Enhanced Memento's Aggregator Framework to Browse the Past Web

Ahmed AlSum
Computer Science Department
Old Dominion University
Norfolk, VA 23508
aalsum@cs.odu.edu

ABSTRACT

Browsing the past Web in an easy, a complete, and consistent way has become an essential need in the recent years. The archived contents are distributed among many systems on different locations, and each one has its own format, protocol, technology for preserving and retrieval. In this research, we propose the "Enhanced Memento Aggregator" framework which is capable of collecting, filtering, ranking the archived copies in a distributed environment. The framework is supported with an easy and interactive user interface.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries

General Terms

Design, Standardization

Keywords

Web Architecture, HTTP, Web Archiving, Digital Preservation

1. INTRODUCTION

The news websites used to cover the important events with details about how, why and when it was happened. So, we are not in danger of "forgetting" the important events like Hurricane Katrina or the Virginia Tech shootings occurred, but we may forget the evolution of these stories. The evolution of the stories and the context in which they were reported are an important part of our cultural memory and the footprints of history. To review the evolution of these stories, you have to go back to the archived copies of the news websites during these periods. Even the Internet Archive has the most, old archived copies, but it does not have all the archived copies in the world, there are different Web archives that carry part of the archived Web.

JCDL'11, June 12-17, 2011, Ottawa, ON, Canada.

The Memento project [28] provides an extension to the "HyperText Transfer Protocol" to browse the web through the time dimension using a new HTTP header "Accept-Datetime" to let the browser asks special resource called "TimeGate" to send the URI near this date instead of the current version. Memento TimeGate is the place that aware with the old versions of this URI. Memento proposed a new tool that is called "Memento Aggregator" which works as a TimeGate for the non-memento complied sites. The aggregator queried different archives for this URI through specific proxies for each one, then it combines the results in one list of the archived copies (mementos) with the content datetime for each copy; this list is called TimeMap. The current design of the Memento Aggregator merges the discovered mementos using the archives' proxies and sorts the results by Memento-Datetime.

In this research, we are going to enhance the Memento Aggregator to be enriched with adding more features like aggregator optimization technique for both the request to filter the queried proxies based on predefined rules, and for the response to check the quality of the archived version. Also, we propose a new query method using non-blocking distributed mechanism which enables the results to be processed faster. The Enhanced framework is supported with a new user-interface that enables that user to take benefits from these features with additional features from the user perspective such as adding his personal TimeGate, and collecting the customer feedback about new archives.

1.1 Motivation

The motivation for this research came from the inability to depend on the results came from the aggregator itself, because it may be misleading in some cases. For example, if you would like to browse the JCDL 2002 conference website (`www.jcdl2002.org`). Figure 1 shows different attempts to find the website. First the user tried to access the URI itself on the current web, but the content was changed to something other than the conference (figure 2(b)); the user decided to go to the Memento Aggregator to get the archived site. The snippet 1 shows the TimeMap for (`www.jcdl2002.org`). If the user selected one of the mementos on 2005, the result will be redirected to another archived host (figure 1(b)) which is available on the archive. If the user selected a memento from 2009, the archived version will be the insurance agent page (figure 1(c)). On the other hand, JCDL used to archive the conference website each

Listening 1: Memento TimeMap for JCDL 2002 www.jcdl2002.org

```
<http://mementoproxy.lanl.gov/aggr/timebundle
    /http://jcdl2002.org>;rel="timebundle",
 <http://jcdl2002.org>;rel="original",
 <http://http://mementoproxy.lanl.gov/aggr/
    timemap/link/http://jcdl2002.org>;
    rel="timemap";
    type="application/link-format",
 <http://mementoproxy.lanl.gov/aggr/
    timegate/http://jcdl2002.org>;
    rel="timegate",
 <http://memento.waybackmachine.org/
    memento/20010819194233/
    http://jcdl2002.org>;
    rel="first memento";
    datetime="Sun, 19 Aug 2001 19:42:33 GMT",
 <http://memento.waybackmachine.org/
    memento/20011216220248/
    http://jcdl2002.org>;
    rel="memento";
    datetime="Sun, 16 Dec 2001 22:02:48 GMT",
...
```

year[1] which appeared on figure 1(d). In this scenario, Memento Aggregator reported 2009 memento as a valid memento for JCDL 2002 website, which is not the expected site, and also the aggregator was not able to capture the valid place for the JCDL past conferences archives because it was not listed on its archives list.

Even if you reach successfully the right archived version of JCDL 2002 and you are able browsing the *site* in the past, it does not mean you are capable of browsing the Web in the past. For example, most conferences attached a link to the program committee member homepages, if you tried to access these homepages which pointed to outside the website, you will probably go to the current version (or 404 if the homepage is no longer existed) not the old version. By using the aggregator in the time travel mode, which allows the aggregator to detect that you are browsing the site on the past, the aggregator automatically forwards the request to the archived version of this homepage instead of the current one. Figure 2 shows two examples of how to access the homepage of one of the members in JCDL 2002 `http://mln.larc.nasa.gov/~mln/` in the normal scenario or using Time Travel mode.

Another drawback for current archives is that you may be able to access different archives which claimed that it has an archive copy for your URI, but you still need to compare manually between the different copies to select the good version. For example [17], if you would like to browse the medical journal Graft[2] which is no longer printed, you can find archived copies on four web archives. figure 3 shows the different copies of Graft Medical magazine, the current page is already redirected to another page that listed the deprecated journals. Internet Archive has an archived copy but when you browse the site, you could not read the pdf because the archive crawled a payment/login page instead of crawling the pdf page. Portico[3] preserved a copy of the Graft journal which has a limited access to their partners. Finally, CLOCKSS[4] has a full archived copy for the Graft journal which is accessible to the public. However, you have three copies of the same URIs but you took the time to rank these copies to select the best one.

[1] `http://jcdl.org/past-event-conf.shtml`
[2] `http://gft.sagepub.com`

1.2 Research Questions

This proposal aims to fulfill the users need to browse the past Web as they do with the current Web. The proposed Enhanced Memento Aggregator framework will handle different users questions such as: what are the available archived versions for this URI? Could I retrieve them? What ranking techniques beside the datetime are possible and useful? Should I trust these archived copies? How to evaluate the aggregator from both a system and a user perspective? What are the services should the aggregator provide and what are the services should the aggregator client provide? Could I have a nice and easy visualization technique?

2. RELATED WORK

In the recent years, there have been various studies about the web archiving. Masanes [12] in 2006 published an overview about the web archiving techniques. Brown [3] in 2006 , provided a practical guide for doing archiving for the internet; the book examined the process of the archiving from selection, collection, storage, and delivery to the user. The book also covered the legal issues and the quality assurance. Shiozaki and Eisenschitz [20] published a survey that was conducted between 16 national libraries designed to clarify how national libraries attempt to justify their web archiving activities.

2.1 Memento Framework

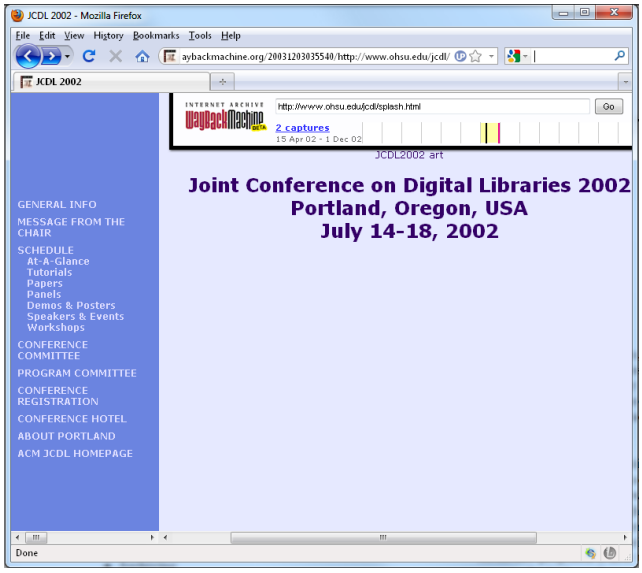
The Memento solution [28, 16, 15, 27, 29] is a protocol-based which aims to achieve a tighter integration between the current and the past web. Memento is an extension for the popular HyperText Transfer Protocol (HTTP) to allow the user to browse the past web as the current web. The term *Memento* (formally URI-M) refers to an archival record of an original resource (URI-R). Memento extendes HTTP content negotiation [7] to be in the datetime dimension, a new HTTP headers proposed a special header "Accept-Datetime" [25]. For example, if users want to retrieve a Memento at a specific datetime, they should issue Accept-Datetime header in the GET request.

```
GET / HTTP/1.1
Host: www.cs.odu.edu
...
Accept-Datetime: Sat, 17 Dec 2005 12:00:00 GMT
...
```

Only the *TimeGate* is capable of negotiation with the web browser in the time dimension. A TimeGate [25] for an Original Resource is a resource that supports negotiation to allow

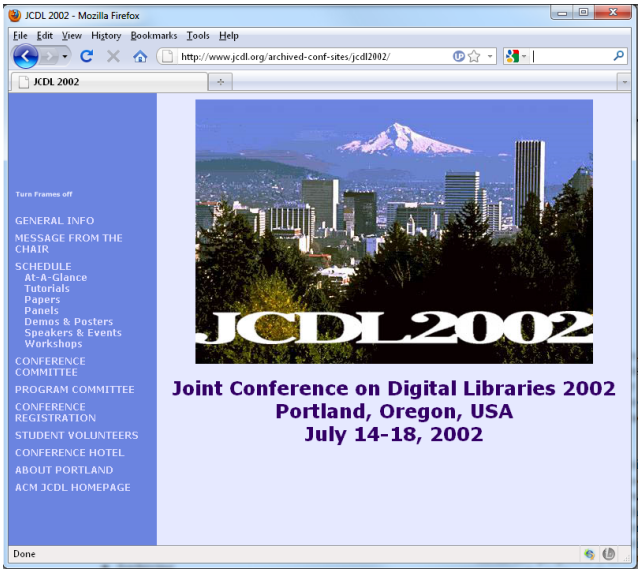
[3] `http://www.portico.org/`
[4] `http://www.clockss.org/`

(a) The current version



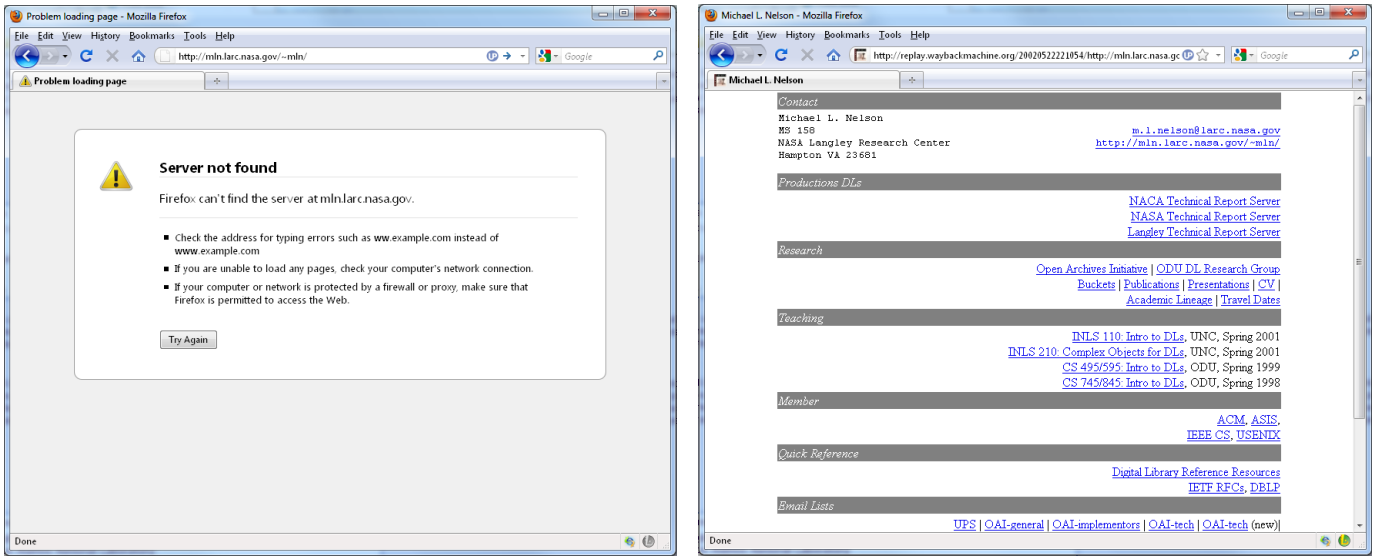
(b) Memento from 2005

(c) Memento from 2009



(d) The right archived version

Figure 1: Snapshots for the JCDL 2002 website (www.jcdl2002.org)



(a) Browse without Time Travel mode (b) Browse with Time Travel mode

Figure 2: Browsing URI outside JCDL 2002 archived site

selective, datetime-based, access to an archived copy of this Original Resource. Multiple TimeGates may exist for any given resource; a *TimeMap* for an Original Resource is a resource from which a list of URIs of Mementos of the Original Resource is available. A TimeMap Aggregator [26] (for simplicity Aggregator) harvests and merges TimeMaps; the the Aggregator exposes its own TimeGates and TimeMaps. The Memento Aggregator provides TimeGates and TimeMaps cross-archives, with finer datetime granularity after merging the different TimeMaps. It can become a shared target for redirection for many web servers that do not have a ready TimeGate. The Memento Aggregator depends on different proxies scripts [18] that can be used to implement by-proxy Memento support for third-party servers such as Web Archives and Content Management Systems.

2.2 Archive Quality

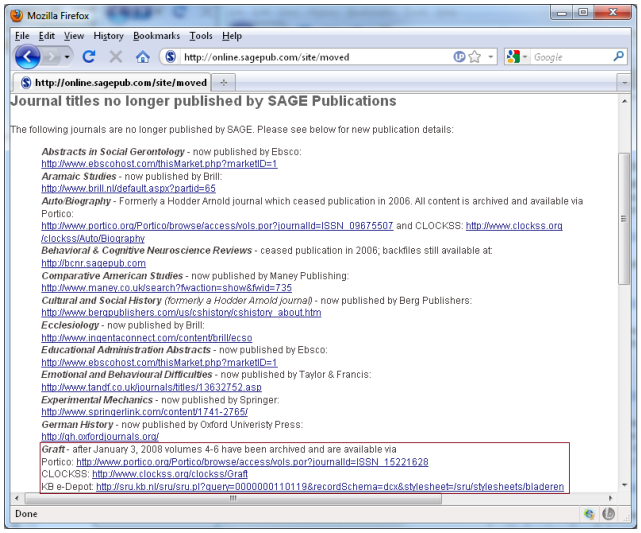
Masanès [11] defined the quality of the Web archive by a) *The Completeness* of material archived and b) *The Coherence* which is the ability to render the original form of the site. Spaniol et al. [21, 5] proposed the SHARC Framework for assessing the data quality in the web archives and for tuning capturing strategies toward coherence and completeness quality measures. The framework proposed two phases, phase one "Single-visit" crawls download every page in a site exactly once to cover the whole site. Phase two "Visit-revisit" revisit the pages after the initial download to cover any intermediate changes during the first crawl. The aim of the second phase is to maximize the "coherence" of the site. The coherence defects could be detected using different visualization techniques [22]. *The Coverage* quality measure [19] of a list of archived copies describes how accurately the archived copies reflects the important versions of a web page in a time interval, it is used to be measured using the freshness of search engine techniques [4].

Hence, most of the quality studies focused on the quality from the archivists perspective by proposing new crawling strategies to enable the Web archive to fulfill these three measures; for increasing the completeness, the breadth-first search crawling [14] could be used, Gomes et al. [6] suggested a technique to detect the duplicates pages before storing in the archive, for the coherence [21, 5], and for the coverage [4]. However, the previous techniques provided mechanisms to assure pre-crawling quality measures, few techniques covered post-crawling and post-delivery quality measures from the user perspective. Brown [3] suggested post-collection testing steps to ensure the quality of the collected pages.

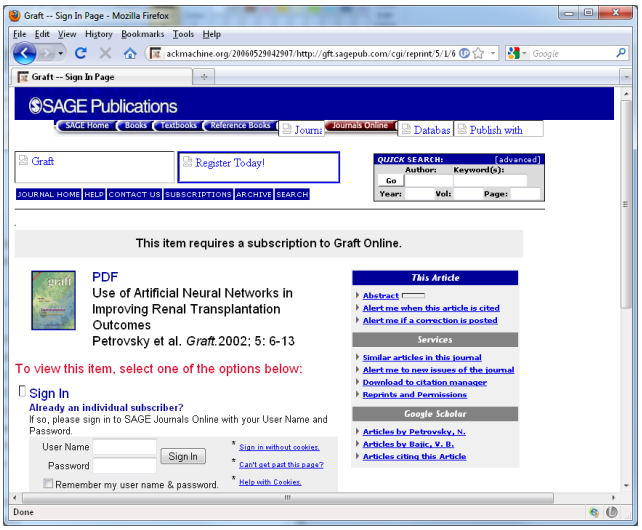
2.3 Browsing the Web archive

There are different techniques to browse the web in the past. Adar et al. [1] proposed "Zoetrope", a system that enables interaction with the historical Web. Even though Zoetrope provides an easy way to browse the past, it can not be applied on the normal archives because it depends on a large number of copies that were preprocessed to be ready for the user interface manipulation. Jatowt et al. [8, 9] proposed different models to browse the past web. Teevan et al. [23] proposed "DiffIE" a browser plug-in that caches the pages a person visits and highlights how those pages have changed when the person returns to them.

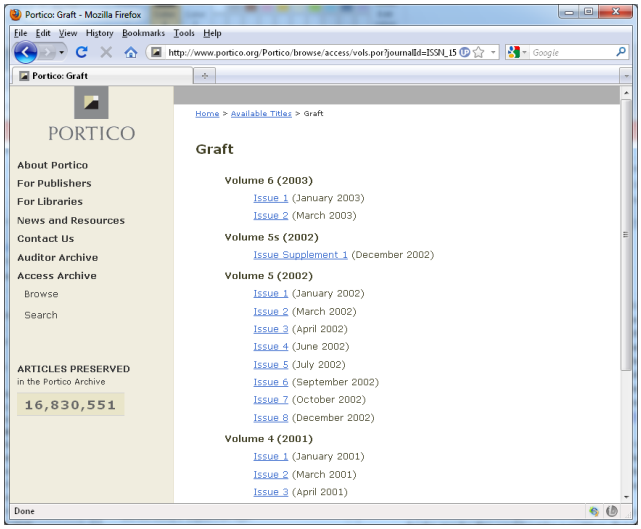
Other services were built to take benefits from the archived pages. Warrick [13] is a utility for reconstructing or recovering a website when a back-up is not available. Warrick will search the Internet Archive, Google, MSN, and Yahoo for stored pages and images and will save them to your filesystem. Synchronicity [10] is a Mozilla Firefox add-on that supports the Internet user in (re-)discovering missing web pages in real time. With the help of Memento, Synchronicity can discover the missing page at its new URI and provides options to retrieve good enough replacement pages in real time.



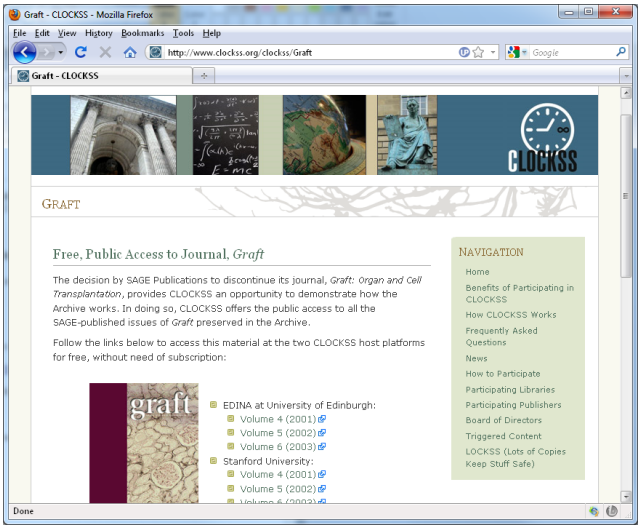
(a) Current Graft website



(b) Graft on Internet Archive



(c) Graft on Portico



(d) Graft on CLOCKSS

Figure 3: Graft Medical Journal on different archives

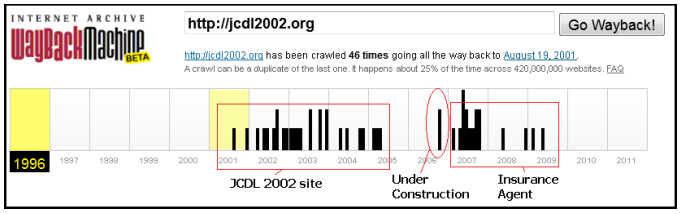


Figure 4: Internet Archive mementos of www.jcdl2002.org

3. ENHANCED WEB ARCHIVING CONCEPTS

3.1 Archive Descriptor

Did you try to search UK Web Archive[5] for The Japan Times[6] website? Do you know any Web Archive that carry web pages around 1996 other-than the Internet Archive? Sometimes when we failed to find an archived copy, we wondered why this Web Archive did not have any single copy for this URI. The truth is that this Web Archive may never have this URI based on the Web Archive characteristics. For each Web Archive, we can determine specific characteristics that may distinguish this archive from the other archives and give an idea about the archive content. For example, the age of the archived copies in the archive and the supported domains for crawling. This description enables the user to exclude the archives that do not have the required URI in specific date/time. We propose an archive descriptor which is a way to communicate a description of the archive content to the users through a pre-defined path on the Web Archive itself (similar to robots.txt) which could be queried easily.

3.2 Quality of the Archived Web Page

In section 2.2, we discussed different quality measurements such as completeness, coverage, and coherence. The coherence is out of scope of our study because we mainly study one URI instead of the whole site. The coverage and page resources completeness measurements will be the focus of our study, additional to the quality of the content. In this module, we are trying to answer the question of "Is it a good memento?"

Initially, it is easier to define the bad memento instead of the good one; figure 4 shows a list of the archived copies of of `www.jcdl2002.org` on the Internet Archive. The figure shows that site was up and running from mid 2001 until mid 2005, after that the domain has been hi-jacked in different perspectives. So, we can consider the mementos from mid 2006 until 2009 are bad mementos. Hence, to rank the mementos quality, we can define the status of the memento into 3 quality levels:

1. Existed, the archive published a memento date which is attached with an existed page (HTTP status 200 or 3xx ends with 200).

2. Valid, there is a memento and the content belongs to the original site.

[5]http://www.webarchive.org.uk
[6]http://www.japantimes.co.jp/

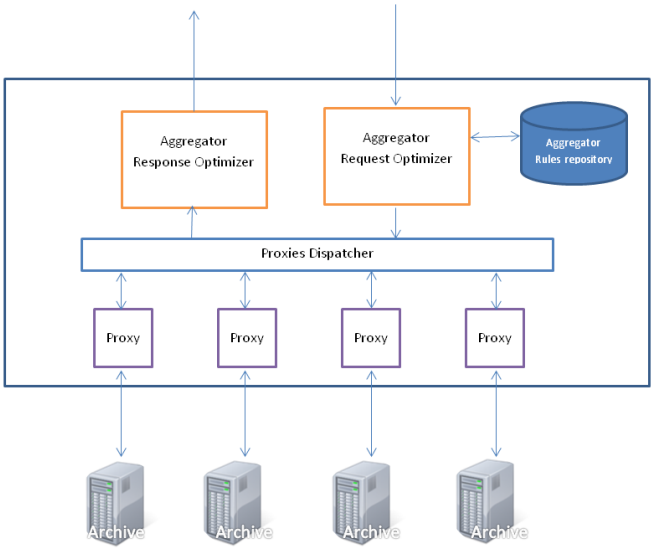


Figure 6: Enhanced Memento Aggregator Framework

3. Completed, there is a memento with text related to the original site, and all the the rest of the pages resources are existed like images and style sheets.

Figure 5 shows four mementos with different levels of quality.

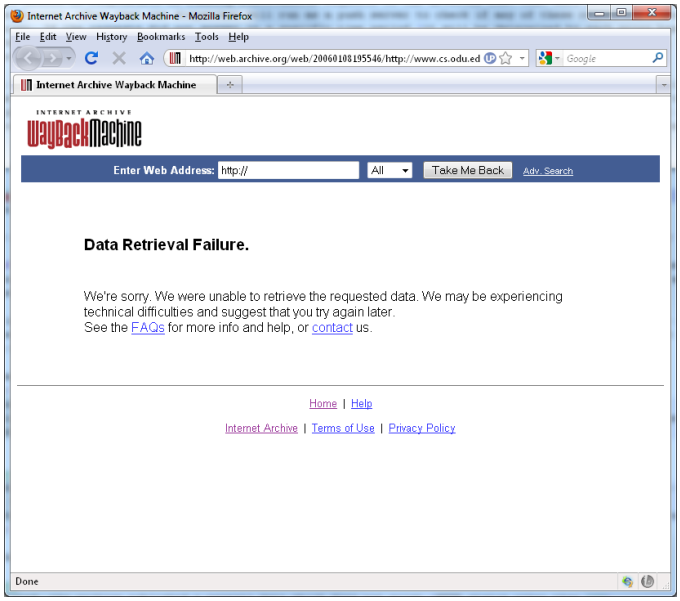
Existed quality level could be determined by using the HTTP HEAD method response code, accept the memento as existed if the HTTP response code is 200 or 3xx ends with 200. *Valid* quality level could be determined by comparing the content of the memento and comparing it with the previous and the next memento in the TimeMap, the percentage of the change should be consistent with the difference of time. The history of IP owner could be used as an indicator for suspicious spam, further techniques will be visited in this part. *Completed* quality level could be determined by checking all the page internal resources for existence.

4. ENHANCED MEMENTO AGGREGATOR FRAMEWORK

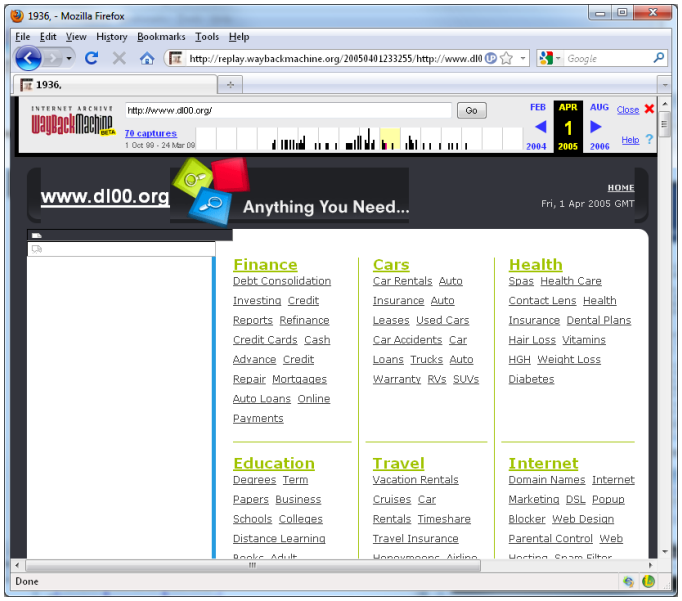
The basic idea for the enhanced memento aggregator framework is similar to the traditional aggregator framework, which is receiving the user URI request, and searching the different proxies for mementos, then aggregating the results into one TimeMap. The enhanced aggregator framework is distinguished in different aspects from the server side; the proposal also contains a new client side user interface that could be used by general users to browse the past web. Figure 6 presents the enhanced aggregator components architecture.

4.1 HTTP Request Headers

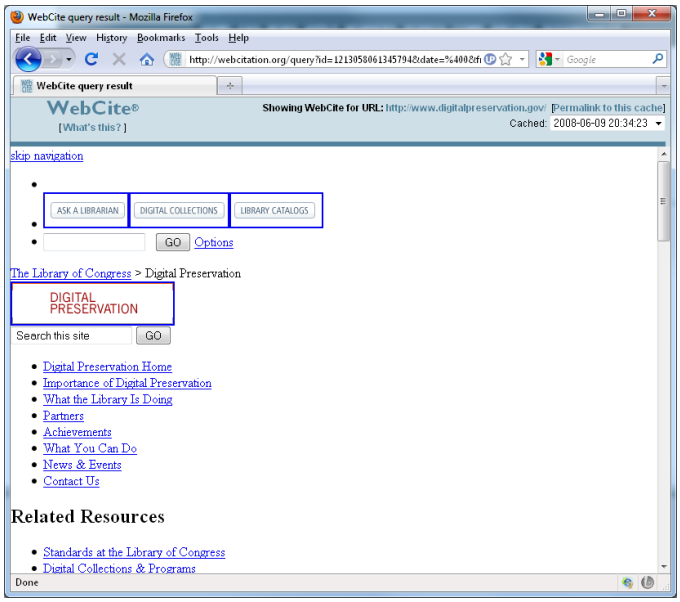
The enhanced memento aggregator headers follow the Memento Internet Draft [25]. The aggregator will focus on "Accept-Datetime" request header that is conveyed in an HTTP GET/HEAD request issued against a TimeGate for an Original Resource, and its value indicates the datetime of the desired past state of the Original. Memento Aggregator



(a) Memento does not Exist



(b) Memento Exists, but not Valid



(c) Memento is Valid, but not Complete

(d) Memento Exists, Valid, and Complete

Figure 5: Mementos with different quality levels

works as a default TimeGate for the non-memento compliance websites.

4.2 Aggregator Request Optimizer

In this module, the aggregator applied predefined rules to select the best proxies to retrieve the mementos. The rules should be concluded based on the archive descriptor, in the case of the absence of the archive descriptor, the archive proxy should contain a description about the designated archive. The rules may be by URL domain (For example, .au domain should not be queried on Library and Archives Canada[7] which is limited to .gc.ca domains), or by Accept-Datetime date (for example, National Archives and Records Administration (NARA[8]) should not be queried for any date before 2006). The rules are pre-edited with the installation of the proxy, and can be edited by external editor. The optimizer may depend on heuristic optimization based on the original page content (if it exists).

The evaluation of the success of this module will be done on quantitative basis. A sample set of URIs will be queried on both the traditional (query all archives each time) and the new one. Both the number of mementos (to measure the coverage), and the processing time (to measure the performance) for each run will be collected. The test will be repeated after that with each type of rules such as: domain rules, datetime rules, or heuristic rules. A combination between these rules will be covered too. The target from this evaluation is to make sure the new optimization technique provided a better performance without affecting the coverage of the mementos.

4.3 Proxy Dispatcher

The proxy dispatcher module is responsible of running simultaneous requests to the archives list that were selected by the aggregator request optimizer. The dispatcher will send asynchronous requests to the proxies, after that the dispatcher will run as a push server to check if any of these requests returned data, then this response will be forwarded to the the aggregator response optimizer. If the requests did not answer in a specific time period (it will be determined by each proxy based on the archive characteristics), it will be terminated. When all the requests have been returned (or terminated), a termination signal will be sent to the aggregator response optimizer. The scalability of this module will be covered by how to install a new proxy to the proxies list on fly. The idea here could be extended to provide the users with a partial TimeMap instead of the full TimeMap in a faster time.

The evaluation of the success of this technique will follow the same quantitative technique. A sample set of URIs will be queried on both the traditional (synchronous and blocking proxies calls), and the proposed asynchronous technique. We will measure the total processing time for the URI sample set, the total processing time for each URI individually, and the processing time for each proxy. From these measurements, we will conclude a total delay for the experiments, a single delay for each URI as the difference between the total processing time of the URI - the longest processing time for the proxy list. Also, the allocated memory space for both techniques will be calculated.

As future work in this module, we aim to give the proxy dispatcher the ability to load balance the requests with others memento aggregators.

4.4 Aggregator Response Optimizer

This module is responsible for assuring the quality of the archived copies, and sorting the results on different way (Memeto-Datetime, mementos quality, or favorite archived based on users preferences). This module will implement the new archive quality measurement to determine a level of the quality of the archived copy from the user perspective.

The evaluation of success of this technique started by creating a baseline for the bad mementos, and creating a list with these mementos with its quality level. Then, we will test the Enhanced Aggregator for each quality level, we will measure how much the aggregator could detect with the right quality level. The performance issue will be considered to detect how long will it take to detect each level.

4.5 Response Format

The response format will follow the Memento Internet draft [25]. We will study different techniques to communicate the quality information to the client. One technique is to use the "Link" http header to propose a new "qualityMap" which is a list of mementos sorted by the quality level. Another technique for the timemap, it is easy to set a new field with the quality level. These aspects will be studied and implemented without affecting the general memento protocol.

4.6 Enhanced Aggregator Client

The current available Memento client which entitled "MementoFox" [24] provides limited capabilities to enter the Time Travel mode. We propose a new Memento Client that could take benefits from the enhanced aggregator. The new client will have new features such as: using different TimeGates in the same time and render the results together, an easy interface that provides the user with the ability to browse the memento timeline. The client will be design to provide to way communications, additional to retrieve and present the data, the users could use the client to contribute to the enhanced aggregator. For example, the user could use the client to suggest new archives which should be included into the TimeGate or to rate the quality of the mementos. The general idea here is how can the aggregator benefit from the client activities.

5. RESEARCH OVERVIEW

5.1 Research Plan

The research covers different aspects from providing all the possible archived copies (functionality), in a reasonable time (performance) and filter the archived copies based on the good one (quality), and finally providing all these features to the user in an ease to use approach (usability). The research will differ based on the feature under the spot. Table 1 shows a timeline for the tasks and the expected plan for the research.

[7] www.collectionscanada.gc.ca/webarchives/
[8] www.archives.gov

Table 1: Research Plan Timeline

Time Frame	Module Contribution and Tasks	Evaluation
04/2011 - 08/2011	**Request Optimizer** • Studying the different Web archives characteristics • Designing and building the rule system • Building the Aggregator Request Optimizer	Select a sample set of URIs, and query both of the original and the enhanced aggregator and measure for both the coverage and performance
09/2011 - 11/2011	**Proxy Dispatcher** • Studying the best distributed implementation technique • Applying deadlock avoidance technique • Building the Proxy Dispatcher component	Run the experiment on a sample set of URIs on the traditional and the enhanced aggregator and measure the performance for each part. Another mathematical analysis will be required after collecting the data
12/2011 - 05/2012	**Aggregator Response Optimizer** • Studying and defining the new quality measurement • Studying techniques to quantify these measurement • Building the Aggregator Response Optimizer • Propose the updates in the response format	We will select a baseline for the bad mementos and run the experiment to calculate the successfulness of the new component to quantify these examples.
06/2012 - 09/2012	**Enhanced Aggregator Client** • Enable the user to select date, and date range to browse the Web during this period • How to visualize the results to the user • How the user could send his feedback to the aggregator	Additional to the quantitivate approach to measure the functionality and performance of the client, a controlled group of people will be asked to evaluate the interface from their perspective.

5.2 Research Methodology

The research depends on a bottom-up approach. First, we started with a detailed study of the available archives [2]. In this paper, we conducted a quantitative study about the percentage of the Web is archived. Also, we are preparing a survey about the current and prospective archives that have proxies in the aggregator to understand the nature of each one. Based on this survey study, we are going to prepare a list with the rules that could be used by Aggregator Request Optimizer; in this part, we will set the format and repository for the rules. Then, the Aggregator Request Optimizer will be implemented and evaluated.

Then, in the proxy dispatcher module, as it is one of the performance bottleneck in the system, it will be designed based on parallel distribution of the requests, and provide immediate return to the upcoming results to the Aggregator Response Optimizer. In this module, we will investigate in different implementation aspects for the distributed processing to reach a better performance with a managed memory consumption and avoid the deadlock drawback.

An abstract study of how to evaluate the Web archive from the archivists perspective and from the end users perspective. In this part, we will formulate the different definitions for each measurements, additional to how we could calculate it. The implementation of these techniques will be embedded in the Aggregator Response Optimizer; The suitable response format will be studied after that.

Finally, the Enhanced Memento Client will be implemented, a new visualization technique of the retrieved results will be presented. Different services will be added one by one to the client implementation and study its functionality and effect. This research aims to provide an easy way to the novice Web users to access the past Web in easy way.

5.3 Contributions of the work

In this dissertation, we propose a new framework for an enhanced model for the Memento Aggregator to collect all the possible archived copies for specific URI. A new Web archive quality measure will be proposed to evaluate the archived copies from the user perspective. We proposed a new standards which is "Aggregator Optimization" to define a set of rules and techniques to enhance the coverage, and quality with an acceptable performance for the archive aggregator. We plan to make the aggregator optimizer as an open standards that is not limited to the Memento protocol only. This work will be supported with a Memento Client which will enable the user to travel through the time transparently.

6. CONCLUSIONS

In this proposal, we have described an enhanced framework for the traditional Memento Aggregator that entitled "Enhanced Memento Aggregator". The new framework is distinguished from the traditional one in the ability to filter the archive queries based on the request URI parameters, the ability to run the query through the proxies through new distributed and asynchronous way, and finally the new framework provides a quality measure for the returned TimeMap response based on quality controls technique to rank the memento. We also proposed a new memento client which enables the user to get benefits from these services, additional to give the user the ability to contribute to the Aggregator proxies.

7. ACKNOWLEDGMENTS

I'm thankful and grateful to my advisor, Dr. Michael L. Nelson for his efforts and guidance in this proposal. Also, I would like to thank Prof. David Rosenthal from Stanford University for his comments about Memento which inspiring me to enhance our model. The Memento work has been partially funded by the Library of Congress.

8. REFERENCES

[1] E. Adar, M. Dontcheva, J. Fogarty, and D. S. Weld. Zoetrope: interacting with the ephemeral web. In *Proceedings of the 21st annual ACM symposium on User interface software and technology*, pages 239–248. ACM, 2008.

[2] S. Ainsworth, A. Alsum, H. SalahEldeen, M. C. Weigle, and M. L. Nelson. How much of the Web is Archived? In *JCDL*, 2011.

[3] A. Brown. *Archiving websites: a practical guide for information management professionals*. Facet, London, first edit edition, 2006.

[4] J. Cho and H. Garcia-Molina. Effective page refresh policies for Web crawlers. *ACM Transactions on Database Systems (TODS)*, 28(4):390–426, 2003.

[5] D. Denev, A. Mazeika, M. Spaniol, and G. Weikum. The sharc framework for data quality in web archiving. *The VLDB Journal*, 20 (to appear):(to appear), March 2011.

[6] D. Gomes, A. L. Santos, and M. J. Silva. Managing duplicates in a web archive. In *Proceedings of the 2006 ACM symposium on Applied computing*, SAC '06, pages 818–825, New York, NY, USA, 2006. ACM.

[7] K. Holtman and A. Mutz. Transparent content negotiation in http, 1998.

[8] A. Jatowt, Y. Kawai, S. Nakamura, Y. Kidawara, and K. Tanaka. Journey to the past: proposal of a framework for past web browser. In *Proceedings of the seventeenth conference on Hypertext and hypermedia*, page 144. ACM, 2006.

[9] A. Jatowt, Y. Kawai, H. Ohshima, and K. Tanaka. What can history tell us?: towards different models of interaction with document histories. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 5–14. ACM, 2008.

[10] M. Klein, M. Aly, and M. L. Nelson. Synchronicity - Automatically Rediscover Missing Web Pages in Real Time. In *In JCDL 2011: Proceedings of the 11th ACM/IEEE-CS joint conference on Digital libraries*, 2011.

[11] J. Masanès. Web Archiving Methods and Approaches: A Comparative Study. *Library Trends*, 54(1):72–90, 2005.

[12] J. Masanès. *Web archiving*. Springer, 2006.

[13] F. Mccown, J. A. Smith, M. L. Nelson, and J. Bollen. Lazy Preservation: Reconstructing Websites by Crawling the Crawlers. In *Proceedings from the 8th ACM International Workshop on Web Information and Data Management (WIDM 2006)*, Nov. 2006.

[14] M. Najork and J. L. Wiener. Breadth-first crawling yields high-quality pages. In *Proceedings of the 10th international conference on World Wide Web*, WWW '01, pages 114–118, New York, NY, USA, 2001. ACM.

[15] M. L. Nelson. 2010-11-15: Memento Presentation at UNC; Memento ID. `http://ws-dl.blogspot.com/2010/11/2010-11-15-memento-presentation-at-unc.html`, Nov.

[16] M. L. Nelson. Memento-Datetime is not Last-Modified. `http://ws-dl.blogspot.com/2010/11/2010-11-05-memento-datetime-is-not-last.html`.

[17] D. S. H. Rosenthal. Memento & the Marketplace for Archiving. `http://blog.dshr.org/2011/01/memento-marketplace-for-archiving.html`, 2011.

[18] R. Sanderson. Memento Tools: Proxy Scripts. `http://www.mementoweb.org/tools/proxy/`, 2010.

[19] R. Schenkel. Temporal Shingling for Version Identification in Web Archives. In C. Gurrin and U. Kruschwitz, editors, *Proceedings of the 32nd European Conference on Information Retrieval (ECIR 2010)*, Milton Keynes, UK, 2010. Springer.

[20] R. Shiozaki and T. Eisenschitz. Role and justification of web archiving by national libraries: A questionnaire survey. *Journal of Librarianship and Information Science*, pages 90–107, 2009.

[21] M. Spaniol, D. Denev, A. Mazeika, G. Weikum, and P. Senellart. Data quality in web archiving. In *Proceedings of the 3rd workshop on Information credibility on the web*, pages 19–26. ACM, 2009.

[22] M. Spaniol, A. Mazeika, D. Denev, and G. Weikum. Catch me if you can: Visual Analysis of Coherence Defects in Web Archiving. In *The 9 th International Web Archiving Workshop (IWAW 2009) Corfu, Greece, September/October, 2009 Workshop Proceedings*, page 1.

[23] J. Teevan, S. T. Dumais, D. J. Liebling, and R. L. Hughes. Changing how people view changes on the web. In *UIST '09: Proceedings of the 22nd annual ACM symposium on User interface software and technology*, pages 237–246, New York, NY, USA, 2009. ACM.

[24] H. Van de Sompel. MementoFox 0.9.2. `https://addons.mozilla.org/en-US/firefox/addon/mementofox/`, 2010.

[25] H. Van de Sompel, M. L. Nelson, and R. Sanderson. HTTP framework for time-based access to resource states. `https://datatracker.ietf.org/doc/draft-vandesompel-memento/`, 2011.

[26] H. Van de Sompel, M. L. Nelson, R. Sanderson, L. Balakireva, S. Ainsworth, and H. Shankar. Memento: TimeMap API for Web Archives. `http://www.mementoweb.org/events/IA201002/slides/memento_201002_TimeMap.pdf`, 2010.

[27] H. Van de Sompel, M. L. Nelson, R. Sanderson, L. L. Balakireva, S. Ainsworth, and H. Shankar. Memento: Updated Technical Details (February 2010). `http://www.slideshare.net/hvdsomp/memento-updated-technical-details-february-2010`.

[28] H. Van de Sompel, M. L. Nelson, R. Sanderson, L. L. Balakireva, S. Ainsworth, and H. Shankar. Memento: Time Travel for the Web. page 14, Nov. 2009.

[29] H. Van de Sompel, R. Sanderson, M. L. Nelson, L. L. Balakireva, H. Shankar, and S. Ainsworth. An HTTP-Based Versioning Mechanism for Linked Data. In *Proceedings of the Linked Data on the Web Workshop (LDOW 2010)*, 2010.