Uncovering the Unarchived Web

Thaer Samar,[1] Hugo C. Huurdeman,[2] Anat Ben-David,[2] Jaap Kamps,[2] and Arjen de Vries[1]

[1] Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
[2] University of Amsterdam, Amsterdam, The Netherlands

{samar|arjen}@cwi.nl, {huurdeman|a.ben-david|kamps}@uva.nl

ABSTRACT

Many national and international heritage institutes realize the importance of archiving the web for future culture heritage. Web archiving is currently performed either by harvesting a national domain, or by crawling a pre-defined list of websites selected by the archiving institution. In either method, crawling results in more information being harvested than just the websites intended for preservation; which could be used to reconstruct impressions of pages that existed on the live web of the crawl date, but would have been lost forever. We present a method to create representations of what we will refer to as a web collection's *aura*: the web documents that were not included in the archived collection, but are known to have existed — due to their mentions on pages that were included in the archived web collection. To create representations of these unarchived pages, we exploit the information about the unarchived URLs that can be derived from the crawls by combining crawl date distribution, anchor text and link structure. We illustrate empirically that the size of the aura can be substantial: in 2012, the Dutch Web archive contained 12.3M unique pages, while we uncover references to 11.9M additional (unarchived) pages.

Keywords

Web Archives; Web Archiving; Web Crawlers; Anchor Text; Web Graph; Information Retrieval

1. INTRODUCTION

Since 1996, web archiving has been performed by international and national heritage institutions such as the Internet Archive and national libraries, in order to preserve digital cultural heritage for future generations. While archiving the entire web remains an impossible task in terms of size and with regards to its ephemerality, parts of the web are being preserved by using crawlers that capture and archive web pages at the time of harvesting. Web archiving crawlers differ in their scope and settings; breadth-first crawls are designed to discover and capture as many pages as possible, while deep crawls are intended to ensure the complete preservation of specific websites. In either way, the use of crawlers for web archiving entails that every web archive is both incomplete and too complete [3, 12]. On the one hand, every web archive is incomplete, since, depending on the settings of the crawler, many pages it encounters are excluded from archiving. On the other hand, every web archive is bigger than its parts, as apart from the intentionally preserved websites, web archives contain additional data — data such as a page's source, its outlinks, the anchor text of these links, and time stamps of the crawl or archive dates. This additional data can be used as handles to establish evidence of pages that existed at the time of the crawl, but were not archived; thereby uncovering parts of the web that were not preserved otherwise, and would have been lost forever otherwise.

In this pilot study, we describe our method to "uncover" the *Unarchived Web* by extracting and aggregating additional information derived from the archived data. We empirically investigate properties of representations of the web of the past that can be recovered, analyzing unarchived pages from the Dutch Web Archive's crawls performed in 2012.

The National Library of the Netherlands (KB), archives a pre-selected (*seed*) list of 5,000 websites [14]. Websites for preservation are selected by the library per categories related to Dutch historical, social and cultural heritage. Each selected website in the seed list has been assigned a UNESCO code corresponding to the category to which it belongs. The archiving crawler is set to completely archive each website on the selected seed list. However, as previously mentioned, the scope of the pages encountered in the archive stretches well beyond the seed list. We distinguish between pages that were intentionally crawled and archived (these are part of the seed list), pages that were unintentionally archived as a result of the crawler's configuration, and pages that are linked from archived pages that were neither crawled nor archived, but are mentioned in the archive. In this study we focus on the latter group of unarchived pages, where we attempt to uncover evidence of their existence and recover representations of their (most likely) content.

2. BACKGROUND

While web archiving theorists acknowledge that the archived web is necessarily incomplete when compared to the live web [3, 12], the motivation to uncover hidden information from the archived web to infer about the web of the past is part of ongoing research. Rauber et al. [15] have analyzed technological features of archived web data, such as operating systems, web servers, file types, scripting languages and link structure

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

ACM 978-1-4503-2257-7/14/07 ...$15.00.
http://dx.doi.org/10.1145/2600428.2609544.

within domains, to infer characteristics of technology maturation over time . Other studies used archived web data to determine the rate of link persistence, or web change over time [8, 9, 16, 17, 7]. The Memento project has used large samples of archived web data to characterize and measure the completeness of the archived web [1, 2, 17]. Memento is an HTTP-based framework that facilitates locating past versions of a given resource, which, through an aggregator of resources from multiple web archives, indicates the location of archived versions of a given resource across multiple web archives [18]. Using the Memento aggregator, the coverage of web archives can be determined following the procedure of [2], albeit only at the level of domains.

In this study we propose an analysis at a page level; and do not stop at just uncovering the missing (unarchived) pages, but also propose to recover a representation of these by using anchor text representations.

Anchor text has previously been used to enrich the representations of web page content, primarily aiming to improve web retrieval. Craswell et al.[4] first experimented with site finding using anchor texts, considering anchor texts as pseudo documents. They considered the anchor text as surrogate documents and indexed these (for ranking by Okapi BM25) — instead of indexing the content of the target pages. They concluded that anchor texts can be more useful than content words for navigational queries. Kraft and Zien showed that anchor texts can produce higher quality query refinement suggestions than content text [11]. Fujii proposed a model for classifying queries into navigational and informational, and use different retrieval methods depending on the query type. The experimental results showed that content of web pages is useful for informational query types, whereas anchor text information and links are useful for navigational query types [6]. Koolen and Kamps concluded that anchor text has added value for ad hoc informational search, and can lead to significant improvements in retrieval effectiveness. They also evaluate some of the factors impacting the success of anchor text, including link density and collection size [10].

In the previous research discussed so far, the anchor text of a page has been considered as a resource that is complementary to the page content, but treated as two independent representations. Dou et al. took the relationships between source and anchor texts into account, and distinguished between links from the same website and links from related sites [5]. Similar in spirit, Meztler et al. set out to overcome the problem of anchor text sparsity by smoothing the influence of anchor text originating from with-in the same domain with what they termed the 'external' anchor text: the aggregated anchor text from all pages that link to a page in the same domain as the page to be enriched [13].

3. METHOD

3.1 Data

The Dutch web archive, created and maintained by the KB, consists of 76,828 compressed ARC files, that were archived in the period from 2009 until 2012, accumulating to about 7 TB of raw data. Each ARC file contains multiple ARC records (the actually archived web content). In total, the collection consists of approximately 148M archived web documents. The analysis presented in this paper uses only the 2012 part of the archive, a collection consisting of approximately 39M archived web documents (see Table 1). The KB also provided us with the seed list of URLs, each annotated with their (hand-assigned) UNESCO codes.

Year	Number of docs
2009	17,014,067
2010	38,157,308
2011	53,604,464
2012	38,865,673
	147,641,512

Table 1: Number of documents per year.

3.2 Link Extraction

In order to uncover the unarchived URLs, we distinguish between three types of URLs found in the archive; 1) URLs that were intentionally crawled (and their content archived) because they are in seed list, 2) URLs that are unintentionally archived, as a result of the crawler's configuration (but not included on the seed list), and 3) the *unarchived URLs*, that are only mentioned in the archived pages but neither crawled nor archived.

The goal of this paper is to investigate the third group, the unarchived URLs. To accomplish this task, we implemented the following pipeline. We first run a MapReduce job that processes all the ARC records (i.e., archived web pages), using JSoup[1] to extract the links from the HTML web pages, and returning for each link its source and target URL, the anchor text, and the UNESCO code. Comparison against the seedlist and UNESCO codes allows us to distinguish between pages that have been archived intentionally or unintentionally.

In a second MapReduce job, we create a temporary index file that lists all the links in the archive with their crawl date. Using a PigLatin script, we finally join these two intermediate result files to create a list of target URLs together with their presence in the archive:

(src, srcUnesco, srcInSeed, target, targetUnesco, targetInSeed, anchorText, archiveData, targetInArchive).

3.3 Link Aggregation

We now proceed to aggregate links by target. Different seeds are harvested at different frequencies; while most sites are harvested only once a year, some sites are crawled more frequently; the most popular online news-site is even harvested daily. Therefore, we first deduplicate the links based on their values for source, target, anchor text, year and a hash of the source's content.

For the ˜12M extracted links that we obtained so far, we decided to distinguish between internal and external links: an internal link has the same domain-name for both source and target (intra-domain), while in an external link the domain-name of the source URL is different from that of the target URL (an inter-domain link). The corpus we analyze in the remainder of the paper consists of all the inter-domain links to pages outside the archive's seed lists. This results in a corpus consisting of 3,205,354 unique target pages without a representation in the current Dutch web archive. We analyze this corpus in section 4.

[1]http://jsoup.org

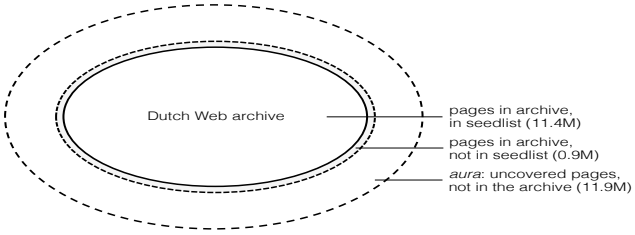

Figure 1: 'Layers' of contents of the Dutch Web Archive (2012)

3.4 Representation Aggregation

Before we discuss our analysis, we first detail the process to create the anchor text representations of the target pages thus obtained. We simply union all the anchor text representations corresponding to links that point to the target page, that existed in the year considered (i.e., 2012). For each target, we then determine the number of unique sources linking to the target page, the number of unique anchor texts used, the number of unique words these anchor texts consisted of, as well as the UNESCO codes for each incoming link.

4. RESULTS

This section provides an estimate of the volume and utility of the extracted representations of the unarchived pages, based on the representations' amount of usable features, including the URL, anchor text, date and UNESCO code. Figure 1 shows the distribution between pages in the archive and seedlist, pages that are archived and not in the seedlist, and pages which are neither in the seedlist nor the archive (the 'aura'). The 2012 web archive collection contains 12,327,673 unique pages, of which 11,395,072 were included on the seedlist (and 932,601 were only included unintentionally). The 'aura' that we uncovered contains an additional (unarchived!) 11,897,662 page representations. In other words, the uncovered web that is only indirectly collected while crawling consists of almost as many pages as the intentionally harvested collection! We now zoom in on the inter-domain links, aggregating link information for 3,205,354 unique pages that are not part of what is considered the web archive. The remainder of this section details a variety of features that can be used to represent the (missing) content of these unarchived pages in a variety of ways.

4.1 Target URLs and timestamps

For all of the 3.2M impressions of unarchived pages mentioned in our corpus, we have immediate access to a basic representation consisting of their URL and estimated timestamp. From the URL, we can derive their domain-names, their level, and their top-level domain (or TLD). A simple slashcount (based on absolute URLs) indicates that 50,41% of these site representations are on the top or first level of their domain (e.g., http://www.example.com or http://www.example.com/forum). This finding is in line with previous research (e.g. [4], [19]) that demonstrated that entry pages of websites often have a higher number of inlinks than other pages of a site.

Table 2 shows the distribution of TLDs in the uncovered web. In earlier work, we generated TLD statistics from the open web crawl CommonCrawl (2012) [2], that consists of over 1.2B webpages, where .com, .de, .org, .net., .uk and .nl

[2]https://github.com/norvigaward/naward15/wiki

TLD	Count	Percentage
com	1,468,946	45.82
nl	1,140,626	35.58
org	165,907	5.17
net	81,866	2.55
jp	71,535	2.23
uk	29,795	0.92
eu	25,996	0.81
be	25,550	0.80
edu	20,449	0.64
de	18,264	0.57
other		5.45

Table 2: Distribution of documents per TLD (2012)

are also appearing in the top 10. Although the regionally focused selection strategy of the Dutch web archive can be detected in the frequency distribution obtained, the similarity between the two distributions does suggest that the structure of the selective national web archive is rather comparable to the structure of a broad web crawl from the same year; an indication of robustness of the data collected.

We also looked at the distribution of domain-names. After basic data cleaning, we observe a total of 348,470 domain-names in the uncovered archive. Of Alexa's top 10 ranking sites in the Netherlands in 2011[3], only nu.nl (a Dutch news aggregator) and Wikipedia are included on the archive's seedlist — most likely as a result of the opt-out crawling process run by the KB. Looking into the uncovered archive however, 6 out of its top 10 domain-names (ordered by the number of occurrences) are also listed as the Alexa top ranking pages in the Netherlands in 2011. We conclude that the uncovered archive may be a valuable resource. Notice however that some of the most influential websites in the Netherlands (according to Alexa) are much less prominent in terms of their occurrences in the Dutch web archive's outlinks; consider for example marktplaats.nl (a large Dutch auction site), live.com and telegraaf.nl (a large Dutch newspaper).

4.2 Target anchor text representation

We now look into the anchor text as an alternative representation, that can be expected to carry (partial) information about the content of the missing pages. The distribution of unique words describing each URL is highly skewed. Out of the 3,205,354 URLs, 1,626,922 (50,75%) were represented by 2 or more unique words, 543,052 URLs (16,94%) by 5 or more unique words, and only 4,08% by 10 or more unique words (see Figure 2).
In other words, only a small fraction of the uncovered URLs can also be recovered by considering their anchor text representation. In future work, we will investigate if taking the text surrounding the anchors into account can help to alleviate this issue.

4.3 Target UNESCO code representation

Pages intentionally harvested for the Dutch web archive are, upon their inclusion in the seed list, annotated using the UNESCO classification. This classification consists of 31 main categories, and the KB assigns 1-3 top level classification codes [14]. For 9.15% of the uncovered pages, the link

[3]https://web.archive.org/web/20110923151640/http://www.alexa.com/topsites/countries/NL

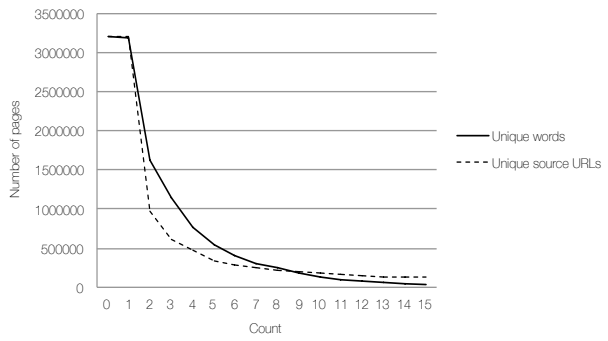

Figure 2: Count of unique source URLs (*dotted line*) and count of unique anchor text words (*solid line*) .

targets refer to a domain-name that has been hand-labelled by the KB.[4] The "31-History and biography", "06-Law and government administration", and "08-Education" are the most prominent categories among these classifications.

4.4 Source URL, timestamp and UNESCO code

Another way of characterizing the pages that are outside of the archive, is by looking at their sources. The source's URL structure (e.g., depth, TLD and words contained in the URL), timestamps and assigned UNESCO-code provide additional clues about the target, possibly enriching page representations. A first foray into this aspect of the representations provided us with insights about the categories of the sources. We checked which UNESCO codes are assigned to the sources of the representations. For 78,34% of all target pages that we uncovered, we can derive at least one category coming from the sources of the links to that page. The top UNESCO categories for these sources are "06-Law and government administration", "23-Art and Architecture", and "31-History and biography". Additional analysis is however necessary to determine whether these categories provide meaningful categorizations of the unarchived target sites.

5. CONCLUSIONS

We have presented a pilot study aimed to uncover the unarchived Web. Our analysis of uncovered URLs (a set we refer to as a web archive's 'aura') extracted from the 2012 Dutch web archive indicates the wealth of data that could help to not just uncover, but also *recover* representations of the unarchived web. The fact that the domain type distribution of uncovered URLs from the selection-based Dutch web archive resembles that of a broad domain web crawl reassures us that we may indeed make inferences about the unarchived web, by using its impressions in the archived web data. While none of the extracted features suffices to generate a rich representation of the unarchived web, a combination of representations may contribute to enriched recovery. Combinations of representations may consist of anchor text, but also the derived structure of source and target sites, assigned categories, or other extractable features. In future research, we hope to determine thresholds on features like the number of unique words or the number of inlinks to a page in relation to the quality of the inferred page representation.

[4]Some of these sites may have opted-out in spite of being originally selected for harvesting, while others have been referred to in the unintentionally crawled pages.

6. ACKNOWLEDGMENTS

This research was supported by the Netherlands Organization for Scientific Research (NWO project # 640.005.001) WebART.

7. REFERENCES

[1] S. Ainsworth, A. Alsum, H. SalahEldeen, M. C. Weigle, and M. L. Nelson. How much of the web is archived? *CoRR*, abs/1212.6177, 2012.

[2] A. Alsum, M. C. Weigle, M. L. Nelson, and H. V. de Sompel. Profiling web archive coverage for top-level domain and content language. In T. Aalberg, C. Papatheodorou, M. Dobreva, G. Tsakonas, and C. J. Farrugia, editors, *TPDL*, volume 8092 of *Lecture Notes in Computer Science*, pages 60–71. Springer, 2013.

[3] N. Brügger. Historical network analysis of the web. *Social Science Computer Review*, 31(3):306–321, 2013.

[4] N. Craswell, D. Hawking, and S. Robertson. Effective site finding using link anchor information. In *SIGIR*, pages 250–257. ACM Press, 2001.

[5] Z. Dou, R. Song, J.-Y. Nie, and J.-R. Wen. Using anchor texts with their hyperlink structure for web search, 2009.

[6] A. Fujii. Modeling anchor text and classifying queries to enhance web document retrieval. In J. Huai, R. Chen, H.-W. Hon, Y. Liu, W.-Y. Ma, A. Tomkins, and X. Zhang, editors, *WWW*, pages 337–346. ACM, 2008.

[7] K. Gyllstrom, C. Eickhoff, A. P. de Vries, and M.-F. Moens. The downside of markup: examining the harmful effects of css and javascript on indexing today's web. In *CIKM*, pages 1990–1994, 2012.

[8] M. Klein and M. L. Nelson. Investigating the change of web pages' titles over time. *CoRR*, abs/0907.3445, 2009.

[9] W. Koehler. Web page change and persistence - a four-year longitudinal study. *JASIST*, 53(2):162–171, 2002.

[10] M. Koolen and J. Kamps. The importance of anchor text for ad hoc search revisited. In F. Crestani, S. Marchand-Maillet, H.-H. Chen, E. N. Efthimiadis, and J. Savoy, editors, *SIGIR*, pages 122–129. ACM, 2010.

[11] R. Kraft and J. Zien. Mining anchor text for query refinement. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 666–674, New York, NY, USA, 2004. ACM.

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

[13] D. Metzler, J. Novak, H. Cui, and S. Reddy. Building enriched document representations using aggregated anchor text. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 219–226, New York, NY, USA, 2009. ACM.

[14] M. Ras. Eerste fase webarchivering, Sept. 2007.

[15] A. Rauber, R. M. Bruckner, A. Aschenbrenner, O. Witvoet, and M. Kaiser. Uncovering information hidden in web archives: A glimpse at web analysis building on data warehouses. *D-Lib Magazine*, 8(12), 2002.

[16] H. SalahEldeen and M. L. Nelson. Losing my revolution: How many resources shared on social media have been lost? In P. Zaphiris, G. Buchanan, E. Rasmussen, and F. Loizides, editors, *TPDL*, volume 7489 of *Lecture Notes in Computer Science*, pages 125–137. Springer, 2012.

[17] R. Sanderson, M. Phillips, and H. V. de Sompel. Analyzing the persistence of referenced web resources with memento. *CoRR*, abs/1105.3459, 2011.

[18] H. Van de Sompel, M. Nelson, and R. Sanderson. Http framework for time-based access to resource states. Technical report, Technical report, Internet Engineering Task Force (IETF), 2011.

[19] T. Westerveld, W. Kraaij, and D. Hiemstra. Retrieving web pages using content, links, urls and anchors. In *TREC*, pages 663–672, 2001.