperlink based shallow crawling might only discover a temporal sub-graph of the holdings.

3. METHODOLOGY

WARC (Web ARChive) files are the de facto standard for web archives to store their output including Domain Name System (DNS) resolutions, Hypertext Transfer Protocol (HTTP) requests, HTTP responses (including headers and payload), and some other things. Each block of various WARC files is then indexed in CDX files. Each entry in the CDX file stores the canonical URI and observation time (Memento-Datetime) as lookup keys and associated data such as the status code, content-type, content digest, record offset in the WARC file, content length, and the WARC file name. The latter three are generally used to locate the capture in a WARC file. These CDX files are just indexes, hence these are significantly smaller than WARC files. Having access to an archive's CDX files gives complete knowledge of its holdings in terms of what URIs it captured, when, and how often. This information is sufficient to build a lightweight profile for the archive.

For the baseline work, in this study I used CDX files from Archive-It (ODU mirror), UK Web Archive, and Stanford Archive to generate profiles, but profile generation is not limited to only CDX processing, it can also be done by sampling URI sets and querying the live archives or by using full-text searching feature provided by some archives. In earlier studies, *URI-M Count* and *URI-R Count* were used to keep track of the amount of holdings under each key, but these are absolute values and do not apply in case of sampling and hard to maintain when merging multiple profiles, so I introduce two relative measures for the same purpose. To keep track of the sum of URI-M counts I use the term "frequency" and to keep track of the number of profiles from where the "frequency" was accumulated I use the term "spread".



(a) Request to all known archives

(b) Response from a few archives

(c) Memento Aggregator with Profile Based Routing

Figure 2: Request-Response Cycle of Memento Aggregator and Various Archives

3.1 Profile Type

Profiles can be of many types based on what attributes of an archive were used to generate them. I will focus on the profiles based on the URI-R, Memento-Datetime, and Content-Language. Each type of the profile can have multiple policies based on how detailed the profiles is, such as, measuring holdings of an archive yearly or monthly where the latter is more detailed than the earlier one.

3.1.1 URI-Key Profile

URI-Key is a term I introduced to describe the keys generated from a URI based on how detailed the profile will be. Figure 3 illustrates a sample *URI-Key* profile serialized in CDXJ format. Initial lines that start with an ! sign describe the metadata about the archive and the profile itself. For example the "type" attribute of the "!meta" key describes the type of the profile ("urikey") followed by the policy used ("H3P1") which means a maximum of three segments from the hostname an a maximum of one segment from the path. Following the metadata lines are the data entries that start with a key followed by a single line JavaScript Object Notation (JSON) block describing the holdings in the archive for each key.

3.1.2 Time Profile

Figure 4 illustrates a sample *Time* profile with the policy that generates keys for every month of the year. This type of profile is good for describing the crawling activity of an archive over time.

3.1.3 Language Profile

Figure 5 illustrates a sample *Language* profile with language keys as per the alpha-2 language codes standard [9]. This type of profile is good for describing the holdings of an archive for any given content language.

3.1.4 Hybrid Profile

Figure 6 illustrates a sample *Hybrid* profile with *URI-Key* as the primary key and the *Time* as the secondary key. In the example I used "H3P1" and "YYYY" policies from the two profile types respectively. This type of profile is good for describing the holdings of an archive for more than one attributes together. For example in the Figure 5 it shows that the holdings under `uk,ac,rpms,)/` are available in two consecutive years where the first year has more activity than the second year. Hybrid profiles are not limited to the combination of only the *URI-Key* and the *Time* attributes, but these can have any two or more attributes together in any order suitable for certain application.

3.2 Profile Merging

To deal with periodic updates of profiles, smaller profiles are generated with new data and these small profiles are merged into the base profile. Without an option to merge smaller profiles to build a large profile gradually, updates will require a complete reprocessing of the entire dataset, including the dataset previously processed. Figure 7 illustrates the process of merging two profiles where "frequency" and "spread" values from the Base and New profiles are summed up in the Merged profile for the key `com,cnn,)/`.

```
!context ["https://oduwsdl.github.io/contexts/archiveprofile"]
!id {"uri": "http://www.webarchive.org.uk/ukwa/"}
!keys ["surt_uri"]
!meta {"name": "UKWA Sample", "type": "urikey#H3P1", "...": "..."}
com,dilos,)/region {"frequency": 14, "spread": 2}
uk,ac,rpms,)/ {"frequency": 124, "spread": 1}
uk,co,bbc,)/images {"frequency": 152, "spread": 3}
```

Figure 3: Sample URI-Key Profile

```
!context ["https://oduwsdl.github.io/contexts/archiveprofile"]
!id {"uri": "http://www.webarchive.org.uk/ukwa/"}
!keys ["year_month"]
!meta {"name": "UKWA Sample", "type": "time#YYYYMM", "...": "..."}
200211 {"frequency": 4508, "spread": 2}
200212 {"frequency": 652, "spread": 1}
200301 {"frequency": 298, "spread": 1}
```

Figure 4: Sample Time Profile

```
!context ["https://oduwsdl.github.io/contexts/archiveprofile"]
!id {"uri": "http://www.webarchive.org.uk/ukwa/"}
!keys ["language"]
!meta {"name": "UKWA Sample", "type": "lang#iso-639-1", "...": "..."}
ar {"frequency": 574, "spread": 2}
en {"frequency": 8736, "spread": 5}
ur {"frequency": 82, "spread": 1}
```

Figure 5: Sample Language Profile

```
!context ["https://oduwsdl.github.io/contexts/archiveprofile"]
!id {"uri": "http://www.webarchive.org.uk/ukwa/"}
!keys ["surt_uri", "year"]
!meta {"name": "UKWA Sample", "type": "urikey/time#H3P1/YYYY", "...": "..."}
uk,ac,rpms,)/ 2002 {"frequency": 854, "spread": 2}
uk,ac,rpms,)/ 2003 {"frequency": 63, "spread": 1}
uk,co,bbc,)/images 2003 {"frequency": 348, "spread": 1}
```

Figure 6: Sample Hybrid Profile

```
# Base Profile:
com,cnn,)/ {"frequency": 30, "spread": 1},
uk,co,bbc,)/ {"frequency": 20, "spread": 1}

# New Profile:
com,cnn,)/ {"frequency": 10, "spread": 1},
com,usatoday,)/ {"frequency": 5, "spread": 1}

# Merged Profile:
com,cnn,)/ {"frequency": 40, "spread": 2},
uk,co,bbc,)/ {"frequency": 20, "spread": 1},
com,usatoday,)/ {"frequency": 5, "spread": 1}
```

Figure 7: Illustration of Profile Merging

Table 2: Publication Plan

#	Title	Target	Status
1	Web Archive Profiling Through CDX Summarization	TPDL15	Published
2	Profiling Web Archives - For Efficient Memento Query Routing	TCDL15	Published
3	Web Archive Profiling Through CDX Summarization	IJDL16	Published
4	Web Archive Profiling Through Fulltext Search	TPDL16	Published
5	Poster: MemGator - A Portable Concurrent Memento Aggregator	JCDL16	Published
6	Object Resource Stream (ORS) and CDX-JSON (CDXJ) Formats	RFC	Progressing
7	Scalable, Maintainable, and Extensible Web Archive Profile Serialization for Efficient Lookup	JCDL18	Planned
8	URI, Time, and Language Profiling from Live Archives via URI Sampling and Fulltex Search	JCDL18	Planned
9	Memento Aggregator Routing Based on Probability Distribution of Memento Availability with Archive Profiles	SIGIR18	Planned
10	Archive X-Ray - Web Archive Profiling for Efficient Memento Aggregation	IJDL18	Planned

4. WORK PLAN

In my work so far, I have completed the baseline research and published a paper on CDX based profiling [3]. I have also worked on fulltext search based profiling [5]. Table 2 lists my tentative plan of publications emerging from this work. I have already completed some research, development, and analysis tasks for my future publications as part of the baseline research. Following is the proposed plan of this research work:

Baseline Profiling Through CDX Files – I have access to some CDX files from various Web archives e.g., Archive-It (ODU mirror), UK Web Archive, and Stanford Archive. I will generate various profiles with different policies of varying level of details based on these CDX files that will serve as the gold standard with which I will compare my sampling methods for evaluation. I have also developed scripts that volunteer archives can run on their servers and send us generated summaries/profiles for further analysis. Providing a means to transmit summaries instead of the CDX files will reduce the bandwidth required as well as it will encourage contribution from the dark and protected archives too. [Feb 2015]

Profile Serialization – I will evaluate various serialization options to find the most suitable mechanism to serialize archive profiles that is expressive, extensible, scalable, and enables efficient lookup. I will provide means to host these profiles publicly, preferably with a revision control system such as GitHub. [Feb 2015]

Fulltext Search Profiling – Web archives that expose a fulltext search interface can be profiled by searching for a set of query terms and inspecting the returned result URIs. I have implemented mechanisms to generate suitable query terms for a given archive and perform fulltext searches to generate an archive profile [5]. [Jan 2016]

Sample URI Dataset – It is not always feasible to generate exhaustive profiles from CDX files. In that case I will rely on profiling via URI sampling. However, generating a URI sample set that is representative of holdings of an archive is difficult. Earlier studies show that different sources of collection URI samples such as DMOZ, archive access logs, Memento Aggregator logs, and URI collection from top query terms are biased in different ways. I will develop methods of combining more than one sources to come up with one or more representative sample URI sets of archive's holdings and will evaluate them against the gold dataset described in the previous step. [Feb 2016]

Multidimensional Profiling – With each query, the Memento Aggregator receives a *URI-R*, it also receives an *Accept-Datetime* header (for TimeGate requests) and sometimes an *Accept-Language* header. These additional pieces of information can be utilized to further filter the archive lookup, hence building Time and Language based profiles along with URI based profiles will be helpful. I will examine the usefulness and cost of generating these additional types of profiles when used individually or combined. [Aug 2016]

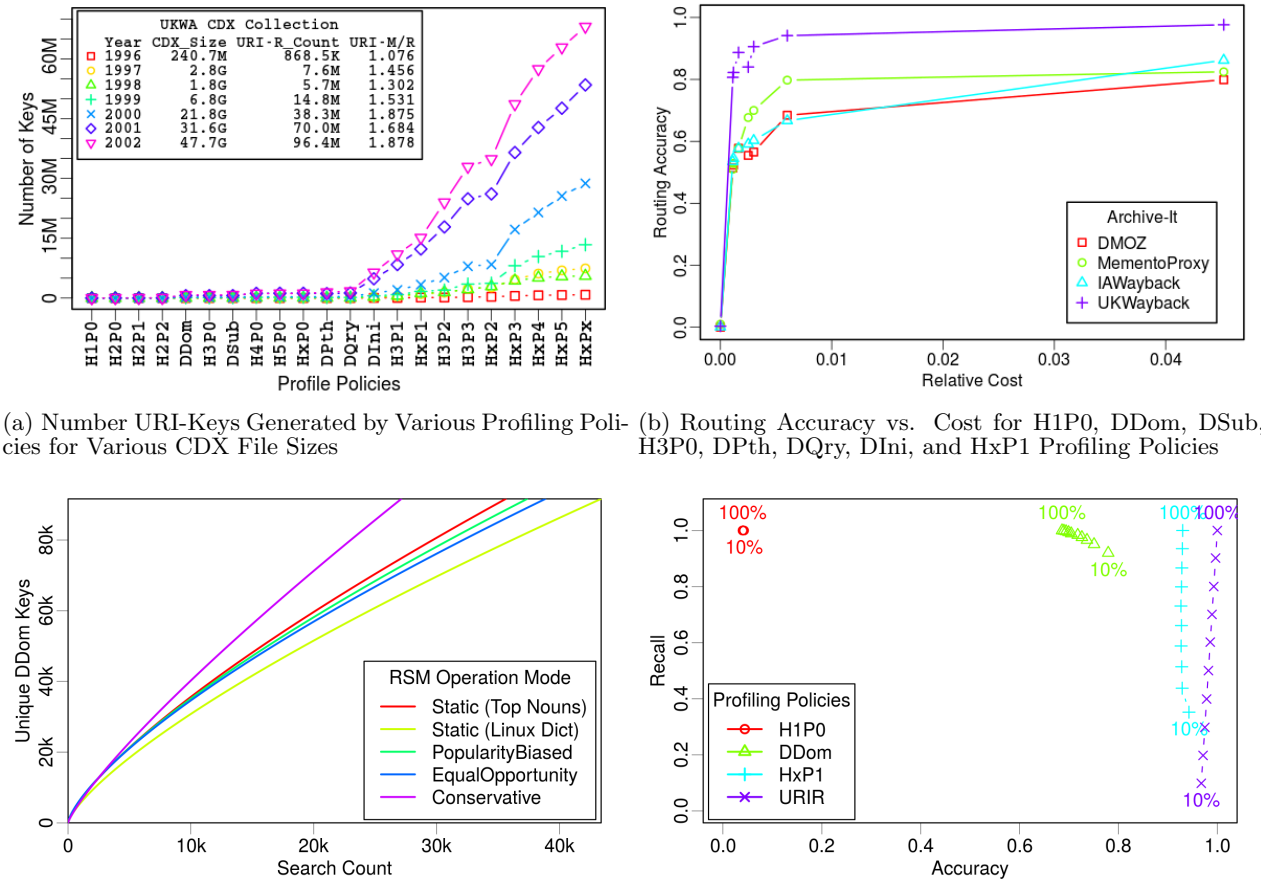
Instrumenting Memento Aggregator – With various types of profiles of various archives in place, I will develop methods to utilize them individually or in combination. I will evaluate various ways to combine the statistical values stored in those profiles to generate probability distribution of availability of Mementos. I will implement a service that takes *URI-R*, *Accept-Datetime*, and *Accept-Language* as inputs and returns a probability distribution of availability of mementos in various archives. An aggregator or any other application can then decide to pick top-K archives from the list and route requests accordingly to aggregate from. [Dec 2016]

5. PRELIMINARY RESULTS

Preliminary results of this work cover analysis of 23 different *URI-Key* profiles of three archives and four sample query sets [3, 4]. The preliminary work establishes relationship among the CDX Size, URI-M Count, URI-R Count, and URI-Key Count (Figure 8(a)). It also analyzes the space and time costs for each profiling policy. Finally, it examines the routing efficiency of each profile with respect to its relative cost with respect to the corresponding complete knowledge profile (Figure 8(b)). I successfully identified about 78% of the URIs that were not present in the archive with less than 1% relative cost as compared to the complete knowledge profile and 94% URIs with less than 10% relative cost without any false negatives. Further, in the fulltext search based profiling analysis experiment I established relationship of search cost with the discovery percentage of the archive holdings (Figure 8(c)). Additionally, I found that we can make routing decisions of 80% of the requests correctly while maintaining about 0.9 recall by discovering only 10% of the archive holdings and generating a profile that costs less than 1% of the complete knowledge profile (Figure 8(d)).

6. EVALUATION PLAN

To evaluate my work, I will generate several profile types with varying policies for different archives. Some profiles will be based on CDX analysis that will provide the gold standard while others will be based on URI sampling or keyword searching. I will generate several URI sample sets using existing datasets (such as DMOZ) and access logs of various web archives. I will then use those gold standard datasets to examine the precision and recall trade-off for every profile against every sample URI set.



(a) Number URI-Keys Generated by Various Profiling Policies for Various CDX File Sizes

(b) Routing Accuracy vs. Cost for H1P0, DDom, DSub, H3P0, DPth, DQry, DIni, and HxP1 Profiling Policies

(c) Searches Needed to Discover Desired Number of Unique Domian Names (DDom Policy URI-Keys)

(d) Accuracy and Recall for Various Profiling Policies as a Function of Percentage of the Archive Holdings Known

Figure 8: Preliminary Results

7. CONCLUSIONS

So far, I have done baseline research to examine the space and routing efficiency trade-offs in different policies for producing URI based profiles of Web archives. I defined the term "URI-Key" to refer to the keys generated from a URI based on various policies that are used to track the distribution of holdings of an archive at different hostname and path depths or different segment counts. I used CDX files from ODU's Archive-It replica, the UK Web Archive, and Stanford Archive for generating profiles, and evaluated the profiles using a query set of four million URIs, created from DMOZ, IA Wayback access logs, UK Wayback access logs, and Memento Aggregator access logs. I can successfully identify about 78% of the sample URIs that were not present in the archive with less than 1% relative cost as compared to the complete knowledge profile and 94% URIs with less than 10% relative cost without any false negatives.

Further, in the fulltext search based profiling analysis experiment I established relationship of search cost with the discovery percentage of the archive holdings. Additionally, I found that we can make routing decisions of 80% of the requests correctly while maintaining about 0.9 recall by discovering only 10% of the archive holdings to generate a profile that costs less than 1% of the complete knowledge profile.

Going forward, I plan to generate more than one type of profiles (such as *Time* and *URI-Key*) and combine their results to improve the routing precision and recall. I plan to study the trade-off between the routing efficiency and recall by utilizing the statistical values stored against each key in the profile. I also plan to create profiles of live archives from non-CDX sources, such as URI and keyword sampling to generate language, time, and hybrid profiles. I plan to implement a service that consumes multiple archive profiles to generate a ranked ordered list of profiles with probabilities of finding mementos of a given URI. As part of the Memento aggregator instrumentation, I created a new Memento aggregator called MemGator [2] which needs to be integrated with the ranked ordered archive list service to utilize profiles. Additionally, for the archive profile serialization, I have introduced a file format called CDXJ [1] that needs some more work to publish it as an RFC.

This work intends to provide a framework to express the holdings of a web archive in a machine as well as human readable manner. The framework will provide with a suite of techniques that not only enables archives to describe themselves, but third parties can also build profiles for any web archive. The nature of these archive profiles and the level of details would vary based on the intended applications and available profiling resources. The primary goal of this work is to make Memento aggregation more efficient. However, there can be many more applications of archive profiles, such as, content classification and meta-searching across archives.

8. ACKNOWLEDGEMENTS

Sawood Alam's advisor is Michael L. Nelson. David S. H. Rosenthal (Stanford), Herbert Van de Sompel (LANL), Lyudmila L. Balakireva (LANL), Harihar Shankar (LANL) provided helpful feedback and contributions. 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. LANL, OldWeb.today, Stanford Archive, and UK National Archive provided useful logs. This work is supported in part by the International Internet Preservation Consortium (IIPC).

9. REFERENCES

[1] S. Alam, I. Kreymer, and M. L. Nelson. Object Resource Stream (ORS) and CDX-JSON (CDXJ) Draft. https://github.com/oduwsdl/ORS, 2015.

[2] S. Alam and M. L. Nelson. MemGator - A Portable Concurrent Memento Aggregator. In *Proceedings of the 16th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '16, pages 243–244, 2016.

[3] 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 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015*, pages 3–14, 2015.

[4] 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. *International Journal on Digital Libraries*, 17(3):223–238, 2016.

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

[6] A. AlSum, M. C. Weigle, M. L. Nelson, and H. Van de Sompel. Profiling Web Archive Coverage for Top-Level Domain and Content Language. In *Proceedings of the International Conference on Theory and Practice of Digital Libraries, TPDL 2013*, pages 60–71, 2013.

[7] 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.

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

[9] Infoterm. ISO 639-1 Language Codes. http://www.infoterm.info/standardization/iso_639_1_2002.php, 2002.

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

[11] ISO 28500. WARC (Web ARChive) file format. http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml, 2009.

[12] W. Meng, C. Yu, and K.-L. Liu. Building Efficient and Effective Metasearch Engines. *ACM Computing Surveys (CSUR)*, 34(1):48–89, 2002.

[13] A. Ntoulas, P. Zerfos, and J. Cho. Downloading Textual Hidden Web Content Through Keyword Queries. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '05, pages 100–109, 2005.

[14] S. Raghavan and H. Garcia-Molina. Crawling the Hidden Web. 2000.

[15] 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.

[16] 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.

[17] A. Sugiura and O. Etzioni. Query Routing for Web Search Engines: Architecture and Experiments. *Computer Networks*, 33(1):417–429, 2000.

[18] T. Tran and L. Zhang. Keyword Query Routing. *Knowledge and Data Engineering, IEEE Transactions on*, 26(2):363–375, 2014.

[19] H. Van de Sompel, M. L. Nelson, and R. Sanderson. HTTP Framework for Time-Based Access to Resource States – Memento. RFC 7089, Dec. 2013.

[20] P. Wu, J.-R. Wen, H. Liu, and W.-Y. Ma. Query Selection Techniques for Efficient Crawling of Structured Web Sources. In *Proceedings of the 22nd International Conference on Data Engineering*, ICDE '06, pages 47–47, 2006.