Automatically Detecting References from the Scholarly Literature to Records in Archives

Tokinori Suzuki[1][0000−0002−4715−6198], Douglas W. Oard[2][0000−0002−1696−0407], Emi Ishita[1][000−0002−1398−8906], and Yoichi Tomiura[1]

[1] Kyushu University, Fukuoka, Japan
[2] University of Maryland, College Park, MD, USA
tokinori@inf.kyushu-u.ac.jp

Abstract. Scholars use references in books and articles to materials found in archives as one way of finding those materials, but present systems for archival access do not exploit that information. To change that, the first step is to find archival references in the scholarly literature; that is the focus of this paper. Several classifier designs are compared using a few thousand manually annotated footnotes and endnotes assembled from a large set of open access papers on history. The results indicate that fairly high recall and precision can be achieved.

Keywords: Archival references · Classification · Test collection.

1 Introduction

Scholars refer to existing materials to support claims in their scholarly work. Citation to books, journal articles, and conference papers can be detected by automated systems and used as a basis for search or bibliometrics (e.g., using citation databases such as Google Scholar, Web of Science, or Scopus). However, there are no comparable databases for citation to the rare, and often unique, unpublished materials in archival repositories. Our goal in this paper is to begin to change that by automating the process of detecting scholarly citation to materials in an archive. We call such citations Archival References (AR).

Our work is motivated by the task of discovering archival content. A recent survey of users of 12 U.S. archival aggregators (e.g., ArchiveGrid, or the Online Archive of California) found that there were a broad range of users for such search services [26]. One limitation of archival aggregation, however, is that it presently relies on sharing metadata that is manually constructed by individual repositories. In the long run, we aim to augment that with descriptions mined from the written text that authors use to cite the archival resources on which they have relied. To do that at scale, we must first automate the process of finding citations that contain archival references. That is our focus in this paper.

Prior studies suggest that that very substantial numbers of archival references exist to be found [3, 18, 22]. As an example, Bronstad [3] manually coded citations in 136 books on history, finding 895 (averaging 6.6 per book) citing archival

Table 1. Examples of citations containing archival references. "Strict" citations contain only archival references; "About" citations also include other accompanying text.

Strict	About	Citation
✓		Roosevelt to Secretary of War, June 3, 1939, Roosevelt Papers, O.F. 268, Box 10; unsigned memorandum, Jan. 6, 1940, ibid., Box 11.
	✓	Wheeler, D., and R. García-Herrera, 2008: Ships' logbooks in climatological research: Reflections and prospects. Ann. New York Acad. Sci., 1146, 1-15, doi:10.1196/annals.1446.006. Several archive sources have been used in the preparation of this paper, including the following: Logbook of HMS Richmond. The U.K. National Archives. ADM/51/3949

repositories. HathiTrust, for example, includes more than 6.5 million open access publications, so we would expect to find millions more archival references there.

As the examples in Table 1 illustrate, archival references differ in important ways from references to published content. Most obviously, conventions used to cite unpublished materials differ from those used to cite published materials [1, 25]. The elements of an archival reference (e.g., repository name, box and folder) are different from the elements for published sources (e.g., journal name, volume and pages). It is also common for archival references to include free-form explanatory text within the same footnote or endnote [7].

In this paper, we aim to begin the process of assembling large collections of archival references by building systems capable of automatically detecting them at large scale. To do this, we have collected documents, automatically detected footnotes and endnotes, annotated some of those "citations" for use in training and evaluation, and compared several classifiers. Our results indicate that automatically detecting archival references is tractable, with levels of recall and precision that we expect would be useful in practical applications.

2 Related Work

Studies of scholars who make use of archival repositories indicate that references in the scholarly literature are among the most useful ways of initially finding the repositories in which they will look. For example, Tibbo reported in 2003 that 98% of historians followed leads found in published literature [24], and Marsh, et al. found in 2023 that of anthropologists 73% did so [16]. There is thus good reason to believe that archival references could be useful as a basis for search. While expert scholars may already know where to look for what they need, search tools that mimic that expert behavior could be useful to novices and itinerant users, who comprise the majority of users in the survey mentioned above [26].

Researchers interested in information behavior and in the use of archives have long looked to citations as a source of evidence, but such studies have almost invariably relied on manual coding of relatively small sets of such references [4, 8, 10–13, 17, 18, 22, 23]. In recent years, rule-based techniques have been applied to detect archival references [2, 3], but we are not aware of any cases in which trained classifiers that rely on supervised machine learning have yet been built.

3 Methods

Here we describe how we assemble documents, find citations in those documents, and decide which citations contain archival references.

3.1 Crawling Documents and Extracting Citations

Our first challenge is to find documents that might include citations that contain archival references. Since we know that historians cite archival sources, we chose to focus on papers in history. We therefore crawled papers with a discipline label of History and a rights label of Open Access by using the public Semantic Scholar API.[1] That API requires one or more query terms. To get a set of query terms, we collected the abstracts of the 2,000 most highly cited papers from Scopus that were published in 2021 with a discipline label of Arts and Humanities. We collected terms in those abstracts sorted in the order of their frequency. Then we issued those terms one at a time to Semantic Scholar, and retrieved PDF files. After repeating this process for some number of keywords, we merged the resulting sets of PDF files. Most of our experiments were run on the 1,204 unique documents that resulted from using the most frequent 5 keywords (the KW5 document set), but we also conducted some experiments with the roughly 13,000 unique documents from using the most frequent 14 keywords (KW14).

We then parsed the documents using GROBID [15], an open-source toolkit for text extraction from academic papers. In the KW5 document set, GROBID found at least one footnote or reference (i.e., at least one citation) in 690 documents. In KW14, GROBID found at least one citation in 5,067 documents.

3.2 Detecting Archival References

For this paper, we built three types of classifiers to detect archival references.

Rule-Based (RB) Classifier. Our RB classifier has a single rule: IF a citation includes any of the strings "Box","Folder", "Series", "Fond", "Container", "Index", "index", "Manuscript", "manuscript", "Collection", "collection", "Library", "library", "Archive", or "archive" THEN it contains an archival reference. Regular expression matching is done without tokenization, lowercasing, or stemming. This is similar to an approach used by Bronstad [3] to search the full text of papers for mentions of repositories. We selected our terms after examining the results from Subset 1 (described below).

Repository Name (RN) Classifier. Our RN classifier looked for one of 25,000 U.S. repository names from the RepoData list [9]. However, across all our experiments RN found only one match that had not also matched a RB classifier term. We did use RN to guide sampling, but we omit RN results for space.

Support Vector Machine (SVM) Classifiers. We experimented with three SVM variants. all using radial basis function kernels, which we found to be better than linear kernels in early experiments. In "SVMterm" the features are

[1] https://www.semanticscholar.org/product/api

frequencies of terms found in citations. Specifically, we tokenized every citation on whitespace or punctuation and removed stopwords. Our tokenizer does not split URLs, so URLs are processed as single term. For our other SVMs, we tokenized each citation using NLTK, used a lookup table to select the pretrained GloVe embedding for each term [20], and then performed mean pooling to create a single embedding per citation. We experimented with both 50 (SVM50) and 300 dimensions (SVM300). We report results for SVM300, which were better than SVM50 with larger training sets. For each SVM we swept C from 1 to 100 by 5 and used the value (20) that gave the best results. We set the gamma for the radial basis function to the inverse of the number of feature set dimensions (e.g., 1/300 for SVM300).

3.3 Sampling Citations for Annotation

We drew five samples from KW5 and one from KW14. One approach ("by document") was to randomly order the documents and then sample citations in their order of occurrence. The other ("by citation") was to randomly order all citations regardless of their source document, and then sample some number of citations from the head of that list. Subsets are numbered in order of their creation. Focusing first on the 59,261 documents in KW5, random selection for Subsets 1 and 6 found 45 archival references among 3,500 sampled citations, a prevalence of 1.3%. This skewed distribution would make it expensive to find enough positive examples for supervised learning, so we turned to system-guided sampling. We merged positive classification results from our RB and RN classifiers to create Subset 2, annotating the first 600 citations (randomized by document). We then trained SVM50 on Subset 2 and used it to guide our draw of Subset 3, manually annotating all 760 citations (randomized by document). To create Subset 4, we first randomly selected and annotated 1,000 of the 59,261 citations (randomized by citation) and then added 259 citations that RB or RN classified as archival references. GROBID found 346,529 citations in the KW14 document set. We randomly sampled 20,000 of those and ran four classifiers on that sample: RB, RN, SVM300, and a BERT classifier (that did not perform well, the description of which we omit for space reasons). We merged and deduplicated positive results from those classifiers, resulting in 880 citations. We call that Subset 5.

3.4 Annotation Criteria and Annotation Process

Our annotation goal was to label whether extracted citations are archival references using two criteria: "Strict" if it included one or more archival references, with no other text; or "About" if it included one or more archival references together with explanatory text. Table 1 shows examples. The first has two archival references, and nothing else, satisfying our Strict criterion. The second has one archival reference and some explanatory text, satisfying our About criterion.

Annotation was done by two annotators. Annotator A1, the first author of this paper (a computer scientist) annotated Subsets 1 through 4 and Subset 6. Before performing any annotation, he examined the citation practice in 207 pages

of endnotes from three published books in history [5, 21, 27] and from one journal article in history [19]. A1's initial annotations of Subsets 1 and 2 were reviewed by the second author of this paper (an iSchool faculty member). A1 reannotated subsets 1 and 2 and then annotated subsets 3, 4, and 6. For time reasons, Subsets 3 and 4 were annotated only by the Strict criterion. Annotation requires some degree of interpretation, so additional research was conducted using Google when necessary (e.g., to see if some unfamiliar word might be a repository name).

Subset 5 was assessed by annotator A2, a Library Science Ph.D. student studying archives. We trained A2 in three phases. First, we demonstrated how to judge whether a citation is an archival reference (by either criterion) using 50 examples from Subset 4. Then A2 annotated 50 more citations from KW14 with the same criteria prevalence. The first three authors then met with A2 to discuss their annotations, and then A2 coded 120 more citations from KW14 We computed Cohen's Kappa [6] between A1 and A2 on those 120 citations as 0.80 (substantial agreement, according to Landis and Koch [14]). Finally, A2 annotated the 880 citations in Subset 5. All our annotations are on GitHub.[2]

4 Results

As measures of effectiveness we report Precision (P), Recall (R) and F_1. Table 2 shows results with Strict+About training. We used two approaches to choosing training and test data. In one, we used separate training and test sets. Because of distributional differences between the training and test sets, this yields conservative estimates for the Recall and Precision that could be obtained in practice with more careful attention to that factor. To avoid distributional differences, we also experimented with training and testing on same subset(s), using five-fold cross-validation. Cross-validation yields somewhat optimistic estimates for Recall and Precision, since that approach eliminates systematic differences in the decisions made by different annotators, and it entirely removes all differences between the distributional characteristics of the training and test sets. Considering results from the two approaches together thus allows us to characterize the range of Precision, Recall and F_1 values that we might expect to see in practice.

Focusing first on the Eval S+A block in Table 2, we see that detecting archival references is not hard. The RB classifier achieves excellent Recall with no training at all, although its Precision is quite poor. Among SVMs, SVM300 does best in every case by F_1. It seems that distributional differences are adversely affecting Recall and Precision when the training and test sets differ (although 95% confidence intervals are about ± 0.2 on the low-prevalence Subset 6 test set). From the Eval: S and Eval: A blocks of Table 2,we see that a classifier with both S and A annotations for training is much better at finding S than A.

As Table 3 shows, removing A from training doesn't help to find more S. Compare, for example, Recall in the second set of experiments in both Tables 2 and 3, both of which were trained and tested on Subset 5. There, training with S+A correctly found more S annotations than did training with only S.

[2] https://github.com/tokinori8/archive-citation-collection

Table 2. Results for classifiers trained with both Strict (S) and About (A) annotations as positive examples. P=Precision, R=Recall, best F_1 bold. Train or test, with number of positive S and A annotations (after removal of any training citations from the test set). Top block: detecting all citations containing archival references; subsequent blocks: same classifiers evaluated only on citations with S or A annotations.

	Train: 1+2 (114S, 22A) Test: Cross-Validation			Train: 5 (243S, 18A) Test: Cross-Validation			Train: 5 (243S, 18A) Test: 6 (7S, 21A)		
Eval: S+A	P	R	F_1	P	R	F_1	P	R	F_1
RB	0.24	1.00	0.39	0.30	1.00	0.47	0.22	0.65	0.33
SVMterm	0.99	0.68	0.80	0.92	0.74	0.82	0.30	0.50	0.38
SVM300	0.94	0.79	**0.85**	0.86	0.81	**0.83**	0.50	0.50	**0.50**
Eval: S	P	R	F_1	P	R	F_1	P	R	F_1
RB	0.20	1.00	0.33	0.28	1.00	0.44	0.11	0.86	0.19
SVMterm	0.92	0.75	0.82	0.91	0.78	**0.83**	0.17	0.64	**0.27**
SVM300	0.86	0.80	**0.83**	0.81	0.81	0.81	0.16	0.64	0.25
Eval: A	P	R	F_1	P	R	F_1	P	R	F_1
RB	0.03	1.00	0.06	0.02	1.00	0.04	0.11	0.50	0.18
SVMterm	0.06	0.31	0.10	0.03	0.56	0.05	0.15	0.50	0.23
SVM300	0.08	0.45	**0.14**	0.05	0.72	**0.08**	0.34	0.45	**0.39**

Table 3. Results for detecting Strict (S) annotations by classifiers trained on only Strict annotations as positive examples. Notation as in Table 2.

	Train: 1+3+4 (110S) Test:Cross-Validation			Train: 5 (243S) Test:Cross-Validation			Train: 5 (243S) Test: 6 (7S)			Train:1+3 (54S) Test: 4 (56S)		
Eval: S	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
RB	0.31	0.71	0.43	0.29	1.00	0.45	0.11	1.00	0.20	0.28	0.63	**0.38**
SVMterm	0.69	0.35	0.46	0.91	0.72	**0.80**	0.20	0.64	0.30	0.06	0.95	0.11
SVM300	0.82	0.58	**0.68**	0.87	0.75	**0.80**	0.36	0.57	**0.44**	0.07	0.95	0.13

5 Conclusion and Future Work

We have shown that archival references can be detected fairly reliably, with F_1 values between 0.5 and 0.83, depending on how well the training and test sets are matched. We have also developed and shared collections that can be used to train and evaluate such systems. Annotator agreement indicates that our Strict and About criteria for characterizing archival references are well defined and replicable. Most archival references satisfy our Strict criterion, and unsurprisingly it is Strict classification decisions where we do best. Experiments with separate training and test sets point to potential challenges from systematic differences in prevalence that result from sampling differences. This work is thus a starting point from which second-generation collections might be built with even better control over prevalence matching between training and test sets, and more robust classification results might be achieved using classifier ensembles. Given our promising results for this archival reference detection task, our next step will be to develop algorithms to segment individual archival references, and then to extract specific elements (e.g., repository name or container).

Acknowledgments This work was supported by JSPS KAKENHI Grant Number JP23KK0005.

References

1. American Psychological Association, et al.: Publication Manual of the American Psychological Association. American Psychological Association (2022)
2. Borrego, Á.: Measuring the impact of digital heritage collections using Google Scholar. Information Technology and Libraries **39**(2) (2020)
3. Bronstad, K.: References to archival materials in scholarly history monographs. Qualitative and Quantitative Methods in Libraries **6**(2), 247–254 (2019)
4. Brubaker, J.: Primary materials used by Illinois state history researchers. Illinois Libraries **85**(3), 4–8 (2005)
5. Carlson, E.: Joe Rochefort's War: The Odyssey of the Codebreaker Who Outwitted Yamamoto at Midway. The Naval Institute Press (2013)
6. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20**(1), 37–46 (1960)
7. David-Fox, M., Holquist, P., Martin, A.M.: Citing the archival revolution. Kritika: Explorations in Russian and Eurasian History **8**(2), 227–230 (2007)
8. Elliott, C.A.: Citation patterns and documentation for the history of science: some methodological considerations. The American Archivist **44**(2), 131–142 (1981)
9. Goldman, B., Tansey, E.M., Ray, W.: US archival repository location data (September 2022), website https://osf.io/cft8r/, visited January 17, 2023
10. Heinzkill, R.: Characteristics of references in selected scholarly English literary journals. The Library Quarterly **50**(3), 352–365 (1980)
11. Hitchcock, E.R.: Materials used in the research of state history: A citation analysis of the 1986 Tennessee Historical Quarterly. Collection Building **10**(1/2), 52–54 (1990)
12. Hurt, J.A.: Characteristics of Kansas history sources: A citation analysis of the Kansas Historical Quarterly. Ph.D. thesis, Emporia Kansas State College (1975)
13. Jones, C., Chapman, M., Woods, P.C.: The characteristics of the literature used by historians. Journal of Librarianship **4**(3), 137–156 (1972)
14. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977)
15. Lopez, P., et al.: GROBID: Generation of bibliographic data. Open source software, https://github.com/kermitt2/grobid, visited February 6, 2023 (2023)
16. Marsh, D.E., St. Andre, S., Wagner, T., Bell, J.A.: Attitudes and uses of archival materials among science-based anthropologists. Archival Science pp. 1–25 (2023)
17. McAnally, A.M.: Characteristics of materials used in research in United States history. Ph.D. thesis, University of Chicago (1951)
18. Miller, F.: Use, appraisal, and research: A case study of social history. The American Archivist **49**(4), 371–392 (1986)
19. Neufeld, M.J., Charles, J.B.: Practicing for space underwater: Inventing neutral buoyancy training, 1963–1968. Endeavour **39**(3-4), 147–159 (2015)
20. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014)
21. Prange, G.W.: At Dawn We Slept. The Untold Story of Pearl Harbor. Penguin Books (1991)
22. Sherriff, G.: Information use in history research: A citation analysis of master's level theses. portal: Libraries and the Academy **10**(2), 165–183 (2010)
23. Sinn, D.: The use context of digital archival collections: Mapping with historical research topics and the content of digital archival collections. Preservation, Digital Technology & Culture **42**(2), 73–86 (2013)

24. Tibbo, H.: Primarily History in America: How U.S. Historians Search for Primary Materials at the Dawn of the Digital Age. The American Archivist **66**(1), 9–50 (2003)
25. University of Chicago Press Editorial Staff: The Chicago Manual of Style. University of Chicago Press (2017)
26. Weber, C.S., Connaway, L., Doyle, B., Langa, L.A., Proffitt, M., Washburn, B., Carbajal, I.A.: Summary of research: Findings from the building a national finding aid network project. Tech. rep., OCLC (2003). https://doi.org/10.25333/7a4c-0r03
27. Yokoi, K.: Global Evolution of the Aircraft Industry and Military Air Power. Nihon Keizai Hyoronsha Ltd. (2016), in Japanese