Internet Archives as a Tool for Research: Decay in Large Scale Archival Records

Hai Nguyen
Department of Computer Science
Rutgers University
New Brunswick, NJ
e-mail: hdn11@cs.rutgers.edu

Matthew S. Weber
Department of Communication
Rutgers University
New Brunswick, NJ
e-mail: matthew.weber@rutgers.edu

Abstract— **Web archiving provides social scientists and digital humanities researchers with a data source that enables the study of a wealth of historical phenomena, from organizational growth, to the dispersion of ideas across the Internet. One of the most notable effort to record the history of the World Wide Web is the Internet Archive (IA) project, which maintains the largest repository of archived data in the world. Understanding the quality of archived data and the completeness of each record of a single website is a central issue for scholarly research, and yet there is no standard record of the provenance of digital archives. Indeed, although present day records tend to be quite accurate, archived Web content deteriorates as one moves back in time. This paper analyzes a subset or archived Web data, measures the degree of degradation in a subset of data, and proposes statistical inference to overcome limitations of the data.**

Keywords-component; formatting; style; styling; insert (key words)

I. Introduction

Large-scale data is increasingly becoming a critical resource for social scientists seeking to explore a wide range of phenomena. For researchers aiming to analyze or reconstruct historical phenomenon from the past three decades, the live Web is an incomplete and inaccurate resource [1]. Web linking traces extracted from Internet archive data have been utilized to explore the effects of hyperlinking on the long-term survival of online newspapers [2, 3]. In addition, archived blog, newspaper and social media content extracted from online sources has been analyzed to examine the memetic transmission of news over time [4]. Moreover, archived social media content is increasingly being utilized to explore social movements and collective action [5]. In one particular case, researchers analyzed Twitter records to explore how activists in the Occupy Wall Street movement coordinated efforts in a new form of digital collective action [6]. In addition to demonstrating the validity of large-scale data sources for research, the findings from this work is leading to the development of new theories, and enhancing our understanding of the world we live within.

This article outlines a particular approach for accessing Internet archive data, referred to as HistoryTracker [7], and utilizing the subsequent output for research in the social sciences. This research is situated at the crossroads of advances in Big Data computing, and the emerging social science field of computational social science [8]. Subsequently, based on initial applications of HistoryTracker in a research context, this article discusses critical barriers to the implementation of large-scale data sources in the social science. In particular, although scholars have lauded the potential of Big Data to transform research, the reality is that large-scale archived data is problematic in that the provenance of the data is often unknown, the accuracy of the archived records is difficult to determine, and the records themselves often contain inaccuracies. Having outlined these core issues, and recognizing that certain issues cannot be overcome, we propose one method for addressing the accuracy of archived Internet records, and suggest a course of action for implementing a corrective measure in future statistical analyses.

II. Web Archives as a Research Resource

The promise of Big Data as a tool for the social sciences has started to come to fruition in recent years. This is due in part to increasing awareness of the variety of data sources that are available, and in part, due to an increase in the number of tools available to access and work with the data. The average life span of a Web page has been calculated at around 3 years [9]; given the billions of Web pages online today, this means that much of the Web is lost to researchers. Archives provide in answer by preserving as much of the Web as is currently possible.

A. *Types of Internet Archives*

There are a wide variety of archived Internet sources that have been made available to the academic community at large, and to public audiences.

Over the past decade, many major libraries – national, regional and local – have shifted resources from traditional print archives to the creation of digital libraries. For instance, the British National Library (http://buddah.projects.history.ac.uk) has funded a number of distinct digital archives, as has Columbia University (see http://hrwa.cul.columbia.edu for an example) and Stanford University, just to name a few (http://library.stanford.edu/projects/web-archiving). Other efforts worth mentioning exist at Bibliothèque nationale de France and Die Deutsche Bibliothek (the German National Library). At a national level in the United States, the Library of Congress has taken a number of steps to preserve Internet resources, including the archiving of Congressional websites, and the construction of an archive of Twitter data. Beyond these collections, the International Internet Preservation Consortium (IIPC) and the Internet Memory Foundation represent two broad efforts to standardize Internet archiving.

Access to corporation-owned social media data has proved particularly problematic, as websites such as Facebook and Twitter restrict crawling for archiving and research purposes. In recent years, a number of corporations have run “data challenges” that have made data available to a subset of research, although access has been limited in scope.[1]

B. *The Internet Archive*

The growth of Internet archives is notable, but the largest single repository in existence continues to be the Internet Archive, a non-profit organization based in San Francisco, CA [10]. As of 2013, the Internet Archive contains more than seven petabytes of data and offers a reliable historical record of Web sites dating from 1995 to the present. The Internet Archive is the largest digital source for historical research pertaining to the Web and its contents over time. Much of the Internet Archive’s data is publically available via the Wayback Machine (www.archive.org) interface to the Internet Archive serves more than 300,000 visitors a day, and more than 200 requests a second. In addition, a vast amount of data is aggregated in reserve. Furthermore, through the Archive-It (www.archive-it.org) program, the Internet Archive now supports the formation of custom archives based on particular institutional needs as a subscription service.

C. *Challenges Associated with Internet Archive Data*

Archived Internet sources are imperfect at best. For instance, one analysis of archived Twitter data covering six major social events from 2009 through 2012 found that after one year approximately 11% of Tweets were no longer recoverable, and after 2 years 27% of Tweets were no longer recoverable. In addition, after 2 years, only 41% of the original content had actually been archived for future use [11]. When users are unable to find content on the “live” Web, they turn to Internet archives. One estimate found that approximately 65 % of requested archived pages from user searches access pages that no longer exist on the live Web [12].

The Internet Archive continues to be the most complete record of archived Internet data, but it is a spotty record at best. The Internet Archive starts in 1996, so the first four years of the Web are missing. The early collections of the Internet Archive relied on donations of data from third parties; the Internet Archive did not begin actively crawling until after 2000. Moreover, dynamic content is difficult to archive, and is often omitted from the Internet Archives records; robots.txt files also prevent the Internet Archive from capturing many records [10, 13].

III. Systems for Access

A. *Research Initiatives to Open Access to Archives*

A number of simultaneous efforts are underway to build new tools to improve research access to archival Internet data. For instance, the L3S Research Center at Leibnitz University in Hannover, Germany, is currently working to build a suite of tools to access, navigate and visualize archived Internet data as part of the Alexandria Project [14, 15]. Elsewhere, the RESAW project led by Aarhus University aims to establish a research infrastructure for archival Internet data, and is helping to support European efforts to archive the Web. Among other ongoing initiatives, the Memento project, started at Old Dominion University, has built new tools for ongoing efforts to archive the Web, including a Google Chrome extension that allows users to create individual archives [16].

B. *HistoryTracker: Hadoop and Internet Archives*

HistoryTracker is a Java-based data extraction tool that is compatible with a number of standard archival data standards. The HistoryTracker project is part of an initiative funded by the National Science Foundation, intended to open access to archival Internet data to social scientists by leveraging advances in computer science. In complement to other initiatives, the HistoryTracker tool is designed to allow users to search through a corpus of archived Internet data, and to extract data based on a number of criteria. The tool works on a link graph basis, examining the connections that exist between webpages, and extracting link list data that specifies the connections that exist between webpages (specifying the date of the link, and the nature of the originating webpage, including the type of content, the size of the file, and any meta-text that may describe the page).

Specifically, HistoryTracker accepts as input the following; (a) a whitelist of URLs against which HistoryTracker will extract matching URLs, and (b) a whitelist of text, against which HistoryTracker will extract webpages that match a Boolean type search comparing the whitelist to any meta-text available within the WAT record.

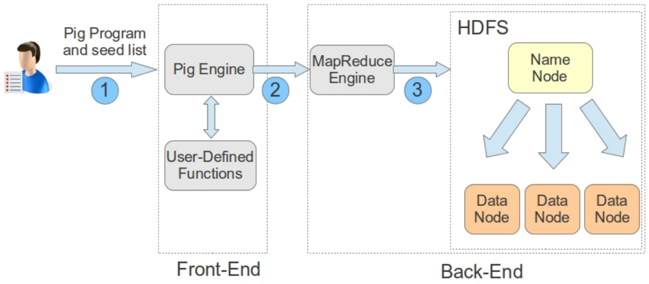
There are two standard Internet archive formats; ARChive file format (ARC) and Web ARChive file format (WARC). The ARC format was developed by the Internet Archive in 1996, and the WARC was subsequently added as an update. The WARC enveloped contains additional metadata that was no present in the original ARC format.[2] A WARC/ARC file may contain multiple archive records. The HistoryTracker tool ingests Web Archive

[1] See, for instance, Twitter’s data challenge (see https://engineering.twitter.com), and a similar effort by LinkedIn (http://economicgraphchallenge.linkedin.com).

[2] See http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml for specification of the WARC format, which is beyond the scope of this manuscript.

Transformation (WAT) files, which are a metadata format that further compresses the WARC/ARC format.[3]

Figure 1. Overview of the HistoryTracker system.



The HistoryTracker application pairs a series of scripts in the PIG programming language with the power of the Hadoop infrastructure to manage the challenges of working with large-scale archival Internet data. In addition, rather than working with the entire corpus of the Internet Archive (which is currently in excess of seven petabytes), the research team chose to extract topical subsets. Topic areas include United States-based media organizations, including blogs and commentary, Occupy Wall Street, .GOV (Senate and House data), and natural disasters (focused on coverage of Hurricane Katrina and Superstorm Sandy). The datasets range in size from 1TB (Hurricane Katrina) to 20TB (US Media). The datasets were created by taking a curated list of seed uniform resource locators (URLs) and extracting all related URLs by crawling out from the seed list.

The datasets provide a testbed for working with Internet Archive data, and additionally each testbed aligns with a current stream of research in the social sciences. Researchers are able to specify extraction based on either of the forms of aforementioned whitelists; the resulting data are outputted in a text file containing the link list structure described above. The result is that HistoryTracker is able to take a large corpus of data, extract a targeted subset, and produce an output format that can be managed by many common statistical packages such as the R framework, STATA or SPSS. The following outlines the technical specifications of the HistoryTracker system, illustrated in Figure 1.

C. *Front-End Design*

Figure 1 illustrates the components of the system. The front-end component receives queries from users and then translates and dispatches queries to the back-end where the queries are processed.

We identified two key requirements for the front-end:

1. The front-end must provide an interface between users and the back-end of the system.
2. The front-end must give users the capability to optimize their query.

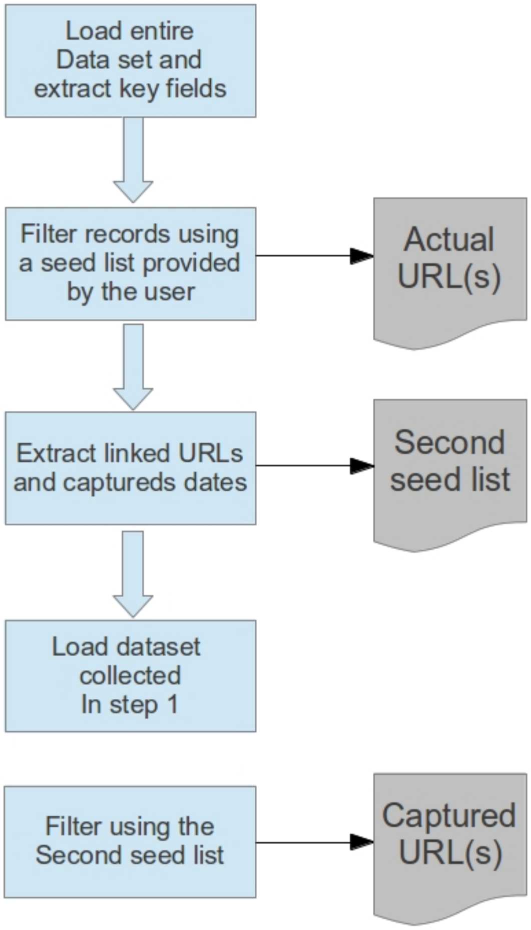
From the above requirements, we choose Apache Pig, an open source platform that supports analysis of large data sets. Users compose their queries using Pig Latin, a high-level dataflow language. Pig then compiles the queries into low-level Map-Reduce jobs.

As shown in Figure 1, in the first step the user starts querying by submitting a program written in Pig Latin language. The user specifies the data set to be processed and a particular date when the sites inside the data set are captured. She also supplies the program with a seed list of URLs. This list consists of URLs that she wants to examine capture quality. By performing early filtering on the seed list and the date, as well as appropriately setting the value of PARALLEL, the user can significantly optimize her queries.

Figure 2 describes the data flows of the Pig program. To enable users to perform the above processing logic, we implement Pig user-defined functions (UDFs) written in Java. The program first extracts important fields from the data set by using one of the UDFs. These fields are shown in Table 1 and are used by Pig in the LOAD operation. The result of the first step is all pairs of URLs and their information that appear in the data set. It next uses one of the UDF to filter the records whose source URLs belong to the seed list provided by the user and the captured date matches the one specified by the user. The seed list is maintained as a hash table to increase the performance of the lookup process. After that the program uses the linked URLs and the captured date of the matched records as a second seed list that is used in next steps.

[3] Please refer to https://webarchive.jira.com/wiki/display/Iresearch/Web+Archive+Transformation+(WAT)+Specification,+Utilities,+and+Usage+Overview for technical specification of the WAT format, which is beyond the scope of this manuscript.

Figure 2. Data flow of the PIG program within HistoryTracker


Load entire Data set and extract key fields
Filter records using a seed list provided by the user
Actual URL(s)
Extract linked URLs and captureds dates
Second seed list
Load dataset collected In step 1
Filter using the Second seed list
Captured URL(s)

The program continues by loading the data set collected in the first step. It then filters that data set for records whose source URLs belong to the second seed list and the captured dates are six months before and after the corresponding dates in the second seed list.

TABLE I. KEY FIELDS FOR EXTRACTION

Fields	Description
srcURL	URL of the site that is captured
dstURL	External links to other sites
Date	The date when the site is captured
Type	The type of link

IV. DATA DEGRADATION IN WEB ARCHIVES

HistoryTracker thus provides a clear path for providing researchers with access to archival Internet data. As HistoryTracker, and other related projects, continue to open up access to these data sources, it is becoming increasingly important that critical issues in terms of data completeness and veracity are addressed.

A central issue in utilizing Internet archives for social science research is that it is difficult to insure that every single page or record is captured. As a result, archived digital trace data are often fragmented. The issue becomes particularly complex when dealing with archived data, as it is difficult to estimate the accuracy of data from prior years as it is difficult to estimate the actual sample size.

Degradation of archived Internet data (Big Data) refers to the fact that archived data sources become less accurate as a user transverses back through time. Prior research has examined aspects of degradation in archived Internet sources. For instance, Spaniol suggests setting crawling strategies that match the degree to which content changes over time in order to insure completeness of an archive crawl [17]. Spaniol and colleagues developed their approach into the archive coherence model. The coherence model focuses on measuring the degree to which a subsection of an archive is complete, and the degree to which that subset captures changes that occur over time. Others have suggested using the size and relative position of a Website in order to determine the importance of archiving a given Website [18]. In part, the lack of alignment regarding archiving strategies compounds the problem, as it is often difficult to determine the provenance of a given data set, and the parameters of a crawl are often not provided (or are not given in a standard format).

In another vein, Brunelle and colleagues [19] crafted a damage measure to indicate the amount of data missing from archived Internet sources. In assessing completeness, they focused on the robustness of a single page record; that is, they measured percent of embedded elements capture in an archive (graphics files, cascading style sheets, and other embedded elements).

The issue with prior approaches is that they focus on coherence as a measure of the completeness of a page, or in the case of the coherence model, the completeness of a subset of pages as they change over time. From a research perspective, this approach has utility, particularly with regards to historical and thematic analyses of content as it evolves over time, as well as linguistic analyses of changes over time.

A. Degradation and Statistical Analysis

The degree of degradation, which we refer to as the degradation factor, leads to significant challenges for statistical analysis of the resulting datasets. Consider, for instance, a use case whereby a researcher seeks to run an standard linear regression on data extracted using the HistoryTracker tool. A core assumption of a linear regression is homocedasticity of the data. Depending on the nature of a given analysis, a change in sampling, or degradation of the data as one travels back in time, potentially opens the door to an increase in the variance of error associated with a dataset. It is equally important to understand the completeness of a crawl, and to be able to defend or explain the sampling method of a given dataset. Thus, by this account, understand the degradation of an archived dataset is an important issue in translation Big Data to large-scale social science research.

B. *Analyzing Degradation of Archived Internet Data*

In order to determine the degradation factor of a large-scale archived Internet data set, we propose examining the link structure in order to measure the completeness of a given crawl of extraction. In this way, we eschew prior approaches that have focused on the completeness of a particular record, or the aggregate completeness of a single time sample. Rather, we focus on understand how many records are actually present within a given extraction.

What follows is an initial assessment of the degradation of a subset of the archival data utilized by the HistoryTracker research team. The aforementioned Pig program was run on a Hadoop cluster consisting of one name node and three data nodes. The program utilized data from the .GOV domain. The web pages in this dataset cover the records for the 109th through the 112th Congresses, across both the senate.gov and house.gov domains. The senate.gov domain records contain 26,965,770 captures representing 8,674,397 unique URLs. The house.gov domain records contain 51,840,777 captures representing 12,410,014 unique URLs.

TABLE II. DEGRADATION SAMPLE FROM SENATE.GOV

Captured Date	Number of captured linked URL(s)	Number of actual linked URL(s)
1/1/09	2	19
10/7/2010	9	22
10/21/2010	85	196
11/18/2010	85	207
11/18/2010	91	215
1/6/2011	77	207

The test case focuses on the senate.gov domain. For this initial analysis, we focused on the root senate.gov URL, and a secondary URL, democrats.senate.gov. We focused on a sample of key records between 2009 and 2011. Table 2 shows a sample of the records from the senate.gov analysis. The degradation analysis was conducted as follows:

1. The target URL was crawled, and key records were extracted utilizing the WAT format.
2. From the WAT record, the outlinks of each archived page were extracted (number of actual linked URLs).
3. A whitelist of outlinks was created, and then re-crawled (number of captured linked URLs).

The resulting dataset provides a basis upon which a degradation curve can be modeled. In order to estimate the degradation curve, we use the number of actual linked URLs (α) and the number of captured URLs (κ), to create a degradation factor (υ) at each time instance, t:

$$\upsilon_t = \alpha_t - \kappa_t$$

For the purposes of the initial estimation of a degradation factor, it was assumed that the time interval was equal between periods, although optimally the time estimation will be more accurate moving forward.

In the case of the senate.gov data, the analysis of υ_t fits to a logarithmic curve, with a strong fit ($R^2 = 0.82$).

$$\varsigma = 3.42 + (75.46 * \ln(t))$$

Where ς is the degradation correction that should be factored into subsequent analysis. The ultimate aim is that the difference, ς, can be used to gauge the acceptability of a given analysis.

A similar analysis reveals results for the democrats.senate.gov URL, given in Table 3, and a linear degradation relationship as follows.

$$\varsigma = 114.53 - (7.43 * t)$$

In the second case, the fit is not as strong ($R^2 = 0.60$); but the relationship is clear.

TABLE III. DEGRADATION SAMPLE FROM DEMOCRATS.SENATE.GOV

Captured Date	Number of captured linked URL(s)	Number of actual linked URL(s)
10/12/2010	40	133
11/3/2010	50	150
1/17/2010	25	135
10/6/2012	16	110
10/23/2012	30	108
1/21/2013	23	81

Interestingly, in this case the records become more complete over time. The degradation factor decreases. This underscores the importance of analyzing the relationship between captured and actual URLs in order to better understand the full scope of the data.

Indeed, the cumulative analysis, however, reveals that there are significant gaps in the data as one crawls out from a single record. Furthermore, degradation appears to increase over time in this small sample, which would make sense given that over time the size of the Internet as a whole (and respective domains) increased, which in turn, crawling capacity did not increase significantly. Ultimately, continued analysis of degradation factors will be an important component for future research.

V. FUTURE DIRECTIONS

As this article has outlined, the promise of archived Internet data for research is significant, particularly within the social sciences. Moreover, it is clear that advances in Big Data analysis, leveraging Hadoop-scale infrastructure, are key to the future of this type of research. The previous discussion and analysis has outlined one particular approach to large-scale research in the social sciences, and has also outlined a number of the challenges associated with this body of work. Future research should continue to explore the challenges associated with historical data analysis.

Acknowledgment

The authors wish to acknowledge support from the National Science Foundation (Grant #1244727), the NetSCI Network Science Research Lab at Rutgers University, as well as the Internet Archive (archive.org) in making this data available for research.

References

[1] P. Wouters, I. Hellsten, and L. Leydesdorff, "Internet time and the reliability of search engines," *First Monday,* vol. 9, 2004.

[2] M. S. Weber, "Newspapers and the Long-Term Implications of Hyperlinking," *Journal of Computer-Mediated Communication,* vol. 17, pp. 187-201, 2012.

[3] M. S. Weber and P. Monge, "Industries in turmoil: Driving transformation during periods of disruption," *Communication Research,* pp. 1-30, 2014.

[4] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM Transactions on Knowledge Discovery from Data,* vol. 1, pp. 1-42, 2007.

[5] W. L. Bennett, "Social movements beyond borders: understanding two eras of transnational activism," *Transnational protest and global activism,* pp. 203-226, 2005.

[6] W. L. Bennett and A. Segerberg, "Digital media and the personalization of collective action: Social technology and the organization of protests against the global economic crisis," *Information, Communication & Society,* vol. 14, pp. 770-799, 2011.

[7] M. S. Weber, "Observing the Web by Understanding the Past: Archival Internet Research," *WWW'14 Companion Proceedings,* 2014.

[8] D. Lazer, A. Pentland, L. A. Adamic, S. Aral, A.-L. Barabasi, D. Brewer*, et al.*, "Computational Social Science," *Science,* vol. 323, pp. 721-723, 2009.

[9] T. Agata, Y. Miyata, E. Ishita, A. Ikeuchi, and S. Ueda, "Life span of web pages: A survey of 10 million pages collected in 2001," in *Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on*, 2014, pp. 463-464.

[10] S. G. Ainsworth, A. Alsum, H. SalahEldeen, M. C. Weigle, and M. L. Nelson, "How much of the web is archived?," presented at the Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries, Ottawa, Ontario, Canada, 2011.

[11] H. M. SalahEldeen and M. L. Nelson, "Losing my revolution: how many resources shared on social media have been lost?," in *Theory and Practice of Digital Libraries*, ed: Springer, 2012, pp. 125-137.

[12] Y. AlNoamany, A. AlSum, M. C. Weigle, and M. L. Nelson, "Who and what links to the Internet Archive," *International Journal on Digital Libraries,* vol. 14, pp. 101-115, 2014.

[13] C. McKay, "Ephemeral to enduring: the Internet Archive and its role in preserving digital media," *Information Technology and Libraries,* vol. 23, p. 3, 2004.

[14] N. R. Asheghi, S. Sharoff, and K. Markert, "Designing and Evaluating a Reliable Corpus of Web Genres via Crowd-Sourcing."

[15] N. K. Tran, A. Ceroni, N. Kanhabua, and C. Niederée, "Back to the Past: Supporting Interpretations of Forgotten Stories by Time-aware Re-Contextualization," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 2015, pp. 339-348.

[16] S. G. Ainsworth and M. L. Nelson, "A Framework for Evaluation of Composite Memento Temporal Coherence," *arXiv preprint arXiv:1402.0928,* 2014.

[17] M. Spaniol, D. Denev, A. Mazeika, G. Weikum, and P. Senellart, "Data quality in web archiving," presented at the Proceedings of the 3rd workshop on Information credibility on the web, Madrid, Spain, 2009.

[18] R. Song, H. Liu, J.-R. Wen, and W.-Y. Ma, "Learning block importance models for web pages," in *Proceedings of the 13th international conference on World Wide Web*, 2004, pp. 203-211.

[19] J. F. Brunelle, M. Kelly, H. SalahEldeen, M. C. Weigle, and M. L. Nelson, "Not all mementos are created equal: Measuring the impact of missing resources," in *Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on*, 2014, pp. 321-330.