Developing a Test Collection for the Evaluation of Integrated Search

Lykke, Marianne; Larsen, Birger; Lund, Haakon; Ingwersen, Peter

Published in:
Advances in Information Retrieval

DOI:
10.1007/978-3-642-12275-0_63

Publication date:
2010

Document Version
Early version, also known as pre-print

Developing a Test Collection for the Evaluation of Integrated Search

Marianne Lykke, Birger Larsen, Haakon Lund and Peter Ingwersen

Royal School of Library and Information Science, Department of Information Interaction and Information Architecture. Birketinget 6, DK-2300 Copenhagen S, Denmark
{mln, blar, hl, pi}@db.dk

Abstract. The poster discusses the characteristics needed in an information retrieval (IR) test collection to facilitate the evaluation of *integrated search*, i.e. search across a range of different sources but with one search box and one ranked result list, and describes and analyses a new test collection constructed for this purpose. The test collection consists of approx. 18,000 monographic records, 160,000 papers and journal articles in PDF and 275,000 abstracts with a varied set of metadata and vocabularies from the physics domain, 65 topics based on real work tasks and corresponding graded relevance assessments. The test collection may be used for systems- as well as user-oriented evaluation.

Keywords: Test collection design; Task-based IR; Integrated search

1 Introduction

As digital libraries offer access to increasingly large and diverse information sources there is recently a move from federated search, where a range of different sources are searched and the results presented for each source, to *integrated search* which instead presents the retrieved items in one, ranked list integrating results from different sources. Integrated search in this meaning is similar to universal search as found in some current web search engines mixing images, video and web results [1].

A main challenge in integrated search is that documents from different sources may be of different types, described on various levels of metadata and vocabularies. If for instance the domain is scientific publications, some documents may be available in full text, with and without metadata description, and some only as metadata records with or without abstracts. As all documents may be potentially relevant, treating all types in the same way in indexing and retrieval may overemphasise some types over others, e.g., resulting in the full text documents being more easily retrieved and higher ranked than documents only being described by metadata.

Evaluating different approaches to integrated search is currently difficult as no test collections exist with sufficiently different document types and comprehensive relevance assessments for each type. An appropriately designed test collection would be valuable and allow, e.g., the design of integrated search algorithms that better identify and rank relevant documents across the different types.

Preprint, published as:
Lykke, M., Larsen, B., Lund, H. & Ingwersen, P. (2010): Developing a Test Collection for the Evaluation of Integrated Search. In: Gurrin, C. & al. eds. *Advances in Information Retrieval, 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010, Proceedings*. Berlin: Springer, p. 627-630. (Lecture Notes in Computer Science ; 5993)

In this poster we describe and analyse a new test collection that we have constructed for this purpose. We describe the development process, and analyse the resulting characteristics of the collection.

2 The integrated search test collection

IR systems evaluation is addressed from two quite different perspectives: the system-driven and the user-oriented perspectives [2]. Our approach is to develop a test collection that support both evaluation perspectives, by using a semi-laboratory/semi-real-life approach, using users' genuine information needs, and non-binary relevance judgements. Our aim has been to facilitate integration of the two perspectives at the study-design level [2], and to seek realism as well as experimental control.

A test collection for experiments with integrated search requires the following as a minimum: a corpus with several different document types, several levels of descriptions, appropriate search tasks, and relevance assessments with adequate amount of relevant documents for each type. In addition, it would be desirable to have documents without copyright restrictions (for acquiring a larger corpus), graded relevance assessments and tasks from users with real tasks/needs (for greater realism).

2.1 Document collection

The scientific domain of physics comprises a realistic case with longstanding traditions for self-archiving of research publications in open access repositories and information sharing between scholarly and professional environments [3]. One of the largest repositories is arXiv.org, containing more than 500,000 papers covering the main areas of physics. We extracted two subsets from arXiv.org:

- 160,000+ full text papers in PDF including separate metadata (courtesy of Tim Brody, www.citebase.org)
- 274,000+ metadata records including abstracts for most documents (harvested using OAI PMH from www.arxiv.org)

The two subsets are very different in nature (see Table 1), the full text documents being much longer (4422 words on average) than the metadata records (272 words on average). In addition, we have added 18,000+ bibliographic book records classified as physics from the Danish national database covering all research and higher education libraries in Denmark (average length 189 words; no abstracts).

2.2 Search tasks and relevance assessments

We extracted 65 natural search tasks from 23 lecturers, PhDs and experienced MSc students from three different university departments of physics. On average each test person provided three search task descriptions, which were captured in online forms via computers located in their own university environment. Prior to describing their tasks they were briefed about the project objectives and the structure and purpose of

the form. After filling the forms they answered an online questionnaire concerning their personal data, domain and retrieval knowledge and experiences.

The search task description form had five questions, in line with [4]: a) *What are you looking for*? b) *Why are you looking for this*? c) *What is your background knowledge of this topic?* d) *What should an ideal answer contain to solve your problem or task?* e) *Which central search terms would you use to express your situation and information need?* Questions b) – c) correspond to questions asked in [4], with b) being about the underlying work task situation or context, and c) about the current knowledge state. Question a) asks about the formulation of the current information need, and d) correspond to the 'Narrative' section in a common TREC topic whilst e) asks for perceived adequate search terms.

The PDFs, abstract-only and library records were downloaded and indexed using the Fedora Generic Search Service with Lucene as search engine (see www.fedora-commons.org). A pool of up to 200 documents per task was retrieved for relevance assessments, separately for each document type and proportional to the corpus distribution where possible. In this way the (longer and more retrievable) PDFs would not be over represented. However, as shown in Table 2, in many tasks there were not enough abstract records to be retrieved, resulting in a higher proportion of PDFs. The searches were carried out manually by the research team as exhaustively as possible, based on the suggested search terms and other tokens in the original task descriptions.

Two months after task creation access to a web-based relevance assessment system was opened, allowing 1) access to the pool of documents to be assessed (sorted randomly within each document type), presented in overview form and with the possibility of opening full text PDFs where applicable; 2) assigning relevance scores according to the following 4-point scale: highly, fairly, marginally and non-relevant [5]. Documents could be re-assessed if the test person chose to. A post-assessment questionnaire on satisfaction with the assessment procedure and search results was filled for each task. Table 2 shows the relevance distributions over document types.

3 Discussion and conclusion

The integrated search test collection has the following central features, Table 1. While there is no great variation in arXiv.org record size (abstracts) the library record size varies more owing to the presence/absence of table-of-contents data.

With reference to Table 1 we observe in line with [4] that the formulation of the information need (a) is short compared to underlying task description (b) and knowledge state (c), but longer than the suggested search terms (e). This phenomenon occurs regardless of the complexity or comprehensiveness of the described search task and knowledge state. Later analyses may reveal degrees of specificity, facets and vocabularies between different sections in the task descriptions, task categories and relevance assessments. Across the three document types 19 of the 65 tasks did not receive assessments covering all three positive degrees of relevance in all the types. By isolating tasks with at least one assessment of 2-3 different positive relevance grades we observe Table 2 that the collection allows for test using four separate sub-collections.

Table 1. Central features of the integrated search test collection.

Features	Number	Mean no. of words per item
PDF items, arXiv.org	160,168	4422*
Abstracts, arXiv.org	274,749	272*
Library Records	18,222	189*
References (Citations)	3,748,555	(23.4 on avg. for 160,168 PDF items only)
Work Task situations	65	104.4*
a) Information need	65	17.7* (13 tasks with 5-10 terms)
b) Work task context	65	35.7*
c) Knowledge state	65	22.2*
d) Ideal information	65	19.3*
e) Search terms	65	9.4* (20 tasks with 3-6 terms)

* minus stop words (using the 318 word list at ir.dcs.gla.ac.uk/resources/linguistic_utils/)

Table 2. Relevance assessment statistics at document level. (*PDF*) are arXiv.org full text documents with metadata, abstract *and* PDF; (*ABS*) are arXiv.org records with metadata and abstract only. (*Meta*) are *all* arXiv.org records, with metadata and abstract, but omitting the PDF field. (+*gr*) implies positive relevance grades for the corresponding number of tasks.

Search Tasks – *Collection types*	High (H)	Fair (F)	Marg. (M)	Non-rel.	Total	Books.	PDF	Abs. (Meta)
Total collect. 65 tasks	337	666	1875	8188	11066	992	5933	4141
Mean (65 tasks)	5.2	10.2	28.8	126	170.2	15.2	91.3	63.7
% (65 tasks)	3.0%	6.0%	16.9%	74.0%	-	9.0%	53.6%	37.4%
PDF, 44 tasks (2-3 +gr)	112	284	947	3213	4556	(← 29.5% H/F/M rel.)		
ABS, 48 tasks (2-3 +gr)	170	249	602	2336	3556	(← 30.4% H/F/M rel.)		
Books, 45 tasks (2-3 +gr)	53	130	241	568	992	(← 42.7% H/F/M rel.)		
Meta, 51 tasks (2-3 +gr)	284	532	1570	6130	8516	(← 28.0% H/F/M rel.)		

The strength of the collection is the amalgamation of realism and control, that the majority of the 65 real-life tasks have a fair number of relevant documents both across documents types and relevance grades (25.9% are relevant to some degree, with 6% and 3% being fairly and highly relevant respectively). The collection can therefore be used for integrated IR tests as intended. Secondly, search simulations of different aspects of information situations and work task contexts can be tested, e.g. to be pooled in a variety of combinations as evidence of the searcher situation.

References

[1] Sullivan, D. (2008): *Google Universal Search: 2008 Edition*. Available: http://searchengineland.com/google-universal-search-2008-edition-13256

[2] Järvelin, K. (2007). An analysis of two approaches in information retrieval: from frameworks to study designs. *JASIST* 58(7), 971-986.

[3] Gómez, N.D. (2004). Physicists' information behaviour: a qualitative study of users. *70th IFLA Council and General Conference IFLA, Buenos Aires, 22-27 August, 2004.*

[4] Kelly, D., & Fu, X. (2007): Eliciting better information need descriptions from users of information search systems. *Information Processing & Management*, 43(1), 30-46.

[5] Sormunen, E. (2002): Liberal relevance criteria of TREC – Counting on negligible documents? In: *Proceedings of SIGIR 2002*. ACM Press, New York, 320-330.