Towards Building a Collection of Web Archiving Research Articles

Brenda Reyes Ayala
Library and Information Science
University of North Texas, Denton, TX 76203
Brenda.Reyes@unt.edu

Cornelia Caragea
Computer Science and Engineering
University of North Texas, Denton, TX 76203
ccaragea@unt.edu

ABSTRACT
The field of Web Archiving exists in a fluid, fragmented, and heterogeneous state. Part of the problem is that this field is relatively new and its literature is scattered across a wide range of journal and conference venues. This makes the state of Web Archiving as a discipline particularly difficult to ascertain. This paper presents an approach to building a collection of articles about the subject. We begin with a small dataset of articles taken from a Web Archiving Bibliography and then proceed to expand it by crawling the Web and collecting additional documents. The crawled documents are then classified using machine learning classification techniques. We show that by extracting the documents' titles and abstracts and representing them using the "bag of words" approach, we are able to accurately identify documents from the Web crawler as documents that are about Web Archiving. We also discuss our results in the context of Web Archiving as an emerging field.

INTRODUCTION
The field of Web Archiving arose to address fears of a Digital Dark Age, caused by the gradual disappearance of digital information (Kuny, 1997). Many institutions to date have implemented Web Archiving programs, a notable example being the Internet Archive, which in 1996 began to capture snapshots of the entire Web with the purpose of preserving them for future generations. Additionally, many national libraries began archiving their own national domains as part of an effort to preserve their digital cultural heritage. University libraries also followed suit, often looking to expand on the strengths of their existing physical collections.

As a nascent field, Web Archiving exists in an uncertain and continuously evolving state. Although there have been several efforts to establish standards and coordinate Web Archiving initiatives, for example, through the foundation of the International Internet Preservation Consortium (IIPC) in 2003, the field remains fluid, fragmented, and heterogeneous, and consequently, so does its literature. Published articles on Web Archives are relatively few compared to older, more established disciplines, and are scattered across a wide range of journals and conferences, including the ACM Web Science Conference (WebSci), the Joint Conference on Digital Libraries (JCDL), and the D-Lib Magazine. For example, a query for the phrase "Web Archiving" in a well-known database such as Web of Science (Thomson Reuters, 2014) returns 27 results, whereas queries for phrases "information retrieval" and "information science" in the same database return 880 and 3,586 results, respectively. Authors who do research in Web Archiving generally do not have official scholarly journals or publication venues, which can provide a sense of the progress or evolution of their field. In short, the state of Web Archiving as a discipline is currently almost impossible to discern. This fact presents a challenge to a researcher interested in understanding the field: *What is the current state of scholarly publication in the field of Web Archiving?*

The current state of a field cannot be ascertained without a corpus of publications in that field that can be examined. To address the above challenge, we pose our main research question: *How do we gather and understand a corpus of Web Archiving research articles, given the scattered nature of the field?* In this paper, we present a process, grounded in information retrieval and machine learning techniques, for gathering a corpus of literature about an emerging field.

RELATED WORK
In the Information Science field, there has been much work done on the subject of exploring and analyzing academic disciplines, usually by making use of bibliometric data. In their prominent study, White and McCain (1998) conducted an extensive domain analysis of the field of Information Science utilizing data from Social Scisearch. They presented a variety of visualizations of the field, such as the most prominent authors, major sub-disciplines, and paradigm shifts over time.

Chen (2006) utilized the Java application CiteSpace II to provide an overview of the trends and patterns in the scientific literature of the research fields of mass extinction and terrorism. More recently, Wang and Tang (2013) mapped the development of the emerging field of open innovation using data from Web of Science and CiteSpace II. We would like to highlight the fact that these research efforts differ from ours in one key factor: the aforementioned authors were working in fields with a strong presence in academic databases and citations indexes. This abundance of bibliometric data and research publications made the task of compiling data and corpus building a substantially less difficult task. This situation is not the case with the field of Web Archiving, and so we

ASIST 2014, November 1-4, 2014, Seattle, WA, USA.

were forced to look for other alternatives of building a research corpus such as employing machine learning techniques for document classification.

Crawling the Web for relevant articles to assemble a dataset seemed like a potentially effective strategy. In the literature, there have been several studies on focused web crawling, a strategy that collects only Web pages that satisfy some specific property, e.g., they belong to a particular topic. Focused crawling first proposed by De Bra et al. is a rich area of research on the Web (Bra et al., 1994), (Junghoo Cho et al., 1998). Chakrabarti et al. (1999) present a discussion on the main components involved in building a focused crawler. Bergmark, Lagoze, and Sbityakov (2002) discuss some of the crawling technologies for building document collections as well as ways to make the crawler highly effective. Batsakis, Petrakis, and Milios (2009) propose state-of-the-art crawlers strategies that use the content of Web pages as well as the link information in order to estimate the relevance of Web pages tied to specific topics. Other works on focused crawling include (Li, Wang, and Du, 2013); (Yang, Kang, and Choi, 2005).

Wu et al. described the evolution of a crawling strategy for CiteSeer^x, which is an academic document search engine (Wu et al., 2012a). CiteSeer^x actively crawls the Web for academic and research documents primarily in Computer and Information Sciences. The authors experimented with using a whitelist (a list of only certain domains that should be crawled) to improve the crawling efficiency of the CiteSeer^x crawler. They found that crawling the whitelist significantly increased the crawl precision by reducing a large amount of irrelevant requests and downloads. In another study, Wu et al. developed a middleware, the Crawl Document Importer (CDI), which selectively imports documents and their associated metadata to the CiteSeer^x crawl repository and database. This middleware provides a universal interface to the crawl database and is designed to support input from multiple open source crawlers and archival formats (Wu et al., 2012b).

Caragea et al. (2014) presented a record linkage approach to building a scholarly big dataset, derived from the CiteSeer^x dataset, which is substantially cleaner than the entire set. More precisely, the authors' approach was to integrate information from an external data source to remove noise in CiteSeer^x that results due to automated techniques used for metadata extraction from Web crawled documents.

In contrast to the above works, we make use of information retrieval and machine learning techniques such as focused crawling and text classification to construct a *scholarly dataset of Web Archiving research articles*. The dataset is available for download to the research community and will particularly be useful to researchers interested in Web Archiving and newcomers to this field. [1]

[1] http://digital.library.unt.edu/ark:/67531/metadc330569/

BUILDING A COLLECTION OF WEB ARCHIVING RESEARCH ARTICLES

In this section, we present our crawling strategy for building a collection of research articles gathered from the Web, that are related to the topic of Web Archiving. We describe the main steps of the crawling process:

1. Compile an initial set of documents related to Web Archiving, which represents the *seed set*.
2. Similarly, compile a set of documents, which are not related to Web Archiving.
3. Train a classifier to accurately discriminate between Web Archiving versus non-Web Archiving documents.
4. Extract the authors from the articles related to Web Archiving in our *seed set* and perform a crawling using these authors' names as well as all their found co-authors as queries that are input to a generic search engine and download other research articles that these authors have published previously.
5. Use the trained classifier in Step 3 to automatically identify the documents related to Web Archiving.

We present further details of these steps in what follows. We start with an initial corpus (our *seed set*) composed of 124 documents about Web Archiving that we extracted from a comprehensive bibliography on the subject (Reyes Ayala, 2013). This bibliography was put together over the course of several months using a variety of methods, such as querying search engines and downloading the publications of prominent authors in the field.

We also gathered a separate corpus of randomly chosen documents from many different disciplines. At the end of this process, we had 124 articles about Web Archiving and 206 randomly chosen articles, for a total of 330 articles. We used a Python library to extract their titles and abstracts. Some documents did not have abstracts, and in such cases, we instead used the document's first 300 words. The motivation for extracting only the title and abstract from a document was that in many cases, documents on the Web are not available as full text, but only as title and abstract. We refer to this set of documents as **Original**.

During the Steps 1 and 2 of the crawling process, we manually labeled these documents using the following labels: documents about Web Archiving were labeled as *positive* or +1, while documents on other topics were labeled as *negative* or -1. Using this labeled dataset, we trained machine learning classifiers to discriminate between the positive and negative documents.

In order to address our research question and discover other documents on the Web that are related to Web Archiving, we employed a *focused crawling* in Step 4. First, from our original small Web Archiving dataset (i.e., our positive *seed set*), we extracted the authors' names and their co-authors. We then crawled the Web for these names in order to extract each author's publications, regardless of its subject of study. We ran several of these crawls, merged the results, and de-duplicated them. The final, merged results from our

crawls contained 3,953 items. We refer to this dataset as **Crawl**. Next, we provide details of our classification task.

Web Archiving Research Paper Identification

We describe our classification task for identifying research articles that are related to the topic of Web Archiving from a collection of documents obtained by crawling the Web. More precisely, our problem can be formulated as follows: given a crawled document, the task is to classify it into one of two classes: Web Archiving articles (the positive class, denoted as +1) and non-Web Archiving articles (the negative class, denoted as -1).

To address this problem, we represented the documents using the commonly used "bag of words" approach for text classification, used in (Mccallum & Nigam, 1998). The "bag of words" approach constructs a vocabulary, which contains all unique words in a collection of documents. A document is then represented as a vector $\underline{x}$ with as many entries as the words in the vocabulary, where an entry i in $\underline{x}$ records the frequency (in the document) of the i^{th} word in the vocabulary, denoted by $\underline{x_i}$. We further represented the documents using *tf-idf* (term frequency-inverse document frequency). The inverse document frequency is given as $\log \frac{N}{df}$. N is the number of documents in the collection, and df is the document frequency of a term in the collection, i.e., the number of documents that contain a particular term. Using these representations, we trained various machine-learning classifiers to classify research papers as Web Archiving or not. These classifiers are Support Vector Machines (SVM), Naïve Bayes Multinomial (NBM) and Logistic Regression (LR) (Bishop, 2006).

Experimental Design

Our experiments are designed around the following research questions:

- What are the units of information (e.g., title, abstract, or both the title and abstract) that most accurately distinguish between documents about Web Archiving and documents about other topics?
- How well do classifiers trained to identify Web Archiving documents perform "in the wild," i.e., on a random sample of documents obtained as a result of a *focused crawling*? More precisely, how well do our classifiers generalize to Web crawled documents?
- What are some of the characteristics of Web Archiving documents obtained by using a focused crawler?

To answer our first question, we extracted the feature representation for each document using three different units of information, the title, the abstract, and both the title and abstract, and trained and compared several classifiers on these feature representations, SVM, NBM and LR. We used the Weka[2] implementation of these classifiers with the default parameters in10-fold cross-validation experiments.

To answer our second question, we evaluated the best resulting classifiers (from the previous experiment) "in the wild." Specifically, by construction, the dataset of 330 examples is fairly balanced, i.e., the number of negative examples is only slightly bigger than the number of positives ones. However, this is not the case in a real-world scenario, where we expect the number of Web Archiving documents to be only a small fraction of the total number of academic documents on the Web. Hence, the performance of a classifier tested using cross-validation on a fairly balanced set would be overestimated. Note that the goal of our previous experiment was to determine the best feature representation and classifier type for our task.

To perform a more realistic evaluation of our classifiers, we randomly sampled a subset of 500 documents directly from the crawl and manually labeled them as positive and negative. We refer to this dataset as **Random**. From this dataset, we extracted the documents' titles and abstracts, and encoded them in the same way as we did for the **Original** dataset. We then ran the same classification experiments using the **Original** dataset for training and the **Random** dataset for testing. Since in our previous experiments the Naïve Bayes classifier yielded the best performance, we used it on the **Random** dataset.

To evaluate the performance of our models, we report the Accuracy and Precision, Recall, and F-score for the positive class, since we are mainly interested in accurately classifying Web Archiving articles. These measures are widely used in Information Retrieval applications.

Finally, to answer our third question, we used the best resulting classifier from the first experiment to predict a label for each of the 3,953 documents obtained from our focused crawler (i.e., the **Crawl** dataset). We extracted the documents' titles and abstracts and encoded them in the same way as before. We characterize the collection in terms of venue popularity, i.e., the venues containing articles on Web Archiving, as well as proficient authors, i.e., authors who published articles in the field of Web Archiving.

RESULTS

The effect of various units of information on the classification performance on the Original dataset.

Table 1 shows the performance of SVM, NBM and LR classifiers on the **Original** dataset, using various information units, i.e., title, abstract, and both title and abstract, to extract the feature representations. As can be seen from the table, using both the title and abstract of documents yields the highest F-score of 0.94 (using NBM) as compared with the settings where we use only the title or only the abstract, that achieve an F-score of 0.91 and 0.79, respectively (also using NBM). NBM achieved the highest performance compared with SVM and LR in terms of F-score and accuracy, although the other classifiers performed well, often with fairly high precision and recall.

[2] http://www.cs.waikato.ac.nz/ml/weka/

Feature/Method	Classifier	Prec	Re	F-score	Acc. (%)
Title/BoW	SVM	0.78	0.73	0.75	84.42
	NB	0.86	0.73	0.79	85.45
	LR	0.90	0.60	0.72	82.42
Abstract/BoW	SVM	0.86	0.82	0.84	87.87
	NB	0.93	0.90	0.91	93.63
	LR	0.92	0.78	0.85	89.39
Title & Abstract (T&A) /BoW	SVM	0.75	**0.96**	0.84	86.67
	NB	0.92	0.94	**0.94**	**95.15**
	LR	**0.95**	0.83	0.89	92.12

Table 1: Results on the Original Dataset.

The performance of the Naïve Bayes Multinomial on the Random dataset

Table 2 shows the performance of the NBM classifier on the **Random** dataset, using both the titles and abstract for each document. T & A in the table means that a document's feature representation is extracted from both its title and abstract, and *tf* stands for term frequency. As can be seen from the table, the accuracy is fairly high (94.4%). However, the precision is very low, although the recall is not very bad. This could be explained by the very small number of positive examples in our 500 random sample.

	Classifier	Prec.	Re.	F-score	Acc.
T&A/*tf*	NBM	0.18	0.75	0.30	94.40%

Table 2. Results on the Random Dataset.

The performance of the Naïve Bayes Multinomial on the Crawl dataset.

Table 3 shows the results of the NBM classifier trained on the **Original** dataset that was used to predict a label for each document in the entire crawled collection of 3,953 documents, i.e., the **Crawl** dataset. As can be seen from the table, there are 216 documents identified as being about Web Archiving. In total, we have 340 (124+216) Web Archiving documents in our built collection.

	Classifier	Docs	WA Docs	Non WA Docs
T&A/*tf*	NBM	3953	216	3737

Table 3: Results on the Crawl Dataset.

An early look at Web Archiving as a Discipline

We took a closer look at this built collection of 340 documents to understand its nature. Specifically, we identified who are the authors with the large number of publications in this set as well as what are the venues with the most number of publication related to Web Archiving. In Tables 4 and 5, we present a list of the top authors, as well as the top venues (conference, journal, magazine, etc.), respectively, as extracted from our Web Archiving document collection. As we can seen from our list of top authors in the field, Web Archiving has a decidedly international and inter-disciplinary character.

Authors come from the United States, Germany, and Denmark, and while some are faculty members in academic institutions, most carry out their work within research institutions. Most authors are situated within the Computer Science discipline, though there are some from Digital Humanities, and Library and Information Sciences.

The venues that tended to publish articles on Web Archiving were mostly in the field of Library and Information Science, such as D-Lib Magazine, the JCDL conference, and International Conference on Preservation of Digital Objects. This Library Science trend contrasts with our above list, which contains mostly Computer Scientists. Though at first this might seem surprising, it seems consistent with White's (2010) assertion that an increasing number of Computer Scientists are making more and more contributions to the field of Library and Information Science.

Rank	Venue Name
1	International Web Archiving Workshop(IWAW) (now discontinued)
2	D-Lib Magazine
3	Joint conference on Digital libraries(JCDL)
4	International Conference on Preservation of Digital Objects(iPres)
5	New Review of Hypermedia and Multimedia
6	The International Journal of Digital Curation
7	The International Federation of Library Associations and Institutions (IFLA) Journal
8	Liber Quarterly
9	Communications of the ACM
10	Library Trends

Table 4. Top Venues for Web Archiving Publications.

CONCLUSIONS AND FUTURE WORK

We presented an approach to building a collection of research articles on the topic of Web Archiving. In particular, we started with a *seed set* of manually annotated Web Archiving research articles and used *focused crawling* based on authors' names extracted from the *seed set* to enlarge our collection.

Name	Discipline	Institution
Nelson, Michael L.	Computer Science	Old Dominion University
Spaniol, Marc	Computer Science	Max-Planck-InstitutfürInformatik
Weikum, Gerhard	Computer Science	Max-Planck-InstitutfürInformatik
McCown, Frank	Computer Science	Harding University
AlSum, Ahmed	Computer Science	Old Dominion University
Sanderson, Robert	Information Science	Los Alamos National Laboratory
Herbert van de Sompel	Library Science/Computer Science	Los Alamos National Laboratory
Brügger, Niels	Digital Humanities	Aarhus University
Marshall, Catherine C.	Digital Humanities	Microsoft Research
Mazeika, Arturas	Computer Science	Max-Planck-InstitutfürInformatik

Table 5. Top Authors in Web Archiving Publications.

In the future, we plan to investigate and explore further ways to enlarge our corpus of articles about Web Archiving. One possible future direction would be to extract the bibliographic references from each of the Web Archiving articles we have collected, since these are likely to also cover the same subject. We could perform a further focused crawler using these documents' titles as queries input to a generic or scholarly search engine. Additionally, further improvement of classification performance would be another interesting direction to pursue along with investigating the use of content and link analysis for improved crawling techniques.

ACKNOWLEDGMENTS

We are grateful to Nancy Reis, Mark Phillips, Cathy Hartman, Helen Hockx-Yu, Herbert van de Sompel, Clement Oury, Peter Stirling, and Ahmed AlSum for their help in constructing the bibliography. This research was supported in part by NSF award #1353418 to Cornelia Caragea. The International Internet Preservation Consortium (IIPC) also provided doctoral support to Brenda Reyes Ayala.

REFERENCES

Batsakis, S., Petrakis, E. G. M. & Milios, E. E. (2009). Improving the performance of focused web crawlers. *Journal of Data Knowledge & Engineering*, 68, 1001-1013.

Bergmark, D., Lagoze, C. & Sbityakov, A. (2002). Focused crawls, tunneling, and digital libraries. In M. Agosti& C. Thanos (Eds.), *ECDL '02 Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries* (pp. 91-106).

Bishop, C. M. (2006).*Pattern recognition and machine learning*. New York: Springer.

Caragea, C., Wu, J., Ciobanu, A., Williams, K., Fernandez-Ramirez, J., Chen, H., Wu, Z., & Giles, C.L. (2014). CiteSeerX: A scholarly big dataset. *Proceedings of the 36th European Conference on Information Retrieval* (pp. 311-322).

Chakrabarti, S., Van den Berg, M., & Dom, B. (1999). Focused crawling: a new approach to topic-specific web resource discovery. *Proceedings of the Eighth International Conference on World Wide Web* (pp. 1623-1640).

Chen, C. (2006).CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science*, *57*(3), 359–77.

Cho, J., Garcia-Molina, H., & Page, L. (1998). Efficient crawling through url ordering. *Proceedings of the seventh international conference on World Wide Web 7*(pp. 161-172).

De Bra, P., Houben, G., & Post, R. (1994). Information retrieval in distributed hypertexts. *Proceedings of the 4th RIAO Conference* (pp. 481–491).

McCallum, A. & Nigam, K. (1998). *A comparison of event models for Naïve Bayes text classification*. Paper presented at ICML/AAAI-98Workshop on Learning for Text Categorization, Madison, WI.

Li, Y., Wang, Y. & Du, J. (2013). E-FFC: An enhanced form-focused crawler for domain-specific deep web databases..*Journal of Intelligent Information Systems*, *40*, 159-184.

Kuny, T. (1997). *A digital dark ages? challenges in the preservation of electronic information*. Presented at the 63rd IFLA (Intl. Federation of Library Associations and Institutions) Council and General Conference, Copenhagen. http://archive.ifla.org/IV/ifla63/63kuny1.pdf

Reyes Ayala, B. (2013). *Web Archiving Bibliography 2013*.UNT DigitalLibrary.http://digital.library.unt.edu/ark:/67531/metadc172362/

Thomson Reuters (2014). Web of Science.[Online database]. Retrieved from http://library.unt.edu

Wang W., & Tang, J. (2013). Mapping development of open innovation visually and quantitatively: A method of bibliometrics analysis.*Asian Social Science, 9*(11), 254.

White, H.D. & McCain, K. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. *Journal of the American Society for Information Science, 49*(4), 327–355.

White, H.D.(2010) Bibliometric overview of information science. InM.J. Bates & M.N Maack (Eds.), *Encyclopedia of library and information sciences* (3rd ed., pp.534 - 545).

Wu, J., Teregowda, P. B., Ramírez, J. P. F., Mitra, P., Zheng, S. & Giles, C. L. (2012a). The evolution of a crawling strategy for an academic document search engine: whitelists and blacklists. In N. S. Contractor, B. Uzzi, M. W. Macy & W. Nejdl (Eds.), *Proceedings of the 3rd Annual ACM Web Science Conference*(pp. 340-343).

Wu, J., Teregowda, P. B., Khabsa, M., Carman, S., Jordan, D., Wandelmer, J. S. P., Lu, X., Mitra, P. & Giles, C. L. (2012b). Web crawler middleware for search engine digital libraries: a case study for citeseerX.. In G. H. L. Fletcher & P. Mitra (Eds.), *Proceedings of the Twelfth Intl. Workshop on Web Information and Data* (pp. 57-64).

Yang, J., Kang, J. & Choi, J. (2005). A focused crawler with document segmentation.. In M. Gallagher, J. M. Hogan & F. Maire (Eds.), *IDEAL'05 Proceedings of the 6th International Conference on Intelligent Data Engineering and Automated Learning* (pp. 94-101