Evaluating Relevant in Context: Document Retrieval with a Twist

Jaap Kamps[1,2] Mounia Lalmas[3] Jovan Pehcevski[4]

[1] Archives and Information Studies, University of Amsterdam
[2] ISLA, Informatics Institute, University of Amsterdam
[3] Department of Computer Science, Queen Mary, University of London
[4] INRIA Rocquencourt, Le Chesnay, France

ABSTRACT

The Relevant in Context retrieval task is document or article retrieval with a twist, where not only the relevant articles should be retrieved but also the relevant information within each article (captured by a set of XML elements) should be correctly identified. Our main research question is: how to evaluate the Relevant in Context task? We propose a generalized average precision measure that meets two main requirements: i) the score reflects the ranked list of articles inherent in the result list, and at the same time ii) the score also reflects how well the retrieved information per article (i.e., the set of elements) corresponds to the relevant information. The resulting measure was used at INEX 2006.

Categories and Subject Descriptors: H.3 [Information Storage and Retrieval]: H.3.3 Information Search and Retrieval; H.3.7 Digital Libraries
General Terms: Measurement, Performance, Experimentation
Keywords: Evaluation, Context, Test collection, XML Retrieval

1. INTRODUCTION

Traditional document retrieval returns atomic documents as answers, and leaves it to users to locate the relevant information inside the document. Focused retrieval, such as practiced at INEX [2], studies ways to provide users with direct access to relevant information in structured documents. INEX 2006 introduced a new retrieval task, Relevant in Context (RiC), that combines article retrieval with XML element retrieval [1]. The RiC task is document or article retrieval with a twist, where not only the relevant articles should be retrieved but also a set of XML elements representing the relevant information within each article. Phrased differently, the system should return the relevant information (captured by a set of XML elements) within the context of the full article.

The task corresponds to an end-user task where focused retrieval results are grouped per article, in their original document order, providing access through further navigational means. This assumes that users consider the article as the most natural unit of retrieval, and prefer an overview of relevance in their context. Interactive experiments at INEX provided support for this task [4]. Moreover, the RiC task corresponds with the assessors' task at INEX, where assessors are asked to highlight the relevant information in a pooled set of articles. The difference is that the INEX assessors can highlight sentences, whereas systems could only return XML elements.

How to evaluate the RiC task? We face two main requirements: i) the score should reflect the ranked list of articles inherent in the result list, and ii) the score should also reflect how well the retrieved information per article corresponds to the relevant information.

Copyright is held by the author/owner(s).
SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

2. RELEVANT IN CONTEXT MEASURE

A submission for the task is a ranked list of articles, with for each article an unranked set of elements covering the relevant material in the article [1].[1] Hence, the evaluation is based on a ranked list of articles, where per article-rank we obtain a score reflecting how well the retrieved set of elements corresponds to the relevant information in the article.

2.1 Score per article

Per retrieved article, the text retrieved by the selected elements is compared to the text highlighted by the assessor. We calculate *Precision* as the fraction of retrieved text (in bytes) that is highlighted; *Recall* as the fraction of highlighted text (in bytes) that is retrieved; and *F-Score* as the combination of precision and recall using the harmonic mean, resulting in a score in [0,1] per article.

More formally, let us assume that the function $\mathsf{rel}(\cdot)$ gives the relevant or highlighted content of an element (or article), and that $\mathsf{ret}(\cdot)$ gives the retrieved content. Furthermore, assume that set operations (union, intersection) are defined for elements (and rel, ret) with the expected behavior in terms of their yield or content. Finally, assume that the function $| \ . \ |$ gives the byte-length of an element or set of elements in terms of their yield or content. Then, for each retrieved article a, Precision and Recall are defined as

$$\mathsf{P}(a) = \frac{|\mathsf{rel}(a) \cap \mathsf{ret}(a)|}{|\mathsf{ret}(a)|} \text{ and } \mathsf{R}(a) = \frac{|\mathsf{rel}(a) \cap \mathsf{ret}(a)|}{|\mathsf{rel}(a)|}$$

if $|\mathsf{rel}(a)| > 0$, and $\mathsf{R}(a) = 0$ otherwise. The combination is the standard F-score,

$$\mathsf{F}(a) = \frac{2 \cdot \mathsf{P}(a) \cdot \mathsf{R}(a)}{\mathsf{P}(a) + \mathsf{R}(a)}.$$

The resulting F-score varies between 0 (article without relevance, or none of the relevance is retrieved) and 1 (all relevant text is retrieved and nothing more), matching the second requirement.

2.2 Ranked list of articles

We have a ranked list of articles, with for each article a score $\mathsf{F}(a) \in [0, 1]$. Hence, we need a generalized measure, and we opt for the most straightforward generalization of precision and recall [3, p.1122-1123]. Over the ranked list of articles, we calculate *generalized Precision* as the sum of F-scores up to an article-rank, divided by the article-rank; and *generalized Recall* as the number of articles with relevance retrieved up to an article rank, divided by the total number of articles with relevance.

More formally, let us assume that for our topic there are in total *Numrel* articles with relevance, and assume that the function

[1] Due to the tree structure of XML documents, only disjoint (non-overlapping) elements were permitted.

Table 1: Significant improvements ($\star$) for MAgP.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1		-	-	-	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
2			-	-	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
3				-	-	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
4					-	-	-	-	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
5						-	-	-	-	-	⋆	-	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
6							-	-	-	-	-	-	-	-	⋆	⋆	⋆	⋆	⋆	⋆
7								-	-	-	-	-	-	⋆	⋆	⋆	⋆	⋆	⋆	⋆
8									-	-	-	-	-	⋆	⋆	⋆	⋆	⋆	⋆	⋆
9										-	-	-	-	⋆	⋆	⋆	⋆	⋆	⋆	⋆
10											-	-	-	⋆	⋆	⋆	⋆	⋆	⋆	⋆
11												⋆	-	-	-	-	⋆	⋆	⋆	⋆
12													-	-	-	-	⋆	⋆	⋆	⋆
13														-	-	-	-	-	⋆	-
14															-	-	-	-	-	-
15																-	-	-	-	-
16																	-	-	-	-
17																		-	-	-
18																			-	-
19																				-
20																				

$\mathsf{relart}(a) = 1$ if article a contains some relevant information, and $\mathsf{relart}(a) = 0$ otherwise. Then, at each article-rank of the list r_{art}, generalized Precision and Recall are defined as

$$\mathsf{gP}(r_{\mathrm{art}}) = \frac{\Sigma_{a=1..r_{\mathrm{art}}} \mathsf{F}(a)}{r_{\mathrm{art}}} \text{ and } \mathsf{gR}(r_{\mathrm{art}}) = \frac{\Sigma_{a=1..r_{\mathrm{art}}} \mathsf{relart}(a)}{\mathit{Numrel}}.$$

These generalized measures are compatible with the standard precision/recall measures, matching the first requirement. Specifically, the generalized Recall definition leads to a very natural interpretation of recall as the fraction of articles with relevance that has been retrieved. In that sense, the *Average generalized Precision* (AgP) for a topic can be calculated by averaging the generalized Precision at natural recall points where generalized Recall increases. When looking at a set of topics, the *Mean Average generalized Precision* (MAgP) is the mean of the Average generalized Precision scores per topic.

3. ANALYSIS

We now look at the ability of the MAgP measure to distinguish between different retrieval systems. We use the top 20 submissions to the INEX 2006 Relevant in Context retrieval task, based on the MAgP score over 111 topics, and label systems by their rank.[2] To determine whether improvements are significant, we use the bootstrap method (one-tailed at significance level 0.05 over 1,000 resamples). Table 1 shows the extent to which higher ranked systems improve significantly over lower ranked system in terms of their MAgP scores. We observe that the MAgP measure is fairly effective at distinguishing between different systems: no less than 112 of the 190 pairwise comparisons are significant.

It is interesting to compare the MAgP measure to standard MAP. First, the recall dimension in MAgP is identical to that used in standard document retrieval MAP (so in fact not generalized). Second, the precision dimension differs only slightly: in case of MAgP, the calculated article score represents a (partial) F-score, while in case of MAP the article score will always be an increment of 1. We compare how the relative systems ranking based on MAgP correlates with the systems ranking based on MAP. That is, we derive an article ranking and evaluate systems using standard MAP (`trec_eval`). The rank correlation (Kendall's tau) between MAP and MAgP is 0.674 over the top 20 official submissions.

Table 2 shows the extent to which higher ranked systems improve significantly over lower ranked system in terms of their MAP scores. Note that systems are labeled by their ranks based on the overall MAgP scores, so the order reflects the differences in the respective rankings. We see a few notable upsets: the system ranked 5th on MAgP is now ranked 1st on MAP, and the system ranked 20th on MAgP is now ranked 6th. This clearly shows that there are important differences between the tasks of document retrieval and RiC. Inspection of the table reveals that no less than 95 of the 190 pairwise comparisons are significant for the MAP measure.

Table 2: Significant improvements ($\star$) for MAP.

	5	1	2	3	6	20	4	13	8	9	7	10	11	12	14	17	16	15	19	18
5		-	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
1			-	-	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
2				-	-	-	-	-	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
3					-	-	-	-	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
6						-	-	-	-	-	-	-	⋆	⋆	-	⋆	⋆	⋆	⋆	⋆
20							-	-	-	-	-	-	-	-	-	⋆	⋆	⋆	⋆	⋆
4								-	-	-	-	-	-	-	-	⋆	⋆	⋆	⋆	⋆
13									-	-	-	-	-	-	-	⋆	-	⋆	⋆	⋆
8										-	⋆	-	-	-	-	-	-	⋆	⋆	⋆
9											-	⋆	-	-	-	-	-	-	⋆	⋆
7												-	-	-	-	-	-	-	-	⋆
10													-	-	-	-	-	-	⋆	⋆
11														⋆	-	⋆	-	-	-	⋆
12															-	⋆	-	-	-	⋆
14																-	-	-	-	⋆
17																	-	-	-	-
16																		-	-	-
15																			-	-
19																				-
18																				

When comparing the numbers of significant differences in Table 1 and Table 2, we see that MAgP is distinguishing more systems. This is perhaps not surprising since we are loosing information in the abstraction toward the article level needed for MAP.

4. CONCLUSIONS

Our main aim was to develop a measure for the evaluation of the Relevant in Context task, which represents a combination of document retrieval with XML element retrieval. We developed a measure that meets two main requirements: i) the score reflects the ranked list of articles inherent in the result list, and at the same time ii) the score also reflects how well the retrieved information per article corresponds to the relevant information.

The RiC task is very similar to the INEX assessors' task, who are highlighting relevant information in a pooled set of articles. Note that, since the assessors can highlight sentences and systems could only return XML elements, it will make it impossible for a system to obtain a perfect score of 1 (although the theoretical maximum will be close to 1). At INEX 2007, systems will be allowed to return arbitrary passages that can be directly evaluated by the MAgP measure, which in turn could enable a system to receive a perfect score when exactly matching the highlighted text.

Acknowledgments INEX is an activity of the Network of Excellence in Digital Libraries. Jaap Kamps was supported by the Netherlands Organization for Scientific Research (NWO) under project numbers 612.066.513, 639.072.601, and 640.001.501), and by the E.U.'s 6th FP for RTD (project MultiMATCH contract IST-033104).

REFERENCES

[1] C. Clarke, J. Kamps, and M. Lalmas. INEX 2006 retrieval task and result submission specification. In N. Fuhr, M. Lalmas, and A. Trotman, editors, *INEX 2006 Workshop Pre-Proceedings*, pages 381–388, 2006.

[2] INEX. INitiative for the Evaluation of XML retrieval, 2007. `http://inex.is.informatik.uni-duisburg.de/`.

[3] J. Kekäläinen and K. Järvelin. Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, 53:1120–1129, 2002.

[4] A. Tombros, B. Larsen, and S. Malik. Report on the INEX 2004 interactive track. *SIGIR Forum*, 39:43–49, 2005.

[2] Note that this contains runs from 10 participants in total, including some very close variants of the same system.