Lost but Not Forgotten: Finding Pages on the Unarchived Web

Huurdeman, H.C.; Kamps, J.; Samar, T.; de Vries, A.P.; Ben David, A.; Rogers, R.A.

Published in:
International Journal on Digital Libraries

DOI:
10.1007/s00799-015-0153-3

Int J Digit Libr (2015) 16:247–265
DOI 10.1007/s00799-015-0153-3

Lost but not forgotten: finding pages on the unarchived web

Hugo C. Huurdeman[1] · Jaap Kamps[1] · Thaer Samar[2] · Arjen P. de Vries[2] · Anat Ben-David[3] · Richard A. Rogers[1]

Received: 2 December 2014 / Revised: 8 May 2015 / Accepted: 10 May 2015 / Published online: 3 June 2015

Abstract Web archives attempt to preserve the fast changing web, yet they will always be incomplete. Due to restrictions in crawling depth, crawling frequency, and restrictive selection policies, large parts of the Web are unarchived and, therefore, lost to posterity. In this paper, we propose an approach to uncover unarchived web pages and websites and to reconstruct different types of descriptions for these pages and sites, based on links and anchor text in the set of crawled pages. We experiment with this approach on the Dutch Web Archive and evaluate the usefulness of page and host-level representations of unarchived content. Our main findings are the following: First, the crawled web contains evidence of a remarkable number of unarchived pages and websites, potentially dramatically increasing the coverage of a Web archive. Second, the link and anchor text have a highly skewed distribution: popular pages such as home pages have more links pointing to them and more terms in the anchor text, but the richness tapers off quickly. Aggregating web page evidence to the host-level leads to significantly richer representations, but the distribution remains skewed. Third, the succinct representation is generally rich enough to uniquely identify pages on the unarchived web: in a known-item search setting we can retrieve unarchived web pages within the first ranks on average, with host-level representations leading to further improvement of the retrieval effectiveness for websites.

Keywords Web archives · Web archiving · Web crawlers · Anchor text · Link evidence · Information retrieval

A. Ben-David: work done while at the University of Amsterdam.

✉ Hugo C. Huurdeman
h.c.huurdeman@uva.nl

Jaap Kamps
kamps@uva.nl

Thaer Samar
samar@cwi.nl

Arjen P. de Vries
arjen@cwi.nl

Anat Ben-David
anatbd@openu.ac.il

Richard A. Rogers
r.a.rogers@uva.nl

[1] University of Amsterdam, Amsterdam, The Netherlands

[2] Centrum Wiskunde en Informatica, Amsterdam, The Netherlands

[3] The Open University, Ra'anana, Israel

1 Introduction

The advent of the web has had a revolutionary impact on how we acquire, share, and publish information. Digital born content is rapidly taking over other forms of publishing, and the overwhelming majority of online publications has no parallel in a material format. However, such digital content is as easily deleted as it is published, and the ephemerality of web content introduces unprecedented risks to the world's digital cultural heritage, severely endangering the future understanding of our era [31].

Memory and heritage institutions address this problem by systematically preserving parts of the Web for future generations in Web archives. As outlined by the International Internet Preservation Consortium, Web archiving involves a "process of collecting portions of the World Wide Web, preserving the collections in an archival format, and then serving the archives for access and use" [15]. Pioneered by the Internet Archive and later joined by national libraries including the

British Library and National Library of the Netherlands, Web archiving initiatives have collectively archived petabytes of Web data. Despite important attempts to preserve parts of the Web by archiving, a large part of the Web's content is unarchived and hence lost forever. It is impossible to archive the entire Web due to its ever increasing size and rapidly changing content. Moreover, even the parts that have been preserved are incomplete at several levels.

Organizations involved in Web archiving employ different strategies to archive the Web, performed by Web crawlers. The first strategy focuses on automatic harvests of websites in large quantities (usually a national domain) and consists of 'breadth-first crawls'. That is, the crawler is set to discover as many links as possible, rather than capturing all deep pages of a domain. For large domains, such crawls may take several months to complete, and due to the breadth-first policy of the crawl, not all deep pages might be preserved. A second strategy is based on selection policies and utilizes 'deep crawls' [7,12,23]. This strategy focuses on capturing complete websites. However, the strategy typically precludes the harvesting of pages outside the selection lists, even if relevant to a country's cultural heritage. In addition, other factors influence the completeness of web archives. First, legal restrictions prevent mass archiving of websites due to privacy, copyright or legislative frameworks in certain countries. Second, several technical limitations impede harvests of certain types of content. Therefore, entire websites, individual pages (e.g. Flash-based pages), and page elements (e.g. embedded Twitter streams) are missing from archives.

The overall consequence is that our web archives are highly incomplete, and researchers and other users treating the archive to reflect the web as it once was may draw false conclusions due to unarchived content. The following is the main research question of this paper: can we recover parts of the unarchived web? This may seem like a daunting challenge or a mission impossible: how can we go back in time and recover pages that were never preserved? Our approach is to exploit the hyperlinked structure of the web and collect evidence of uncrawled pages from the pages that were crawled and are part of the archive.

We show empirically that it is possible to first *uncover* the existence of a substantial amount of unarchived pages from evidence in the web archive, and second to *recover* significant parts of the unarchived web by reconstructing representations of these pages from the links and anchor text in the crawled pages. We refer to the recovered web documents as the web archive's *aura*: the web documents which were not included in the archived collection, but are known to have existed—references to these unarchived web documents appear in the archived pages.

The paper investigates the following research questions:

RQ1 Can we uncover a significant fraction of unarchived web pages and web sites based on references to them in the Web archive?

We exploit the link structure of the crawled content to derive evidence of the existence of unarchived pages. In addition, we quantify and classify pages, domains, and hostnames in the unarchived aura of the archive.

RQ2 How rich are the representations that can be created for unarchived web pages?

We build implicit representations of unarchived web pages and domains, based on link evidence, URL words, and anchor text and investigate the richness (or sparseness) of the descriptions in the number of incoming links and the aggregated anchor text. We further break this down over unarchived homepages and other pages.

RQ3 Can we create richer representations of web sites based on aggregating page-level evidence from pages sharing the same hostname?

Here, we look at the additional value of aggregating anchor text and URL words at the host level, in effect creating site-level representations. We investigate the quantity and richness of these additional representations and compare this with page-level representations of homepages.

RQ4 Are the resulting derived representations of unarchived pages and web sites useful in practice? Do they capture enough of the unique page content to make them retrievable amongst millions of other page representations?

As a critical test, we study the effectiveness of the derived representations of unarchived home pages and deep pages in a known-item search setting. Only if the derived page and site-based representations sufficiently characterize the unique page's content, do we have a chance to retrieve the page within the top search results.

The remainder of the paper is organized as follows: we first introduce related work (Sect. 2), followed by a description of the experimental setup (Sect. 3). Next, we look at the results of our analysis, characterizing the actual contents of the Dutch web archive and the *aura* of unarchived pages around the archive (Sect. 4). We study the potential richness of generated representations of unarchived web pages (Sect. 5). Next, we take a look at the comparative richness of aggregated host-level representations of unarchived websites (Sect. 6). The generated representations are evaluated by their utility to retrieve the page or host in a known-item search sce-

nario (Sect. 7). We conclude by discussing the implications of our findings and we outline future work (Sect. 8).

2 Background

In this section, we discuss related work, which falls in two broad areas. First, we discuss related research in web archiving and web preservation. Second, we discuss previous literature on search, based on link evidence and anchor text.

2.1 Web archives and web preservation

Experts in the web archiving community discuss the shortcomings of web archiving crawlers in terms of the content they fail to capture [23]. Some websites are intentionally excluded, breadth-first crawls might not capture deeper pages of a website, and selective crawlers exclude sites beyond the scope of the selection policy. However, as argued by Day [7], in most cases, even the websites that meet selection guidelines on other criteria may not be captured in their entirety, as pages might be missing due to crawling errors. Even the captured pages may contain errors: pages might be deprived of certain elements. For example, crawlers often fail to capture embedded elements, such as JavaScript and Flash [7,12,23]. Some authors have looked at the impact of missing resources on the display of web pages. Brunelle et al. [4] have proposed a damage rating measure to evaluate archive success. Using this measure, they showed that the Internet Archive is missing an increasing number of important embedded resources over the years.

Hence, the limits of web archives' crawlers may result in partial and incomplete web archives. This incompleteness can impede the value of web archives, for example in the context of scholarly research [3,13]. However, crawlers do register additional information about web content they encounter. This additional information includes server-side metadata of harvested pages (such as timestamps and HTML response codes), and information embedded in pages (for instance their hyperlinks and associated anchor text). Rauber et al. [29] have recognized the wealth of additional information contained in web archives which can be used for analytical purposes. Gomes and Silva [10] used data obtained from the domain crawl of the Portugese web archive to develop criteria for characterizing the Portugese web.

The Memento project has expanded the scope of analysis of archived web data beyond the boundaries of a single archive to profile and analyze coverage of archived websites across different web archives. Memento [32] is an HTTP-based framework which makes it possible to locate past versions of a given web resource through an aggregator of resources from multiple web archives. In a recent study, AlSum et al. [1] queried the Memento aggregator to profile and evaluate the coverage of twelve public web archives. They found that the number of queries can be reduced by 75 % by only sending queries to the top three web archives. Here, coverage (i.e. whether a resource is archived and in which archive its past versions are located) was calculated based on the HTTP header of host-level URLs.

Instead of just measuring what is missing, we will try to uncover significant parts of the unarchived web, by reconstructing representations of the unarchived web pages and websites using URL and anchor text evidence.

2.2 Link evidence and anchor text

One of the defining properties of the Internet is its hyperlink-based structure. The web's graph structure is well studied and also methods to use this structure have been widely applied, especially in the context of web retrieval (for example PageRank [27]). The links which weave the structure of the web consist of destination URLs and are described by anchor text. Aggregating anchor text of links makes it for example possible to create representations of target pages. Techniques based on the graph structure of the web and anchor text have been widely used in web retrieval. In this paper, we mainly focus on the use of anchor text.

Craswell et al. [5] explored the effectiveness of anchor text in the context of site finding. Aggregated anchor texts for a link target were used as surrogate documents, instead of the actual content of the target pages. Their experimental results show that anchor texts can be more effective than content words for navigational queries (i.e. site finding). Work in this area led to advanced retrieval models that combine various representations of page content, anchor text, and link evidence [16]. Fujii [9] presented a method for classifying queries into navigational and informational. Their retrieval system used content-based or anchor-based retrieval methods, depending on the query type. Based on their experimental results, they concluded that content of web pages is useful for informational query types, while anchor text information and links are useful for navigational query types. Contrary to previous work, Koolen and Kamps [19] concluded that anchor text can also be beneficial for ad hoc informational search, and their findings show that anchor text can lead to significant improvements in retrieval effectiveness. They also analyze the factors influencing this effectiveness, such as link density and collection size.

In the context of web archiving, link evidence and anchor text could be used to locate missing web pages, of which the original URL is not accessible anymore. Martinez-Romo and Araujo [22] studied the problem of finding replacements of broken links (missing pages) in a web page. They constructed queries based on terms selected from anchor text pointing to the broken link and expanded these queries by

adding information from the web page containing the broken link. Then, they submitted the constructed queries to a standard search engine to retrieve candidate replacements for the missing page. Klein and Nelson [17] computed lexical signatures of lost web pages, using the top *n* words of link anchors, and used these and other methods to retrieve alternative URLs for lost web pages. The versatility and potential use of anchor text is further exemplified by Kraft and Zien [21], who show that anchor text can also be used to generate query suggestions and refinements.

Following Kleinberg [18], Dou et al. [8] took the relationships between source pages of anchor texts into account. Their proposed models distinguish between links from the same website and links from related sites, to better estimate the importance of anchor text. Similarly, Metzler et al. [24] smoothed the influence of anchor text which originates from within the same domain, using the 'external' anchor text: the aggregated anchor text from all pages that link to a page in the same domain as the target page. Their proposed approach also facilitates overcoming anchor text sparsity for pages with few inlinks. Along the same line, Broder et al. [2] used site-level information for improving web search results. They achieved this by creating two indices: a URL index based on the page content, and a site index representing the entire website. They introduced two approaches for creating site-based representations. The first site-level representation was created by concatenating the content text of all pages from a site, or from a sample of pages. The second site-representation was created by aggregating all anchor text of external links pointing to pages in the site. Their experimental evaluation showed that the combination of page and site-level indices is more effective for web retrieval than the common approach of only using a page-level index. Their results also indicate that site-level anchor text representations perform better than site-level representations based on concatenated content text.

Another aspect of anchor text is its development over time: often single snapshots of sites are used to extract links and anchor text, neglecting historical trends. Dai and Davison [6] determined anchor text importance by differentiating pages' inlink context and creation rates over time. They concluded that ranking performance is improved by differentiating pages with different in-link creation rates, but they also point to the lack of available archived resources (few encountered links were actually available in the Internet Archive).

Our approach is inspired by the previous results on various web centric document representations based on URL and incoming anchor text, typically used in addition to representations of the page's content [5,16,20,25]. We focus on the use case of the web archive, which is different from the live web given that we cannot go back and crawl the unarchived page and hence have to rely on these implicit representations exclusively. It is an open question whether the resulting derived representations—based on scant evidence of the pages—is a rich enough characterization to be of practical use. The current paper builds on earlier work on uncovering unarchived pages [30] and the recovery and evaluation of unarchived page descriptions [14]. Parts of the page-level results were presented in [14], now extended with the analysis of site-level representations to overcome anchor text sparsity at the page level [2].

3 Experimental setup

This section describes our experimental setup: the dataset, the Hadoop-based link extraction methods, and the way the links were aggregated for analysis.

3.1 Data

This study uses data from the Dutch web archive at the National Library of the Netherlands (KB). The KB currently archives a pre-selected (seed) list of more than 5000 websites [28]. Websites for preservation are selected by the library based on categories related to Dutch historical, social, and cultural heritage. Each website on the seed list has manually been categorized by the curators of the KB using a UNESCO classification code.

Our snapshot of the Dutch web archive consists of 76,828 ARC files, which contain aggregated web objects. A total number of 148 M web objects has been harvested between February 2009 and December 2012, resulting in more than 7 Terabytes of data (see Table 1). In our study, we exclusively focus on the content crawled in 2012 (35.7 % of the total data harvested between 2009 and 2012). Additional metadata is available in separate documents, including the KB's selection list, dates of selection, assigned UNESCO codes, and curators' annotations. Table 2 shows a summary of the types of content crawled in 2012, with their counts and percentages. It shows that the majority of content types crawled in 2012 are HTML-based textual content (65.3 %) and images in various formats (22.9 %).

In our extraction, we differentiate between four different types of URLs found in the Dutch web archive:

1. URLs that have been archived intentionally as they are included in the seedlist,

Table 1 Crawled web objects per year in the Dutch web archive

Year	Number of web objects
2009–2011	108,775,839
2012	38,865,673
	147,641,512

Table 2 Types of web objects crawled in 2012 (MIME-types)

MIME-type	Count	%
Text/html	25,380,955	65.3
Image/jpeg	6,518,954	16.8
Image/gif	1,222,480	3.1
Image/png	1,171,585	3.0
Application/pdf	816,746	2.1
Text/plain	642,282	1.7
Text/xml	488,569	1.3
Rss/xml	483,858	1.2
Other	2,140,244	5.5

2. URLs that have been unintentionally archived due to the crawler's configuration,
3. unarchived URLs, discovered via the link structure of the archive, of which the parent domain is included in the seedlist (which we will refer to as the *inner aura*)
4. unarchived URLs, discovered via the link structure of the archive, of which the parent domain is not on the seedlist (which we will refer to as the *outer aura*).

Section 4 describes these four types of archived and unarchived URLs in more detail.

3.2 Link extraction

We created our dataset by implementing a specific processing pipeline. This pipeline uses Hadoop MapReduce and Apache Pig Scripts for data extraction and processing.

The first MapReduce job traversed all archived web objects contained in the archive's ARC files. Web pages with the following properties were processed:

– crawled in 2012
– having the MIME-type *text/html*, and
– having at least one anchor link including a destination URL and (non-empty) anchor text.

From all pages with these properties, each anchor link found in the page was extracted using JSoup. As we focus on links to textual content, only '`a`' anchors were extracted and other references to embedded content, such as '`script`' links, embedded images (via the '`img`' tag), and embedded '`iframe`' content were ignored. We keep the *source URL* (which is the page URL), *target URL* (the URL of the page that the link is pointing to), and the *anchor text* of the link (the textual description of a link). Links without anchor text were discarded. This extracted link information is combined with basic ARC metadata about the source page (e.g. crawldate). In addition, other relevant information is added, such as the hashcode (MD5) of the source page, the occurrence of the source and target page on the KB's seedlist, and assigned UNESCO classification codes.

A second MapReduce job built an index of all URLs in the Dutch web archive, with their associated crawldate. Using this index, we performed lookups to validate whether or not a target URL found in the link information exists in the archive in the same year. Our final output format for extracted links contains the following properties:

(*sourceURL, sourceUnesco, sourceInSeedProperty, targetURL, targetUnesco, targetInSeedProperty, anchorText, crawlDate, targetInArchiveProperty, sourceHash*).

In our study, we look at the content per year. While some sites are harvested yearly, other sites are captured biannually, quarterly or even daily. This could result in a large number of links from duplicate pages. To prevent this from influencing our dataset, we deduplicated the links based on their values for year, anchor text, source, target, and (MD5) hashcode. The hashcode is a unique value representing a page's content and is used to detect if a source has changed between crawls. We keep only links to the same target URL with identical anchor texts if they originate from unique source URLs.

In our dataset, we include both inter-server links, which are links between different servers (external links), and intra-server links, which occur within a server (site internal links). We also performed basic data cleaning and processing: removing non-alphanumerical characters from the anchor text, converting the source and target URLs to the canonicalized sort-friendly format known as SURT, removing double and trailing slashes, and removing *http*(*s*) prefixes (see http://crawler.archive.org/apidocs/org/archive/util/SURT.html).

3.3 Link aggregation

We combine all incoming links and anchor text to an aggregated page-level representation. In this process, we create a representation that includes the target URL, and grouped data elements with source URLs, anchor texts, and other associated properties. Using another Apache Pig script, we performed further processing. This processing included tokenization of elements such as anchor text and URL words. This, combined with other processing, allowed us to include counts of different elements in our output files, for example, the unique source sites and hosts, unique anchor and URL words, and the number of links from seed and non-seed source URLs. We also split each URL to obtain separate fields for TLD (top-level domain), domain, host, and filetype. To retrieve correct values for the TLD field, we matched the TLD extension from the URL with the official IANA list of all TLDs, while we matched extracted filetype extensions of each URL with a list of common web file formats.

We also aggregate all page-level evidence within the same hostname to an aggregated site-level representation. A site-

level representation aggregates evidence for all pages of a given website, or a sample thereof (see for instance [2]). In this paper, we consider all pages sharing the same parent *host* as a 'site'. We adapted our Pig scripts to aggregate link evidence for all pages under each host, including anchor text and incoming URLs. We carried out similar processing steps as for the page-level representations. For later analysis, we saved up to 100 URLs of pages under each host.

The final aggregated page and site-level representations containing target URLs, source properties, and various value counts were subsequently inserted into MySQL databases (13 and 0.5 M rows), to provide easier access for analysis and visualization via a web application.

4 Expanding the web archive

In this section, we study RQ1: Can we uncover a significant fraction of unarchived web pages and web sites based on references to them in the Web archive? We investigate the contents of the Dutch web archive, quantifying and classifying the unarchived material that can be uncovered via the archive. Our finding is that the crawled web contains evidence of a remarkable number of unarchived pages and websites, potentially dramatically increasing the coverage of the web archive.

4.1 Archived content

We begin by introducing the actual archived content of the Dutch web archive in 2012, before characterizing the unarchived contents in the next subsection. As introduced in Sect. 3.2, we look at the pages in the archive with a *text/html* MIME-type. Here, we count the unique text-based web pages (based on MD5 hash) in the web archive's crawls from 2012, totaling in 11,041,113 pages. Of these pages, 10,158,586 were crawled in 2012 as part of the KB's seedlist (92 %). An additional 882,527 pages are not in the seedlist but included in the archive (see Table 3). As discussed in Sect. 2.1, each 'deep' crawl of a website included in the seedlist also results in additional ('out of scope') material being harvested, due to crawler settings. For example, to correctly include all embedded elements of a certain page, the crawler might need to harvest pages beyond the predefined seed domains. These unintentionally archived contents amount to 8 % of the full web archive in 2012. Close dependencies exist between the chosen crawl settings and the resulting harvested material, the details of which are beyond the scope of this paper. To avoid influence of artificial effects caused by variations in these crawl settings, we chose only one year for our evaluation instead of multiple years, giving us a relatively stable setting to work with.

Table 3 Unique archived pages (2012)

	On seedlist (%)	Not on seedlist (%)	Total
Pages	10,158,586 (92.0)	882,527 (8.0)	11,041,113

Table 4 Unique archived hosts, domains and TLDs

	On seedlist (%)	Not on seedlist (%)	Total
Hosts	6157 (14.2)	37,166 (85.8)	43,323
Domains	3413 (10.1)	30,367 (89.9)	33,780
TLDs[a]	16 (8.8)	181 (100)	181

[a] Since the values for the TLDs overlap for both categories, percentages add up to more than 100 % (same for Table 8)

Table 5 Coverage in archive

Mean page count	On seedlist	Not on seedlist
Per host	1650	24
Per domain	2976	29
Per TLD	634,912	4876

We can take a closer look at the contents of the archive by calculating the diversity of hosts, domains, and TLDs contained in it. Table 4 summarizes these numbers, in which the selection-based policy of the Dutch KB is reflected. The number of hosts and domains is indicative of the 3876 selected websites on the seedlist in the beginning of 2012: there are 6157 unique hosts (e.g. papierenman.blogspot.com) and 3413 unique domains (e.g. okkn.nl).

The unintentionally archived items reflect a much larger variety of hosts and domains than the items from the seedlist, accounting for 37,166 unique hosts (85.8 %), and 30,367 unique domains (89.9 % of all domains). The higher diversity of the non-seedlist items also results in a lower coverage in terms of number of archived pages per domain and per host (see Table 5). The mean number of pages per domain is 2976 for the domains included in the seedlist, while the average number of pages for the domains outside of the seedlist is only 29.

According to the KB's selection policies, websites that have value for Dutch cultural heritage are included in the archive. A more precise indication of the categories of websites on the seedlist can be obtained by looking at their assigned UNESCO classification codes. In the archive, the main categories are Art and Architecture (1.3 M harvested pages), History and Biography (1.2 M pages) and Law and Government Administration (0.9 M pages) (see Table 6). The pages harvested outside of the selection lists do not have assigned UNESCO codes. A manual inspection of the top 10 domains in this category (35 % of all unintentionally har-

Table 6 Categories of archived pages in KB seedlist (top 10)

Inner aura		Count	%
1	Art and architecture	1,328,114	13.0
2	History and biography	1,174,576	11.5
3	Law and government administration	910,530	8.9
4	Education	803,508	7.9
5	Sport, games and leisure	712,241	7.0
6	Sociology, statistics	694,636	6.8
7	Political science	663,111	6.5
8	Medicine	590,862	5.8
9	Technology and industry	469,132	4.6
10	Religion	377,680	3.7

vested pages) shows that these are heterogeneous: 3 websites are related to Dutch cultural heritage, 2 are international social networks, 2 websites are related to the European Commission, and 3 are various other international sites.

4.2 Unarchived content

To uncover the unarchived material, we used the link evidence and structure derived from the crawled contents of the Dutch web archive in 2012. We refer to these contents as the web archive's *aura*: the pages that are not in the archive, but whose existence can be derived from evidence in the archive.

The unarchived *aura* has a substantial size: there are 11 M unique pages in the archive, but we have evidence of 10.7 M additional link targets that do not exist in the archive's crawls from 2012. In the following sections, we will focus on this aura and differentiate between the *inner aura* (unarchived pages of which the parent domain is on the seedlist) and the *outer aura* (unarchived pages of which the parent domain is not on the seedlist). The inner aura has 5.5 M (51.5 %) unique link targets, while the outer aura has 5.2 M (48.5 %) unique target pages (see Fig. 1; Table 7).

Like the number of pages, also the number of unique unarchived hosts is quite substantial: while *in* the archive there are 43,323 unique hosts, we can reveal a total number of 490,836 hosts in the unarchived aura. There is also a considerable number of unique domains and TLDs in the unarchived contents (see Table 8).

Table 7 Unarchived *aura* unique pages (2012)

	Inner aura (%)	Outer aura (%)	Total
Pages	5,505,975 (51.5)	5,191,515 (48.5)	10,697,490

Table 8 Unarchived unique hosts, domains and TLDs

	Inner aura (%)	Outer aura (%)	Total
Hosts	9039 (1.8)	481,797 (98.2)	490,836
Domains	3019 (0.8)	369,721 (99.2)	372,740
TLDs	17 (6.6)	259 (100)	259

Table 9 Unarchived *aura* coverage (2012)

Mean page count	Inner aura	Outer aura
Per host	609	10
Per domain	1823	14
Per TLD	323,881	20,044

The tables above also show the difference between the *inner* and *outer* aura. The outer aura has a much larger variety of hosts, domains, and TLDs compared to the inner aura (Table 8). On the other hand, the coverage in terms of the mean number of pages per host, domain, and TLD is much greater in the inner aura than the outer aura (see Table 9). This can be explained by the fact that the pages in the inner aura are closely related to the smaller set of domains included in web archive's seedlist, since they have a parent domain which is on the seedlist.

Finally, to get an overview of the nature of the unarchived resources, we have matched the link targets with a list of common web file extensions. Table 10 shows the filetype distribution: the majority consists of URLs without an extension (http), html, asp, and php pages for both the inner and

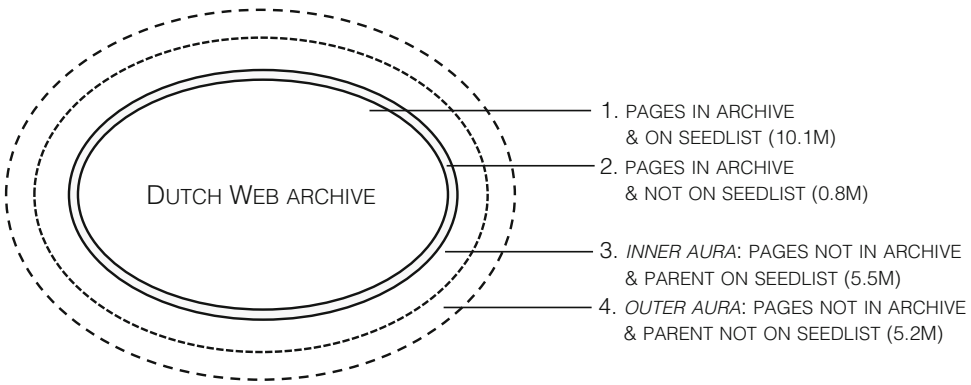

Fig. 1 'Layers' of contents of the Dutch web archive (2012)

Table 10 Unarchived *aura* filetypes

Inner aura	Count	%	Outer aura	Count	%
http	4,281,750	77.77	http	3,721,059	71.68
html	351,940	6.39	php	585,024	11.27
php	321,095	5.83	html	582,043	11.21
asp	38,0964	6.92	asp	181,963	3.51
pdf	70,371	1.28	jpg	30,205	0.58

outer aura. Only a minority of references are other formats, like pdfs and non-textual contents (e.g. jpg files in the outer aura). This suggests that the majority of references to the unarchived aura which we extracted from the web archive points to textual web content.

As we observe a large fraction of URLs without an extension, we performed an additional analysis to shed more light on the included filetypes. We retrieved HTML status codes and MIME-types from the 'live' web, for a random sample of 1000 unarchived URLs in the inner aura, and 1000 URLs in the outer aura. For the inner aura, 596 of all URLs from the sample were available, while the remaining 404 were not accessible (resulting in 30× and 404 status codes). Of the resolvable URLs, 580 (97.3 %) were of the MIME-type *text/html*, while only 16 URLs (2.7 %) led to other filetypes (i.e. images, videos or PDFs). For the outer aura, 439 URLs were still accessible, and 561 of all URLs could not be retrieved. Of the URLs which resolved, 412 (93.8 %) were of the *text/html* type, while only 27 (6.2 %) led to other MIME-types. Hence, there is clear evidence that the uncovered URLs in both the inner and outer aura are predominantly text-based pages.

4.3 Characterizing the "aura"

Here, we characterize unarchived contents of the archive based on the top-level domain distribution and the domain coverage.

From the top-level domains (TLDs) we derive the origins of the unarchived pages surrounding the Dutch web archive. Table 11 shows that the majority of unarchived pages in the inner aura (95.69 %) have Dutch origins. The degree of .nl domains in the outer aura is lower, albeit still considerable, with 31.08 % of all 1.8 M pages. The distribution of TLDs in the outer aura seems to resemble the TLD distribution of the open web. Even though the regional focus of the selection policy of the Dutch web archive is apparent in the distribution of the top 10, the comparison does provide indications that the outer aura is more comparable to the full web. The prominence of the .jp TLD can be explained by the fact that some Japanese social networks are included in the unintentionally harvested pages of the Dutch archive.

Another way to characterize the unarchived contents of the Dutch web is by studying the distribution of the target domain names. This distribution is quite distinct in the two subsets of the aura: while the inner aura contains many specific Dutch domains, as selected by the KB (e.g. noord-hollandsarchief.nl and archievenwo2.nl), the outer aura contains a much more varied selection of sites, which include both popular international and Dutch domains (e.g. facebook.com and hyves.nl), and very specific Dutch sites potentially related to Dutch heritage (e.g badmintoncentraal.nl).

To get more insights into the degree of popular websites in the unarchived aura, we compare the domains occurring in the aura against publicly available statistics of websites' popularity. Alexa, a provider of free web metrics, publishes online lists of the top 500 ranking sites per country, on the basis of traffic information. Via the Internet Archive, we retrieved a contemporary Alexa top 500 list for sites in the Netherlands (specifically, http://web.archive.org/web/20110923151640/alexa.com/topsites/countries/NL). We counted the number of domains in Alexa's top 100 that occur in the inner and outer aura of the Dutch archive (summarized in Table 12). The inner aura covers 7 domains of the Alexa top 100 (including Dutch news aggregator *nu.nl* and *wikipedia.org*), while the outer aura covers as much as 90 of the top 100 Alexa domains, with a considerable number of unique target pages. For these 90 domains, we have in total 1,227,690 URL references, which is 23.65 % of all unarchived URLs in the outer aura of the archive. This means that we have potentially many representations of the most popular websites in the Netherlands, even though they have not been captured in the selection-based archive itself.

In addition to the discussed Alexa rankings, the assigned UNESCO classification codes provide indications of the categories of pages in the archive (Table 13). 98.39 % of the pages in the inner aura can be categorized using UNESCO codes, since their parent domain is on the seedlist. The most frequently occurring classifications match the top categories in the archive: History and Biography (e.g. noord-hollandsarchief.nl) and Art and Architecture (e.g. graphicdesignmuseum.nl), as previously summarized in Table 6. A few categories have different positions though: the Religion (e.g. baptisten.nl) and General (e.g. nu.nl) categories are more frequent in the inner aura than in the archive, and the opposite holds true for Law and Government Administration (e.g. denhaag.nl).

For the domains in the outer aura, virtually no UNESCO categorizations are available (only for 0.04 % of all pages), since they are outside of the scope of the selection policy in which these classifications codes are hand-assigned. Therefore, we generated a tentative estimate of the categories of target pages by counting the UNESCO categories of source

Table 11 TLD distribution

Inner aura		Count	%	Outer aura		Count	%
1	nl	5,268,772	95.7	1	com	1,803,106	34.7
2	com	130,465	2.4	2	nl	1,613,739	31.1
3	org	52,309	1.0	3	jp	941,045	18.1
4	net	44,348	0.8	4	org	243,947	4.7
5	int	8127	0.2	5	net	99,378	1.9
6	Other	1954	0.1	6	eu	80,417	1.6
				7	uk	58,228	1.1
				8	de	44,564	0.9
				9	be	43,609	0.8
				10	edu	29,958	0.6

Table 12 Coverage of most popular Dutch domains (*Alexa position*)

Inner aura	Count (K)	Outer aura	Count (K)
nu.nl (*6*)	74.2	twitter.com (*9*)	266.7
wikipedia.org (*8*)	17.4	facebook.com (*3*)	227.0
blogspot.com (*15*)	3.5	linkedin.com (*7*)	184.9
kvk.nl (*90*)	2.2	hyves.nl (*11*)	125.6
anwb.nl (*83*)	1.7	google.com (*2*)	106.4

Table 13 Top 10 of inner aura categories (*rank in archive*)

Inner aura		Count	%
1	History and biography (*2*)	2,444,188	44.4
2	Art and architecture (*1*)	609,271	11.1
3	Religion (*10*)	567,604	10.3
4	Education (*4*)	235,529	4.3
5	Political science (*7*)	233,095	4.2
6	General (*12*)	190,727	3.5
7	Law and government administration (*3*)	187,719	3.4
8	Sports, games and leisure (*5*)	132,576	2.4
9	Technology and industry (*9*)	114,926	2.1
10	Medicine (*8*)	108,874	2.0

hosts. Consider, for example, the host onsverleden.net ('our history'), of which the pages together receive inlinks from 14 different hosts. Eight of these hosts are categorized as 'Education', six have 'History and biography' as their category, and two are part of 'Literature and literature history'. Therefore, we have an indication that the topic of the target site is likely related to education and history. This is validated by a manual check of the URL in the Internet Archive (as the URL is not available in the Dutch web archive): the site is meant for high school pupils and contains translated

Table 14 Outer aura categories derived from link structure (top 10)

Outer aura		Count	%
1	N/a	1,582,543	29.1
2	Art and architecture	582,122	10.7
3	Education	409,761	7.5
4	General	406,011	7.5
5	History and biography	405,577	7.5
6	Political science	362,703	6.7
7	Law and government administration	360,878	6.6
8	Sociology and statistics	292,744	5.4
9	Medicine	160,553	3.0
10	Commerce	117,580	2.2

historical sources. For each host in the aura, we then chose the most frequently occurring category. This resulted in the count-based ranking in Table 14. 29.1 % of all pages in the outer aura cannot be categorized (due to, e.g. inlinks from archived sites which are not on the seedlist). Of the remaining 70.9 %, the pages have similar categories as the pages in the *inner aura*, but different tendencies, for instance a lower position for the *History and Biography* category (position 5 with 7.5 % of all pages).

Summarizing, in this section we have quantified the size and diversity of the unarchived websites surrounding the selection-based Dutch web archive. We found it to be substantial, with almost as many references to unarchived URLs as pages in the archive. These sites complement the sites collected based on the selection policies and provide context from the web at large, including the most popular sites in the country. The answer to our first research question is resoundingly positive: the indirect evidence of lost web pages holds the potential to significantly expand the coverage of the web archive. However, the resulting web page representations are different in nature from the usual representations based on web page content. We will characterize the web page representations based on derived descriptions in the next section.

5 Representations of unarchived pages

In this section, we study RQ2: How rich are the representations that can be created for unarchived web pages? We build implicit representations of unarchived web pages and domains, based on link evidence, URL words and anchor text, and investigate the richness (or sparseness) of the resulting descriptions in the number of incoming links, aggregated anchor text and URL words. We break this down over unarchived home pages and other pages. Our finding is that the anchor links and text descriptions have a highly skewed distribution: popular pages such as home pages have more terms, but the richness tapers off quickly.

5.1 Indegree

In general, a representation which we can generate for a target page can potentially be richer if it includes anchor text contributed from a wider range of source sites, i.e. has a higher indegree. Therefore, we look at the number of incoming links for each target URL of the uncovered archive in Fig. 2. It reflects a highly skewed distribution: all target representations in the outer aura have at least 1 source link, 18 % of the collection of target URLs has at least 3 incoming links, and 10 % has 5 links or more. The pages in the inner aura have a lower number of incoming links than the pages in the outer aura. To check whether this is related to a higher number of intra-server (internal site) links, we also assessed the types of incoming links.

We differentiate between two link types that can be extracted from archived web content: intra-server links, pointing to the pages in the same domain of a site, and inter-server links, that point to other websites. Table 15 shows the distribution of link types to unarchived URLs. The majority of unarchived URLs in the inner aura originate from the same source domain (i.e. a site on the seedlist), while the degree of intra-server links pointing to unarchived URLs in the outer aura is much smaller. There are very few link targets with both intra-server and inter-server link sources in the inner and outer aura.

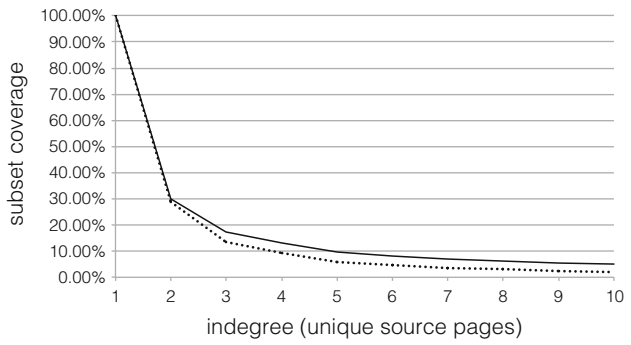

Fig. 2 Number of unique source pages (based on MD5 hash) compared to subset coverage (*dotted line* inner aura, *solid line* outer aura)

Table 15 Link types of unarchived URLs

	Inner aura	%	Outer aura	%
Intra-server	5,198,479	94.4	2,065,186	39.8
Inter-server	289,412	5.3	3,098,399	59.7
Both	18,084	0.4	27,930	0.5

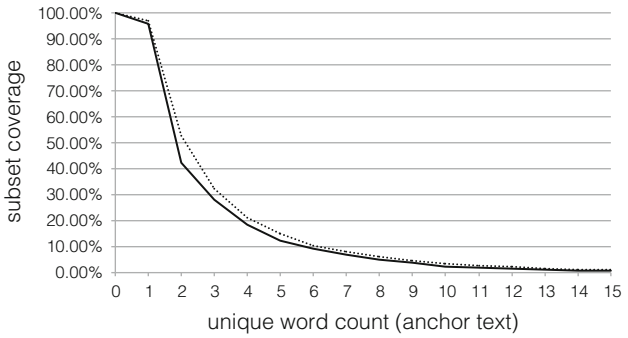

Fig. 3 Number of unique anchor words in the anchor text representation compared to subset coverage (*dotted line* inner aura, *solid line* outer aura)

5.2 Anchor text representations

A key influence on the utility of possible representations of unarchived pages is the richness of the contributed anchor text. In the aggregated anchor text representations, we counted the number of unique words in the anchor text. Figure 3 shows the number of unique words compared to subset coverage. Like the previous distribution of incoming source links, the distribution of unique anchor text is rather skewed. While 95 % of all target URLs in the archive have at least 1 word describing them, 30 % have at least 3 words as a combined description, and around 3 % have 10 words or more (though still amounting to 322,245 unique pages). The number of unique words per target is similar for both the inner and outer aura.

5.3 URL words

As the unique word count of page representations is skewed, we also looked at other sources of text. One of these potential sources are the words contained in the URL.

For instance, the URL aboriginalartstore.com.au/aboriginal-art-culture/the-last-nomads.php can be tokenized and contains several unique words which might help to characterize the page. Specifically, we consider alphanumerical strings between 2 and 20 alphanumerical characters as words. Hence, this URL contains 10 words: ‘aboriginalart-

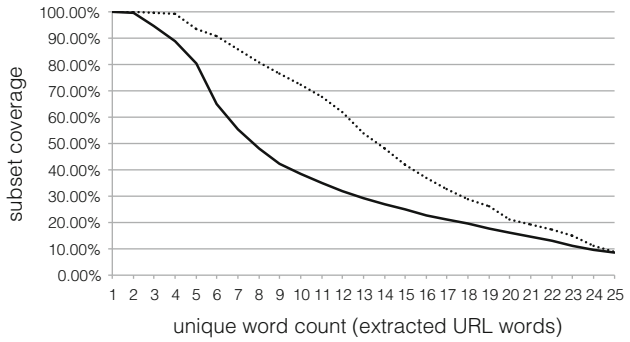

Fig. 4 Number of unique words in the URL representation compared to subset coverage (*dotted line* inner aura, *solid line* outer aura)

Table 16 Target structure distribution

Slash count	Inner aura	%	Slash count	Outer aura	%
0	3765	0.1	0	324,782	6.3
1	373,070	6.8	1	921,719	17.8
2	587,416	10.7	**2**	**1,543,129**	**29.7**
3	662,573	12.0	3	535,293	10.3
4	**1,098,947**	**20.0**	4	417,361	8.1
5	535,564	9.7	5	284,237	5.5

Bold values indicate most frequent slash count

store', 'com', 'au', 'aboriginal', 'art', 'culture', 'the', 'last', 'nomads,' and 'php'.

Figure 4 indicates the number of words contributed for each subset. It shows a difference between the inner and outer aura: the pages in the inner aura have more URL words than those in the outer aura. One likely reason is that the URLs for the pages in the inner aura are longer and, therefore, contribute more words.

To obtain a better view of the distribution of pages at different site depths, we also looked at the slashcount of absolute URLs (see Table 16). This analysis shows that the pages in the outer aura are mainly located at the first levels of the site (i.e. homepage to third level). The links towards the inner aura, however, are pointing to pages that are deeper in the hierarchy, probably because 94.4 % of this subset consists of intra-site link targets (links within a site).

5.4 Homepage representations

As mentioned in Sect. 2.2, anchors have been used for homepage finding, since links often refer to homepages. A homepage, as defined by the Merriam-Webster dictionary, is "the page typically encountered first on a Web site that usually contains links to the other pages of the site". To verify to what extent our dataset contains homepages, we looked at whether a homepage is available for each captured host in the outer aura. A basic way to calculate this is to equate homepages with the entry-level pages of hosts in the unarchived aura. Hence, we counted all unarchived target URLs consisting of only a hostname, resulting in 324,807 captured entry-level pages for the outer aura of the archive. In other words, 67.0 % of all hosts have their homepage captured in our dataset. Another way is to count pages having a slash-count of 0, but also counting additional pages with higher slashcounts using manual string-based filters (e.g. URLs including '/index.html'), yielding homepages for 336,387 hosts (69.8 %).

This can be important from a preservation and research perspective, since homepages are essential elements not only of websites, but also for the representations that we can generate from the link evidence, because homepages tend to have a higher indegree and more available anchor text. In our dataset, this is for instance reflected in the higher average number of anchor words (2.71) for the homepages as compared to the non-homepages (2.23 unique words). Here, we looked at the specific homepages available in the dataset, but not at the lower pages in the hierarchy. In Sect. 6 we look at the potential added value of aggregating link evidence from all pages under a host, creating site-level representations.

Table 17 Sample aggregated anchor and URL words

(A) Vakcentrum [domain]	(B) Nesomexico [non-domain]
Vakcentrum.nl (6)	Mexico (3)
Detailhandel (2)	Government (1)
Zelfstandige (2)	Overheid (1)
Ondernemers (2)	Mexican (1)
Levensmiddelen (2)	Mexicaanse (1)
Brancheorganisatie (1)	Beurzen (1)
httpwwwvakcentrumnl (1)	Nesomexico (1)
Vgl (1)	Scholarship (1)
Vereniging (1)	Programmes (1)

5.5 Qualitative analysis

Finally, we provide some concrete examples of representations that we can create for target URLs in this dataset. We first look at a homepage from our evaluation sample: vakcentrum.nl, a Dutch site for independent professionals in the retail sector. It has 142 inlinks from 6 unique hosts (6 different anchor text strings), resulting in 14 unique words. Table 17A displays 9 of the unique words (excluding stopwords). They provide a basic understanding of what the site is about: a branch organization for independent retailers in the food sector.

For other non-homepage URLs it is harder to represent the contents based on the anchor text alone. Take,

for example knack.be/nieuws/boeken/blogs/benno-barnard, a page that is not available on the live web anymore. It only has 2 anchor text words: 'Benno' and 'Barnard'. From the URL, however, we can further characterize the page: it is related to news ('nieuws'), books ('boeken') and possibly is a blog. Hence, we have discovered a 'lost' URL, of which we can get a (basic) description by combining evidence. Other non-homepage URLs have a richer description, even if the source links only originate from 1 unique host. For example nesomexico.org/dutch-students/study-in-mexico/study-grants-and-loans is a page that is not available via the live web anymore (3 incomplete captures are located in the Internet Archive). The anchor text, originating from utwente.nl (a Dutch University website), has 10 unique words, contributed from 2 unique anchors. In Table 17B the anchor words are shown. The URL words can enrich the representation, providing an indication of the page's content together with the anchor text. Of course, the richness of potential descriptions varies for each recovered target URL. The number of unique words in both anchor text and URL can serve as a basic estimate of the utility of a representation.

Summarizing, the inspection of the richness of representations of unarchived URLs indicates that the incoming links and the number of unique words in the anchor text have a highly skewed distribution: for few pages we have many descriptions which provide a reasonable number of anchors and unique terms, while the opposite holds true for the overwhelming majority of pages. The succinct representations of unarchived web pages are indeed very different in nature. The answer to our second research question is mixed. Although establishing the existence of 'lost' web pages is an important result in itself, this raises doubts whether the representations are rich enough to characterize the page's content. Therefore, we investigate in the next section whether aggregations of unarchived pages at the host level will improve the richness (and utility) of derived representations.

6 Representations of unarchived websites

In this section, we study RQ3: Can we create richer representations of web sites based on aggregating page-level evidence from pages sharing the same hostname? We build aggregated representations of unarchived contents at the host level. We use the combined anchor text and URL words to enrich representations of unarchived websites in the outer *aura*. Our finding is that the site-level representations' distributions of indegree and number of words are richer than the distributions of (home) page representations: the indegree and number of unique words are substantially higher. However, the introduction of pages from different site depths potentially introduces some noise.

6.1 Rationale and method

From a perspective of preservation and research, entry pages are important pages to capture in web archives. We saw in the previous section that the indegree and number of unique words for the homepages (defined as entry pages at the host level) are slightly higher than for other pages of a site. However, these succinct representations might not always be rich enough to characterize full unarchived websites. To amend this, we now focus on generating richer representations for hosts in the Dutch web archive.

We focus on the outer aura of the archive, since generating site-level representations for hosts in this subset is potentially more valuable than for the inner aura of the archive (as the main contents are already in the archive). We aggregate pages of unarchived websites at the *host* level (i.e. we create separate representations for zorg.independer.nl, forum.independer.nl etc.), because this results in more fine-grained representations than aggregation at the *domain* level (i.e. aggregating *.independer.nl under one representation).

In the previous section we found homepages for 324,807 out of 481,797 detected hosts in the outer aura of the web archive. Here, we take this a step further and create site-level representations for each host: this representation consists of aggregated evidence such as incoming URLs, incoming anchor words, unique sources pages, and so forth. We aggregate this information for all uncovered pages under a given host. For example, a site-level representation of 3fm.nl, a Dutch radio station, contains information from pages under 3fm.nl/nieuws, 3fm.nl/dj, 3fm.nl/megatop50, plus all other child pages of the uncovered host. Taken together, these might provide a richer description of an unarchived website, reducing anchor text sparsity. In case there is no entry-level page captured in the outer aura, we still aggregate the deeper pages of a site under a certain host (e.g. fmg.ac/projects/medlands and other non-entry pages are aggregated under fmg.ac). This way, we generated 481,305 site-level representations for uncovered hosts in the outer aura of the archive.

6.2 Comparisons

Table 18 shows a comparison between page-level homepage representations and aggregated site-level representations. The mean indegree for the site representations is more than two times higher than for the page-based representations. In addition, the mean number of anchor text words and URL words are substantially higher, as well as the combined unique words from both sources. The mean overlap between anchor text and URL words is 0.48 words for the page-based representations, and 4.03 words for the site-based representations, indicating higher similarity between anchor and URL words for the site-based representations. We now look more

Table 18 Richness of host representations at the page-level (entry pages) and site-level (aggregated pages)

Representations	Count (K)	Indegree	Anchor words	URL words	Combined uniq words
Page-level	325	24.14	2.71	2.32	4.56
Site-level	481	56.31	10.06	11.82	17.85

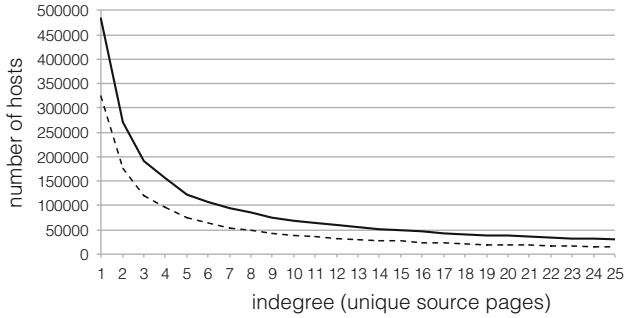

Fig. 5 Indegree of site-level representations (*solid line*) versus page-level homepage representations (*dotted line*)

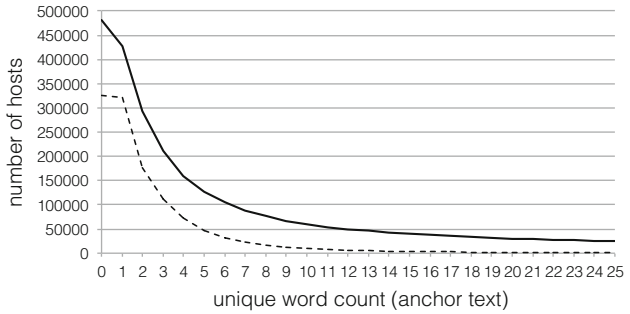

Fig. 6 Number of unique words in the anchor text for site-level representations (*solid line*) versus page-level homepage representations (*dotted line*)

in detail at the involved indegree, anchor text, and URL word distributions.

6.3 Indegree, anchor text, and URL words

Figure 5 shows the indegree distribution of site-level anchor text representations in the outer aura. This indegree is based on the number of inlinks from unique source pages (based on the MD5 hash) to all pages of a given host. The graph also includes the indegree for the page-level representations of hosts in the outer aura. We see that the site-based approach results in more representations of hosts, with a higher number of unique inlinks. For example, there are 123,110 site representations (25.6 %) with at least 5 inlinks, compared to only 74,092 page-based representations of entry-level pages with at least 5 inlinks (22.8 %). The higher indegree could contribute to a richer representation, also in terms of unique anchor text words.

In Fig. 6, the number of unique anchor text words for each site-based anchor text representation is summarized.

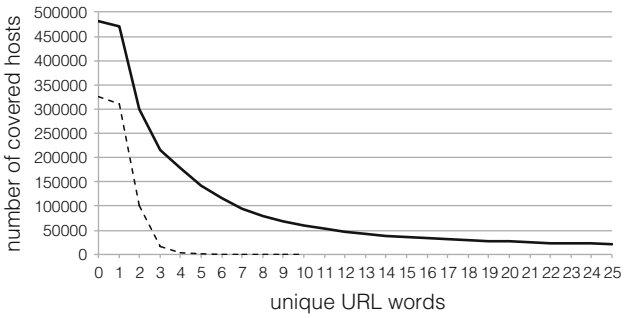

Fig. 7 Number of unique words in the URL of site-level (*solid line*) versus page-level homepage representations (*dotted line*)

The number of unique words is still skewed, but much richer than the page-based representations. There are 126,218 site-level representations with at least 5 anchor words (26.2 % of all site-level representations), while at the page-level there is only available evidence for 46,795 entry pages of hosts with 5 anchor words or more (14.4 % of all page-level homepage representations).

Figure 7 shows the distribution of the number of URL words for homepage and site representations. Naturally, the URL word count for the page representations is quite small: representations of entry pages usually have short URLs. For the site representations, we aggregate the tokenized words from up to 100 URLs per host. We observe the value of aggregating URL words at the host level: substantially more words are available for website representations.

The higher number of available anchor text and URL words means that generated representations are potentially richer, as more words can shed light on more facets of a site. The question is if these words also improve the potential to correctly characterize a site, since they are contributed from different pages and sections of a site.

6.4 Qualitative analysis

So far, we have seen indications of the added value of site-level representations to reduce anchor text sparsity. Take, for example, webmath.com, a website to solve math problems. For the homepage of this site, we have 1 unique incoming anchor, contributing only one unique word ("webmath"). However, as the previous sections have shown, we can also generate representations for websites by aggregating all anchor text for a given host. In the case of webmath.com, we then have 10 unique incoming anchors from 4 unique

Table 19 Aggregated anchor words webmath.com

(A) Page-level anchors	(B) Host-level anchors	
Webmath (1)	Math (9)	Webmath (4)
	Algebra (2)	Geometry (2)
	Calculus (2)	Web (2)
	General (2)	Everyone (2)
	Plots (2)	Stepbystep (1)

source hosts. These together contribute 15 unique words, which characterize the site in a better way than the single word for the homepage (see Table 19). On the other hand, this does not necessarily apply to all websites: for some hosts, adding aggregated anchor text introduces noise. For example, the Centre for European Reform (cer.org.uk) has 5 unique anchor words, including "centre" and "reform". Aggregating anchor text evidence from all pages under this host results in 30 unique words, contributed from 17 unique hosts. These words include words useful for the description of the site, e.g. "european", and "future". However, there are also many words added to the site representation, such as "Nabucco", and "India", that might not be suitable to characterize the whole site. The question that follows from our qualitative analysis is whether the site-based representations are actually specific enough to characterize a given site's content, as noise might be added when aggregating anchor text evidence on the site level.

Summarizing, we looked at the impact of host-level aggregations of link evidence on the richness of generated representations. Our analysis showed a significant increase of unique anchor words available for each site, potentially overcoming anchor text sparsity. In our qualitative analysis we saw examples of not only improved representations, but also of added noise caused by the aggregation. Hence the answer to our third research question is still mixed. Although establishing the existence of 'lost' web pages and websites is an important result in itself, the resulting representations are sparse, and may not be rich enough to characterize their unique content. We decide to investigate this in the next section.

7 Finding unarchived pages and sites

In this section, we study RQ4: Are the resulting derived representations of unarchived pages and web sites useful in practice? Do they capture enough of the unique page content to make them retrievable amongst millions of other page representations? We focus on the retrieval of unarchived web pages based on their derived representations in a known-item search setting and compare page-level with host-level representations. Our finding is that the succinct representation is generally rich enough to identify pages on the unarchived web: in a known-item search setting we retrieve these pages within the first ranks on average. The aggregation of pages at the host-level leads to further improvement of the retrieval effectiveness.

7.1 Evaluation setup

To evaluate the utility of uncovered evidence of the unarchived web, we indexed representations in the *outer aura* of the archive. Hence, we indexed the unarchived pages, detected via the archive's link structure, whose parent domain is *not* on the seedlist. These representations consist of unarchived URLs, aggregated anchor text and URL words of unarchived pages and hosts. We indexed these documents using the Terrier 3.5 IR Platform [26], utilizing basic stopword filtering and Porter stemming. We indexed three sets of representations:

- page-level representations for all 5.19 M unarchived URLs
- page-level representations for 324.807 homepages (entry pages of hosts)
- aggregated site-level representations for 324.807 unarchived hosts

For each set of representations, we created three indices. The first index of every category uses only the aggregated anchor words (*anchT*). The second index (*urlW*) uses other evidence: the words contained in the URL. Non-alphanumerical characters were removed from the URLs and words of a length between 2 and 20 characters were indexed. Finally, the third index for each set of representations consists of both aggregated anchor text and URL words (*anchTUrlW*).

To create known-item queries, a stratified sample of the dataset was taken, consisting of 500 random non-homepage URLs, and 500 random homepages. Here, we define a non-homepage URL as having a slashcount of 1 or more, and a homepage URL as having a slashcount of 0. These URLs were checked against the Internet Archive (pages archived in 2012). If no snapshot was available in the Internet Archive (for example because of a *robots.txt* exclusion), the URL was checked against the live web. If no page evidence could be consulted, the next URL in the list was chosen, until a total of 150 queries per category was reached. The consulted pages were used by two annotators to create known-item queries. Specifically, after looking at the target page, the tab or window is closed and the topic creator writes down the query that he or she would use for refinding the target page with a standard search engine. Hence the query was based on their recollection of the page's content, and the annotators were completely unaware of the anchor text representation

(derived from pages linking to the target). As it turned out, the topic creators used 5–7 words queries for both homepages and non-homepages. The set of queries by the first annotator was used for the evaluation ($n = 300$), and the set of queries by the second annotator was used to verify the results ($n = 100$). We found that the difference between the annotators was low: the average difference in resulting MRR scores between the annotators for 100 homepage queries in all indices was 8 %, and the average difference in success rate was 3 %.

For the first part of our evaluation (Sect. 7.2), we ran these 300 homepage and non-homepage queries against the *anchT*, *urlW* and *anchTUrlW* page-level indices created in Terrier using its default InL2 retrieval model based on DFR and saved the rank of our URL in the results list. For the second part of our evaluation (Sect. 7.3), the 150 homepage queries were run against page-level and site-level indices of hosts in the outer aura of the archive.

To verify the utility of anchor, URL words and combined representations, we use the Mean Reciprocal Rank (MRR) for each set of queries against each respective index.

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{Q} \frac{1}{\text{rank}_i} \tag{1}$$

The MRR (1) is a statistical measure that looks at the probability of retrieving correct results. It is the average over the scores of the first correct result for each query (calculated by $\frac{1}{\text{rank}}$). We also compute the success rate at rank 10, that is, for which fraction of the topics do we actually retrieve the correct URL within the first 10 ranks.

We used unarchived pages uncovered from the Dutch web archive, that are either available in the Internet Archive, or still available on the live web, in order to have the ground truth information about the page's content. This potentially introduces bias—there can be some difference between the pages that still are active, or have been archived, and those that are not—but the URLs did not suggest striking differences. Out of all randomly chosen homepages surveyed, 79.9 % were available via either the Internet Archive or the live web. However, this was not the case for the non-homepages (randomly selected pages with a slash count of 1 or more), as only 49.8 % could be retrieved via the Internet Archive or the live web. The underlying reasons that many URLs could not be archived include restrictive robots.txt policies (e.g. Facebook pages), contents specifically excluded from the archive (e.g. Twitter accounts and tweets), but also links pointing to automatically generated pages (e.g. LinkedIn 'share' links). The unavailability of URLs strengthens the potential utility of generated page representations, for example, via aggregated anchor text, since no page evidence can be retrieved anymore.

Table 20 Mean reciprocal rank (MRR)

MRR	Queries	AnchT	UrlW	AnchTUrlW
Homepages	150	0.327	0.317	0.489
Non-homepages	150	0.254	0.384	0.457
Combined	300	0.290	0.351	0.473

7.2 Page-based representations

This section contains the first part of our evaluation, focusing on page-based representations of unarchived content. We use the indices of the 5.19 M page representations in the outer aura of the archive, combined with the 150 homepage and 150 non-homepage known-item queries.

7.2.1 MRR and success rate

MRR scores were calculated for the examined homepages and non-homepages to test as to what extent the generated representations suffice to retrieve unarchived URLs. The final results of the evaluation based on MRR are summarized in Table 20. We found that the MRR scores for the homepages and non-homepages are quite similar, though some differences can be seen. Using the anchor text index, the homepages score higher than the non-homepages, possibly because of the richer representations available for these homepages. The scores for the URL words index are naturally higher for the non-homepages: they have longer URLs and, therefore, more words that could match the words used in the query. Finally, we can see that the combination of anchor and URL words evidence significantly boosts the retrieval effectiveness: the MRR is close to 0.5, meaning that in the average case the correct result is retrieved at the second rank.

We also examined the success rate, that is, for which degree of the topics do we actually retrieve the correct URL within the first 10 ranks? Table 21 shows that again there is some similarity between the homepages and non-homepages. The homepages score better using the anchor text index than the non-homepages: 46.7 % can be retrieved. On the other hand, the non-homepages fare better than the homepages using the URL words: 46.0 % of the non-homepages is included in the first 10 ranks. Again, we see that combining both representations results in a significant increase of the success rate: we can retrieve 64.0 % of the homepages, and 55.3 % of the non-homepages in the first 10 ranks.

The MRR scores indicate that anchor text in combination with tokenized URL words can be discriminative enough to do known-item search: the correct results can usually be retrieved within the first ranks. Second, the success rates show that by combining anchor text and URL word evidence, 64 % of the homepages, and 55.3 % of the deeper pages can

Table 21 Success rates (target page in top 10)

Success@10	Queries	AnchT (%)	UrlW (%)	AnchTUrlW (%)
Homepages	150	46.7	39.3	64.0
Non-homepages	150	34.7	46.0	55.3
Combined	300	40.7	42.7	59.7

Table 22 Division based on indegree of unique hosts

Indegree	Pages	Word cnt	MRR anchT	Homepage (%)
1	251	2.9	0.29	42.6
2	28	3.8	0.19	82.1
3	12	4.5	0.29	100
4+	9	7.3	0.49	88.9

be retrieved. This provides positive evidence for the utility of these representations.

The performance on the derived representations is comparable to the performance on regular representations of web pages [11]. Here we used a standard retrieval model, without including various priors tailored to the task at hand [20].

7.2.2 Impact of indegree

We now examine the impact of the number of unique inlinks on the richness of anchor text representations at the page level. For example, the homepage Centre for European Reform (cer.org.uk) receives links from 3 unique hosts: portill.nl, europa-nu.nl, and media.europa-nu.nl, together contributing 5 unique anchor words, while the page actionaid.org/kenya has 1 intra-server link from actionaid.org, contributing only 1 anchor word. For the combined 300 topics (domains and non-domains together), we calculated the mean unique word count, the MRR, and the degree of homepages in the subset.

The results in Table 22 show that an increase in the number of inlinks from unique hosts results in a rise of the mean word count. However, it also illustrates the skewed distribution of our dataset: the majority of pages (251 out of 300) have links from only one source host, while a much smaller set (49 out of 300) have links from 2 or more unique source hosts. The table also provides evidence of the hypothesis that the homepages have more inlinks from unique hosts than non-homepages: at an indegree of 2 or more, the homepages take up more than 80 % of the set of pages. We can also observe from the data that the MRR using the anchor text index in our sample is highest when having links from at least 4 unique hosts.

7.3 Site-based representations

The encountered importance of a higher indegree and unique word count for representing unarchived pages encouraged us to experiment with other approaches to improve representations. This second part of our evaluation compares the retrieval effectiveness of entry page representations of hosts (at the page-level) with aggregated representations of all pages under a certain host (at the site-level). We use the indices created using these two sets of 324.807 representations in combination with the 150 known-item homepage queries.

7.3.1 MRR and success rate

Table 23 summarizes the MRR scores for the page and site-level representations. The score for the site-level index based on anchor text is 4 % higher than the page-level anchor text index. In our derived site-level representations, we aggregated up to 100 URLs and tokenized the URL words. The value of this approach is seen in the MRR score for the site-level *urlW* index: the MRR score rises with 5 %. Finally, the combined representations (*AnchTUrlW*) work best, with a higher MRR rating than both *AnchT* and *UrlW* indices alone. The MRR score is improved almost 6 % by aggregating the anchor text and URL word evidence, showing the value of site-level aggregation of link-based evidence.

To get a better insight into which queries perform better and worse, we look more in detail at the differences in performance for all 150 homepage queries across the page-level and site-level AnchTUrlW indices (Table 24). For 21 topics, the site-based representations fare better, and for 14 topics, the page-based representations fare better. Hence we see some evidence for both the improvement of site representations and for the potential introduction of noise (influencing the topics that did not perform better). Another striking observations is that the scores for 115 of the topics remain the same.

Table 23 Site representations: mean reciprocal rank (MRR)

MRR	#Q	AnchT	UrlW	AnchTUrlW
Page-level	150	0.435	0.393	0.590
Site-level	150	0.452	0.412	0.626

Table 24 MRR score comparison for homepage queries in site-level (*srAnchTUrlW*) and page-level (*plAnchTUrlW*) anchor text indices

Site-level	Better	Same	Worse	Page-level
	21	115	14	

Table 25 Coverage of URLs by site representations and associated counts, mean number of words, and MRR

Covered URLs	Repr cnt	Page-rep words	Site-rep words	Page-rep MRR	Site-rep MRR
1	103	4.3	4.3	0.62	0.64
2	22	4.8	7.0	0.47	0.52
3	6	7.1	10.7	0.44	0.57
4+	19	7.3	36.9	0.62	0.66

Table 26 Site representations: success rates (target page in top 10)

Success@10	#Q	AnchT (%)	UrlW (%)	AnchTUrlW (%)
Page-level	150	56.0	46.0	74.7
Site-level	150	55.3	46.7	74.7

The reason might be related to the skewed distribution of our dataset: for some hosts we might have few captured URLs which could be used in a site-based representation.

Therefore, we now look at the number of URLs available per host, as this might influence the richness of representations. Table 25 shows that for the majority of target hosts (103 out of 150 known-item queries), there is only one URL in our dataset. For 47 hosts, 2 or more URLs are available. In the table, we also see that the mean number of unique words increases when the number of covered URLs increases. For the 47 hosts with two or more covered URLs, the MRR values for the site-level representations are clearly higher than for the page-level representations.

Finally, we look at the success rates (the degree of topics with the correct URL in the first 10 results). The success rates in Table 26 show a different outcome than the MRR score comparison: page-level and site-level representations score remarkably similar. There is a slightly lower success rate for the site-level anchor text index, and a slight improvement for the URL words index. Similar scores might be caused by the skewed distribution of the dataset. As we have seen in Table 25, for 103 out of 150 hosts in our evaluation set we have just 1 captured URL. In those cases, aggregating URLs by host does not increase success rates.

Summarizing, we investigated whether the derived representations characterize the unique content of unarchived web pages in a meaningful way. We conducted a critical test cast as a known-item finding task, requiring to locate unique pages amongst millions of other pages—a true needle-in-a-haystack task. The outcome of the first part of our evaluation is clearly positive: with MRR scores of about 0.5, we find the relevant pages at the second rank on average, and for the majority of pages the relevant page is in the top 10 results. The second part of our evaluation compared page-level representations with site-level representations. We found that using site-level representations improves retrieval effectiveness for homepage queries with 4–6 %, while the success rates remain stable. Hence, the answer to our fourth research question is again positive: we can reconstruct representations of unarchived web pages that characterize their content in a meaningful way.

8 Discussion and conclusions

Every web crawl and web archive is highly incomplete, making the reconstruction of the lost web of crucial importance for the use of web archives and other crawled data. Researchers take the web archive at face value and equate it to the web as it once was, leading to potentially biased and incorrect conclusions. The main insight of this paper is that although unarchived web pages are lost forever, they are not forgotten in the sense that the crawled pages may contain various evidence of their existence.

In this article, we proposed a method for deriving representations for unarchived content, using the evidence of the unarchived and lost web extracted from the collection of archived web pages. We used link evidence to first *uncover* target URLs outside the archive, and second to *reconstruct* basic representations of target URLs outside the archive. This evidence includes aggregated anchor text, source URLs, assigned classification codes, crawl dates, and other extractable properties. Hence, we derived representations of web pages and websites that are not archived, and which otherwise would have been lost.

We tested our methods on the data of the selection-based Dutch web archive in 2012. The analysis presented above first characterized the contents of the Dutch web archive, from which the representations of unarchived pages were subsequently uncovered, reconstructed and evaluated. The archive contains evidence of roughly the same number of unarchived pages as the number of unique pages included in the web archive—a dramatic increase in coverage. In terms of the number of domains and hostnames, the increase of coverage is even more dramatic, but this is partly due to the domain restrictive crawling policy of the Dutch web archive. Whereas this is still only a fraction of the total web, using the data extracted from archived pages we reconstruct specifically those unarchived pages which once were closely interlinked with the pages in the archive.

The recovery of the unarchived pages surrounding the web archive, which we called the 'aura' of the archive, can be used for assessing the completeness of the archive. The recovered pages may help to extend the seedlist of the crawlers of selection-based archives, as these pages are potentially

relevant to the archive. Additionally, representations of unarchived pages could be used to enrich web archive search systems and provide additional search functionality. Including the representations of pages in the outer aura, for example, is of special interest as it contains evidence to the existence of top websites that are excluded from archiving, such as Facebook and Twitter. This is supported by the fact that only two years since the data were crawled, 20.1 % of the found unarchived homepages and 45.4 % of the non-home pages could no longer be found on the live web nor the Internet Archive.

However, given that the original page is lost and we rely on indirect evidence, the reconstructed pages have a sparse representation. For a small fraction of popular unarchived pages we have evidence from many links, but the richness of description is highly skewed and tapers off very quickly—we have no more than a few words. This raises doubts on the utility of the resulting representations: are these rich enough to distinguish the unique page amongst millions of other pages? We addressed this with a critical test cast as a known item search in a refinding scenario. As it turns out, the evaluation of the unarchived pages shows that the extraction is rather robust, since both unarchived homepages and non-homepages received similar satisfactory MRR average scores: 0.47 over both types, so on average the relevant unarchived page can be found in the first ranks. Combining page-level evidence into host-level representations of websites leads to richer representations and an increase in retrieval effectiveness (an MRR of 0.63). The broad conclusion is that the derived representations are effective and that we can dramatically increase the coverage of the web archive by our reconstruction approach.

Hence, it would be valuable to integrate derived representations into search systems for web archives. The representations of unarchived web content are relatively sparse, which may lead to low rankings for unarchived representations in comparison to fully archived content. There are two standard approaches to avoid this. First, a search index for archived results could be combined with a second search index for unarchived results, and archived and unarchived results can be interleaved, either by round robin or a score weighted interleaving, ensuring highly ranked archived and unarchived results will appear together. A second approach is the use of an anchor text index for ranking both archived and unarchived results. The use of the same type of evidence for ranking both types of content also facilitates the concurrent display of both archived and unarchived results in web archive search systems.

There are some limitations to the method as described in this study that could be addressed in follow up work. The first concerns the aggregation of links by year, which may over-generalize timestamps of the unarchived pages and, therefore, decrease the accuracy of the representation. Further study is needed on the right window, or weighted representations, taking into account estimates of the volatility or dynamic nature of the websites and web pages at hand. Second, we used data from a selective archive, whose crawler settings privilege selected hostnames and are instructed to ignore other encountered sites. This affects the relative distribution of home pages and non-homepages, both in the archive as well as in the unarchived pages. Hence, the exact impact of the crawling strategy remains an open problem. It would be of interest to determine which crawling strategies provide the best starting point for reconstructing the associated unarchived web. Third, our initial results in this paper are based on straightforward descriptions of pure anchor text and URL components and standard ranking models. In follow-up research we will examine the effect of including further contextual information, such as the text surrounding the anchors, and advanced retrieval models that optimally weight all different sources of evidence. Fourth, the recovered representations are rather skewed and hence most of the uncovered pages have a relatively sparse representation, while only a small fraction has rich representations. We addressed this by generating site-level representations, but advanced mixture models at various levels of representation, and advanced weighting schemes treating the observed evidence as a sample from a larger population, can further enrich the representations.

Acknowledgments We thank the reviewers for their comments that helped shape this article, and gratefully acknowledge the collaboration with the Dutch Web Archive of the National Library of the Netherlands, in particular René Voorburg. This research was supported by the Netherlands Organization for Scientific Research (NWO CATCH program, WebART project #640.005.001). The link extraction and analysis work was carried out on the Dutch national e-infrastructure with the support of SURF Foundation.

References

1. AlSum, A., Weigle, M., Nelson, M., Van de Sompel, H.: Profiling web archive coverage for top-level domain and content language. Int. J. Digit. Libr. **14**, 149–166 (2014). doi:10.1007/s00799-014-0118-y
2. Broder, A.Z., Gabrilovich, E., Josifovski, V., Mavromatis, G., Metzler, D., Wang, J.: Exploiting site-level information to improve web search. In: CIKM, ACM, pp. 1393–1396 (2010). doi:10.1145/1871437.1871630
3. Brügger, N.: Web history and the web as a historical source. Zeithist. Forsch. **9**, 316–325 (2012)
4. Brunelle, J.F., Kelly, M., SalahEldeen, H., Weigle, M.C., Nelson, M.L.: Not all mementos are created equal: measuring the impact of missing resources. In: DL, IEEE, pp. 321–330 (2014)

5. Craswell, N., Hawking, D., Robertson, S.: Effective site finding using link anchor information. In: SIGIR, pp. 250–257. ACM, New York (2001)
6. Dai, N., Davison, B.D.: Mining anchor text trends for retrieval. In: ECIR, LNCS, vol. 5993, pp. 127–139. Springer, Berlin (2010)
7. Day, M.: Preserving the fabric of our lives: a survey of web. In: ECDL, LNCS, vol. 2769, pp. 461–472. Springer, Berlin (2003)
8. Dou, Z., Song, R., Nie, J.Y., Wen, J.R.: Using anchor texts with their hyperlink structure for web search. In: SIGIR, pp. 227–234. ACM, New York (2009)
9. Fujii, A.: Modeling anchor text and classifying queries to enhance web document retrieval. In: WWW, pp. 337–346. ACM, New York (2008)
10. Gomes, D., Silva, M.J.: Characterizing a national community web. ACM Trans. Intern. Technol. **5**, 508–531 (2005)
11. Hawking, D., Craswell, N.: Very large scale retrieval and web search. In: TREC: Experiment and Evaluation in Information Retrieval, Chapter 9. MIT Press, Cambridge (2005)
12. Hockx-Yu, H.: The past issue of the web. In: Web Science, p. 12. ACM, New York (2011)
13. Hockx-Yu, H.: Access and scholarly use of web archives. Alexandria **25**, 113–127 (2014)
14. Huurdeman, H.C., Ben-David, A., Kamps, J., Samar, T., de Vries, A.P.: Finding pages in the unarchived web. In: DL, IEEE, pp. 331–340 (2014)
15. International Internet Preservation Consortium (2014) Web Archiving Why Archive the Web? http://netpreserve.org/web-archiving/overview. Accessed 2014-12-01
16. Kamps, J.: Web-centric language models. In: CIKM, pp. 307–308. ACM, New York (2005)
17. Klein, M., Nelson, M.L.: Moved but not gone: an evaluation of real-time methods for discovering replacement web pages. Int. J. Digit. Libr. **14**(1–2), 17–38 (2014)
18. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM **46**(5), 604–632 (1999). doi:10.1145/324133.324140
19. Koolen, M., Kamps, J.: The importance of anchor text for ad hoc search revisited. In: SIGIR, pp. 122–129. ACM, New York (2010)
20. Kraaij, W., Westerveld, T., Hiemstra, D.: The importance of prior probabilities for entry page search. In: SIGIR, pp. 27–34. ACM, New York (2002)
21. Kraft, R., Zien, J.: Mining anchor text for query refinement. In: WWW, pp. 666–674. ACM, New York (2004). doi:10.1145/988672.988763
22. Martinez-Romo, J., Araujo, L.: Analyzing information retrieval methods to recover broken web links. In: ECIR, LNCS, vol. 5993, pp. 26–37. Springer, Berlin (2010)
23. Masanès, J.: Web Archiving. Springer, Berlin (2006)
24. Metzler, D., Novak, J., Cui, H., Reddy, S.: Building enriched document representations using aggregated anchor text. In: SIGIR, pp. 219–226. ACM, New York (2009). doi:10.1145/1571941.1571981
25. Ogilvie, P., Callan, J.P.: Combining document representations for known-item search. In: SIGIR, pp. 143–150. ACM, New York (2003)
26. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A high performance and scalable information retrieval platform. In: Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006) (2006)
27. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: bringing order to the web. In: Technical Report 1999-66. Stanford University, Stanford (1999)
28. Ras, M.: Eerste fase webarchivering. In: Tech. rep., Koninklijke Bibliotheek, The Hague (2007)
29. Rauber, A., Bruckner, R.M., Aschenbrenner, A., Witvoet, O., Kaiser, M.: Uncovering information hidden in web archives: a glimpse at web analysis building on data warehouses. D-Lib Mag. **8**(12) (2002)
30. Samar, T., Huurdeman, H.C., Ben-David, A., Kamps, J., de Vries, A.: Uncovering the unarchived web. In: SIGIR, pp. 1199–1202. ACM, New York (2014). doi:10.1145/2600428.2609544
31. UNESCO (2003) Charter on the preservation of digital heritage (article 3.4). http://portal.unesco.org/en/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html. Accessed 1 Dec 2014
32. Van de Sompel, H., Nelson, M., Sanderson, R.: RFC 7089 - HTTP framework for time-based access to resource states—Memento. In: RFC, Internet Engineering Task Force (IETF) (2013). http://www.rfc-editor.org/rfc/rfc7089.txt. Accessed 1 Dec 2014