UNIVERSITY OF COPENHAGEN

Københavns Universitet

Evaluation in Context

Jaap, Kamps; Lalmas, Mounia; Larsen, Birger

Publication date:
2009

Document Version
Early version, also known as pre-print

Citation for published version (APA):
Jaap, K., Lalmas, M., & Larsen, B. (2009). Evaluation in Context. Paper presented at European Conference on Digital Libraries, Corfu, Greece.

Download date: 17. Feb. 2019

Evaluation in Context*

Jaap Kamps[1,2], Mounia Lalmas[3], and Birger Larsen[4]

[1] Archives and Information Studies, University of Amsterdam
[2] ISLA, University of Amsterdam
[3] Department of Computing Science, University of Glasgow
[4] Information Studies, Royal School of Library and Information Science

Abstract. All search happens in a particular context—such as the particular collection of a digital library, its associated search tasks, and its associated users. Information retrieval researchers usually agree on the importance of context, but they rarely address the issue. In particular, evaluation in the Cranfield tradition requires abstracting away from individual differences between users. This paper investigates if we can bring some of this context into the Cranfield paradigm. Our approach is the following: we will attempt to record the "context" of the humans already in the loop—the topic authors/assessors—by designing targeted questionnaires. The questionnaire data becomes part of the evaluation test-suite as valuable data on the context of the search requests. We have experimented with this questionnaire approach during the evaluation campaign of the INitiative for the Evaluation of XML Retrieval (INEX). The results of this case study demonstrate the viability of the questionnaire approach as a means to capture context in evaluation. This can help explain and control some of the user or topic variation in the test collection. Moreover, it allows to break down the set of topics in various meaningful categories, e.g. those that suit a particular task scenario, and zoom in on the relative performance for such a group of topics.

1 Introduction

The history of information retrieval (IR) is a showcase of theoretical progress going hand-in-hand with experimental evaluation. The scientific evaluation of IR systems is rooted on the Cranfield experiments [4], and the main thrust in recent years has been the Text REtrieval Conference (TREC) and its various regional and task-specific counterparts such as CLEF, NTCIR, and INEX. The Cranfield tradition of test collection development tries to abstract away from individual differences between users [15]. Yet at the same time, it has been known for a long that individual differences are one of the greatest sources of variation in relevance judgments and system failures [3, 13]. In fact, even within the test collections built in the Cranfield tradition, the "topic effect" or "user effect" is the largest source of variation [2]. Nonetheless, the overwhelming success of experimental IR can be interpreted as a clear signal that the test collection abstraction is effective for evaluating document retrieval.

Digital libraries researchers are addressing search tasks of increasing complexity that require finer-grained judgments than standard document retrieval. Examples are

* This research was partly funded by DELOS (an EU network of excellence in Digital Libraries) through the INEX initiative for the evaluation of XML retrieval.

information pin-pointing tasks like Question Answering and XML Retrieval. For these tasks, it is more than plausible that individual differences have a much greater impact. As Sparck Jones [14], p.15 puts it:

> *TREC needs to engage, more positively and fully, with context and the nature of the whole setup information-seeking task T rather than just the experimental task X. ... I believe that it is important for TREC's system-building participants to be encouraged to work forward from a fuller knowledge of the context, rather than limiting their attention to the attenuated form of the context that the D * Q * R environment normally embodies, and recovering what may be distinctive about this – and hence somewhat indicative of significant features of the larger context – by working backwards from system results.*

The "user effect" is keeping test collection builders in a double-bind: On the one hand, building a stable and reusable test collection requires abstracting from task and user differences. On the other hand, more realistic search tasks such as those occurring in digital libraries requires bringing some of the user and task context into consideration.

The main research question of this paper is to investigate if we can bring some of the user's context into the Cranfield paradigm. Our approach is the following: rather than fitting a retrieval task and its evaluation to a particular context, we will attempt to record the "context" of the humans already in the loop during the construction of a test collection: the topic authors/assessors. Recall that, as mentioned above, individual difference are by no means outlawed in the Cranfield tradition—they are the greatest source of variation in standard search tasks investigated at TREC and other evaluation forums. By designing targeted questionnaires, we can record salient features of the topic authors and their topics of request, and of the assessors and their judging behavior. The resulting questionnaire data becomes part of the evaluation test-suite as valuable contextual data. This can help explain and control some of the user or topic variation in the test collection. Moreover, it will allow to break down the set of topics in various meaningful categories, and zoom in on the relative performance for such a group of topics. This allows for the testing of a larger variety of research questions as the tests can be restricted to the appropriate subset of topics.

Our aims are closely related to those of the TREC HARD track (2003–2005) and its continuation in the ciQA task at the TREC QA track (2006–2007). The HARD tracks [1] investigated whether retrieval systems could improve by i) query metadata that better described the information need, ii) interaction with the searcher through clarification forms, and iii) passage level retrieval and relevance judgments. Here the query metadata—consisting of fields like familiarity, genre, geography, subject, and related text—is most closely related. The main focus of the HARD track (and its continuation) was on user system interaction focusing on additional information that can be directly used by retrieval systems, whereas our main focus is on recording the broader context of the topics and assessments to provide insight in the constructed test collection, and to aid further analysis.

We have conducted an exploratory experiment with our questionnaire approach during the 2007 evaluation campaign of the INitiative for the Evaluation of XML Retrieval. Research in XML retrieval attempts to take advantage of the structure of explicitly

marked up documents to provide more focused retrieval. This is believed to be of benefit when users are searching large documents, such as those often contained in digital libraries. The task of XML retrieval is a much more complicated one than standard document retrieval. Not only must XML element retrieval systems be able to identify relevant content; in addition a suitable granularity of the returned elements must be decided on. As a consequence the creation of test collections for XML retrieval is a notable challenge in itself. During INEX 2007, topic authors completed a questionnaire immediately after submitting the final version of their topics, and assessors completed a questionnaire after finishing the judging of their topics. The topic author questionnaire consisted of 19 questions about the topic familiarity, the type of information requested and expected, results presentation, and the use of structured queries. The assessor questionnaire consisted of 13 questions about the topic of request, the meaning of their relevance judgments. We will focus here on the assessor questionnaire, and investigate the value of the context recorded by the questionnaires to help answer some important questions underlying focused retrieval in structured documents.

The rest of this paper is structured as follows: Section 2 presents some background on the INEX initiative, its search tasks, and some of the main underlying questions. Section 3 presents the questionnaire and presents an analysis of the main results. Section 4 discusses the test collection in context, where the questionnaire is related to the rankings of retrieval systems. In Section 5, we end by discussing our results and by drawing some conclusions.

2 Ad hoc Retrieval at INEX

In 2002 until 2004, assessors judged pools of retrieved elements, although presented in their article context, using complex two-dimensional judgments [10]. These judgments where based on exhaustivity (basically topical relevance, whether the element contained enough relevant information) and specificity (whether the element contained no excess non-relevant information), both judged on a 4-point scale. The complexity of the assessments also led to complex measures, having to deal with a range of problems [9]. In 2005, this was substituted for by a much simpler assessment system, in which assessors are asked to highlight all, and only, the relevant text [10]. This greatly simplified the assessment process and obviated the need for complex rules to make assessments consistent over partly overlapping XML elements. This made the assessors tasks an intuitive one, leading to eliciting more natural judgments. However, from the start there was a lively discussion on the meaning of the highlighted passages: Are assessors highlighting relevant text according to some global criterion based on the narrative? Or are assessors highlighting the most relevant text according to the local article context?

This led to the introduction of two search tasks at INEX 2006: Relevant in Context and Best in Context, and the elicitation of a separate Best-entry-point judgment. The first task corresponds to an end-user task where focused retrieval answers are grouped per document, in their original document order, providing access through further navigational means. This assumes that users consider documents as the most natural units of retrieval, and prefer an overview of relevance in their original context.

Relevant in Context (RiC) This task asks systems to return non-overlapping relevant document parts clustered by the unit of the document that they are contained within.

An alternative way to phrase the task is to return documents with the most focused, relevant parts highlighted within.

The second task is similar to Relevant in Context, but asks for only a single best point to start reading the relevant content in an article.

Best in Context (BiC) This task asks systems to return a single document part per document. The start of the single document part corresponds to the best entry point for starting to read the relevant text in the document.

The Relevant in Context Task is evaluated against the text highlighted by the assessors, whereas the Best in Context Task is evaluated against the best-entry-points. For both tasks, mean average generalized precision (MAgP) is used [8]. This is a MAP-like measure where the score per document varies between 0 and 1. Specifically, for Relevant in Context the score per document is determined by how well the retrieved text corresponds to the highlighted text, and for Best in Context the score per document depends on the distance between the retrieved entry point and the assessor's best entry point. In the paper, we will focus on these two tasks.

3 Questionnaires at INEX 2007

We designed and used topic creator and assessor questionnaires in the INEX 2007 evaluation campaign [5]. An IR test collection consists of a collection of documents, a set of search topics, and relevance judgments. For INEX 2007, the document collection is an XML'ified version of the English Wikipedia. Search requests or topics are authored (and also judged) by the INEX participants. At the INEX 2007 ad hoc track, a total of 130 topics was used and a total of 107 topics was assessed. Directly after submitting a candidate topic, the topic author was presented with a new page containing a questionnaire consisting of 19 questions and an open space for comments on the questionnaire. For 107 of the 130 ad hoc topic, this topic questionnaire is available.[5] Directly after assessing a topic, the judge was presented with a questionnaire consisting of 13 questions and an open space for comments. These 13 questions dealt with various issues on the topic of request, the meaning of the highlighted passages, and the meaning of the best-entry-point (BEP). In the rest of this paper, we will discuss the responses to the assessor questionnaire. We restrict the analysis to the 91 topics for which we have both a topic creator and assessor questionnaire (and judgments).

3.1 Topic of request

We first discuss the responses to questions about the topic of request. Table 1 shows question C1. Almost 60% of the topics have been assessed by the original topic author. Table 2 shows question C2. The majority of judges is familiar with the subject matter of the topic at hand, although there are still 5% of the topics where assessors venture into unfamiliar terrain. Table 3 shows question C3. It is reassuring that the majority of the topics was easy to judge. Table 4 shows question C4. For over 80% of the topics, the Wikipedia is an obvious resource to look for information.

[5] Some topics were derived from other sources, such as the topics used at the INEX 2006 Interactive track, and hence we do not have a topic creator questionnaire for these topics.

Table 1. (C1) *Did you submit this topic to INEX?*

Answer	Freq.	Perc.
no	37	41%
yes	54	59%

Table 2. (C2) *How familiar were you with the subject matter of the topic?*

Answer		Freq.	Perc.
Not familiar	1	5	5%
	2	13	14%
	3	32	35%
	4	24	26%
Very familiar	5	17	19%

Table 3. (C3) *How hard was it to decide whether information was relevant?*

Answer		Freq.	Perc.
Very easy	1	16	18%
	2	40	44%
	3	21	23%
	4	14	15%
Very difficult	5	0	0%

Table 4. (C4) *Is Wikipedia an obvious source to look for information on the topic?*

Answer	Freq.	Perc.
no	16	18%
yes	75	82%

Table 5. (C5) *Can a highlighted passage be (check all that apply):*

Answer	Freq.	Perc.
a single sentence	69	76%
a single paragraph	86	95%
a single (sub)section	77	85%
a whole article	62	68%

Table 6. (C6) *Is a single highlighted passage enough to answer the topic?*

Answer		Freq.	Perc.
None of them is	1	11	12%
	2	17	19%
	3	30	33%
	4	28	31%
All of them are	5	5	5%

Table 7. (C7) *Are highlighted passages still informative when presented out of context?*

Answer		Freq.	Perc.
None of them is	1	2	2%
	2	16	18%
	3	22	24%
	4	41	45%
All of them are	5	10	11%

Table 8. (C8) *How often does relevant information occur in an article about something else?*

Answer		Freq.	Perc.
Never	1	9	10%
	2	35	38%
	3	27	30%
	4	20	22%
Always	5	0	0%

3.2 Highlighted Passages

We now show the responses to questions about the meaning of highlighted passages: Table 5 shows question C5. It is clear that there is no fixed unit of retrieval, for almost 1/4 of the topics a relevant passage cannot be a sentence, and for almost 1/3 of the topics a relevant passage cannot be a whole article. Table 6 shows question C6. There is also no consensus on whether a single passage could suffice as an answer. Table 7 shows question C7. Again no consensus on whether an isolated passage is still informative. Table 8 shows question C8. Relevant passages frequently occur in articles about something else, which support the motivation behind focused retrieval. Table 9 shows

Table 9. (C9) *How well does the total length of highlighted text correspond to the usefulness of an article?*

Answer		Freq.	Perc.
Never	1	3	3%
	2	20	22%
	3	36	40%
	4	25	27%
Always	5	7	8%

Table 10. (C10) *Which of the following two strategies is closer to your actual highlighting: (I) the best passages, (II) all relevant text?*

Answer		Freq.	Perc.
I: best	1	12	13%
	2	22	24%
	3	5	5%
	4	30	33%
II: relevant	5	22	24%

Table 11. (C11) *Can a best entry point be (check all that apply):*

Answer	Freq.	Perc.
the start of a highlighted passage	84	92%
the sectioning structure containing the highlighted text	55	60%
the start of the article	51	56%

Table 12. (C12) *Does the best entry point correspond to the best passage?*

Answer		Freq.	Perc.
Never	1	1	1%
	2	15	16%
	3	26	29%
	4	31	34%
Always	5	18	20%

Table 13. (C13) *Does the best entry point correspond to the first passage?*

Answer		Freq.	Perc.
Never	1	2	2%
	2	21	23%
	3	21	23%
	4	29	32%
Always	5	18	20%

question C9. There is no evidence for assumption that the length of a highlighted passage corresponds to its usefulness, an assumption that has been repeatedly proposed for evaluation measures at INEX. Table 10 shows question C10. There is a remarkable division over the two assessment strategies: the strategy I highlighting the "best passages" is chosen almost as frequently as the strategy II highlighting "relevant passages." Since the particular strategy will have an impact on the resulting assessments, where a judge using strategy I will regard less text as relevant than a judge using strategy II, this can have a large potential impact on the ranking of systems.

3.3 Best Entry Points

We now show the responses to questions about the meaning of the best entry point. Table 11 shows question C11. For almost all topics, the best-entry point can be the start of the highlighted passage (C11), but other types of BEPs occur [11]. Table 12 shows question C12. In the majority of cases, the BEP corresponds to the best passage. Table 13 shows question C13. Again, in the majority of cases the BEP corresponds to the first passage. The responses to C11-C13 may be in part explained by vast majority of relevant articles (4,581 out of 6,491) having only a single highlighted passage [5].

3.4 Relations

We now analyze the relation between responses to different questions in the questionnaire. Table 14 show the relations between pairs of questions in the questionnaire. Since

Table 14. Relationship between answers for pairs of questions (chi-square test at percentiles 0.95 and 0.99).

	C1	C2	C3	C4	C5				C6	C7	C8	C9	C10	C11			C12
					sen	par	sec	art						pas	sec	art	
C2	0.99																
C3	-	0.95															
C4	0.95	-	-														
C5 sen	-	-	-	-													
C5 par	-	-	-	-	-												
C5 sec	-	-	0.95	-	-	-											
C5 art	-	-	-	-	-	0.95	0.99										
C6	-	-	-	-	-	-	-	-									
C7	-	-	-	-	-	-	-	-	0.99								
C8	-	-	-	-	-	-	-	-	-	-							
C9	-	-	-	-	-	-	-	-	-	-	-						
C10	-	-	-	-	-	0.99	0.99	0.99	-	-	-	0.99					
C11 pas	-	-	-	-	0.95	-	-	-	-	-	-	-	-				
C11 sec	-	-	0.95	-	-	0.99	0.95	-	-	-	-	-	0.95	-			
C11 art	-	-	-	-	-	-	0.99	0.99	-	-	-	-	0.99	0.95	0.99		
C12	-	-	0.99	-	0.95	0.99	0.95	-	-	-	-	-	-	-	-	-	
C13	-	0.95	-	-	-	0.99	-	0.95	0.95	0.99	-	-	-	-	0.99	-	-

most questions give nominal answers (e.g., yes/no) we use a chi-square test. In particular, we have found that there are a number of relations worth considering. There is an interesting, but not unexpected, relation between being a topic author (C1) and being more familiar with the topic at hand (C2), and also a relation between familiarity (C2) and ease of judging (C3). This confirms the importance of having topics assessed by the original topic author. The judging strategy, judging the best or all relevant passages (C10), is indeed clearly affecting judging behavior: it is related to the granularity of highlighted passages (C5) and to the choice of BEP (C11). There is also a relation with the correspondence between the amount of highlighted text and its usefulness (C9), which holds for assessors highlighting all relevant text, but not for assessors highlighting only the best passages. Finally, the choice of BEPs as best passages (C12) is related to relevant passages being smaller units of the document structure (sentences, paragraphs, (sub)sections), and the choice of BEPs as first passages (C13) is related to relevant passages being complete answers (a single passage is an answer, and still informative in isolation).

Perhaps the most striking observation is that there is such great variety in the responses of the topic authors. This suggests that there are distinct search tasks underlying XML retrieval. This also gives support to the decision at INEX to define a number of distinct search tasks, thus allowing the study of alternative search scenarios for digital libraries. There is rich contextual information in XML retrieval, and the questionnaires provide a means to extract it. But what is the relative importance of the contextual information? In the following section, we will investigate this in terms of system effectiveness.

4 Test Collection in Context

In Section 3, we reported the responses to the questionnaire as a survey amongst assessors. The outcomes exemplify the wide range of what we can consider contextual

Table 15. (C5 article): Can a highlighted passage be: a whole article?

			Relevant in Context task											Best in Context task									
Answer	Freq.	%	Overall rank of top 10										Tau	Overall rank of top 10										Tau
yes	62	68.13	1	4	*2*	8	*5*	6	3	**10**	*7*	12	0.9301	**1**	2	3	4	8	*9*	6	5	14	15	0.8938
no	29	31.87	6	*7*	1	*2*	4	*5*	18	3	*22*	14	0.7725	5	*25*	*22*	*18*	12	11	*10*	*21*	16	7	0.7175

information regarding the topics and their assessments. In this way, the questionnaire data also becomes part of the evaluation test-suite constructed during INEX 2007. Our conjecture is that the questionnaire data can provide valuable contextual data on the topics of request and their topic authors. In this section, we start to explore how this additional data can be used.

We investigate how context affects the system ranking—what systems turn out to be effective?—by looking at the relative ranking of systems for a subset of the topics corresponding to a particular context. We analyze the official submissions to the INEX 2007 ad hoc retrieval track's Relevant in Context and Best in Context tasks. There were in total 66 valid runs submitted for the Relevant in Context task, and 71 valid runs for the Best in Context task. For each question in the questionnaire, we can break down the set of topics over the answer categories allowing us to investigate the performance of systems for a particular context. That is, for each of the subsets of the topics we can calculate how the 66 or 71 retrieval systems are ranked, and compare the resulting ranking to the ranking over all topics (i.e., the official outcomes over 104 assessed topics). Below, we will discuss some of the questions in detail. For each answer category, we calculate the rank-correlation (we use Kendall's tau) between the score over all topics, and the score over the selected topics corresponding to a particular answer category.[6] We will show the ten best performing systems for both Relevant in Context (RiC) and Best in Context (BiC) for topics corresponding to each of the answer categories.

Table 15 shows the results of question C5 (only the "article" part) over the 91 topics for which we have both the questionnaires and assessments. Each answer category (column 1) selects a number of topics (columns 2 and 3). In columns 4–13 (Relevant in Context) and columns 15–24 (Best in Context) we show the ten best performing systems for this subset of the topics. The systems are labeled with their system rank over the whole topic set (i.e., based on the official scores). That is, in columns 4 and 15 we find the best scoring system for the subsets, which in case of the subset of 62 topics with response "yes" is also labeled "1" (for both tasks) and hence the best overall system. Over the 29 topics with response "no", the best RiC system was ranked 6th over all topics, and the best BiC system was ranked 5th over all topics. Columns 14 and 25 show the rank correlation between the ranking over all topics and the ranking based on the subset of the topics. We see that the ranking over the "yes" topics corresponds well with the overall ranking (rank correlations around 90%). The ranking over the "no" topics shows remarkable upsets. To aid the analysis we have highlighted two types of runs. We show runs retrieving only whole articles in **bold**, since they were remarkably effective

[6] Since all topics in the subset are necessarily included also in the whole topic set, the subset scores will automatically approximate the overall scores depending on the size of the subset. This makes it difficult to compare the rank correlations for different subsets of the topics, especially if they contain different numbers of topics.

Table 16. (C7): Are highlighted passages still informative when presented out of context?

			Relevant in Context task		Best in Context task	
Answer	Freq.	%	Overall rank of top 10	Tau	Overall rank of top 10	Tau
1+2 (none)	18	19.78	6 9 8 14 4 1 *22* *26* 3 *5*	0.8079	5 *25* 2 3 4 6 *22* *9* 15 14	0.7328
3	22	24.18	1 *2* *5* *7* 4 3 6 **10** 9 8	0.8424	**1** 2 5 11 8 *21* *9* 13 6 *10*	0.7859
4+5 (all)	51	56.04	**10** 1 4 *2* 6 *7* *5* 8 12 3	0.8937	**1** 3 4 8 2 **17** 15 14 *9* *10*	0.9155

Table 17. (C10): Which of the following two strategies is closer to your actual highlighting: (I) the best passages, (II) all relevant text?

			Relevant in Context task		Best in Context task	
Answer	Freq.	%	Overall rank of top 10	Tau	Overall rank of top 10	Tau
1+2 (I)	34	37.36	6 1 *2* *7* 4 8 3 11 *5* 14	0.8844	5 2 13 *9* **1** *22* 4 3 *10* 12	0.8334
3	5	5.49	8 *13* 6 **10** 12 21 *5* 11 14 *16*	0.6942	*9* **1** 8 5 *20* 27 13 6 **31** 37	0.5670
4+5 (II)	52	57.14	1 4 *2* *5* 3 *7* 6 8 **10** 9	0.8918	**1** 3 4 2 8 15 14 11 6 *10*	0.8777

for BiC. We show runs based on the structured (or CAS, content and structure) query in *italics*, since they seemed not to improve over standard keyword query runs. A structured query contains references to the document structure and generally suggests the retrieval of particular XML elements, and participants could use either the structured query or a standard keyword query. We see, especially for BiC, that many of the runs that perform well on the "no" topics use such structured queries. Again for BiC, the run performing best over all topics, and over the "yes" topics, is always retrieving the start of the article as BEP—a strategy not particularly effective for the "no" topics. The rank correlation between the two subsets of the topics selected by responses "yes" and "no" is 0.7268 for RiC, and only 0.6483 for BiC.

Table 16 discusses C7, whether highlighted passages are still informative out of context (1=None of them is, 5=All of them are). We have collapsed the 5-point scale to a three point scale, to have a sufficient number of topics per answer category. The topics where passages are self-contained, response "4+5 (all)," correlate the best with the overall ranking. On these topics, runs retrieving whole articles (indicated in bold) are effective for BiC, but also for RiC. The system ranking for "1+2 (none)" shows remarkable upsets, especially from runs using the structured query (indicated in italics). The effectiveness of article retrieval seems counter-intuitive, and is is partly due to the Wikipedia's structure, where individual entries are exclusively focused on a single topic, and are often relatively short [7].

Table 17 discusses C10, about the two assessor strategies highlighting the best information, or all relevant information (1=strategy I, 5=strategy II). Strategy II is leading to fully highlighted articles, and this is favorable for article retrieval strategies (indicated in bold). Despite the radical differences between the strategies, the effect on the system rankings is not dramatic—the system-rank correlation between "1+2 (I)" topic and "4+5 (II)" topics is 0.7855 (RiC) and 0.7481 (BiC). This also suggests that the retrieval techniques effective for finding the best information are also effective for finding all relevant information, and the other way around.

Table 18 discusses C11 (article part), whether the best entry point can be the start of the article. For BiC, the "yes" topics resemble the overall ranking closely for the

Table 18. (C11 article): Can a best entry point be: the start of the article?

			Relevant in Context task		Best in Context task	
Answer	Freq.	%	Overall rank of top 10	Tau	Overall rank of top 10	Tau
yes	51	56.04	*5* 6 1 8 4 **10** *2* 3 *7* 12	0.8993	**1** 2 3 4 8 6 5 *9* 14 15	0.8801
no	40	43.96	1 *2* *7* 4 6 3 11 *5* 18 8	0.8751	*18* 5 *10* 12 11 *9* **1** 3 4 7	0.8358

Table 19. (C12): Does the best entry point correspond to the best passage?

			Relevant in Context task		Best in Context task	
Answer	Freq.	%	Overall rank of top 10	Tau	Overall rank of top 10	Tau
1+2 (never)	16	17.58	4 *5* 1 3 *2* *7* 8 *13* 6 9	0.7837	**1** *9* 8 2 **17** *20* 11 *28* 5 7	0.7787
3	26	28.57	8 4 **10** 1 *2* 6 *7* *5* 3 12	0.8294	**1** 8 3 4 **17** *9* 14 15 6 11	0.8117
4+5 (always)	49	53.85	1 *2* 6 4 *7* *5* 3 8 14 11	0.9152	5 2 13 3 4 6 *10* **1** *9* 12	0.8922

Table 20. (C13): Does the best entry point correspond to the first passage?

			Relevant in Context task		Best in Context task	
Answer	Freq.	%	Overall rank of top 10	Tau	Overall rank of top 10	Tau
1+2 (never)	23	25.27	*2* 1 *5* *7* 3 4 6 11 9 8	0.9105	*21* *10* *18* 12 7 11 13 3 4 **1**	0.8036
3	21	23.08	6 8 14 9 4 3 *7* 15 *5* **23**	0.8396	2 5 4 3 14 15 **1** 23 16 6	0.8455
4+5 (always)	47	51.65	4 1 **10** *2* 8 *5* 6 12 *7* *13*	0.8443	**1** 8 **17** *9* 6 2 5 *20* 3 4	0.8406

top runs, and the "no" topics favors systems using the structured query (similar to the breakdown over C5 in Table 15). For RiC the ranking of the top runs is more similar for the ranking over the "no" topics. This suggests that the RiC and BiC tasks do have a different nature. Note also that the article run labeled "1" for BiC is exactly identical to the article run labeled "10" for RiC. The system rank correlation between "yes" and "no" topics is 0.7893 for RiC, and 0.7465 for BiC.

Table 19 shows C12, whether the BEP corresponds to the best passage (1=Never, 5=Always). For BiC, article retrieval (indicated in bold) seems particularly effective for the "1+2 (never)" and "3" topics. For RiC, the overall ranking is more resembling the ranking over "4+5 (always)" topics, again suggesting differences between the two tasks.

Table 20 shows C13, whether the BEP corresponds to the first passage (1=Never, 5=Always). For both tasks, article retrieval is effective for "4+5 (always)," which makes sense since the first passage will be closer to the start of the article. Also, for both tasks the structure query is effective for "1+2 (never)," which makes sense since elements matching the structural constraints of the query need not occur early in the article. Again, there is a divergence between the overall ranking of RiC resembling the "1+2 (never)" topics, and the overall ranking of BiC resembling the "4+5 (always)" topics.

The main general conclusion is that context matters for the relative ranking of systems: we see varying levels of agreement between the ranking over all topics, and the ranking on subsets of the topics sharing particular context.

5 Discussion and conclusions

One of the greatest achievements of the field of IR is the development of a rigorous methodology to evaluate retrieval effectiveness [4]. The Cranfield approach as continued by TREC, CLEF, NCTIR and INEX has served us very well: virtually all progress in IR owes directly or indirectly to test collections built within the Cranfield paradigm.

The Cranfield tradition of test collection development tries to abstract away from individual differences between topic authors and assessors [15]. However, more complex search tasks that are a closer approximation of real-world information seeking in action, such as those prevalent in digital libraries, seem to require, in contrast, that some of the user's context is taken into account [6, 14].

This paper experimented with a new approach: rather than fitting a search task and its evaluation to a particular context, we record the "context" of the humans already in the loop: the topic authors/assessors. In particular, we investigated the use of a dedicated questionnaires to elicit and record salient aspects of the topic author's and assessor's context. The questionnaire data helps control the construction of a test collection. Moreover, the questionnaire data becomes part of the evaluation test-suite as contextual data of the search topics and their topic authors and assessors. There is a risk that extensions to Cranfield will limit the reusability of the resulting test collection. The extension proposed in this paper keeps the original test collection (documents, topics, relevance judgments) completely intact. In fact, the extension seems more likely to increase reusability, e.g., by allowing researchers to analyse their system's performance over topics that best match their intended search context.

How does the questionnaire data improve the evaluation in comparison with the traditional "bag of topics"? On the one hand, the questionnaires can be used to control the test collection building. The responses also give an overview of the composition of the topic set, and highlights sources of divergence. This can be crucial during the selection of candidate topics to be assessed, or when combining test collections from different years. On the other hand, the questionnaires can be used directly by participants to investigate which contextual aspects impacted their system's performance (and how). We have shown this in Section 4. It is important to note that we do not envision the main system comparison to change, such comparison requires the common ground provided by the entire topic set, and the ultimate aim is to have a systems that performs well on all topic types. But in addition to this, the contextual data gathered is facilitating deeper analysis of what worked and what not and why it worked and why not—especially in the case of failure analysis, this can be insightful.

It is also clear that the questionnaires are no a panacea. Interpreting the data is hard: What does it mean precisely when for n% of the topics a certain response is given? What does it mean precisely if my system performs well for a particular subset of topics? This may not be immediately clear, although the contextual data in the questionnaire will help focus on "interesting" contextual aspects and subsets of the topics, and will at least give a good hint about the interpretation. This is partly because we are venturing in new terrain, and cannot compare to data from earlier years or to other well-understood data.

Evaluating a comprehensive search systems, such as a digital library, is a complex and difficult undertaking [12]. All searches happen in a particular context—such as the particular collection of the digital library, its associated search tasks, and its associated users. Evaluation should take relevant parts of this context into account. The most striking observation overall is that there is such great variety in the responses of the assessors—much greater than we expected beforehand. Perhaps equally surprising is that the system rankings are nonetheless reasonably robust—much more robust than we expected beforehand. This can also be interpreted as strong support that Cranfield is

working: despite the differences in context (or noise in terms of Cranfield) the system rankings are remarkably stable with significant system rank correlations above 0.6 for all reasonably sized subsets of topics (also between different subsets of topics). That is, to a large extent we find ourselves in the same position as Zobel [16], despite all the upsets detailed in Section 4, there is also broad agreement on separating the good systems from the bad systems.

Acknowledgements Jaap Kamps was supported by the Netherlands Organization for Scientific Research (NWO, grants # 612.066.513, 639.072.601, and 640.001.501). Mounia Lalmas position is funded by Microsoft Research/Royal Academy of Engineering.

References

[1] J. Allan. HARD track overview in TREC 2005: High accuracy retrieval from documents. In *The Fourteenth Text REtrieval Conference (TREC 2003)*. National Institute of Standards and Technology. NIST Special Publication 500-255, 2004.

[2] D. Banks, P. Over, and N.-F. Zhang. Blind men and elephants: Six approaches to TREC tasks. *Information Retrieval*, 1:7–34, 1999.

[3] C. Buckley. Why current IR engines fail. In *Proceedings of the 27th Annual International ACM SIGIR Conference*, pages 584–585. ACM Press, 2004.

[4] C. W. Cleverdon. The Cranfield tests on index language devices. *Aslib*, 19:173–192, 1967.

[5] N. Fuhr, J. Kamps, M. Lalmas, S. Malik, and A. Trotman. Overview of the INEX 2007 ad hoc track. In *Focused access to XML documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2007)*, pages 1–23. Springer, 2008.

[6] P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer, 2005.

[7] J. Kamps, M. Koolen, and M. Lalmas. Locating relevant text within XML documents. In *Proceedings of the 31th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 847–849. ACM Press, 2008.

[8] J. Kamps, J. Pehcevski, G. Kazai, M. Lalmas, and S. Robertson. INEX 2007 evaluation measures. In *Focused access to XML documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2007)*, pages 24–33. Springer, 2008.

[9] G. Kazai, M. Lalmas, and A. P. de Vries. The overlap problem in content-oriented XML retrieval evaluation. In *Proceedings of the 27th Annual International ACM SIGIR Conference*, pages 72–79. ACM Press, 2004.

[10] B. Piwowarski, A. Trotman, and M. Lalmas. Sound and complete relevance assessments for XML retrieval. *ACM Transactions in Information Systems*, 27(1), 2008.

[11] J. Reid, M. Lalmas, K. Finesilver, and M. Hertzum. Best entry points for structured document retrieval: Parts I & II. *Information Processing and Management*, 42:74–105, 2006.

[12] T. Saracevic. Digital library evaluation: Toward evolution of concepts. *Library Trends – Special issue on Evaluation of Digital Libraries*, 49(2):350–369, 2000.

[13] T. Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *JASIS*, 26:321–343, 1975.

[14] K. Sparck Jones. What's the value of TREC – is there a gap to jump or a chasm to bridge? *SIGIR Forum*, 40:10–20, 2006.

[15] E. M. Voorhees. The philosophy of information retrieval evaluation. In *Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001*, pages 355–370. Springer, 2002.

[16] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314. ACM Press, 1998.