Evaluating Focused Retrieval Tasks

Jovan Pehcevski[1] and James A. Thom[2]

[1]AxIS Project Team
INRIA Rocquencourt, France
jovan.pehcevski@inria.fr

[2]School of Computer Science and Information Technology
RMIT University, Melbourne, Australia
james.thom@rmit.edu.au

Overview

- Focused retrieval
- Taxonomy of text retrieval tasks
- Evaluation framework
- Fidelity tests
- Discussion
- Q&A session

Focused retrieval

- Focused retrieval, including question answering, passage retrieval, and XML element retrieval, investigates ways to provide users with direct access to relevant information in retrieved documents
- Evaluating focused retrieval is a challenging task
  - different retrieval techniques typically produce answers of various sizes and granularity
  - there is a need for common evaluation framework where different aspects of focused retrieval can be consistently measured and compared

INEX

- The INitiative for the Evaluation of XML retrieval (INEX) has studied different aspects of focused retrieval since 2002
  - by mainly considering XML element retrieval techniques that can effectively retrieve information from structured document collections
  - by introducing different focused retrieval tasks, such as the *in context* tasks
  - by using a highlighting assessment procedure to gather relevance assessments for the retrieval topics

In context tasks

- *Relevant in context*: retrieve relevant documents, and identify the set of non-overlapping document parts representing the relevant information within each document
- *Best in context*: retrieve relevant documents, and identify the best entry point for starting to read the relevant information within each document
- The *in context* tasks correspond to end-user tasks, where focused retrieval answers are grouped per document, in their original document order
  - interactive experiments and user studies carried out within and outside INEX provide support for these tasks
  - the tasks loosely correspond to the INEX highlighting assessment procedure

Evaluation

- How to evaluate the *in context* tasks of focused retrieval?
- Two main requirements
  - the score should reflect the ranked list of documents inherent in the result list
  - the score should also reflect how well the retrieved information per document corresponds to the relevant information
- We want to use measures that directly exploit the INEX highlighting assessment procedure, and that are:
  - simple and easy to interpret
  - natural extensions of the well-established measures used in traditional information retrieval

Evaluation ...

Robertson's compatibility argument

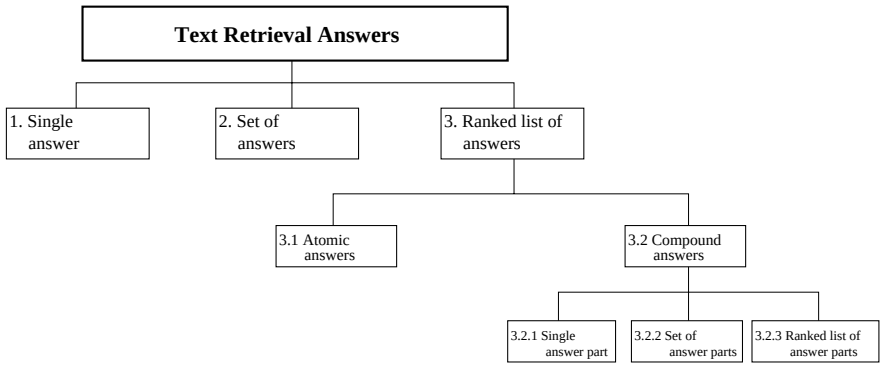
"*[...] there is a strong compatibility argument for researchers to use the same methods as each other unless there is very good reason to depart from the norm.*"

S.E. Robertson. Evaluation in information retrieval. In *ESSIR Proceedings*, p. 81–92, 2001.

Taxonomy of text retrieval tasks

- We present a taxonomy of text retrieval tasks based on the structure of the *answers* required by a task
- We also discuss some *assumptions* associated to a task, which model what users actually prefer
- These assumptions, together with the answer structure, define a retrieval task and influence how it should be evaluated

Answers



Assumptions

- Basic assumption: *Users want to see as much relevant information as possible with as little irrelevant information as possible*
- This basic assumption is not sufficient to determine how best to evaluate most text retrieval tasks
- We need to make further assumptions about what users actually prefer. For example:
  - A1: *Users consider all answers to be equally useful*
  - A2: *Users consider longer more detailed answers to be more useful than shorter answers*

Evaluation framework

- We present an evaluation framework for the *in context* tasks of focused retrieval
- The framework focuses on the compound answers shown in the taxonomy
- The evaluation of the *in context* tasks:
  - calculates scores for ranked lists of documents, where
  - the score per document reflects how well the retrieved information corresponds to the relevant information in the document

Score per document

- Three scores per document $S(d)$ could be calculated, depending on whether a single answer part, a set of answer parts, or a ranked list of answer parts are retrieved from the document
- We focus on the case where a set of non-overlapping answer parts is retrieved
- For a returned document, the text identified by the selected set of retrieved parts is compared to the text highlighted by the assessor

Score per document ...

- We calculate the following:
  - Precision $P(d)$, as the fraction of retrieved text (in characters) that is highlighted for the document
  - Recall $R(d)$, as the fraction of highlighted text (in characters) that is retrieved for the document
  - F-Score $F(d)$, as the combination of precision and recall using their harmonic mean
- We use the F-score as an appropriate document score for the case where a set of non-overlapping answer parts is retrieved:

$$S(d) = F(d)$$

Scores for ranked list of documents

- Over the ranked list of documents, we calculate the following:
  - generalized Precision $gP[r]$, as the sum of document scores up to a document-rank r, divided by the rank r:

$$gP[r] = \frac{\sum_{j=1}^{r} S(d_j)}{r} \quad (1)$$

Scores for ranked list of documents ...

- generalized Recall $gR[r]$, as the number of relevant documents retrieved up to a document-rank r, divided by the total number of relevant documents (modelling assumption A1):

$$gR[r] = \frac{\sum_{j=1}^{r} rel(d_j)}{Nrel} \quad (2)$$

- Average generalized Precision AgP (modelling assumption A1):

$$AgP = \sum_{r=1}^{|\mathcal{D}|} \frac{1}{Nrel} \cdot rel(d_r) \cdot gP[r] \quad (3)$$

Scores for ranked list of documents ...

- generalized Recall $gR'[r]$, as the amount of relevant text retrieved up to a document-rank r, divided by the total amount of relevant text highlighted for the topic (modelling assumption A2):

$$gR'[r] = \frac{\sum_{j=1}^{r} rsize(d_j)}{Trel} \quad (4)$$

- Average generalized Precision AgP' (modelling assumption A2):

$$AgP' = \sum_{r=1}^{|\mathcal{D}|} \frac{rsize(d_r)}{Trel} \cdot rel(d_r) \cdot gP[r] \quad (5)$$

Scores for ranked list of documents ...

- Traditional information retrieval (IR) measures:
  - Precision $P[r]$, as the fraction of retrieved relevant documents up to a document-rank r:

$$P[r] = \frac{\sum_{j=1}^{r} rel(d_j)}{r} \quad (6)$$

  - Average Precision AP, as the average of the precisions calculated at natural recall levels:

$$AP = \sum_{r=1}^{|\mathcal{D}|} \frac{1}{Nrel} \cdot rel(d_r) \cdot P[r] \quad (7)$$

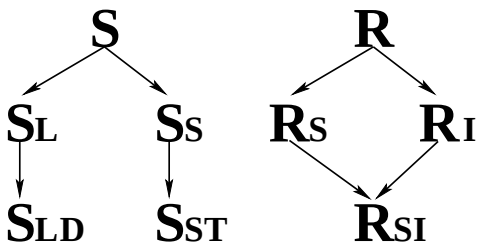
Fidelity tests

- Fidelity tests are designed to assess whether evaluation measures indeed measure what they are supposed to measure
- Simulated runs constructed in a controlled way are typically used to determine the fidelity of an evaluation measure
- Depending on the retrieval task, the best retrieval performance should be achieved by using the right (and desired) answer granularity, while preserving a reasonable relative ordering of the other simulated runs

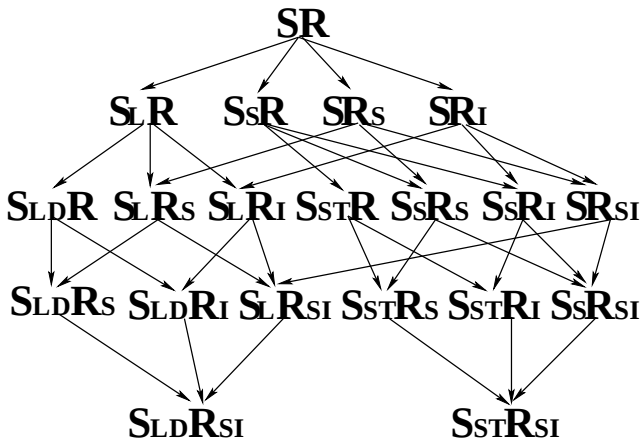
Simulated runs

- For the *in context* retrieval tasks, there are two dimensions that we need to consider within the overall space of possible runs:
  - runs with different amounts of relevant and non-relevant information in the set of passages/elements returned for each document (dimension S)
  - runs with different rankings of the documents (dimension R)
- For a given evaluation measure these two dimensions may interact in unexpected ways

Expected orderings per dimension



Expected orderings for the S-R space



Experimental results

- We use version 5.0 of the INEX 2006 relevance assessments, which contains a set of judgements for 114 topics from INEX 2006
- We analyse the run performances to separately investigate the expected orderings on each of the two dimensions, as well as on the complete S-R space
- We compare scores obtained with the three overall performance measures (*AgP*, *AgP'*, and *AP*)

Investigating the two dimensions

- Expected run orderings for the S dimension (different sets of answer parts returned for a document)
  - correctly captured by both *AgP* and *AgP'*, but not by *AP*
  - information is lost in the abstraction toward the document level needed for *AP*
- Expected run orderings for the R dimension (different document rankings)
  - correctly captured by *AgP*
  - the swap of the first two document ranks without inserting a non-relevant document at the top is not captured by *AP*
  - the swap of the first two document ranks after inserting a non-relevant document at the top is not captured by *AgP'*

Investigating the S-R space

- Expected run orderings for the S-R space
  - correctly captured by *AgP*
  - four notable disagreements between *AgP* and *AgP'* when comparing run pairs that insert non-relevant document at the top of their rankings
  - there are cases where the mean absolute performance differences obtained by *AgP'* are much larger than those obtained by *AgP*

Investigating the S-R space ...

Run ordering	AgP Diff (%)	AgP >	AgP ==	AgP <	AgP p	AgP′ Diff (%)	AgP′ >	AgP′ ==	AgP′ <	AgP′ p
$SR \rightarrow S_L R$	+20	112	2	0	2.2e-16	+17	112	2	0	2.2e-16
$SR \rightarrow S_S R$	+13	112	2	0	2.2e-16	+8	112	2	0	2.2e-16
$SR \rightarrow SR_S$	0	0	114	0	—	0	0	114	0	—
$SR \rightarrow SR_I$	+10	114	0	0	2.2e-16	+24	114	0	0	2.2e-16
$S_L R \rightarrow S_{LD} R$	+26	113	1	0	2.2e-16	+16	113	1	0	2.2e-16
$S_L R \rightarrow S_L R_S$	+0.07	52	13	49	0.6023	+0.5	52	13	49	0.2962
$S_L R \rightarrow S_L R_I$	+9	114	0	0	2.2e-16	+20	114	0	0	2.2e-16
$S_S R \rightarrow S_{ST} R$	+41	114	0	0	2.2e-16	+45	114	0	0	2.2e-16
$S_S R \rightarrow S_S R_S$	+0.07	43	29	42	0.4146	+0.5	43	29	42	0.0963
$S_S R \rightarrow S_S R_I$	+9	114	0	0	2.2e-16	+22	114	0	0	2.2e-16
$SR_S \rightarrow S_L R_S$	+20	112	2	0	2.2e-16	+17	112	2	0	2.2e-16
$SR_S \rightarrow S_S R_S$	+13	112	2	0	2.2e-16	+9	112	2	0	2.2e-16
$SR_S \rightarrow SR_{SI}$	+10	114	0	0	2.2e-16	+22	114	0	0	2.2e-16
$SR_I \rightarrow S_L R_I$	+18	112	2	0	2.2e-16	+14	112	2	0	2.2e-16
$SR_I \rightarrow S_S R_I$	+12	112	2	0	2.2e-16	+7	112	2	0	2.2e-16
$SR_I \rightarrow SR_{SI}$	0	0	114	0	—	−2	0	0	114	5.9e-13

Investigating the S-R space ...

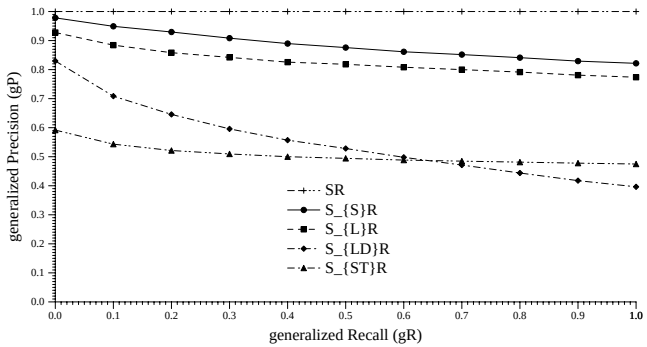
Run ordering	AgP Diff (%)	AgP >	AgP ==	AgP <	AgP p	AgP′ Diff (%)	AgP′ >	AgP′ ==	AgP′ <	AgP′ p
$S_{LD} R \rightarrow S_{LD} R_S$	+0.7	67	8	39	0.0004	+3	67	8	39	5.9e-05
$S_{LD} R \rightarrow S_{LD} R_I$	+7	114	0	0	2.2e-16	+18	114	0	0	2.2e-16
$S_L R_S \rightarrow S_{LD} R_S$	+27	113	1	0	2.2e-16	+18	113	1	0	2.2e-16
$S_L R_S \rightarrow S_L R_{SI}$	+9	114	0	0	2.2e-16	+18	114	0	0	2.2e-16
$S_L R_I \rightarrow S_{LD} R_I$	+24	113	1	0	2.2e-16	+14	113	1	0	2.2e-16
$S_L R_I \rightarrow S_L R_{SI}$	+0.03	52	13	49	0.6023	−1	25	0	89	2.4e-06
$S_{ST} R \rightarrow S_{ST} R_S$	+0.1	60	0	54	0.4904	+1	60	0	54	0.2141
$S_{ST} R \rightarrow S_{ST} R_I$	+5	114	0	0	2.2e-16	+11	114	0	0	2.2e-16
$S_S R_S \rightarrow S_{ST} R_S$	+41	114	0	0	2.2e-16	+45	114	0	0	2.2e-16
$S_S R_S \rightarrow S_S R_{SI}$	+9	114	0	0	2.2e-16	+20	114	0	0	2.2e-16
$S_S R_I \rightarrow S_{ST} R_I$	+36	114	0	0	2.2e-16	+34	114	0	0	2.2e-16
$S_S R_I \rightarrow S_S R_{SI}$	+0.03	43	29	42	0.4146	−1	12	0	102	1.9e-09
$SR_{SI} \rightarrow S_L R_{SI}$	+18	112	2	0	2.2e-16	+14	112	2	0	2.2e-16
$SR_{SI} \rightarrow S_S R_{SI}$	+12	112	2	0	2.2e-16	+7	112	2	0	2.2e-16
$S_{LD} R_S \rightarrow S_{LD} R_{SI}$	+6	114	0	0	2.2e-16	+15	114	0	0	2.2e-16
$S_{LD} R_I \rightarrow S_{LD} R_{SI}$	+0.4	67	8	39	0.0004	+0.05	46	0	68	0.8790
$S_L R_{SI} \rightarrow S_{LD} R_{SI}$	+24	113	1	0	2.2e-16	+15	113	1	0	2.2e-16
$S_{ST} R_S \rightarrow S_{ST} R_{SI}$	+5	114	0	0	2.2e-16	+10	114	0	0	2.2e-16
$S_{ST} R_I \rightarrow S_{ST} R_{SI}$	+0.05	60	0	54	0.4896	−1	48	0	66	0.0189
$S_S R_{SI} \rightarrow S_{ST} R_{SI}$	+36	114	0	0	2.2e-16	+35	114	0	0	2.2e-16

Discussion

- We use our findings to motivate a discussion about the following research topics:
  - the comparison between passage and element retrieval
  - the usefulness of focused and traditional document retrieval in identifying relevant information
  - the importance of modelling appropriate evaluation assumptions for a retrieval task

Passage versus element retrieval

- Perfect retrieval for the *relevant in context* task can only be achieved when retrieving all the highlighted passages within a document, in their exact size



Passage versus element retrieval ...

- The absolute performance difference between the passage run and our best simulated element run was 13%, which shows that no element run can achieve perfect retrieval
- One explanation for this could be that there is an inherent bias of the INEX highlighting assessment procedure towards passage retrieval
- How can passage and element retrieval be sensibly compared?

Passage versus element retrieval ...

- If there is an inherent bias towards passages, then this should be taken into account when comparing these two types of retrieval
- Two different sub-tasks could be identified to allow for sensible comparison:
  - *Passage retrieval sub-task*, where the retrieval answers are passages and it makes sense to compare whether element retrieval techniques help in identifying more relevant passages
  - *Element retrieval sub-task*, where the retrieval answers are XML elements and it makes sense to compare whether passage retrieval techniques help in identifying more relevant elements.

Focused versus traditional document retrieval

- Traditional IR measures, such as *AP*, cannot fully capture the level of detail required by focused retrieval
- More specifically, the *AP* measure partially captures the different ordering of documents in the result list, but it does not capture how well the retrieved information per document corresponds to the relevant information
- The average generalized precision *AgP* measure is able to fully capture both evaluation aspects, which makes it more useful than *AP* in measuring the retrieval performance
- On the INEX 2006 test collection, *AgP* is able to distinguish more significant performance differences than *AP*

Modelling evaluation assumptions

- Assumptions A1 and A2 are of particular importance for *in context* retrieval tasks, as it is not entirely clear which of the two assumptions should be preferred for evaluation
- Our fidelity tests demonstrate that the *AgP'* measure (based on assumption A2) is not entirely measuring what it is supposed to measure, and that the *AgP* measure (based on assumption A1) correctly captures the expected run orderings
- An argument for assumption A2 is that it motivates the preference given to more exhaustive answers in the evaluation

Modelling evaluation assumptions ...

- It may be possible that the current *AgP'* definition (shown in Equation 5) is not correctly modelling assumption A2
- Fixing this definition requires further investigation, which might be solved in one of these two ways:
  - interpolated average generalized precision could be used instead of the current non-interpolated definition
  - the current non-interpolated *AgP'* definition could be re-defined as follows:

$$AgP' = gR'[|\mathcal{D}|] \cdot \frac{\sum_{r=1}^{|\mathcal{D}|} rel(d_r) \cdot gP[r]}{\sum_{r=1}^{|\mathcal{D}|} rel(d_r)} \quad (8)$$

Modelling evaluation assumptions ...

- A more fundamental challenge, however, relates to the user preference of the two evaluation assumptions
- Would users regard a focused and more concise answer as more useful than a lengthy exposition?
- Or would they indeed perceive the answer that contains more relevant (and possibly repeating) information as more useful?
- We believe that it may be possible to determine the answers to these and similar questions either via user experiments or by questioning assessors about how they valued the answers for their topics

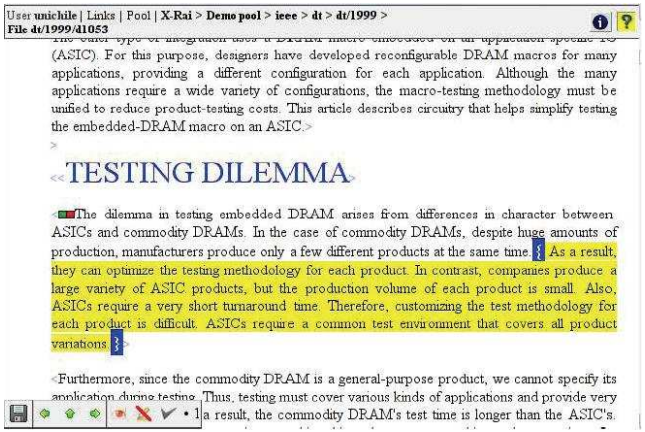
Questions?

Greetings from Versailles!

Appendix A: INEX highlighting assessment procedure

- Since 2005, a highlighting assessment procedure is used at INEX to gather relevance assessments
- Assessors are asked to highlight sentences representing the relevant information in a pooled set of documents
- The assessment tool automatically computes the relevance of the judged document parts (including full documents) as the ratio of highlighted to fully contained text
- The relevance values are drawn from a continuous [0,1] relevance scale

INEX highlighting assessment procedure ...



Appendix B: HiXEval evaluation scenarios

- An XML retrieval system may return a series of smaller elements that belong to a larger fully highlighted element, with the goal to boost the performance scores (overall and at selected rank cutoffs)
- We use two scenarios that allow us to perform a more detailed analysis of this (possibly undesirable) evaluation behaviour

HiXEval evaluation scenarios ...

- Let us assume that two systems, System A and System B, respectively retrieve the following ranked lists of elements:

Rank	System A	System B
1	`/article[1]/bdy[1]/sec[1]`	`/article[1]/bdy[1]/sec[1]/p[1]`
2	`/article[1]/bdy[1]/sec[2]`	`/article[1]/bdy[1]/sec[1]/p[2]`
3	`/article[1]/bdy[1]/sec[3]`	`/article[1]/bdy[1]/sec[1]/p[3]`

Scenario 1

- The recall-base contains only one fully highlighted section, which consists of three fully highlighted paragraphs:

```
<element path="/article[1]/bdy[1]/sec[1]"
size="99" rsize="99"/>
<element path="/article[1]/bdy[1]/sec[1]/p[1]"
size="33" rsize="33"/>
<element path="/article[1]/bdy[1]/sec[1]/p[2]"
size="33" rsize="33"/>
<element path="/article[1]/bdy[1]/sec[1]/p[3]"
size="33" rsize="33"/>
```

Scenario 2

- The recall-base contains two fully highlighted sections, and the first section consists of three highlighted paragraphs:

```
<element path="/article[1]/bdy[1]/sec[1]"
size="99" rsize="99"/>
<element path="/article[1]/bdy[1]/sec[1]/p[1]"
size="33" rsize="33"/>
<element path="/article[1]/bdy[1]/sec[1]/p[2]"
size="33" rsize="33"/>
<element path="/article[1]/bdy[1]/sec[1]/p[3]"
size="33" rsize="33"/>
<element path="/article[1]/bdy[1]/sec[2]"
size="99" rsize="99"/>
```


HiXEval performance scores

System	HiXEval measure			
	P@3	*R*@3	*F*@3	*AP*
Scenario 1				
A	0.33	**1.00**	0.50	**1.00**
B	**1.00**	**1.00**	**1.00**	**1.00**
Scenario 2				
A	0.67	**1.00**	**0.80**	**1.00**
B	**1.00**	0.50	0.67	**0.50**

Table: HiXEval performance scores for the two evaluation scenarios, obtained with rank cutoff and overall performance measures. Best scores under each HiXEval measure are shown in bold.

HiXEval performance scores ...

- The desired trade-off between retrieving as much relevant information as possible and not retrieving a substantial amount of non-relevant information is correctly captured by *AP* and *F*@*r* (to some extent), but not by *P*@*r* and *R*@*r*
- We currently normalise over the number of elements retrieved (the rank cutoff *r*), and not over the user effort required to reach that rank cutoff
- Future work: normalise over the amount of text returned instead of over the number of elements retrieved (that is, calculate Precision/Recall at number of characters read)
- How to determine the exact cutoff values?