Extraction and Analysis of Referenced Web Links in Large-Scale Scholarly Articles

Ke Zhou
University of Edinburgh
Ke.Zhou@ed.ac.uk

Richard Tobin
University of Edinburgh
richard@inf.ed.ac.uk

Claire Grover
University of Edinburgh
Claire.Grover@ed.ac.uk

ABSTRACT

In this paper we report on a sub-task undertaken as part of Hiberlink, a project which is examining the phenomenon of reference rot within scholarly works. In our sub-task we aim to quantify and understand the nature of occurrence of links to web resources referenced from papers in very large-scale scholarly collections. We first introduce the challenges involved in extracting links from scholarly articles and develop and evaluate the accuracy of a set of link extraction systems. Secondly, five collections containing millions of scholarly articles with different characteristics (across different disciplines, time periods and publication types) are studied and we demonstrate that web resources are widely cited in scholarly publications and should be an important concern for digital preservation.

Categories and Subject Descriptors: H.3.7 [Information Systems: Information Storage and Retrieval]: Digital Libraries
Keywords: Link Extraction, Scholarly Data, Digital Preservation

1. INTRODUCTION

Citation of sources is fundamental to scholarly discourse. Traditionally such sources relate to statements made by other scholars and to the evidence from which these statements are drawn, i.e. published articles or books. Now, in the digital age, web-based scholarly endeavour has greatly enlarged the range of scholarly artifacts that are being published and referenced. Many of these are resources created as part of research activity such as software, datasets, websites, presentations, blogs, videos, etc. as well as scientific workflows and ontologies. The real-time nature of the web enables immediate access to those resources and dramatically increases the speed of knowledge dissemination.

The work reported here forms part of a broader project, Hiberlink[1] [3], which aims to quantify the extent to which referenced web resources cited at the time of publication can be accessed later on. Here we present the first step towards the Hiberlink research by profiling the way in which web resources are increasingly widely referenced in the scholarly world. We mainly aim to investigate the research question (RQ): *What is the extent of web-based links (URLs) that are referenced by scholarly works?*

The contributions of this paper are two-fold: (1) We provide an extensive evaluation of a variety of link extraction tools. We show that this task is quite challenging and our prototype performs well. (2) We conducted a large-scale study analysing and quantifying occurrences of referenced links extracted from scholarly articles. We show that a large number of scholarly articles contain links.

2. LINK EXTRACTION AND ANALYSIS

Challenges Accurately extracting web links from scholarly articles is not trivial. While some URLs are as simple as *http://foo.bar/*, others are considerably more complex and require knowledge of allowed protocols (ftps://foo.bar/ should be rejected).

Extraction is especially challenging when working with the most prevalent format, PDF, since this is mainly designed for presentation, rather than storing information in a structured fashion. There are two main challenges: (1) line breaks appearing within URLs (e.g. “http://en.wikipedia.org/" and “wiki/Url" appear in two separate lines in the PDF); (2) the use of images to represent characters in PDFs, in particular, an underscore character quite frequently occurs as an image in URLs.

Approach Our link (URL) extraction system consists of three steps: (1) Converting PDF into XML using the command line tool “pdftohtml -xml"[2]; (2) Fixing URL line breaks and underscore images within the PDF text in order to more accurately extract URLs; and (3) Extracting links/URLs from the XML file using regular expression matching.

For (2), line breaks in URLs can lead to extraction of a string that does not correspond to the actual URL (e.g. only part of it) or to an inability to detect one. Therefore, we apply a conservative strategy to fix frequently occurring (manually observed) error patterns, i.e. we only fix them when we are confident that no new false-positives are introduced. For example, we concatenate “http://" at the end of the first line to the remaining part of the URL at the start of the second line. In addition, by applying heuristics we recognise and convert images standing in for the underscore character as part of the process of converting PDF to XML.

For (3), working with XML files, there are two sources of URLs. One source is the explicit “a href" links annotated by authors or publishers, while the other source is links mentioned in the text. The former is a more reliable source (although mistakes/typos from the authors can occur) while the latter relies heavily on the performance of regular expression matching to identify URLs. We utilize regular expression matching for both sources of URLs.

Extracting URLs using a regular expression (regex) is not new

[1]The Hiberlink project (*http://www.hiberlink.org/*) is supported by the Andrew W. Mellon Foundation. We would like to thank our project partners from EDINA and Los Alamos National Laboratory Research Library for their useful feedback.

[2]http://pdftohtml.sourceforge.net/.

978-1-4799-5569-5/14/$31.00

Table 1: The characteristics and statistics of five large-scale scolarly article collections.

Statistics and Results/Collections [6]	arXiv	Citeseer	PMC	ETDs	Elsevier	all
(a). Subject	physics, statistics	information science	biology, medical	all subjects,	all subjects	-
(b). Publication Period	1997-2013	1994-2012	1997-2012	1997-2012	1997-2013	1994-2013
(c). # of docs	456,049	1,312,134	494,785	87,229	674,789	3,024,986
(d). percentage of docs with links	18.4%	23.6%	29.3%	16.4%	19.9%	22.7%
(e). total # of links extracted	723,326	1,966,739	557,432	121,995	287,061	3,656,553

and has been extensively investigated. Bynens[3] set up a challenge to collect possible regular expressions for matching URLs. To help with testing the regular expressions, he posted a collection of both positive (36) and negative (39) examples, that is, strings that should be accepted as proper URLs or rejected. A total of 12 responses (from each participant) were collected for the challenge and the provided answers range in length (median values 38 to 1,347) and accuracy (0.56 to 1) as measured on a training set. There are also other regexes that are developed in industrial settings (e.g. Twitter[4]) and the regex[5] used in a previous study [2] by the Los Alamos Hiberlink team. We have extensively tested all of these in extracting links in scholarly works.

Test Collection Based Evaluation To measure the performance of different URL extraction systems, our evaluation approach follows the standard procedure for test collection-based evaluation [1]. Firstly we construct ground-truths by asking annotators to manually extract links from the PDF documents. Then we compare the URLs extracted using our module with the ground-truth, using three standard metrics to evaluate system performance, i.e. precision, recall and f-measure [1].

To our knowledge, there is no standard test collection for this task. Therefore, we created one for this purpose. We select arXiv (Table 1) as the basis for our test collection since it is one of the largest collections we have (with half a million scholarly articles) and it covers different disciplines (computer science, physics, astronomy, etc.) and a wide range of publication times (from 1997 to 2013). We believe that it is comprehensive in representing the challenges in URL extraction. Our annotation process was as follows: we randomly sampled 1,000 PDFs from arXiv. Annotators were then instructed to carefully examine the whole PDF document and manually extract all the links inside. They were free to use any search function necessary (e.g. searching the string “http" or “www"). Ultimately, in our test collection, 21.6% (216 out of 1,000) of the scholarly articles were annotated as containing links and in total, 433 links were found.

We show the detailed evaluation results in Table 2. As our baseline we use the link extraction system [2] used by our partners in their previous study. From the results we can observe that some regexes perform better than others and the best one (*Spoon Library*) outperforms the baseline significantly, with respect to both precision and recall. In addition, we found that fixing line breaks and underscore images (*Spoon Library(l)*) in the PDFs helps to further improve the accuracy of link retrieval. The best performing system (*Spoon Library(l)*) performs well with an f-measure score of 0.80, achieving a fairly high precision of 0.83 and recall of 0.78. We use this as our prototype for further link extraction. After conducting error analysis, we observe three areas where our extraction system could be further improved. Firstly, there are link errors in the PDF files arising from mistyping of the links. Secondly, our prototype sometimes fails to recognize the correct "end of URL" (e.g. parenthesis in the URL). Thirdly, some errors are due to line breaks or spaces in links which our processing did not fix in PDFs.

Table 2: Evaluation of Different Link Extraction Systems

System	Precision	Recall	F-measure
Sanderson 2011 [2]	0.53	0.54	0.54
Jeffrey Friedl	0.19	0.18	0.18
mattfarina	0.32	0.36	0.34
krijnhoetmer	0.33	0.38	0.35
gruber	0.43	0.50	0.46
rodneyrehm	0.40	0.61	0.48
gruber_v2	0.45	0.51	0.48
scottgonzales	0.42	0.66	0.51
stephenhay	0.51	0.57	0.54
cowboy	0.51	0.58	0.54
Twitter	0.69	0.63	0.66
imme emosol	0.65	0.70	0.67
diegoperini	0.78	0.76	0.77
Spoon Library	0.80	0.75	0.77
Spoon Library (l)	**0.83**	**0.78**	**0.80**

Link Analysis We apply the best performing link extraction system described above (i.e. our prototype) to five large-scale scholarly collections (shown in Table 1). After URL extraction, we normalize and filter for incorrect extractions and deduplicate the sets of links using the same approach as in the baseline work [2]. The five collections consist of a total of more than three million scholarly works across a long time span and across various subjects. To our knowledge, this is the largest study in this field. From Table 1 (d) to (e), we can observe two important findings: firstly, we found that a large proportion of the scholarly documents (ranging from 16.4% to 29.3%) contain web links and on average over all the collections, there are more than 22.7% of documents containing links. In particular, 29.3% of the PMC collection represents the highest density of links and from a close manual examination, we believe this is because there are more links annotated either by the authors or the publishers. In addition, we can also observe that on average, each scholarly article contains more than one extracted link.

3. CONCLUSIONS

This paper has developed a URL extraction benchmark to accurately extract URLs from scholarly articles and examined the occurrence of web links in the scholarly world through analyzing five vast collections of scholarly literature. We demonstrated that web resources are widely cited in scholarly publications and must therefore be an important concern for digital preservation. Future work includes extending the analysis to quantify the extent to which the referenced web resources can still be accessed.

4. REFERENCES

[1] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[2] R. Sanderson, M. Phillips, and H. Van de Sompel. Analyzing the persistence of referenced web resources with memento. *arXiv preprint arXiv:1105.3459*, 2011.

[3] R. Sanderson, H. Van de Sompel, P. Burnhill, and C. Grover. Hiberlink: Towards time travel for the scholarly web. In DPRMA ’13, pp. 21–21. ACM, 2013.

[3] See http://mathiasbynens.be/demo/url-regex for detailed regular expressions.

[4] https://dev.twitter.com/docs/tco-url-wrapper/how-twitter-wrap-urls.

[5] http://daringfireball.net/2010/07/improved_regex_for_matching_urls.

[6] The source of all the collections used can be found in http://bit.ly/1nMRqCF.