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]

[1] University of Amsterdam, Amsterdam, the Netherlands, {huurdeman, kamps, r.a.rogers}@uva.nl
[2] Centrum Wiskunde & Informatica, Amsterdam, the Netherlands, {samar, arjen}@cwi.nl
[3] The Open University, Ra'anana, Israel, anatbd@openu.ac.il

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.

1 Introduction

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.

We propose a method for deriving representations for unarchived content, by using the evidence of the unarchived web extracted from the collection of archived web pages. We use link evidence to firstly *uncover* target URLs outside the archive, and secondly 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 derive representations of web pages and websites that are not archived, and which otherwise would have been lost.

2 Unarchived Web Representations

We tested our methods on the data of the selection-based Dutch web archive in 2012. The analysis first characterizes 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.

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 their utility: are these rich enough to distinguish the unique page amongst millions of other pages?

We address this with a critical test cast as a known-item search in a refinding scenario. The evaluation 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).

3 Discussion and Conclusions

We investigated the recovery of the unarchived pages surrounding the web archive, which we called the 'aura' of the archive. 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. This is supported by the fact that only two years since the data was crawled, 20% of the found unarchived homepages and 45% of the non-home pages could no longer be found on the live web nor the Internet Archive.

The unarchived web pages can be used for assessing the completeness of the archive. The recovered pages help to extend the seedlist of the crawlers of selection-based archives, as the pages are potentially relevant to the archive. Additionally, representations of unarchived pages can 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.

Acknowledgments This is an extended abstract of [2] and [1]. Funded by NWO (CATCH program, WebART project # 640.005.001).

4 References

[1] H. C. Huurdeman, A. Ben-David, J. Kamps, T. Samar, and A. P. de Vries. Finding pages on the unarchived web. In *IEEE/ACM Joint Conference on Digital Libraries, JCDL 2014, London, United Kingdom, September 8-12, 2014*, pages 331–340. IEEE, 2014. http://dx.doi.org/10.1109/JCDL.2014.6970188.

[2] H. C. Huurdeman, J. Kamps, T. Samar, A. P. de Vries, A. Ben-David, and R. A. Rogers. Lost but not forgotten: finding pages on the unarchived web. *Int. J. on Digital Libraries*, 16:247–265, 2015. http://dx.doi.org/10.1007/s00799-015-0153-3.

DIR'15 November 27, 2015, Amsterdam, The Netherlands.