Evaluating Focused Retrieval Tasks

Jovan Pehcevski, James A. Thom

Evaluating Focused Retrieval Tasks

Jovan Pehcevski
AxIS project team
INRIA-Rocquencourt, Le Chesnay, France
jovan.pehcevski@inria.fr

James A. Thom*
School of Computer Science and IT
RMIT University, Melbourne, Australia
james.thom@rmit.edu.au

ABSTRACT

Focused retrieval, identified by question answering, passage retrieval, and XML element retrieval, is becoming increasingly important within the broad task of information retrieval. In this paper, we present a taxonomy of text retrieval tasks based on the structure of the answers required by a task. Of particular importance are the *in context* tasks of focused retrieval, where not only relevant documents should be retrieved but also relevant information within each document should be correctly identified. Answers containing relevant information could be, for example, best entry points, or non-overlapping passages or elements. Our main research question is: How should the effectiveness of focused retrieval be evaluated? We propose an evaluation framework where different aspects of the *in context* focused retrieval tasks can be consistently evaluated and compared, and use fidelity tests on simulated runs to show what is measured. Results from our fidelity experiments demonstrate the usefulness of the proposed evaluation framework, and show its ability to measure different aspects and model different evaluation assumptions of focused retrieval.

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, In Context, Test collection, XML Retrieval

1. INTRODUCTION

Traditional information retrieval (IR) typically returns whole documents as answers, and leaves it up to users to locate the relevant information within each retrieved document. Focused retrieval [22], including question answering [23], passage retrieval [1, 2, 6, 24], and XML element retrieval [16], investigates ways to provide users with direct access to relevant information in retrieved documents. Evaluating focused retrieval is a challenging task since different retrieval techniques typically produce answers of various sizes and granularity, which calls for a common evaluation framework where different aspects of focused retrieval can be consistently measured and compared.

The INitiative for the Evaluation of XML retrieval (INEX) has studied different aspects of focused retrieval since 2002, by considering XML element retrieval techniques that can effectively retrieve information from structured document collections [16]. Since 2005, a highlighting assessment procedure is used at INEX to gather relevance assessments for the INEX retrieval topics [15]. In this procedure, assessors from the participating groups are asked to highlight sentences representing the relevant information in a pooled set of documents. An assessment program then computes the relevance of the judged elements (including whole documents) as the ratio of highlighted to fully contained text, where the element relevance values are drawn from a continuous scale in the range 0 to 1.

INEX 2006 introduced two new retrieval tasks, *relevant in context* and *best in context*, that combine document retrieval with XML element retrieval [4]. The *relevant in context* task is document retrieval with a twist, where not only the relevant documents should be retrieved, but also a set of non-overlapping XML elements representing the relevant information within each document should be correctly identified. The *best in context* task is similar, except that here systems are asked to return only one element per document, which corresponds to the best entry point for starting to read the relevant information in the document.

These two *in context* tasks correspond to end-user tasks 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 context. Moreover, the *in context* tasks loosely correspond to the assessment procedure used at INEX 2006, with the difference that the INEX assessors highlighted sentences whereas the systems only returned XML elements.

Interactive experiments at INEX [21], along with user studies carried out within and outside INEX [3, 9, 13], have also confirmed the usefulness of grouping the retrieved elements by their contained documents. The need for element grouping is mainly motivated by the fact that users not only want to locate more focussed information within a document, but they also want to "see what the document is" [3]. These findings justify the inclusion of the *in context* retrieval tasks at INEX, and highlight their importance in focused retrieval. In Section 2, we present a taxonomy for text retrieval tasks based on the structure of the answers required by a task, and discuss how it covers the *in context* tasks of focused retrieval.

How to evaluate the *in context* tasks of focused retrieval? There are two main requirements [10]: i) the score should reflect the ranked list of documents inherent in the result list, and ii) the score should also reflect how well the retrieved information per document corresponds to the relevant information. In Section 3, we propose an evaluation framework where different aspects of the *in context* focused retrieval tasks can be consistently evaluated and compared. To measure the extent to which text retrieval systems return relevant information, we design evaluation measures that consider the amount of highlighted text in relevant documents [17, 18]. Our proposal is motivated by the need to use measures that are simple

*This work was undertaken while James Thom was visiting the AxIS team at INRIA in 2007.

SIGIR 2007 Workshop on Focused Retrieval
July 27, 2007, Amsterdam, The Netherlands

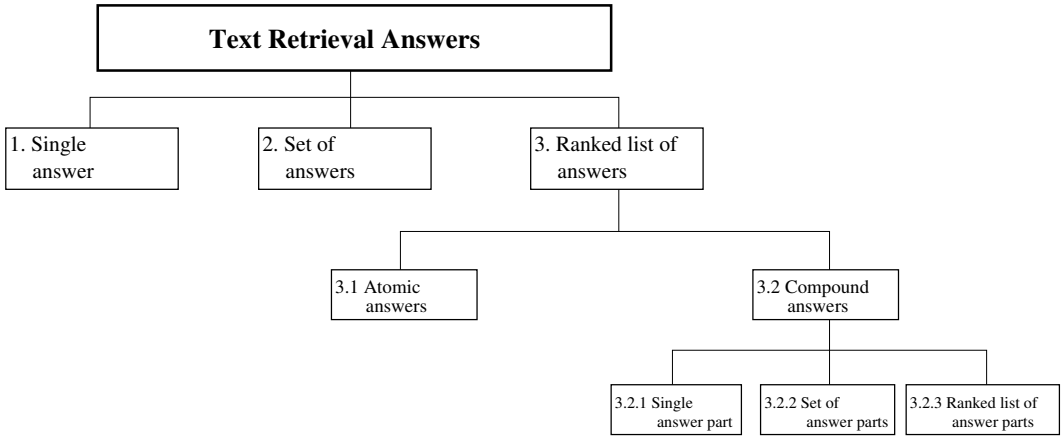

Figure 1: Taxonomy of text retrieval answers

and easy to interpret [7] and that are natural extensions of the well-established measures used in traditional information retrieval [20].

Since a variety of evaluation measures can be used to evaluate retrieval effectiveness, it is essential to carry out tests to determine whether they measure what are they intended to measure, and whether the reported evaluation scores can be trusted. Accordingly, two important tests are used to qualify the *evaluation* of evaluation measures: *fidelity* and *reliability* [23]. Simulated runs constructed in a controlled way are typically used to determine the *fidelity* of an evaluation measure [5, 11, 19]. In XML retrieval, these runs contain various granularity of elements in their answer lists (such as ideal elements, full document elements, or leaf elements). A measure successfully passes the fidelity test if the obtained evaluation scores demonstrate that the best retrieval performance is indeed achieved when using the right (and desired) answer granularity, while preserving a reasonable relative ordering of the other simulated runs. The results from our fidelity tests shown in Section 4 demonstrate the usefulness of the proposed evaluation framework, and its ability to measure different aspects and model different evaluation assumptions of focused retrieval.

We conclude this paper with our discussions in Section 5, where we use our findings to reflect on the comparison between passage and element retrieval, the usefulness of focused and traditional document retrieval in identifying relevant information, and the importance of choosing appropriate evaluation assumptions.

2. A TAXONOMY OF RETRIEVAL TASKS

In this section, we present a taxonomy of text retrieval tasks based on the structure of the answers required by a task. We only consider tasks where non-overlapping answers are allowed. We also discuss some assumptions about what users want; these assumptions, together with the answer structure, define a retrieval task and influence how it should be evaluated.

Answers

In text retrieval answers can include either or both documents (or equivalently document identifiers) and excerpts of documents. The excerpts could be passages (identified by start and end positions) or in the case of XML retrieval, elements (identified by XPath expressions). Furthermore, depending on the retrieval task, answers may be a single result, an unordered set of results, or a ranked list of results. This leads us to a partial taxonomy of tasks based on answers as shown in Figure 1. For each type of answer in the taxonomy (such as an atomic answer or a compound answer), we describe one or more text retrieval tasks that can be used to generate that particular answer. The taxonomy parts are explained as follows.

1. Single answer

   For tasks where the user is only interested in one document (or excerpt of a document) as an answer, such as in Google's "I'm Feeling Lucky™".

2. Set of answers

   For Boolean retrieval tasks where the user is interested in finding all matching documents (or excerpts).

3. Ranked list of answers

   3.1 Atomic answers

   For tasks where the answers are a ranked list of documents, such as a list of web pages found by a search engine, or a ranked list of elements as retrieved for the INEX *thorough* or *focused* tasks [4], or a ranked list of passages for the TREC *question answering* task [23].

   3.2 Compound answers

   For *in context* tasks where the result of a query is a ranked list of answers (usually documents) and clustered for each answer in the list, further information (answers parts) needs to retrieved from the document. These could be:

   3.2.1 Single answer part, such as the best entry point returned in the INEX *best in context* task [4] or text snippets returned as document summaries by search engines.

   3.2.2 Set of answer parts, such as the elements returned in the INEX *relevant in context* task [4] (in 2007 INEX will allow passages as well as elements).

   3.2.3 Ranked list of answer parts. It is conceivable that the answer parts could be returned as a sub-list of ranked elements, which could be represented by using a document heat-map.

This paper is concerned with evaluation of the last group of tasks, which are considered in more detail in the taxonomy. These are the *in context* tasks that are based on compound answers. Specifically, we consider the *relevant in context* task where the result of a query is a ranked list of answers documents, and, for each document in the answer list, a set of passages or elements is returned.

Assumptions

In defining a text retrieval task it is also necessary to define the assumptions about what the user is wanting to see. We make the following basic assumption about all text retrieval tasks:

Users want to see as much relevant information as possible with as little irrelevant information as possible. Such an assumption is the basis of methods for evaluating the effectiveness of information retrieval systems based on recall and precision.

This basic assumption is not sufficient to determine how best to evaluate most text retrieval tasks. For this, we need to make further assumptions about what users actually prefer, which for example we may choose to test via user experiments. These assumptions may depend on the type of retrieval task, as illustrated by the following examples.

1. *Users do not want to see the same (or similar) answers more than once.* This motivates the work behind evaluating aspectual retrieval [14], and influences the way commercial search engines present answers.

2. *Users want the shortest and the most complete answer.* This might be motivated by a question answering task where an answer needs to be seen in isolation, and it need not be required to provide any context.

3. *Users consider longer more detailed answers to be more useful than shorter answers.* This models users that prefer documents containing longer passages with more relevant information over documents containing shorter passages.

4. *Users consider all answers to be equally useful.* This models users that place equal value on each relevant document, and here documents with longer relevant passages are considered as equally useful as those with shorter passages.

The last two assumptions are likely to depend on the task. We explore these assumptions in more detail for the *relevant in context* task later in this paper.

3. EVALUATION FRAMEWORK

In this section, we describe an evaluation framework for the *in context* tasks of focused retrieval. The framework focuses on compound answers given in the taxonomy shown in Figure 1. The evaluation of the *in context* tasks calculates scores for ranked lists of documents, where per document we obtain a score reflecting how well the retrieved information corresponds to the relevant information in the document.

Score per document

Three different scores per document can 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 retrieved document, the text identified by the selected set of retrieved parts is compared to the text highlighted by the assessor [17, 18]. More formally, let d be the retrieved document, and let p be a part (element or passage) that belongs to $\mathcal{P}_d$, the set of retrieved parts from document d. Let $rsize(p)$ be the amount of highlighted relevant text contained by p (if there is no highlighted text, $rsize(p) = 0$). Let $size(p)$ be the total amount of text contained by p, and let $Trel(d)$ be the total amount of highlighted relevant text for the document d.

We calculate the following:

- Precision, as the fraction of retrieved text (in characters) that is highlighted:

$$P(d) = \frac{\sum_{p \in \mathcal{P}_d} rsize(p)}{\sum_{p \in \mathcal{P}_d} size(p)} \quad (1)$$

  The $P(d)$ measure ensures that, to achieve a high precision value for the document d, the set of retrieved parts for that document needs to contain as little non-relevant information as possible.

- Recall, as the fraction of highlighted text (in characters) that is retrieved:

$$R(d) = \frac{\sum_{p \in \mathcal{P}_d} rsize(p)}{Trel(d)} \quad (2)$$

  The $R(d)$ measure ensures that, to achieve a high recall value for the document d, the set of retrieved parts for that document needs to contain as much relevant information as possible.

- F-Score, as the combination of precision and recall using their harmonic mean, resulting in a score in [0,1] per document:

$$F(d) = \frac{2 \cdot P(d) \cdot R(d)}{P(d) + R(d)} \quad (3)$$

For retrieved non-relevant documents, all the above scores evaluate to zero: $P(d) = R(d) = F(d) = 0$.

We use the F-score as an appropriate document score for the case where a set of answer parts is retrieved: $S(d) = F(d)$. The resulting $S(d)$ score varies between 0 (document without relevance, or none of the relevance is retrieved) and 1 (all relevant text is retrieved without retrieving any non-relevant text).

Scores for ranked list of documents

We have a ranked list of documents $\mathcal{D}$, and for each document we have a document score $S(d_r) \in [0, 1]$, where d_r is the document retrieved at rank r ($1 \leq r \leq |\mathcal{D}|$). Hence, we need generalized evaluation measures, and we utilise the most straightforward generalization of precision and recall [12]. More formally, let us assume that for a retrieval topic there are in total $Nrel$ documents with relevance, and let us also assume that the function $rel(d_r) = 1$ if document d_r contains relevant information, and $rel(d_r) = 0$ otherwise. Let $rsize(d_r)$ be the amount of highlighted relevant text contained by d_r (if there is no highlighted text, $rsize(d_r) = 0$), and let $Trel$ be the total amount of highlighted relevant text for the retrieval topic (calculated across the $Nrel$ relevant 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:

$$gP[r] = \frac{\sum_{j=1}^{r} S(d_j)}{r} \quad (4)$$

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

$$gR[r] = \frac{\sum_{j=1}^{r} rel(d_j)}{Nrel} \quad (5)$$

The generalized Recall definition, as given in Equation 5, follows the assumption that each document with relevance is treated as equally relevant, and thus as equally useful to retrieve (Assumption 4). However, since the documents in the answer list are ranked in a descending order of their estimated likelihood of relevance, an alternative (and equally plausible) assumption would be that documents with more highlighted relevant text should be considered to be more relevant (and therefore more useful to retrieve) than documents with less highlighted text (Assumption 3). To model this evaluation assumption, we use the alternative generalized Recall definition shown in Equation 6:

$$gR'[r] = \frac{\sum_{j=1}^{r} rsize(d_j)}{Trel} \quad (6)$$

These generalized measures are compatible with the standard precision/recall measures used in traditional information retrieval. Specifically, the Average generalized Precision for a retrieval topic can be calculated by averaging the generalized Precisions at natural recall points where generalized Recall increases (the generalized Precision of non-retrieved relevant documents is 0).

A consequence of introducing two generalized Recall definitions ($gR[r]$ and $gR'[r]$) is that two Average generalized Precision definitions need to be respectively used in calculating the overall performance score: AgP, which uses $gR[r]$ and is shown in Equation 7; and AgP', which uses $gR'[r]$ and is shown in Equation 8.

$$AgP = \sum_{j=1}^{|\mathcal{D}|} \frac{1}{Nrel} \cdot rel(d_j) \cdot gP[j] \quad (7)$$

$$AgP' = \sum_{j=1}^{|\mathcal{D}|} \frac{rsize(d_j)}{Trel} \cdot rel(d_j) \cdot gP[j] \quad (8)$$

When looking at a set of topics, the Mean Average generalized Precision ($MAgP$ or $MAgP'$) is simply the mean of the Average generalized Precision scores per topic.

Traditional IR measures

The traditional IR measures treat each retrieved document as either relevant or not, and therefore assign a binary score per document: $S(d_r) = rel(d_r)$. Over the ranked list of documents, we use the following traditional IR measures:

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

$$P[r] = \frac{\sum_{j=1}^{r} rel(d_j)}{r} \quad (9)$$

- Recall ($R[r]$), as the fraction of relevant documents retrieved up to a document-rank r (which is the same as the generalized Recall definition given in Equation 5), and

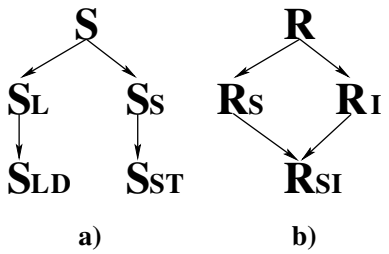

Figure 2: Expected orderings for runs of a dimension of the S–R space: (a) different sets of parts (b) different document rankings

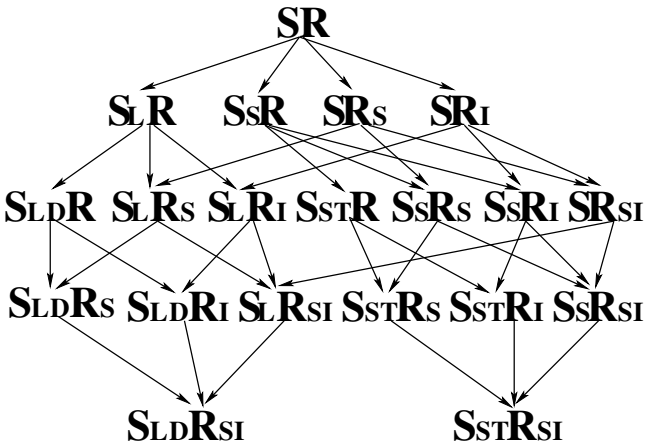

Figure 3: Expected orderings for runs of the S–R space

- Average Precision (AP), as the average of the precisions calculated at natural recall points:

$$AP = \sum_{j=1}^{|\mathcal{D}|} \frac{1}{Nrel} \cdot rel(d_j) \cdot P[j] \quad (10)$$

For a set of topics, the Mean Average Precision (MAP) is simply the mean of the Average Precision scores per topic.

4. FIDELITY TESTS

Fidelity tests should be designed to assess whether evaluation measures indeed measure what they are supposed to measure. In testing the fidelity of evaluation measures for *in context* retrieval, where there are sets of passages/elements returned for each document in the ranked list, 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, and
- runs with different rankings of the documents.

For a given evaluation measure these two dimensions may interact in unexpected ways.

Simulated runs and expected orderings

We designed the following suite of simulated runs that took the two dimensions into account.

Run	$gP[r]$			$gR[r]$			$MAgP$	$gR'[r]$			$MAgP'$	MAP
	1	2	10	1	2	10		1	2	10		
$\mathbf{SR}$	1.0000	1.0000	0.9763	0.0419	0.0838	0.3722	1.0000	0.2403	0.3661	0.7194	1.0000	1.0000
$\mathbf{SR_S}$	1.0000	1.0000	0.9763	0.0419	0.0838	0.3722	1.0000	0.2403	0.3661	0.7194	1.0000	1.0000
$\mathbf{SR_I}$	0.0000	0.5000	0.8833	0.0000	0.0419	0.3420	0.8954	0.0000	0.2403	0.6952	0.7647	0.8954
$\mathbf{SR_{SI}}$	0.0000	0.5000	0.8833	0.0000	0.0419	0.3420	0.8954	0.0000	0.1258	0.6952	0.7838	0.8954
$\mathbf{S_LR}$	0.8584	0.8506	0.7830	0.0419	0.0838	0.3722	0.7976	0.2403	0.3661	0.7194	0.8314	1.0000
$\mathbf{S_LR_S}$	0.8427	0.8506	0.7830	0.0419	0.0838	0.3722	0.7969	0.1258	0.3661	0.7194	0.8262	1.0000
$\mathbf{S_LR_I}$	0.0000	0.4292	0.7136	0.0000	0.0419	0.3420	0.7113	0.0000	0.2403	0.6952	0.6289	0.8954
$\mathbf{S_LR_{SI}}$	0.0000	0.4213	0.7136	0.0000	0.0419	0.3420	0.7110	0.0000	0.1258	0.6952	0.6428	0.8954
$\mathbf{S_{LD}R}$	0.7935	0.7280	0.5664	0.0419	0.0838	0.3722	0.5352	0.2403	0.3661	0.7194	0.6719	1.0000
$\mathbf{S_{LD}R_S}$	0.6624	0.7280	0.5664	0.0419	0.0838	0.3722	0.5278	0.1258	0.3661	0.7194	0.6422	1.0000
$\mathbf{S_{LD}R_I}$	0.0000	0.3968	0.5241	0.0000	0.0419	0.3420	0.4700	0.0000	0.2403	0.6952	0.4931	0.8954
$\mathbf{S_{LD}R_{SI}}$	0.0000	0.3312	0.5241	0.0000	0.0419	0.3420	0.4664	0.0000	0.1258	0.6952	0.4926	0.8954
$\mathbf{S_SR}$	0.9578	0.9489	0.8693	0.0419	0.0838	0.3722	0.8687	0.2403	0.3661	0.7194	0.9194	1.0000
$\mathbf{S_SR_S}$	0.9400	0.9489	0.8693	0.0419	0.0838	0.3722	0.8680	0.1258	0.3661	0.7194	0.9140	1.0000
$\mathbf{S_SR_I}$	0.0000	0.4789	0.7905	0.0000	0.0419	0.3420	0.7742	0.0000	0.2403	0.6952	0.6966	0.8954
$\mathbf{S_SR_{SI}}$	0.0000	0.4700	0.7905	0.0000	0.0419	0.3420	0.7739	0.0000	0.1258	0.6952	0.7117	0.8954
$\mathbf{S_{ST}R}$	0.4942	0.4715	0.4518	0.0419	0.0838	0.3722	0.4589	0.2403	0.3661	0.7194	0.4722	1.0000
$\mathbf{S_{ST}R_S}$	0.4488	0.4715	0.4518	0.0419	0.0838	0.3722	0.4578	0.1258	0.3661	0.7194	0.4660	1.0000
$\mathbf{S_{ST}R_I}$	0.0000	0.2469	0.4118	0.0000	0.0419	0.3420	0.4093	0.0000	0.2403	0.6952	0.3578	0.8954
$\mathbf{S_{ST}R_{SI}}$	0.0000	0.2243	0.4118	0.0000	0.0419	0.3420	0.4088	0.0000	0.1258	0.6952	0.3642	0.8954

Table 1: Performance scores for simulated runs of the S–R space, obtained with different measures using the 114 INEX 2006 topics. The runs are grouped in five clusters, depending on the answer parts retrieved ($\mathbf{S}$, $\mathbf{S_L}$, $\mathbf{S_{LD}}$, $\mathbf{S_S}$, $\mathbf{S_{ST}}$).

The first dimension for the simulated runs covers the set of elements/passages returned for each document. We considered five different sets:

$\mathbf{S}$ the set of non-overlapping passages that are highlighted as relevant by the assessor;

$\mathbf{S_L}$ for each passage in $\mathbf{S}$ return the smallest element containing the passage, that is an element which is larger than (or equal in size to) the passage;

$\mathbf{S_{LD}}$ return the whole document;

$\mathbf{S_S}$ for each passage in $\mathbf{S}$ return the largest non-overlapping elements fully contained within the passage, that is one or more elements which are smaller than (or one element equal in size to) the passage; and

$\mathbf{S_{ST}}$ for each passage in $\mathbf{S}$ return the smallest elements fully contained within the passage that do not contain any sub-elements.

The expected ordering of these runs is shown in Figure 2(a).

The second dimension for the simulated runs covers different document rankings. We considered four different rankings:

$\mathbf{R}$ in order of decreasing relevant information from the document containing the most relevant information (that is the most text highlighted as relevant by an assessor) to the document containing the least;

$\mathbf{R_S}$ same as ranking $\mathbf{R}$ but with the first two documents swapped;

$\mathbf{R_I}$ same as ranking $\mathbf{R}$ but with a document containing no relevant information inserted at the start of the list; and

$\mathbf{R_{SI}}$ same as ranking $\mathbf{R_S}$ but after swapping the first two documents, a document containing no relevant information is inserted at the start of the list.

The expected ordering of these runs is shown in Figure 2(b). This ordering is based on the evaluation measure addressing the assumption that users want longer more detailed answers in preference to shorter answers.

As we are interested in how these two dimensions interact, we combine the runs in an $\mathbf{S}$–$\mathbf{R}$ space as shown in Figure 3, which gives the expected ordering of the various combinations of the two dimensions.

For example, the run $\mathbf{SR}$ corresponds to returning as answers the documents in the order from the document with the most text highlighted as relevant to the document with the least text highlighted as relevant ($\mathbf{R}$), and for each document only returning as answer parts those passages corresponding to all the highlighted text ($\mathbf{S}$). This run $\mathbf{SR}$ should be perfect retrieval (under most assumptions), and no other run should perform better than $\mathbf{SR}$ for any topic (even though for some assumptions they may perform as well as this run).

As other examples, the runs $\mathbf{S_{ST}R_{SI}}$ and $\mathbf{S_{LD}R_{SI}}$ correspond to returning the following as answers: a document containing no relevant text, followed by the document containing second highest amount of relevant, followed by remaining documents in order of most to least highlighted text ($\mathbf{R_{SI}}$). In the $\mathbf{S_{ST}R_{SI}}$ run each document in the list contains as parts of an answer only the (too small) elements within the highlighted passages, that is elements with no other elements nested within them ($\mathbf{S_{ST}}$). In the $\mathbf{S_{LD}R_{SI}}$ run the whole document is returned as the only answer part ($\mathbf{S_{LD}}$). As illustrated in Figure 3, of all the runs we consider, we would expect one or both of these two runs to be the worst performing.

Experimental results

We now present experimental results for the simulated runs of the $\mathbf{S}$–$\mathbf{R}$ space. We use version 5.0 of the INEX 2006 relevance assessments, which contains a set of judgements for 114 topics from INEX 2006.

Table 1 shows performance scores obtained with different evaluation measures on the 114 INEX 2006 topics. We base our analysis

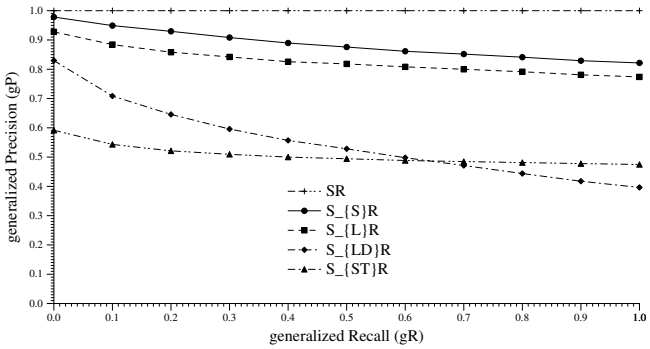

Figure 4: Evaluation of the overall performance of five simulated runs of the S-R space, using a fixed document ranking (R). The graph shows values for interpolated generalized precision (gP) at 11pt generalized recall (gR).

on the results obtained with the three overall performance measures (*MAgP*, *MAgP'*, and *MAP*), although results obtained with the three rank cutoff measures (*gP[r]*, *gR[r]*, and *gR'[r]*) are also reported. The runs are grouped in five clusters, depending on the answer parts retrieved ($\mathbf{S}$, $\mathbf{S_L}$, $\mathbf{S_{LD}}$, $\mathbf{S_S}$, or $\mathbf{S_{ST}}$).

Several observations can be drawn from these results.

First, when analysing the performance differences of runs with a fixed document ranking, we aim at separately investigating the first dimension of the $\mathbf{S}$–$\mathbf{R}$ space (different sets of parts). The expected orderings for this dimension are correctly captured by both *MAgP* and *MAgP'*, but not by *MAP*. This is perhaps not surprising, since we are losing information in the abstraction toward the document level needed for MAP. Figure 4 shows an 11 point interpolated recall/precision graph plots for five simulated runs containing different sets of parts. Our initial expectations are confirmed: the passage run $\mathbf{S}$ results in perfect retrieval, and no other element run performs better than this run; returning $\mathbf{S_L}$ elements that fully contain the highlighted passages results in better performance than returning whole documents ($\mathbf{S_{LD}}$); and returning larger fully highlighted elements ($\mathbf{S_S}$) results in better performance than returning smaller fully highlighted elements ($\mathbf{S_{ST}}$). Although we did not initially speculate about the expected ordering between $\mathbf{S_S}$ and $\mathbf{S_L}$, both Figure 4 and the scores in Table 1 show that, for the INEX 2006 topic set, returning larger fully highlighted elements ($\mathbf{S_S}$) seems to be a better retrieval strategy than returning elements that fully contain the highlighted passages ($\mathbf{S_L}$).

Second, when analysing the performance differences of runs in each cluster, we aim at separately investigating the second dimension of the $\mathbf{S}$–$\mathbf{R}$ space (different document rankings). As expected, we observe that the first run of each cluster, which ranks documents in a descending order of their contained relevant information ($\mathbf{R}$), either outperforms or performs as well as the other runs in the same cluster, irrespective of the overall performance measure used. The case of inserting a non-relevant document at the top of the ranking ($\mathbf{R}$ versus $\mathbf{R_I}$ and $\mathbf{R_S}$ versus $\mathbf{R_{SI}}$) is also correctly captured by the three measures; however, the swap of the first two document ranks ($\mathbf{R}$ versus $\mathbf{R_S}$) is correctly captured only by *MAgP* and *MAgP'*, but not by *MAP*. We also observe a (somewhat unexpected) behaviour for the *MAgP'* measure when comparing $\mathbf{R_I}$ with the $\mathbf{R_{SI}}$ document ranking. Our initial expectation was that the $\mathbf{R_I}$ ranking would perform at least as good as its swapped counterpart $\mathbf{R_{SI}}$, which is indeed correctly captured by *MAgP* and *MAP*. However, for all but the third $\mathbf{S_{LD}}$ cluster *MAgP'* captures the exact opposite performance behaviour. These results therefore suggest that *MAgP'* is not as reliable as *MAgP*, which seems to correctly capture the expected run orderings for the second (as well as the first) dimension of the $\mathbf{S}$–$\mathbf{R}$ space.

Last, in order to reflect the interaction between the two dimensions in the $\mathbf{S}$–$\mathbf{R}$ space, we perform a per-topic analysis to investigate whether the expected run orderings (shown in Figure 3) are correctly captured by the two overall performance measures, *AgP* and *AgP'*. Table 2 shows the results of this analysis. For an expected run ordering (a row in the table), we report the following values: mean absolute difference between the run performances (*Diff*, in percentage); the number of topics (of the total 114 INEX 2006 topics) where the first run performs better ($>$), is equal to ($==$), or performs worse ($<$) than the second run; and the actual t-test p values used to check if the mean absolute performance differences are statistically significant. The general trend among these results is clear: *AgP* is capable of correctly capturing the expected run orderings of the simulated runs in the $\mathbf{S}$–$\mathbf{R}$ space, where for each comparison among the run pairs (the rows in the table), the first run performs better or as good as the second run. We also observe four notable disagreements between *AgP* and *AgP'* when comparing run pairs that insert non-relevant document at the top of their rankings (the rows containing negative *AgP' Diff* numbers for mean absolute performance differences). As discussed previously, *AgP'* fails to correctly capture the expected run orderings after a non-relevant document is inserted at the top of the ranking.[1] However, we also observe that there are cases where the mean absolute performance differences obtained by *AgP'* are much larger than those obtained by *AgP*, which is especially true when comparing $\mathbf{R}$–>$\mathbf{R_I}$ and $\mathbf{R_S}$–>$\mathbf{R_{SI}}$ run orderings. This suggests that, even though the fidelity tests demonstrate that it is not as capable as *AgP* at capturing the expected behaviour, there may be cases where the *AgP'* measure is likely to be more sensitive than *AgP* at distinguishing between different retrieval approaches.

5. DISCUSSION AND CONCLUSIONS

In this section, we use our findings from the previous section 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; and the importance of modelling appropriate evaluation assumptions for a retrieval task.

Passage versus element retrieval

The results of our fidelity tests in Section 4 demonstrate that 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. The absolute difference in *MAgP* scores between the passage and our best simulated element run was 13%, which shows that no element run can achieve perfect retrieval (although the score achieved by the perfect element run could be higher than the one achieved by our best element run). One explanation for this could be that there is an inherent bias of the highlighting assessment procedure towards passage retrieval, since assessors are allowed to highlight sentences which could span across or even be contained within element boundaries.

How can passage and element retrieval be sensibly compared? If there is an inherent bias towards passages, then this should be taken into account when comparing these two types of retrieval.

[1] Although *AgP'* may correctly capture the expected run orderings when a non-relevant document is inserted after the first highly ranked document.

Run ordering	AgP					AgP'				
	Diff (%)	$>$	$==$	$<$	p	*Diff* (%)	$>$	$==$	$<$	p
$\mathbf{SR}$–>$\mathbf{S_L R}$	+20	112	2	0	2.2e-16	+17	112	2	0	2.2e-16
$\mathbf{SR}$–>$\mathbf{S_S R}$	+13	112	2	0	2.2e-16	+8	112	2	0	2.2e-16
$\mathbf{SR}$–>$\mathbf{SR_S}$	0	0	114	0	—	0	0	114	0	—
$\mathbf{SR}$–>$\mathbf{SR_I}$	+10	114	0	0	2.2e-16	+24	114	0	0	2.2e-16
$\mathbf{S_L R}$–>$\mathbf{S_{LD} R}$	+26	113	1	0	2.2e-16	+16	113	1	0	2.2e-16
$\mathbf{S_L R}$–>$\mathbf{S_L R_S}$	+0.07	52	13	49	0.6023	+0.5	52	13	49	0.2962
$\mathbf{S_L R}$–>$\mathbf{S_L R_I}$	+9	114	0	0	2.2e-16	+20	114	0	0	2.2e-16
$\mathbf{S_S R}$–>$\mathbf{S_{ST} R}$	+41	114	0	0	2.2e-16	+45	114	0	0	2.2e-16
$\mathbf{S_S R}$–>$\mathbf{S_S R_S}$	+0.07	43	29	42	0.4146	+0.5	43	29	42	0.0963
$\mathbf{S_S R}$–>$\mathbf{S_S R_I}$	+9	114	0	0	2.2e-16	+22	114	0	0	2.2e-16
$\mathbf{SR_S}$–>$\mathbf{S_L R_S}$	+20	112	2	0	2.2e-16	+17	112	2	0	2.2e-16
$\mathbf{SR_S}$–>$\mathbf{S_S R_S}$	+13	112	2	0	2.2e-16	+9	112	2	0	2.2e-16
$\mathbf{SR_S}$–>$\mathbf{SR_{SI}}$	+10	114	0	0	2.2e-16	+22	114	0	0	2.2e-16
$\mathbf{SR_I}$–>$\mathbf{S_L R_I}$	+18	112	2	0	2.2e-16	+14	112	2	0	2.2e-16
$\mathbf{SR_I}$–>$\mathbf{S_S R_I}$	+12	112	2	0	2.2e-16	+7	112	2	0	2.2e-16
$\mathbf{SR_I}$–>$\mathbf{SR_{SI}}$	0	0	114	0	—	−2	0	0	114	5.9e-13
$\mathbf{S_{LD} R}$–>$\mathbf{S_{LD} R_S}$	+0.7	67	8	39	0.0004	+3	67	8	39	5.9e-05
$\mathbf{S_{LD} R}$–>$\mathbf{S_{LD} R_I}$	+7	114	0	0	2.2e-16	+18	114	0	0	2.2e-16
$\mathbf{S_L R_S}$–>$\mathbf{S_{LD} R_S}$	+27	113	1	0	2.2e-16	+18	113	1	0	2.2e-16
$\mathbf{S_L R_S}$–>$\mathbf{S_L R_{SI}}$	+9	114	0	0	2.2e-16	+18	114	0	0	2.2e-16
$\mathbf{S_L R_I}$–>$\mathbf{S_{LD} R_I}$	+24	113	1	0	2.2e-16	+14	113	1	0	2.2e-16
$\mathbf{S_L R_I}$–>$\mathbf{S_L R_{SI}}$	+0.03	52	13	49	0.6023	−1	25	0	89	2.4e-06
$\mathbf{S_{ST} R}$–>$\mathbf{S_{ST} R_S}$	+0.1	60	0	54	0.4904	+1	60	0	54	0.2141
$\mathbf{S_{ST} R}$–>$\mathbf{S_{ST} R_I}$	+5	114	0	0	2.2e-16	+11	114	0	0	2.2e-16
$\mathbf{S_S R_S}$–>$\mathbf{S_{ST} R_S}$	+41	114	0	0	2.2e-16	+45	114	0	0	2.2e-16
$\mathbf{S_S R_S}$–>$\mathbf{S_S R_{SI}}$	+9	114	0	0	2.2e-16	+20	114	0	0	2.2e-16
$\mathbf{S_S R_I}$–>$\mathbf{S_{ST} R_I}$	+36	114	0	0	2.2e-16	+34	114	0	0	2.2e-16
$\mathbf{S_S R_I}$–>$\mathbf{S_S R_{SI}}$	+0.03	43	29	42	0.4146	−1	12	0	102	1.9e-09
$\mathbf{SR_{SI}}$–>$\mathbf{S_L R_{SI}}$	+18	112	2	0	2.2e-16	+14	112	2	0	2.2e-16
$\mathbf{SR_{SI}}$–>$\mathbf{S_S R_{SI}}$	+12	112	2	0	2.2e-16	+7	112	2	0	2.2e-16
$\mathbf{S_{LD} R_S}$–>$\mathbf{S_{LD} R_{SI}}$	+6	114	0	0	2.2e-16	+15	114	0	0	2.2e-16
$\mathbf{S_{LD} R_I}$–>$\mathbf{S_{LD} R_{SI}}$	+0.4	67	8	39	0.0004	+0.05	46	0	68	0.8790
$\mathbf{S_L R_{SI}}$–>$\mathbf{S_{LD} R_{SI}}$	+24	113	1	0	2.2e-16	+15	113	1	0	2.2e-16
$\mathbf{S_{ST} R_S}$–>$\mathbf{S_{ST} R_{SI}}$	+5	114	0	0	2.2e-16	+10	114	0	0	2.2e-16
$\mathbf{S_{ST} R_I}$–>$\mathbf{S_{ST} R_{SI}}$	+0.05	60	0	54	0.4896	−1	48	0	66	0.0189
$\mathbf{S_S R_{SI}}$–>$\mathbf{S_{ST} R_{SI}}$	+36	114	0	0	2.2e-16	+35	114	0	0	2.2e-16

Table 2: Comparison of AgP and AgP' scores of expected run orderings in the S–R space, using the 114 INEX 2006 topics. For each expected run ordering, a row shows the mean absolute performance difference (*Diff*), the number of topics where the first run performs better ($>$), is equal to ($==$), or performs worse ($<$) than the second run, and the t-test p value.

Accordingly, two different sub-tasks could be identified that allow a sensible comparison between passage and element retrieval:

- A *passage retrieval sub-task*, where the retrieval answers are passages and it makes sense to compare whether element retrieval techniques (based on the underlying XML structure) help in identifying more relevant passages; and
- An *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 [8].

The evaluation measures proposed in this paper could be consistently used for evaluation of both sub-tasks.

Focused versus traditional document retrieval

The results of our fidelity tests in Section 4 demonstrate that the traditional IR measures, such as *MAP*, cannot fully capture the level of detail required by focused retrieval. More precisely, although the *MAP* score correctly reflects the different ordering of documents in the result list, it still does not reflect how well the retrieved information per document corresponds to the relevant information. On the other hand, we demonstrated that our proposed mean average generalized precision measure (*MAgP*) is able to fully capture both evaluation aspects, which makes it more useful than *MAP* in measuring the retrieval performance.

In a separate study, Kamps et al. [10] have used the top 20 run submissions in the INEX 2006 *relevant in context* task to compare the correlation of relative system rankings based on *MAgP* with that of *MAP*, and the extent to which the two measures are capable at distinguishing between different retrieval approaches. The rank correlation (Kendall's tau) between *MAP* and *MAgP* was found to be 0.6740 over the top 20 official submissions, while when comparing the numbers of significant differences, *MAgP* was able to distinguish more performance differences than *MAP* (112 versus 95 of the total 190 pairwise comparisons).

Modelling evaluation assumptions

In Section 2 we have listed several assumptions which are typically used in evaluating different text retrieval tasks. Assumption 3 (users consider longer more detailed answers to be more useful than shorter answers) and Assumption 4 (users consider all retrieved answers to be equally useful) are of particular importance for *in context* retrieval tasks, as it is not entirely clear which of the two assumption should be preferred for evaluation of the *in context* tasks. We have modelled these two assumptions with the two generalized recall definitions and their corresponding average generalized precision definitions, shown in Equations 5 to 8 in Section 3. However, our fidelity tests in Section 4 have demonstrated that the *AgP'* measure, based on Assumption 3, is not entirely measuring what it is supposed to measure, and that the *AgP* measure, based on Assumption 4, correctly captures the expected run orderings.

An argument for Assumption 3 is that it also motivates the preference given to more exhaustive answers in some evaluations, and one could argue whether the *AgP'* definition, shown in Equation 8, is really correctly modelling this assumption. However, fixing this definition requires further investigation, which might be solved in one of these two ways: first, a definition for interpolated average generalized precision could be used instead of the current non-interpolated definition; and second, the current non-interpolated *AgP'* definition could be re-defined as follows:

$$AgP' = gR'[|\mathcal{D}|] \cdot \frac{\sum_{j=1}^{|\mathcal{D}|} rel(d_j) \cdot gP[j]}{\sum_{j=1}^{|\mathcal{D}|} rel(d_j)} \qquad (11)$$

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? Currently, we do not have exact answers to these questions. 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.

Acknowledgements We thank the anonymous reviewers for providing useful comments on a draft of this paper.

REFERENCES

[1] J. Allan. HARD track overview in TREC 2003 high accuracy retrieval from documents. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, pages 24–37, 2004.

[2] J. Allan. HARD track overview in TREC 2004 high accuracy retrieval from documents. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, 2004.

[3] S. Betsi, M. Lalmas, A. Tombros, and T. Tsikrika. User expectations from XML element retrieval. In *Proceedings of the ACM-SIGIR International Conference on Research and Development in Information Retrieval*, pages 611–612, Seattle, USA, 2006.

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

[5] N. Gövert, N. Fuhr, M. Lalmas, and G. Kazai. Evaluating the effectiveness of content-oriented XML retrieval methods. *Information Retrieval*, 9(6):699–722, 2006.

[6] W. Hersh, A. Cohen, P. Roberts, and H. Rekapalli. TREC 2006 genomics track overview. In *Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006)*, 2006.

[7] D. Hiemstra and V. Mihajlovic. The simplest evaluation measures for XML information retrieval that could possibly work. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology*, pages 6–13, Glasgow, UK, 2005.

[8] W. Huang, A. Trotman, and R. A. O'Keefe. Element retrieval using a passage retrieval approach. In *Proceedings of the 11th Australian Document Computing Symposium (ADCS 2006)*, pages 80–83, Brisbane, Australia, 2006.

[9] J. Kamps and B. Sigurbjörnsson. What do users think of an XML element retrieval system? In *Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval, INEX 2005*, volume 3977 of *Lecture Notes in Computer Science*, pages 411–421, 2006.

[10] J. Kamps, M. Lalmas, and J. Pehcevski. Evaluating Relevant in Context: Document retrieval with a twist. In *Proceedings of the ACM-SIGIR International Conference on Research and Development in Information Retrieval*, Amsterdam, The Netherlands, 2007 (to appear).

[11] G. Kazai, M. Lalmas, and A. de Vries. Reliability tests for the XCG and inex-2002 metrics. In *Advances in XML Information Retrieval: Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004*, volume 3493 of *Lecture Notes in Computer Science*, pages 60–72, 2005.

[12] J. Kekäläinen and K. Järvelin. Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, 53(13):1120–1129, 2002.

[13] H. Kim and H. Son. Users interaction with the hierarchically structured presentation in XML document retrieval. In *Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval, INEX 2005*, volume 3977 of *Lecture Notes in Computer Science*, pages 422–431, 2006.

[14] E. Lagergren and P. Over. Comparing interactive information retrieval systems across sites: the TREC-6 interactive track matrix experiment. In *Proceedings of the ACM-SIGIR International Conference on Research and Development in Information Retrieval*, pages 164–172, Melbourne, Australia, 1998.

[15] M. Lalmas and B. Piwowarski. INEX 2006 relevance assessment guide. In *INEX 2006 Workshop Pre-Proceedings*, pages 389–395, 2006.

[16] S. Malik, G. Kazai, M. Lalmas, and N. Fuhr. Overview of INEX 2005. In *Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval, INEX 2005*, volume 3977 of *Lecture Notes in Computer Science*, pages 1–15, 2006.

[17] J. Pehcevski. *Evaluation of Effective XML Information Retrieval*. PhD thesis, RMIT University, Melbourne, Australia, 2006. `http://www.cs.rmit.edu.au/~jovanp/phd.pdf`.

[18] J. Pehcevski and J. A. Thom. HiXEval: Highlighting XML retrieval evaluation. In *Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval, INEX 2005*, volume 3977 of *Lecture Notes in Computer Science*, pages 43–57, 2006.

[19] B. Piwowarski and G. Dupret. Evaluation in (XML) information retrieval: Expected precision-recall with user modelling (EPRUM). In *Proceedings of the ACM-SIGIR International Conference on Research and Development in Information Retrieval*, pages 260–267, Seattle, USA, 2006.

[20] S. Robertson. Evaluation in information retrieval. In *European Summer School on Information Retrieval (ESSIR)*, volume 1980 of *Lecture Notes in Computer Science*, pages 81–92, 2001.

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

[22] A. Trotman and S. Geva. Passage retrieval and other XML-retrieval tasks. In *Proceedings of the SIGIR 2006 Workshop on XML Element Retrieval Methodology*, pages 43–50, Seattle, USA, 2006.

[23] E. Voorhees. Overview of the TREC 2003 question answering track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, pages 54–68, 2004.

[24] C. Wade and J. Allan. Passage retrieval and evaluation. Technical report, CIIR, University of Massachusetts, Amherst, 2005.