Searching Unseen Sources for Historical Information: Evaluation Design for the NTCIR-18 SUSHI Pilot Task

Douglas W. Oard[1], Tokinori Suzuki[2], Emi Ishita[2] and Noriko Kando[3]

[1]University of Maryland, College Park, MD USA
[2]Kyushu University, Fukuoka, Japan
[3]National Institute of Informatics, Tokyo, Japan

Abstract

In evaluation of ranked retrieval, the usual assumption is that the documents to be searched can be indexed before the query is received and the search is performed. The NTCIR-18 SUSHI Pilot Task, by contrast, models the case in which only a small sample of the documents to be searched can be indexed before the query is received. This task model arises in the context of searching within large archives of paper documents, for example. The stark difference in what can be indexed before the query is received has consequences for both task design and evaluation design, both of which are discussed in this paper.

Keywords

Information retrieval, Archival access, Evaluation

1. Introduction

Information retrieval has generally been modeled on the idea of the library catalog. We have some collection of materials, we can index those materials in some way, and then in response to a query we can suggest the materials that the searcher might want to see. Archives,[1] however, are different. Archives collect unique historical materials, and because those materials are unique, and thus irreplaceable, archives typically must place much greater emphasis on acquisition and preservation than on access. Access is not ignored, but it must be done within stringent resource constraints. It is thus not practical to describe (or digitize) many of the individual docuements in an archival collection. Instead, archivists typically describe collections at higher levels of aggregation, such as folders, boxes, or segments of the collection that go by names such as record groups, series, or fonds.

This situation creates challenges for searchers, in part because different parts of an archive often are arranged differently. This happens because archivists can economize on the effort needed to arrange the materials in a collection by taking advantage of the original order in which the materials were organized when they were in active use [1]. For example, materials on space exploration might be originally arranged by program (Mercury, Gemini, Apollo, Shuttle, ...), while materials on diplomacy might have been organized by country (Sweden, Uganda, Japan, ...). Further down in the organization, we might find the diplomacy materials organized by topic (agriculture, education, economy, ...), while the space exploration materials might be organized by function (design, testing, contract management, ...). The Swedish agriculture records might then be further organized by author (Kissinger, Smith, Kennan, ...), whereas the Apollo design materials might be organized by component (space suit, thruster, radio, ...). And so on all the way down. Well, actually not all the way down, since the description process simply must stop before getting to the level of individual documents. After all, the U.K. National Archives has

EMTCIR '24: The First Workshop on Evaluation Methodologies, Testbeds and Community for Information Access Research, December 12, 2024, Tokyo, Japan
oard@umd.edu (D. W. Oard); tokinori@inf.kyushu-u.ac.jp (T. Suzuki); ishita.emi.982@m.kyushu-u.ac.jp (E. Ishita); kando@nii.ac.jp (N. Kando)
0000–0002–1696–0407 (D. W. Oard); 0000–0002–4715–6198 (T. Suzuki); 000–0002–1398–8906 (E. Ishita); 0000-0002-2133-0215 (N. Kando)

[1]The term "archive" is often used in computer science to mean a collection (e.g., a zip archive). In this paper, by contrast, we use "archive" to name a type of information institution. According to the Merriam Webster dictionary, an archive is "a place in which public records or historical materials (such as documents) are preserved."

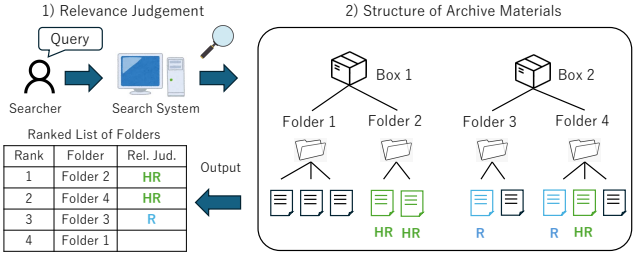

Figure 1: Creation of relevance judgments for folders based on the max judgments for judged documents in that folder. R: Relevant, HR: Highly Relevant.

about 14 billion printed pages, which if put in a single stack would stretch halfway around the world. Nobody, and indeed no group of 100 people, could ever hope to look at all of that, much less to describe all of it at the level of individual documents. But even that is just a tiny part of the problem; there are about 26,000 archives in the United States alone [2], many with nowhere near the level of funding (relative to the size of their collections) as the U.K. or U.S. National Archives. Simply put, nobody will ever see all of this stuff.

The net result of this situation is to shift a greater burden onto those who want to find things in an archive. Searchers must learn where collections that might contain what they want are stored, they must know how those collections are organized, they must (given the paucity of digitization) then travel to an archive, request access to containers (e.g., boxes) that might contain what they are looking for, and then look through those materials. This is a process in which time is measured in weeks, and costs might be measured in hundreds or even thousands of U.S. dollars. Per search.

Our goal in the NTCIR-18 SUSHI[2] Pilot Task[3] is to begin to reduce the time and expense of finding materials in an archive. Our focus is on materials that are on physical media (e.g., paper or microfilm), and on materials that, like the vast majority of most archives (e.g., like 97.6% of the U.S. National Archives) have not yet been digitized, or even described at the level of individual documents. We seek to do that by supporting the development of automated systems that can learn from a very limited number of examples that have been digitized, and from whatever metadata at higher levels of aggregation might exist, to suggest where in a collection a searcher might most productively look. SUSHI is designed to support that research by developing test collections that model the real problem, and that do so in a way that supports insightful and affordable evaluation.

2. The Folder Ranking Task

The principal task in the NTCIR-18 SUSHI Pilot Task is Folder Ranking. Figure 1 displays the input (a query) and the output (a ranked list of folders) for this task. Specifically, given a query and an unsorted list of all folders in the collection, use the metadata describing each of those folders, together with a sparse sample of digitized documents and document-level metadata from some documents in some of those folders, rank the folders a searcher might most want to see. Table 1 illustrates some of the

[2]Searching Unseen Sources for Historical Information
[3]https://sites.google.com/view/ntcir-sushi-task/

NARA File Name	Record Group	Series Name	Box Name	Folder Name	Title
pol18braz_10-08-65_rio753.pdf	General Records of the Department of State	Subject-Numeric Files	1941	POL 18 BRAZ 04/01/1965	Telegram from Rio to State Concerning Negrao de Lima Interview in O Globo (Rio 753): 10/8/1965
pol26braz_10-09-63_rio787.pdf	General Records of the Department of State	Subject-Numeric Files	3838	POL 26 BRAZ 02/01/1963	Memo from Ralph J. Burton to Cleo McNelly Concerning The Student Christian Movement in New England:
pol7braz_08-05-71_rio5556_xr.pdf	General Records of the Department of State	Subject-Numeric Files	2130	POL 7 BRAZ 01/01/1971	Telegram from State to Rio Concerning Geneva for Stevenson (Rio 5556): 8/5/1971

Table 1
Examples of some folder-level and document-level metadata in our Folder Ranking Task test collection.

metadata that is available in our Folder Ranking Task test collection. The sparse digitized sample in our dry run test collection includes five documents per box, with one document sampled from each of the five largest folders in each box (although the 21 boxes with fewer than five folders have more than one sample drawn from some folders). For actual final Folder Ranking Task, we plan to explore a mixture of even sampling (the same number of folders per box) and uneven sampling (with more samples from some boxes than from others). Both approaches can be useful. For early experiments drawing the same number of samples from each box can help achieve better control over experimental conditions. In real archives, however, digitization and description are both unevenly applied, and uneven sampling can help to characterize the additional challenges that such a situation produces.

2.1. Evaluation for the NTCIR-18 Pilot Task

In order to simulate the real task, we have assembled a collection that we can easily sample and that we can easily judge for relevance. This is a collection of 31,682 U.S. State Department documents from 1,337 folders in 124 boxes. What makes the collection easily judged is that it is fully digitized, and that we have topical metadata for every document. Table 1 shows examples of that metadata, which (like the documents) is from the the US National Archives.

Since a system under test will know which folders exist, we operationalize the idea of "searching well" as ranking those folders well. We measure how well a system constructs that ranking using NDCG@5 as the principal evaluation measure. A cutoff at 5 corresponds to about a half hour's work by someone who is actually looking at physical documents in an archive. We estimate this from the fact that a folder contains an average of $31682/1337 \approx 24$ documents, together with out expectation that a skilled searcher could recognize a relevant document in 15 seconds (skilled users of archives flip through documents very quickly). Obtaining the boxes that contain those folders might take another hour or two, but requesting new boxes could be interleaved with examining results from prior requests.

In the NTCIR-18 SUSHI Pilot Task we have two ways of getting the topics on which queries are based. The first approach, used for the dry run, was to randomly select a "query document" that participating systems did not see at training time, and then to use the title metadata for that document as the query.[4] We then treat any document with the same title metadata (from anywhere in the collection, not just from the known query document) as being relevant. This is an extended variant of known-item retrieval Note that participating systems can't see those titles (because systems can only see document-level metadata for training documents, and we don't treat training documents as relevant). So systems must perform some sort of inference in order to rank folders without ever having seen a single one of the relevant documents anywhere in the collection [3, 4].

The approach we used to create the dry run test collection can be useful as a basis for initial system development, but exact matching on document titles is at best a weak proxy for true human relevance judgments. We therefore need a second approach in which actual people make those judgments. For our final evaluation, we plan to rely instead on human assessors, preferably graduate students with a background in history or library science. The assessors will initially create search topics in the traditional title/description/narrative format,[5] based on their understanding of the collection's content. They will then check to see if at least a few relevant documents actually exist, using a full-collection document-level search system that we have built using PyTerrier. With this system, assessors can issue queries (either free form, or copied form a topic field), rank documents using that query based on one of several ways of indexing the collection (e.g., OCR-only, Title-only, or both),[6] skim the folder label and title for every document in the resulting ranked list, selectively view PDF scans of individual documents, search within a document for any term, and record their tentative relevance judgments for

[4]This is actually a bit oversimplified. From initial experiments, we learned that we also need to limit the length of the title, because some titles are so long as to be unrealistic surrogates for a human-issued query. We therefore control the query length by first assembling all titles in the collection, making separate randomly-ordered lists for all unique 2, 3, 4, and 5-word titles, and then having two annotators each select 25 of each that look to them like realistic queries.

[5]In our dry run collection we use the same topic format, but for the dry run all three topic fields are identical.

[6]Other ways include using folder labels to expand document text, or using GPT summaries of OCR text.

any documents that they encounter during the topic development process. Once they have finalized a topic, we will save their tentative relevance judgments so that they can later finalize those judgments when performing relevance assessment.

The relevance assessment process will then be performed in the same way, using the same system, but with enough time allocated for more careful searching, a process known as interactive search and judgment [5]. During relevance assessment we will also ask the assessors to review the tentative relevance judgments that they had created during topic development. We separate the topic development and relevance judgment processes both for convenience (we need the topics sooner) and to encourage assessors to treat the topics as well defined and immutable during the relevance assessment process.

In this first year of the task we don't plan to use pooling to build assessment pools because participating systems will have ranked lists of folders, but relevance judgments are made not on folders but on individual documents (documents which the participating systems never saw, and thus could not have ranked). In future evaluations we may consider assessment processes that could benefit from folder pooling (e.g., allocating some assessor time to searching pooled folders more thoroughly). For the NTCIR-18 SUSHI Pilot Task we won't use folder pooling as a part of our assessment process, but we will look at what folder pooling would have produced in order to see if the density of relevant folders is markedly higher than random selection, and if it is we might employ folder poling in the future.

Because the systems to be evaluated produced ranked lists of folders, we must map our document-level relevance judgments to folder-level relevance judgments in some way. As Figure 1 illustrates, for the NTCIR-18 SUSHI Pilot Task we will aggregate document-level judgments to folder-level judgments by simply using as the folder's judgment the highest judgment for any judged document in that folder. The resulting relevance judgments can then be used directly to compute folder-level nDCG@5, or binarized to compute, for example, Mean Average Precision (MAP).

2.2. Future Evaluation Design Issues

That's as far as we expect to be able to get for the NTCIR-18 SUSHI Pilot Task, but the task design raises several other important evaluation design issues. Here we highlight three of those questions.

First, our use of nDCG@5 simplifies the goal perhaps more than we might like. All evaluation involves model building, and all models are simplifications of reality [6]. But we might productively complexify our evaluation measure in two ways. First, we might switch from a one-and-done measurement approach to one based on the density of relevant documents in a folder. In our present approach, systems get no more credit for finding a folder with five relevant documents than for finding a folder with just one. It is probably more realistic to have some extra credit for highly ranking a folder with a larger number of relevant documents, and perhaps to get somewhat more credit for finding folders with fewer documents that have to be looked through (for any given number of relevant documents in the folder). A cost model based on the discovery rate of relevant documents would be one formulation that could address both factors. For this, we might also look for inspiration to the evaluation measures that were designed for the INEX Retrieval In Context task, where the time required to examine ranked elements from hierarchically structured content was the focus (in that case, the time required to examine ranked passages that had been extracted from documents) [7].[7]

The situation is, however, not really even that simple. In the U.S. National Archives, for example, searchers request access not to folders, but to the boxes that contain the folders they want to see. All else equal, we would therefore prefer to find highly ranked folders that happen to be in the same box (or in nearby boxes, since for practical reasons archivists are often equally happy to fetch a short sequence of boxes that are stored together). There's probably no end to how much we could complexify this (e.g., how about constructing the shortest path through an archive to pick up some set of boxes that contain folders that together contain some given number of relevant documents?). We are not yet ready to commit to a new measure, but we are able to see that we likely will ultimately want one.

Second, confidence intervals and testing for significant differences is a bit more complex in this environment than in a typical ranked retrieval evaluation. The reason for this is that we need to account

[7]Thanks to an anonymous reviewer for this suggestion!

not just for random variations from the choice of query, but also random variations from our choice of the training set. In our dry run we can see the effect of the training set because we run half our queries with one randomly sampled training set and half with another. But when we compute confidence intervals, we ignore the training set variation and (for convenience) compute the confidence intervals only over the queries. In the future we will likely want to use somethign along the lnes of an ANOVA in an effort to tease apart topic and trainign set effects [8].

Third, our present approach is vulnerable to the common criticism of classic information retrieval test collections that they are not typically designed to characterize cross-collection differences. This may be a minor sin when we might expect BM25 or BERT to work about as well on one English news collection as another, but in SUSHI we are seeking to model a real situation in which different collections can have vastly different metadata structures. Because just searching the folder metadata is an obvious baseline against which to compare, we need to pay attention to these differences in what metadata is available. And, of course, real archival collections are not all equally amenable to OCR—there are handwritten collections, photograph collections, collections written entirely in hieroglyphics or cuneiform, and (in one memorable case) a collection that consisted of nothing but x-ray film containing images of fish skeletons. We're not going to be able to explore that entire space of possible collections in finite time, but that's not the key concern here. Rather, the question is how best to explore *any* of it.

To see why that canbe challenging, it may help to articulate what a collection must have. First, it must have content for which we can create topics and for which we can perform relevance judgments. So the fish skeletons are off the table. Second, we must know at least which box contained each item (e.g., each document), and we would love to also know which folder in that box contained each item. Third, we really want to have scanned images of all the documents. This third one seems non-negotiable – we tried going over to the U.S. National Archives and doing relevance judgments for the top-ranked box for one query. It was a 10 minute drive from our office, but once we got there it took three hours just to get the box. So doing large numbers of relevance judgments by requesting and then examining paper doesn't seem like a scaleable solution.

It is not hard to find reasonably large collections of digitized materials, and it is not hard to find large collections that have good box and folder metadata. But it is harder than you might expect to find both together (and even harder if you initially prefer to omit handwritten materials and photographs). Because most archives do not yet make both content and metadata available through an API, building relationships with institutions that have something close to what we need will be key to gaining access to the collections we need, and for getting approval to share those collections broadly with other researchers.

3. The Archival Reference Detection Task

The sizes of our training sets in the Folder Ranking Task are designed to model the sparsity of existing document-level metadata in real collections. For example, sampling an average of 5 documents per box closely approximates the actual fraction of the U.S. National Archives collection that has document-level description. One way of improving the potential for inferring where to look for materials that might match a query is to increase the number of documents for which document-level metadata is available. That is the goal of the Archival Reference Detection Task. Given a text of endnote or footnote, the task is to determine whether that footnote or endnote contains any references to archival materials.

The key insight that motivates our Archival Reference Detection Task is that when scholars cite materials from archives in their published work, that creates an additional source of document-level description that we could use in search tasks as well. We can use these descriptions in two ways. Most directly, we can parse the archival reference to extract document-level metadata such as a document title and location information (e.g., which archive, which series, which box, and perhaps even which folder). For example:

- 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

In the first example, we can see the collection name "Roosevelt Papers," document descriptions, and some box numbers. As the second example illustrates, scholars sometimes also package descriptive text together with an archival reference in the same footnote or endnote. When present, that could potentially serve as a useful free-form document-level description of some identifiable document in an archive. We could also potentially use the content at and near the point where the corresponding citation was made in the main body of a paper as a free-form description not only of what the cited document contains, but also of how that document's content might be useful (in at least one context).

In our early work on detecting archival references in papers on History [9], we found that 45 of 3,500 automatically extracted footnotes or endnotes were references to archival materials, a prevalence of about 1.3%. From this we can estimate that if we wish to collect 10,000 archival references, we would need to run Archival Reference Detection on about a million documents. We thus chose one million documents as our target collection size for the Archival Reference Detection Task. This is a classification task in which systems are asked to return a binary decision indicating whether each footnote or endnote includes an archival reference. For footnotes or endnotes that are classified by a system as an archival references, a system can optionally also elect to extract the archival reference (i.e., to segment the archival reference from any descriptive text that the author had packaged with it).

3.1. Evaluation for the NTCIR-18 Pilot Task

We provided a dry run test collection for the Archival Reference Detection Task that we had manually annotated for the presence or absence of archival references in our previous research. That collection contains 1,836 footnotes or endnotes from open-access papers in the field of History that we obtained using the Semantic Scholar API.[8] Annotations were performed by two annotators, one of whom was a Ph.D. student with expertise in the use of cultural heritage materials. Cohen's Kappa for agreement on a subset of that collection, with one of us as the second annotator, was 0.8 [9], which Landis and Koch would characterize as substantial agreement [10]. From this we conclude that the manual annotation task is tractable, at least at small scale.

We therefore began a larger crawl of Semantic Scholar for use in the Archival Reference Detection Task. Because this is a highly unbalanced binary classification task, we will evaluate participating systems on that larger collection using the F_1 measure (the harmonic mean of precision and recall). To control evaluation costs, we will compute F_1 using a stratified sample. Our stratification will be based on the number of participating systems that classified each footnote or endnote as an archival reference. For example, we will sample footnotes or endnotes that are classified as archival references by every participating system most densely, and we will very sparsely sample footnotes or endnotes that are not classified as archival references by any system. We plan to evaluate the optional task of segregating an archival reference from any text that is packaged with it in the same footnote or endnote using the Jaccard coefficient. We will compute this Jaccard coefficient on characters, dividing the number of characters that both the system and the annotator believe are in the archival reference by the number of characters that either the system or the annotator believe are in the archival reference (i.e., the intersection over the union).

To create these annotations, we plan to hire annotators who are graduate students in history or in some related discipline. Because the scholarly papers from which we extract footnotes and endnotes will be in English, we will further require reading fluency in English. Based on our earlier experience with human annotation, we expect that (after training) annotators will be able to classify about two footnotes or endnotes per minute. We thus expect each annotator to produce about 1,000 annotations per week. We will therefore design our sampling to select about 5,000 footnotes or endnotes for annotation. We

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

will then subsample from the annotations marked as archival references by an annotator and ask that annotators further segment the archival reference from any text that is packed with it in the same footnote or endnote. In our experience, such packaging is relatively infrequent, occurring perhaps 10% of the time, so we expect this second annotation process to go fairly quickly. Overall, we expect that annotation will require about one month, although to guard against unexpected delays we perform assessment in batches that are all sampled in a way that would allow estimation (with broader confidence intervals) even if annotation of some of the later batches can not be completed in the available time.

3.2. Future Evaluation Design Issues

Our goal in the Archival Reference Detection Task is to begin the process of finding and using archival references by first finding them. In future editions of the task, we can then extend the goal to include not just classification and segmentation, but also extraction of specific fields (such as title, archive and box) and extraction of associated descriptive text from the main body of the paper that cited this archival reference. Once that has been done, we could progress to extrinsic evaluation, measuring the benefits to an actual search task of having a broader set of document-level metadata (and other forms of document-level description) on which to base its inference. This will ultimately require new test collections for the search task, since our present Folder Ranking Task test collection draws content from too narrow a subset of the full archival universe to be useful for evaluating the impact of footnotes or endnotes that could reference anything in any archive anywhere.

Our initial experience with archival reference detection points to two other potential challenges. One is that our present approach to stratified sampling is better suited to characterizing what has been found than it is to characterizing what has been missed. This is a natural consequence of the class skew in the classification task. In future shared task evaluations, we might want to consider the use of active learning as a way of better exploring the range of cases that all participating systems are missing [11].

A second challenge is that materials in some archives are quite clearly receiving more attention in the scholarly literature than materials in other archives. For example, we see anecdotally that materials in the U.K. National Archives have been cited much more often (in the small sets that we have examined to date) than are materials in the U.S. National Archives, despite the holdings of those two institutions being of similar sizes. More careful study of this skewed distribution seems to be called for, and of course we can expect that the large-scale results from the Archival Reference Detection Task could serve as one useful basis for such a study.

4. Building a Research Community

Several years ago, one of us wrote a thought piece in which we identified factors that might lead to a decline in the demand for shared task information retrieval evaluations [12]. Without rehashing the complete argument, the basic idea was that many of the important affordances of shared task evaluations can now also be achieved in other ways, and that some of those way have advantages in cost, friction, lead time, or scalability. There was, however, one exception to that prognostication, and that was the value of shared tasks for building research communities. With that in mind, we are pleased to now have six registered teams (as of late October).[9] Looking toward the future, we also know some people who are working on other aspects of archival access, and we are familiar with an earlier metadata-focused cultural heritage task run at CLEF,[10] We've tried to spread the word in both of those communities. We have also tried to lower barriers to staring on the SUSHI task by making baseline systems available that potential participants can easily modify without having to develop code anew for the rather complex data handling that is needed to fully specify the training and test conditions in the Folder Ranking Task.

[9]We also note, however, that two of those teams include task organizers
[10]http://ims.dei.unipd.it/data/chic/

5. Conclusion

In this paper we have described the design of the NTCIR-18 SUSHI pilot task, and we have identified some new evaluation questions that emerge from that work. Although SUSHI is already an NTCIR-18 Pilot Task, there are still many issues around evaluation design, community building, and scaling up the size of our test collection that we feel could benefit from further discussion at this workshop.

Acknowledgements

This work has been supported in part by the Japan Society for the Promotion of Science KAKENHI Grant Number 23KK0005 and the National Institute of Informatics Open Collaborative Research 2024 (24S0505).

References

[1] G. Wiedeman, The historical hazards of finding aids, The American Archivist 82 (2019) 381–420.
[2] B. Goldman, E. M. Tansey, W. Ray, US archival repository location data, 2023. Website https://osf.io/cft8r/, visited October 3, 2024.
[3] D. W. Oard, Known by the company it keeps: Proximity-based indexing for physical content in archival repositories, in: International Conference on Theory and Practice of Digital Libraries, 2023, pp. 17–30.
[4] T. Suzuki, D. W. Oard, E. Ishita, Y. Tomiura, Searching for physical documents in archival repositories, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, pp. 2614–2618.
[5] M. Sanderson, H. Joho, Forming test collections with no system pooling, in: Proceedings of the 27th annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2004, pp. 33–40.
[6] G. E. Box, Science and statistics, Journal of the American Statistical Association 71 (1976) 791–799.
[7] P. Arvola, J. Kekäläinen, M. Junkkari, Expected reading effort in focused retrieval evaluation, Information Retrieval 13 (2010) 460–484.
[8] N. Ferro, M. Sanderson, How do you test a test? a multifaceted examination of significance tests, in: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, 2022, pp. 280–288.
[9] T. Suzuki, D. W. Oard, E. Ishita, Y. Tomiura, Automatically detecting references from the scholarly literature to records in archives, in: Proceedings of the 25th International Conference on Asia-Pacific Digital Libraries, Springer, 2023, pp. 100–107.
[10] J. Landis, G. Koch, The measurement of observer agreement for categorical data, Biometrics (1977).
[11] G. V. Cormack, M. R. Grossman, Evaluation of machine-learning protocols for technology-assisted review in electronic discovery, in: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, 2014, pp. 153–162.
[12] D. W. Oard, The future of information retrieval evaluation: NTCIR's legacy of research impact, in: Evaluating Information Retrieval and Access Tasks, Springer, 2021.