On the Need for and Provision for an 'IDEAL' Scholarly Information Retrieval Test Collection

Birger Larsen[1] and Christina Lioma[2]

[1] Dept. Communication, University of Aalborg Copenhagen, Denmark
[2] Dept. Computer Science, University of Copenhagen, Denmark

birger@hum.aau.dk, c.lioma@di.ku.dk

Abstract. Much about the potential, strengths and challenges of applying bibliometric techniques to scholarly Information Retrieval (IR) can be learned from conducting user studies with scholars and scientists, e.g. when interacting with bibliometrically enhanced IR prototypes or operational systems. However, most of IR research is still carried out as laboratory studies using test collections. As there is a lack of appropriate scholarly test collections, we argue in this position paper that one or more well-crafted test collections with scholarly documents is needed in order to engage the IR community whole-heartedly in Bibliometric IR research, and to facilitate foundational and high-quality work in this area. Based on the experiences gained from creating the iSearch test collection (Lykke et al. 2010 - http://itlab.dbit.dk/~isearch/) we reflect on the properties of an ideal scholarly IR test collection and then examine three possible ways of realising the creation of such a collection. The first considers the possibilities for basing the test collection on the ever more readily available open access collections of scholarly documents; the second examines the possibilities of setting up alliances with proprietary content producers or aggregators; and the third considers collaboration with major university libraries and their content.

Keywords. Scholarly IR; Test Collections; Citation-Based Search

1 Introduction

From the birth of Information Retrieval (IR) research in the Cranfield experiments (Cleverdon, Mills & Keen, 1966), IR has been engaged in improving access to scholarly information. The foundational Cranfield experiments were carried out using scholarly information, and only later the focus shifted to news, government and web documents, in particular after the start of the Text REtrieval Conferences (TREC) in 1992. The Bibliometric IR (BIR) workshops and activities in the Digital Libraries community show a continued interest in scholarly IR, but on a small scale and with relatively few researchers involved. At the same time there is a plethora of ideas on how to improve scholarly IR, e.g. using various citation-based measures and links for back-end retrieval and front-end interaction and navigation. Some of this research has resulted in bibliometrically enhanced IR prototypes or operational systems (e.g. CiteSeerX[1]). Much about the potential, strengths and challenges of applying bibliometric techniques to scholarly IR can be learned from conducting user studies with scholars and scientists when interacting with such systems. However, scholarly IR is still not a task that engages IR researchers widely. We argue in this position paper that one or more well-crafted test collections with scholarly documents is needed in order to engage the IR community whole-heartedly in BIR research, and to facilitate foundational and high-quality work in this area.

The notion of test collections was established with the Cranfield II experiments. A test collection consists of documents, queries (i.e. information needs; also called 'topics'), and relevance assessments of the documents in relation to the queries (Cleverdon, Mills & Keen, 1966)[2]. Test collection-based research has several important limitations (see e.g. Borlund, 1998) but still has prime importance in IR research for a number of reasons: First, test collections facilitate studying the relative merits and performance of different IR methods by providing a common experimental framework within which all can be tested. Second, test collections are re-usable and enable cost-effective testing of many hypotheses without further involvement of human assessors once the collections are created. Third, test collections are often shared across research groups creating standard benchmarks that can aid in progressing beyond state-of-the-art. In domains where no test collections are available, IR research is hampered because the back-end methods cannot be easily optimised before being subjected to user testing and user studies. Also, different methods cannot be easily compared because of the lack of a common benchmark, and thus their relative merits are hard to assess.

Currently, research in scholarly IR is based on relatively small and purpose-built test collections. Ritchie and colleagues for instance, custom-built a test collection of 9084 documents with extracted citation information from the area of linguistics (Ritchie, 2008). This is in stark contrast to current test collections e.g., in TREC that include tens or hundreds of millions of documents (Collins-Thompson et al., 2014). In addition to Ritchie et al.'s work and the original Cranfield test collections, a number of test collections have been based on scholarly documents over the years (see Table 1 below for an overview). To our knowledge the largest publicly available scholarly test collection with citation information is the iSearch collection[3] (Lykke et al. 2010) with 434,817 full text physics documents and 3.7 million internal citations from arxiv.org, 65 topics and relevance assessments. Despite being the largest publically available scholarly test collection, iSearch is a small test collection for TREC standards. The problem with such small test collections is that research results may not scale well to larger datasets or other domains, and there is a risk that properties of the collection affect results heavily.

[1] http://citeseerx.ist.psu.edu/index

[2] In this paper we refer to collections of documents as '*test collections*'. We refer to collections of documents without topics and/or relevance assessments as '*document collections*'.

[3] http://itlab.dbit.dk/~isearch/

Table 1. Overview of 1) existing scholarly test collections with details on the number and nature of documents, number of topics, and if references are included as well as a citation index, and 2) document collections without topics and relevance assessments that might be used in future scholarly test collections.

Collection	#documents	#topics	references	citations
Cranfield II Cleverdon, Mills & Keen (1966)	1,400 (no full text)	225	yes	yes
CACM Fox (1983)	3,204 (no full text)	63	(yes)	yes
Cystic Fibrosis Shaw et al. (1991)	1,239 (no full text)	100	yes	yes

INEX 2002-2005
Gövert & Kazai (2003);
Fuhr et al. (2004);
Malik et al. (2005; 2006)
	12,107 (2002-2004)
	16,819 (2005)
	(full text XML)
		24 (2002)
		36 (2003)
		39 (2004)
		47 (2005)
			(in XML)	no

TREC Genomics Track skynet.ohsu.edu/trec-gen/	525,938 (2003) 4.5 million (2004-2005) (no full text 2003-2005) 162,259 (2006-2007) (full text HTML)	50 (2003) 50 (2004) 50 (2005) 28 (2006) 50 (2007)	(Some can be extracted through PubMed?)	no
ACL (Ritchie 2008)	9084 (full text + citation contexts)	82	yes	yes
iSearch (Lykke et al., 2010)	434,817 (PDF/LaTex from arXiv.org)	65	12+ million	3.7 million
CORE core.ac.uk	25 million (full text PDFs?)	no	(in PDFs?)	no
CiteSeerX csxstatic.ist.psu.edu/about/data	6 million (full text PDFs)	no	yes	yes
PubMed OA www.ncbi.nlm.nih.gov/pmc/tools/openftlist/	1,1 million (full text PDFs)	no	(in PDFs)	(through PubMed)
Rexa www.rexa.info	380,000 (full text) 6 million (non full text)	no	yes?	yes?
BioMedCentral old.biomedcentral.com/about/datamining	278,502 (full text XML)	no	(in XML)	no
ACL-ARC (Bird et al., 2007)	10,921 (full text PDFs)	no	yes	38,767

We therefore argue that one or more well-crafted test collections with scholarly documents is needed in order to progress significantly and to interest and engage a wider range of IR researchers. As the review below shows, current test collections are insufficient for this. There is significant progress to be made in the area of bibliometrically enhanced IR in terms of models, methods and algorithms, and high quality scholarly test collections can facilitate work on these. The time for this is ripe: there is growing interest in bibliometrically enhanced IR research and development; many new possibilities offered by advances in Machine Learning and Natural language Processing can be ported to BIR tasks; and large amounts of scholarly documents are becoming available in formats that can be used to build larger and better scholarly test collections. Based on the motivation above and our experiences from creating the iSearch test collection, we address the following two questions:

1) What are the ideal properties of a scholarly test collection, and
2) How could scholarly IR test collections be created practically?

2 Challenges and ideal properties of scholarly test collections

Test Collection Size & Span
Challenge: Even though the production of scholarly documents is increasing, the total number of scholarly publications is small relative to documents appearing on the web e.g. in social media or news portals. This practically means that very large scholarly test collections (e.g. with 10s or 100s of millions of documents) are unlikely to be realised in the near future. **Ideal Property:** We posit that an ideal scholarly test collection should contain a *minimum critical mass* of documents, which is reasonably large to attract IR researchers used to working on large collections. In addition, the time span of the documents should cover at least a decade of publications so that citations can accumulate over time.

Domain-Specific Dissemination & Retrieval
Challenge: Different academic fields and specialities may have widely differing publication, citation and retrieval practices. This means that BIR methods may not always be transferable across domains. For instance, the prime dissemination venue can be journals in some domains and conferences in others; similarly, citation aging is shorter in some domains (where advances become outdated very quickly, e.g. computer science) than others. This affects the IR practices across domains too, introducing an added complexity in retrieval tasks across domains. **Ideal Property:** Documents from several different domains should be included (to capture and study different publication and citation behaviour), and at the same time sufficient amounts of documents from each domain must be included. Clearly specified user models focussing on a single or a few tasks should be defined and form the basis of the test collection.

Realistic queries and relevance assessments
Challenge: Creating realistic queries and relevance assessments would require the collaboration of active scientists. These are however notoriously busy and might be hard

to engage, even if funding were available to pay them for their time. It is also unlikely that assessors with sufficient skills could be crowdsourced. Finally, because both the queries and the scholarly documents may be much more complex, scholarly assessors are likely to need more time to compose queries and assess the relevance of each document – compared to e.g. news or web documents. **Ideal Property:** Active scientists should be involved in creating queries and in assessing pooled documents for relevance – and be given ample time to do so. Their behaviour should inform user models focussing on a single or a few tasks.

Pooling

Challenge: Text-based retrieval models are likely not to identify many of the documents that can be retrieved by citation-based retrieval approaches (because they exploit the citation network rather text-based features). Therefore the pooling (where documents to be assessed for relevance are selected) needs to include both text-based and citation-based retrieval runs to ensure that the test collection can be used for both. **Ideal Property:** Pooling should include a wide range of text-based IR models as well as several types of citation-based retrieval approaches. The pool depth may therefore need to be somewhat larger than in other test collections.

Format

Challenge: Random crawls of scholarly documents may not be sufficiently topically focussed to ensure that there are enough citations among the documents. The reason for this is that many central documents in the network are not available on the open web, e.g. scholarly publications from commercial publishers. In addition, obtaining large amounts of scholarly documents in a useful format consistent across all documents is a major challenge. **Ideal Property:** In order to fully allow the exploration of bibliometrically enhanced IR, the full text of the documents should be available (e.g. for citation context analysis) as well as the bibliographic references. Preferably, consistent document metadata should be available as well as the references in parsed form and matched into a citation index (i.e. listing any internal links among documents in the collection).

3 Three ways of creating academic test collections

3.1 Basing test collections on Open Access document collections

Many interesting scholarly publications are published by professional organisations or commercial publishers who may be reluctant to make a large number of these available in full text for research purposes in a test collection. Even though open access publishing is increasing it is unlikely that many of the core publications of any given field can be included in a test collection. A major challenge in creating an academic test collection is obtaining a sufficiently large number of documents in full text and with reference and citation information. As shown in Table 1 several Open Access collections have now passed the 1 million mark and it is worth considering basing academic test collections on some of these. Below we discuss how scholarly test collections can be developed from the following Open Access initiatives.

The **CORE** (COnnecting REpositories) service aggregates "…open access research outputs from repositories and journals worldwide and make them available to the public."[4] CORE attempts to extract metadata and the reference list of each publication. CORE is already active in the Digital Libraries community organising a workshop series on 'Mining Scientific Publications' and regularly releases large data dumps of harvested documents. The current collection includes more than 25 million documents and is one of the largest publicly available academic document collections. The datasets are distributed by CORE for download on their website[5]. Two main challenges are involved in creating an academic test collection based on a CORE data dump: 1) *Creation of topics and relevance assessments*. Ideally active researchers need to be recruited and commit time to create topics and carry out relevance assessments. Given some coordination, this could be organised by the researchers who are interested in working with the test collection. This was the approach taken in INEX where the researchers themselves created and assessed topics in their own area or recruited people from their network to do so. CORE has the advantage that it covers many areas and thus it might be easier to find researchers willing to participate. The organisation could take place under one of the IR evaluation campaigns, e.g. TREC, CLEF, NTCIR, FIRE etc. if an academic track was accepted by one of these. Topic creation and relevance assessment can be handled remotely as in INEX (Piwowarski, Trotman & Lalmas, 2008). 2) *Extracting references and matching citations*. As the CORE content is very heterogeneous, consistent and reliable reference and citation extraction will be a challenge. Some techniques exist for achieving this (e.g. as used in CiteSeerX), but due to the many potential links and the very scale of this heterogeneous collection this may be a major computational challenge which preferably should be undertaken before topic creation is begun. CORE includes content from several of the other collections listed in Table 1, including CiteSeerX and PubMed OA.

CiteSeerX has at present crawled more than 6 million documents and regularly releases datasets via Amason S3. Topically CiteSeerX is more focussed with an emphasis on computer science. This may make it harder to recruit topic authors, but has the advantage that many IR researchers could be involved in topic creation making the process discussed for CORE viable also for CiteSeerX. CiteSeerX has the advantage that references have already been extracted and matched into citations, and also offers additional resources like citing citation contexts[6].

PubMed OA has at present more than 1.1 million full text OA PDFs from different medical sub-disciplines. These come with high quality metadata from PubMed and has an API access to extracted references and any citing documents within the dataset. A major challenge with PubMed OA might be to recruit active medical researchers willing to create and assess topics. PubMed OA spams many different medical specialities which on one hand can be an advantage (providing diversity as discussed above), and on the other hand might turn out to be a problem (if there is not enough critical mass in

[4] https://core.ac.uk/about

[5] https://core.ac.uk/intro/data_dumps

[6] http://csxstatic.ist.psu.edu/about

any given subfield to ensure enough relevant documents and a coherent citation network).

3.2 Alliances with publishers or content holders

As noted in Section 2, many interesting and central scholarly publications are published by commercial publishers or professional organisations, who may not be willing to let the publications they control be part of scholarly test collections to be freely shared among researchers. To realise scholarly test collections with such content it would be necessary to form alliances with such publishers or organisations. As access to the full text of the documents is likely to be a stumbling block maybe a cloud-based solution similar to the one set up for the VICERAL project may be feasible (Hanbury et al., 2013). Here “the algorithms are brought to the data” and retrieval runs executed in a secure cloud where the IR researchers cannot access the original documents. While this will lead to some restrictions, clearly documented datasets and procedures as well as possibilities for executing own code on the documents can still facilitate interesting work. The solution proposed and tested in the VICERAL project also has a number of advantages in terms of reproducibility and smaller requirements for local processing and storage capacity.

Another group of content holders with vast collections of scholarly documents include academic search engines (e.g. Google Scholar and Microsoft Academic Search) and academic social networking sites (e.g. Medeley and ResearchGate). These may also be reluctant to provide direct access to the publications they have crawled or that have been uploaded by their users, and a similar cloud-based solution may also work in relation to these.

If such alliances can be made topic authors and assessors still need to be recruited – probably along similar lines as discussed above in Section 3.1. These will need secure access to the full text of the document to construct their topics and do relevance assessments.

3.3 Collaboration with major university libraries

An alternative to the approaches outlined above would be to involve major university and academic libraries in creating scholarly test collections. Academic libraries have vast collections of scholarly documents and their main role is to provide access to these documents. They may be interested in contributing actively in creating scholarly test collections in order to improve their services. The proposal below is based on the following premise: while any particular major academic library will have some documents that are unique, access to most of the documents will also be offered by other major academic libraries. Therefore a group of, say, 5-10 large academic libraries could collaborate on recruiting topics authors and assessors among their patrons. Each library would recruit a manageable number of active researchers, maybe from among those that ask for help in identifying literature for their research. The researchers would create a topic based on their current research, and pooling would consist of librarians searching each of their own library collections to identify potentially relevant documents. The resulting documents for each topic would be pooled across libraries and sent for assessment to the topic author. A disadvantage of this approach is that only bibliographic records and not the full text of the documents can be distributed (due to copyright restrictions) along with the topics and relevance assessments. Instead this would have to be accessed through an academic library. Note that not all libraries will have all the assessed documents. This may not be a serious disadvantage if the overlap between libraries turns out to be large, and if the pools are deep and contain a stratified sample across participating libraries.

The advantage of the proposal is that it draws on the active participation of academic libraries and their patrons – who both may have a direct interest in the outcomes.

4 Conclusion

Test collections have long been driving research and innovation in many areas of mainstream but also specialised IR. Bibliometric IR (BIR) has not benefitted yet from organised initiatives to create test collections targeted to its needs. We argue that there is a need to do so and we outline several challenges of such an initiative. We conclude with three practical suggestions of moving forward, namely by considering open access services, alliances with publishers, and collaborations with university libraries.

5 References

1. Lykke, M., Larsen, B., Lund, H. & Ingwersen, P. (2010). *Developing a test collection for the evaluation of integrated search*. ECIR, p. 627-630.
2. Cleverdon, C.W, Mills, J. & Keen, M. (1966): *Factors determining the performance of indexing systems; Volume 1: Design*. The College of Aeronautics, Cranfield, England, 1966.
3. Borlund, P. (2003): The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research,* 8(3), paper no. 152.
4. Ritchie, A. (2008): Citation Context Analysis for Information Retrieval. University of Cambridge. (PhD thesis - http://research.microsoft.com/apps/pubs/default.aspx?id=73161)
5. Collins-Thompson, K., Macdonald, C., Bennett, P., Diaz, F., Voorhees, E. (2014): TREC 2014 Web Track Overview. In: *The Twenty-Third Text REtrieval Conference (TREC 2014) Proceedings*. (NIST Special Publication: SP 500-308).
6. Fox, E. (1983): *Characteristics of Two New Experimental Collections in Computer and Information Science Containing Textual and Bibliographic Concepts.* Cornell University: Computing Science Department (Technical Report TR 83-561).
7. Shaw, W. I. M., Wood, J. B., Wood, R. E., & Tibbo, H. R. (1991): The cystic fibrosis database: Content and research opportunities. *Library and Information Science Research*, *13*(4), 347-366.
8. Gövert, N. & Kazai, G. (2003): Overview of the INitiative for the Evaluation of XML retrieval (INEX) 2002. In: Proceedings of INEX 2002.
9. Fuhr, N., Malk, S. & Lalmas, M. (2004): Overview of the INitiative for the Evaluation of XML Retrieval (INEX) 2003. In: Proceedings of INEX 2003.

10. Malik, S., Lalmas, F. & Lalmas, M. (2005): Overview of INEX 2004. In: *Advances in XML Information Retrieval: Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004, Dagstuhl Castle, Germany, December 6-8, 2004, Revised Selected Papers*. Belin: Springer, p. 1-15.
11. Malik, S., Kazai, G. Lalmas, M. & Fuhr, N. (2006): Overview of INEX 2005. In: *Advances in XML Information Retrieval and Evaluation: 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005, Dagstuhl Castle, Germany, November 28-30, 2005. Revised Selected Papers*. Berlin: Springer, p. 1-15.
12. Bird, S., Dale, R., Dorr, B., Gibson, B., Joseph, M., Kan, M.-Y., Lee, D., Powley, B., Radev, R. & Tan, Y. F. (2008): The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In:. Marrakesh, Morocco.
13. Piwowarski, B., Trotman, A., Lalmas, L. (2008): Sound and complete relevance assessment for XML retrieval. *ACM Transactions on Information Systems, 27(1),* 37 pages.
14. Hanbury, A., Müller, H., Langs, G. & Menze, B. H. (2013): Cloud–Based Evaluation Framework for Big Data. In: The Future Internet: Future Internet Assembly 2013: Validated Results and New Horizons. Berlin: Springer, p. 104-114.