Computer Assisted Appraisal in Web Archives

An Annotated Bibliography

Ed Summers ehs@pobox.com

In 2008 Google estimated that it had 1 trillion unique URLs in its index (Alpert & Hajaj, 2008). When I looked today (7 years later) the Internet Archive's home page announced that it has archived 438 billion Web pages. It's an astounding achievement, but the Web has certainly grown many times over since 2008. Also, it's important to note the difference in terminology: *URL* versus *Web page*. A Web page has a unique URL, or address, but the content of a Web page can change over time. Capturing the record of documents as they change over time is essential for Web archives. So by design, there are many duplicate URLs included in the 438 billion Web pages that the Internet Archive has collected. If you ignore the duplicates and the fact that the Web has grown, it looks like Internet Archive has archived 37% of the Web. But if you consider the growth of the Web and the duplicates that are present in the archive this estimate is far too high.

As more and more information is made available on the Web how do archivists decide what to collect, and when? The Internet Archive's Heritrix bots walk from link to link on the Web archiving what they can. Members of the International Internet Preservation Consortium (IIPC) run their own crawls of specific parts of the Web: either country domains like the .uk top-level-domain, or specific websites that have been deemed within scope of their collection development policy. These policies inform the appraisal of whether particular Web content is deemed worth adding to an archive. Archivists are aware of how these appraisal decisions shape the archive over time, and by extension also shape what we know of our past. Appraisal, or deciding what to save, and what not to save, is difficult in the face of so much information.

This annotated bibliography provides a view into the emerging field of computer assisted appraisal in Web archives. How can computers assist archivists in the selection of content for archiving? Similarly, how can archivists guide the appraisal and crawling of Web content? There are two primary themes that emerge in this review: identification and evaluation. This review is not meant to be complete, but rather to be suggestive of a field of study at the intersection of archival and computer science.

1. Finding Content

The following papers discuss ways of discovering relevant content on the Web. Particular attention has been paid to approaches that incorporate social media into appraisal decisions.

Jiang, J., Yu, N., & Lin, C.-Y. (2012). FoCUS: Learning to crawl Web forums. *WWW 2012 Companion.*

As more content goes on the Web researchers of all kinds are increasingly interested in analyzing Web forums in order to extract structured data, question/answer pairs, product reviews and opinions. Forum crawling is non-trivial because of paging mechanisms that can result in many duplicate links (different URLs for the same document) which can consume large amounts of time and resources. For example the researchers found that 47% of URLs listed in sitemaps and feeds were duplicates in a sample of 9 forums.

Jiang et al. detail a procedure for automatically detecting the structure of forum websites, and their URL types, in order to guide a Web crawler. The research goal is to save time, and improve coverage compared to breadth-first and other types of crawlers. The process is to automatically learn Index-Thread-Page-Flipping (ITF) regular expressions for identifying the types of pages in Web forums, and then use these patterns during the Web crawl.

The researchers studied the structure of 40 different Web forum software platorms to find common patterns in page layout/structure as well as URL and page types. For example, timestamps on pages in chronological and reverse chronological order are good indicators of thread and index pages respectively. Also, paging elements can be identified by noticing links with longer than usual URLs combined with short numeric anchor text. A training set for four different web forums was fed into a Support Vector Machine classifier, which was then used to generate ITF regular expressions for each site.

To analyze their procedure they selected nine different types of forum software and ran three types of crawlers over each: a generic crawler, an entry point crawler and a structure driven crawler. The measured effectiveness and coverage were reported for each combination. Experimantal results found that the structure driven crawler significantly outperformed the other types of crawlers. The authors note that these results have bearing on other types of similarly structured sites such as question/answer sites and blogs. They also hope to improve the 97% coverage by handling JavaScript paging mechanisms which were present in 2% of the forums tested.

On the surface this paper doesn't seem to have much to do with automated appraisal in Web archives. But the authors demonstrate that attention to the detail and structure of websites can improve efficiency and accuracy in document collection. Forums, blogs and question/answer sites are very common on the Web, and represent unique and high value virtual spaces where actual people congregate and share opinions on focused topics. As such they are likely candidates for appraisal, especially in social science and humanities focused Web archives. The ability to automatically identify forums on particular topics as part of a wider web crawl could be a significantly important feature when deciding where to focus archiving resources. In addition, this work presents important heuristics for identifying duplicate content, which is important for knowing what not to collect, as we will see later in Kanhabua, Niederée, & Siberski (2013).

Gossen, G., Demidova, E., & Risse, T. (2015). ICrawl: Improving the freshness of web collections by integrating social web and focused web crawling. In *Proceedings of the Joint Conference on Digital Libraries*. Association for Computing Machinery. Retrieved from http://www.l3s.de/~gossen/publications/gossen_et_al_jcdl_2015.pdf

This paper draws upon a significant body of work into focused Web crawling and specifically work done as part of the ARCOMEM project. Gossen et al. provide an important analysis of how the integration of social media streams, in this case Twitter, can significantly augment the freshness and relevance of archived Web documents. In addition they provide a useful description of their system and its open source technical components to help others to build on their work.

The analysis centers on measuring relevancy and freshness of archived content in two contemporary Web crawls related to the Ebola outbreak and the conflict in the Ukraine. In each case four different Web crawls were run: unfocused, focused, Twitter-based and integrated. Each type of crawl begins with a seed URL and wanders outwards collecting more results. The focused crawl uses their own link prioritization queue to determine which pages get collected first. The Twitter based crawler simply crawls whatever URLs that are mentioned in relevant tweets from the Twitter API. The integrated crawler is a combination of the focused and Twitter based crawlers, and represents the main innovation of this paper.

The results show that the Twitter based search is able to return the freshest results, with the integrated crawler coming in second. However the integrated crawler performed best at returning the most relevant results. Freshness on the Web is difficult to measure since it involves knowing when a page was first published, and there is not consitent metadata for that. The researches devised some hueuristics for determining creation time, and eliminated pages from the study for which freshness couldn't be determined. The relevancy measure is also used by the prioritization queue, so in some ways I am concerned that relevancy was only measuring itself. But it is interesting that relevancy was improved while factoring in the Twitter stream. One area of related research that could build on this work is how feedback from archivists or curators could influence the system.

Yang, S., Chitturi, K., Wilson, G., Magdy, M., & Fox, E. A. (2012). A study of automation from seed URL generation to focused web archive development: The CTRnet context. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 341–342). Association for Computing Machinery.

This paper introduces the idea of using social media streams, in this case Twitter, to determine a list of seed URLs to archive in time sensitive situations such as natural disasters and other crises. In these time sensitive situations it is difficult for archivists to build a list of potential seed URLs to harvest, because of the large amount of new content being published on the Web in a very short time period. The goal was to prototype and test a system that could run with minimum human intervention. This was the first reference I could find of using Twitter in this way to augment web archiving, which was discussed more fully in Gossen et al. (2015).

The authors created a prototype Python Django application that manages the workflow of relevant tweets, to URL extraction, to web crawling with Heritrix, and the data extraction. The external service TwapperKeeper was used to collect the Twitter data, which is no longer available as a service today. Details about the data extraction did not seem to be included in the paper. The study used 5 different contemporary events to study the precision of the system: the Virginia Tech shooting, a Measles outbreak, a typhoon in the Philippines, violence in Sudan, and Emergency Preparedness. The results showed that precision varied depending on the type of query used. In some cases a query picked up unrelated Web content because it was unintentionally broad. The paper mentioned a filtering component for reducing spam, but did not discuss it in detail. It also gives precision results without really discussing ther method for measuring it. But there was a poster that accompanied the paper, so perhaps these details could be found there.

The paper does a nice job of introducing a new idea (social media streams in Web archiving) and sets the stage for future work in terms of how to filter out spam and measure precision. Similar to Gossen et al. (2015) it hints at future work that could integrate a archivist or curator who can influence the direction of the crawl as part of the process.

Pereira, P., Macedo, J., Craveiro, O., & Madeira, H. (2014). Time-aware focused web crawling. In *Advances in Information Retrieval* (pp. 534–539). Springer.

As discussed in Gossen et al. (2015), determining the time a Web page was published, or its freshness can be surprisingly difficult. When you view a Web page it is most likely being served directly from the originating Web server on the Internet, or perhaps from an intermediary cache that has a sufficiently recent copy. However it is useful to be able to determine the age of a page, especially when ordering search results, and also for appraising a given web page in an archival setting.

Pereira discusses a technique for crawling the Web in a time-aware way. Most previous work on focused web crawling has involved topic analysis (the text in the page and its similarity to the desired topic). This paper details a process for determining the age of a given Web page (temporal segmentation), and then integrating those results into a Web crawler's behavior (temporal crawling).

The paper describes an experiment that compares the results of crawling two topics (World War 2 and September the 11th) by crawling outwards from Portuguese Wikipedia pages, using two different techniques: no time restriction and a time restriction. The results indicate that the crawl with a time-restriction performs significantly better over time, however the shape of the results is different for each topic.

The authors admit that the results are preliminary, and that their project is a proof of concept. Unfortunately the authors don't appear to provide any source code for their prototype. It would be interesting to compare the time-based crawling with more traditional topic-based crawling, and perhaps consider a hybrid approach that would allow both approaches to be used in a single crawl.

2. Evaluating Content

Once content is identified and retrieved there are a set of factors that can be considered to help inform a preservation decision about the content. Metrics that can be generated without significant human intervention are important to highlight, as are systems that allow interventions from an archivist to shape the appraisal process.

Lyle, J. A. (2004). Sampling the umich. edu domain. In *4th International Web Archiving Workshop, Bath, UK* (Vol. 2). Retrieved from http://iwaw.europarchive.org/04/Lyle.pdf

This paper is part historical overview of sampling in traditional archival appraisal, and part a study of sampling in the Web archive records for the umich.edu domain. Lyle provides an excellent overview of passive and active appraisal methods, and how they've been employed to help shape archival collections. Active appraisal largely came about as the result of an over abundance of records in the post World War 2 era. However a partial shift back to passive appraisal was observed as electronic records became more prevalent, storage costs plummeted, and it became conceivable to think of collecting everything. In addition it became possible to automatically crawl large amounts of Web content given the structure of the World Wide Web. At the same time there was a movement towards active appraisal, where archivists became more involved with record creation, to insure that electronic documents use particular formats and have standard metadata.

Lyle also discusses the benefits and drawbacks of several different types of document sampling methods: purposive, systemic, random and mixed-mode sampling. The intent is to use these methods on records of high evidential value, but not on records of high informational value. The distinction between informational and evidential value introduced by Schellenberg isn't clearly made, and Lyle questions whether information documents are always more valuable to the record than evidential documents.

In the second half of the paper Lyle documents the results of a Web crawl of the umich.edu domain performed by the Internet Archive. This focused crawl was performed for the purposes of this study, and identified four million URLs. Only 87% of these URLs were working (not broken links) and almost half were deemed to be duplicates. An analysis of different types of sampling was performed on the resulting 1.5 million documents using information from crawl logs: size of the document, size of URL. The study looked specifically at bias in the sample results, and found that stratified random sampling worked best; although details of how the bias in results was ascertained was not discussed.

In the discussion of the results Lyle surmises that sampling is a useful way to get an idea of what sub-collections are present in a large set of Web documents, rather than a criteria for accessioning itself. He notes that to some extent the whole process was a bit suspect, since the crawl itself potentially had inherent bias: the chosen entry point, the algorithm for link discovery, and the structure of the graph of documents. The author's specific conclusions are somewhat unclear but indicate that more work is needed to study sampling in Web archives. Sampling culd be a useful appraisal tool to discover the shape of collections, but is not an explicit mechanism for determining whether to preserve or destroy a particular document. The design of such a sampling tool that could inform appraisal decisions and be integrated with Web crawl results could be a valuable area of future work.

Kenney, A. R., Nancy, M., Botticelli, P., Entlich, R., Lagoze, C., & Payette, S. (2002). Preservation risk management for web resources: Virtual remote control in cornell's project prism. *D-Lib Magazine*, *8*(1). Retrieved from http://www.dlib.org/dlib/january02/kenney/01kenney.html

This paper examines uses the technique of risk management to identify factors that can be used in the appraisal of Web documents. These factors center around the document itself, the document's immediate context in the Web (its links), the website that the document is a part of, and the institutional context that the website is situated in. Some of the document and contextual factors are reflected in later work by Banos, Kim, Ross, & Manolopoulos (2013) such as format, standards, accessibility and metadata. A particularly interesting metric mentioned by the authors are monitoring factors such as inbound and outbound links over time, and the shape of the website graph. These measures of change can be used to determine the rate at which it is being maintained. The assumption being that a site that is not being maintained is more of a preservation risk.

A general move is made in this paper to reposition archivists from being custodians of content to being active managers of digital objects on the network. This effort seems to be worthwhile, especially if it is sustained. The discussion would have benefited from references to the existing literature on post-custodial archives which was available at the time. One somewhat discordant part of the paper is that the two project links figured prominently at the top of the article do not go to the PRISM project page, which is still available. Also, in hindsight the criticisms of the Internet Archive seem overly dismissive. The paper would have been better situated in terms of opportunities for establishing a community of practice.

Kanhabua, N., Niederée, C., & Siberski, W. (2013). Towards concise preservation by managed forgetting: Research issues and case study. In *Proceedings of the 10th International Conference on Preservation of Digital Objects, iPres* (Vol. 2013). Retrieved from http://l3s.de/~kanhabua/papers/iPRES2013-Managed_Forgetting.pdf

Appraisal is often thought about in terms of what artifacts to preserve or save for the future. But implicit in every decision to save is also a decision not to forget. Consequently, it's also possible to look at appraisal as decisions about about what can be forgotten. In this paper Kanhabua and her colleagues at the L3S Research Center investigate processes for making these types of decisions, or *managed forgetting* which is materialized in the form of *forgetting actions* such as aggregation, summarization, revised search, ranking behavior, elimination of redundancy and deletion.

The article provides a useful entry point into the literature about human memory in the field of cognitive psychology. It also highlights several jumping off points for HCI discussions about designing systems and devices for managing memory. But the primary focus of the paper is on the interaction between information management systems and archival information systems: the first which is used to access information, and the second being the stores of content that can be accessed.

In order to describe how the act of forgetting is present in these systems the authors used historical snapshots of public bookmarks available by the BibSonomy social bookmarking project. The 15 BibSonomy snapshots taken at different periods of time provide a view into when users have chosen to bookmark a particular resource, as well as when that resource has been deleted. Their analysis determined that there was a correlation between a users delete ratio and the number of bookmarks they created, but not between the users delete ratio and the total number of bookmarks they possessed.

The paper admits that they are still in a very early phase of research into the idea of *managed forgetting*. I think the paper does a nice job of articulating why this way of looking at appraisal matters, and provides an example of one possible study that could be done in this area. I think it would have been useful to discuss a little bit more about how the choice of BibSonomy as a platform to study could have potentially influenced (but not invalidated) the results. It would be interesting to take another social bookmarking site like Pinboard or Digg and see if a similar correlation holds. The implications of managed forgetting for building digital preservation and access systems seems like a very viable area of research, and I hope to see more of it.

Banos, V., Kim, Y., Ross, S., & Manolopoulos, Y. (2013). CLEAR: A credible method to evaluate website archivability. *International Journal on Digital Libraries*, 1–23.

CLEAR stands for Credible Live Evaluation of Archive Readiness which is a process for measuring *website archivability* of a particular Web document. The paper provides a method for generating an archivability score based on a set of five *archivability facets*: accessibility, standards compliance, cohesion, performance and metadata. The authors created a working prototype called ArchiveReady that you can find on the Web and use to evaluate Websites manually or automatically with their API.

The motivation for the work on CLEAR was traced back to previous work in New Zealand on the Web Curator Tool (WCT) and in the UK on the Web At Risk project which made quality assurance part of the archiving process. Quality assurance was found to be particularly time consuming, which consequently slowed down the work of timely processing. Banos et al.'s goal with CLEAR is to provide a measure of archivability that allows archivists to select a quality threshold under which Web content would be greenlighted for accession into an archive.

The paper includes useful details about the technical system: Python, Flask, Backbone and MySQL for the Web application, Redis for managing parallel processing, and JHOVE for file identification. In addition the precise formula for generating the CLEAR metric was clearly described. An analysis of CLEAR results compared with quality assurance results from a human curator would have been useful. It would be interesting to see if automated and human appraisal decisions are correlated, and also what likely threshold values could be set to.

SalahEldeen, H. M., & Nelson, M. L. (2013). Carbon dating the web: Estimating the age of web resources. In *Proceedings of the 22nd International Conference on World Wide Web Companion* (pp. 1075–1082). International World Wide Web Conferences Steering Committee. Retrieved from http://arxiv.org/abs/1304.5213

As Gossen et al. (2015) also discusses, it is often important to identify when a document was first added to the Web. The age of Web documents, or their freshness, is important for digital library research, as well as for making informed appraisal decisionns Yang et al. (2012). Determining the age of Web documents can be difficult when the page itself lacks an indicator of when it was created. Metadata such as the Last-Modified HTTP header are not typically reliable as a source for create date since publishers often change it to encourage reindexing by search engines , and to influence cache behavior. Therefore alternative methods need to be invented.

SalahEldeen show how trails of references, citations and social media indicators can be used to estimate the creation time for a Web document. The papers describes, and also demonstrates (through a prototype application) a mechanism for estimating Web page creation time using backlink discovery (Google), social media sharing of the document (Twitter using the Topsy API), archival versions available (Memento Aggregator API) and URL shortening services (Bitly).

The authors tested their system by creating a corpus of 1,200 Web documents from popular media sites with clearly marked creation times, and then used their algorithm to guess the creation time. The results showed that they were able to determine the correct creation date in 75% of the cases. However the Google backlinks and Bitly short URLs had little effect on the result. The determining factors were the archival snapshots available from Web archives and the references found in Twitter. Future areas of research would be to identify other potential social media indicators such as ones from Facebook, Instagram and Twitter. Do these exhibit similar behavior? Also, it would be interesting to see how well the process works when using a baseline of comparison of pages that are not from large media outlets, and may not be as well represented in Twitter.

Brunelle, J. F., Kelly, M., SalahEldeen, H., Weigle, M. C., & Nelson, M. L. (2014). Not all mementos are created equal: Measuring the impact of missing resources. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 321–330). IEEE Press.

Brunelle et al. observe that the nature of the Web has changed significantly in the last 15 years. Specifically there is an increasing amount of dynamic content being made available with JavaScript and content from Web service APIs. This content has historically been difficult to archive because of the asymmetry between the technology used to archive content (the Web crawler) and the technology used for accessing archived content (the Web browser). Their study aims to measure the degree to which 1) this hypothesis is true and 2) the degree to which this impacts the experience of using archived content over time.

The first part of the study uses two sets of 1000 URLs: one being Bitly URLs found in Twitter, and the other being a sample of URLs found in ArchiveIt collections. These two sets of URLs were deemed to be quite different in terms of their source and type. A measure of URL complexity was used to characterize the URLs from each source, which showed that URLs obtained through Twitter were significantly more complex that those from ArchiveIt. Three different archiving tools (wget, Heritrix and WebCite) were then used to archive the URLs, and then results were compared using an instance of the Wayback Machine. The Wayback Machine ran on a server disconnected from the Internet in order to highlight potential leakage (URLs that targeted the live Web instead of the archive). A headless Web browser (PhantomJS) was used to measure the types of requests and their results. Results showed that the Twitter dataset is much more difficult to archive, and that this is the result of reliance on JavaScript for complete rendering.

The second part of the study looks at the combined set of Twitter and ArchiveIt URLs, and identifies ones that are available for the 2005-2012 time period. Mementos, or snapshots of the pages, were then retrieved from the Internet Archive and the number of requests coming from HTML vs JavaScript was measured. The authors were able to show that between 2005-2012 there was a 14.7% increase in JavaScript use. More striking was the finding that over the same period the number of missing resources due to JavaScript rose from 39% to 73.1%.

The format of this paper was somewhat hard to digest in that it really felt like two separate studies in one. The results were significant both for Web archive crawlers that must integrate JavaScript execution in order to be create full fidelity websites. In addition the study highlighted the need for easy to use curator tools that help identify leakage in the Web archive content. Ideally there should be a solution which does not require the archivist to run a local Web archive server (Wayback Machine) with the Web archive data held locally. The implications for Web publishers were also significant, if archivability and accessiblity of their web content is of interest. Another avenue to explore would be an *archivability* metric that could be derived through an analysis of the page, which could be useful when appraising content from a Web crawl.

References

Alpert, J., & Hajaj, N. (2008, July). We knew the web was big. Retrieved from https://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html