Retrieval Evaluation on Focused Tasks

Besnik Fetahu
Graduate School of Computer Science,
Saarland University
Saarbrücken, Germany
bfetahu@mmci.uni-saarland.de

Ralf Schenkel
Saarland University
Saarbrücken, Germany
schenkel@mmci.uni-saarland.de

ABSTRACT

Ranking of retrieval systems for focused tasks requires large number of relevance judgments. We propose an approach that minimizes the number of relevance judgments, where the performance measures are approximated using a Monte-Carlo sampling technique. Partial measures are taken using relevance judgments, whereas the remaining part of passages are annotated using a generated relevance probability distribution based on result rank. We define two conditions for stopping the assessment procedure when the ranking between systems is stable.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.4 Systems and Software: Performance Evaluation

Keywords

Information Retrieval, Evaluation, Ranking, Algorithms

1. INTRODUCTION

We consider the problem of comparing retrieval systems for focused tasks, where not complete documents, but focused fragments of documents are retrieved. For building existing benchmark collections such as INEX [3], relevance judgments on the level of passages are created by human annotators or, recently, using crowdsourcing [1, 5]. A large number of documents are considered for assessment for each topic (500-750 at INEX), which makes the assessment procedure very work intensive (and expensive if done with crowdsourcing). Recent results [6] indicate that much smaller pools yield largely similar rankings of systems, but still in the order of 50 documents per topic need to be assessed.

For document-level retrieval, an interesting idea to reduce assessment effort was proposed in [2], which proposes to determine a minimal set of documents (MTC) that have the highest impact on average precision at a given rank. However, focused tasks are often not evaluated with rank-based quality metrics, but with recall-based measures such as interpolated precision (iP) [4] evaluated at a fixed set of recall levels. Unlike precision at a rank, iP is not monotonic in the number of assessments, i.e., the iP of a run can decrease when additional relevant passages are identified even outside the run. Figure 2 shows an example for this, where an initial iP[0.01] of 1.0 (1) first drops to 0.7 (2) when another relevant passage is found (and recall grows), and then increases again (3) with the next relevant passage.

Copyright is held by the author/owner(s).
SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

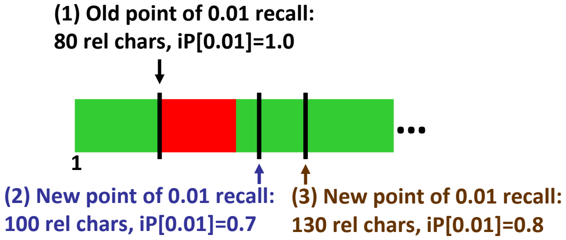
(1) Old point of 0.01 recall:
80 rel chars, iP[0.01]=1.0
...
1
(2) New point of 0.01 recall:
100 rel chars, iP[0.01]=0.7
(3) New point of 0.01 recall:
130 rel chars, iP[0.01]=0.8

Figure 1: Example for non-monotonicity of iP

In this poster we extend the idea of MTC towards focused retrieval and recall-based metrics, and show that comparable system rankings can be computed with much less assessments that currently done.

2. EVALUATION METRICS

We use ***interpolated precision (iP)*** at different recall levels, a standard evaluation measure for Focused Tasks [4]. Here, the recall base is the set of all relevant characters for a topic (collected during assessment), and each rank of the result list is assigned recall (fraction of all relevant characters retrieved up to this rank) and precision (ratio of retrieved relevant to all retrieved characters up to this rank). The metric is evaluated at recall levels $x \in \{0, 0.01, \ldots, 1.0\}$, and the value at recall level x is the highest precision at any rank with recall at least x. Additionally, average precision AiP over all recall points and mean average precision $MAiP$ over all topics are defined. Please see [4] for the full formal definitions.

3. OUR METHOD

We now present our method that, given a set of runs $A, B, \ldots, N$ for a set of topics, selects passages from these runs to assess and finally builds a ranking of the runs.

3.1 Document Selection

We want to assess only the passages that have most impact on system ranking. To achieve this we first create, a pool of unique passages retrieved from all systems on all topics. We then compute the impact on iP of assessing each passage for each run where it appears (considering other non-assessed passages as non-relevant). If a passage appears in multiple systems ($A \ldots N$) its weight is constructed by the pairwise difference of each individual weight (as the precision of a system up to that point) in all systems, shown here for some system c (n is the number of systems in which the passage does not appear). The highest weight is assigned to a passage (from the passage weights computed at other systems):

$$\Delta iP = |iP_A - iP_B| + \cdots + |iP_{N-1} - iP_N| + n * iP_c$$

We then select a number k of passages for each topic, sorted by the weight defined above. This is repeated until all systems are marked as stable by one of the stopping conditions.

3.2 Approximated Metrics

We now explain how to compute an approximated value for interpolated precision when only a subset of all passages have been assessed. A first alternative would be to consider *partial iP*, where unassessed passages are set to not relevant, but this is too simplistic since many passages at early ranks will be unassessed in the initial rounds of our method. Instead, we use Monte Carlo sampling to compute an estimated iP value. Here, the relevance of unassessed passages is considered a random variable, with a probability of relevance based on the rank on which they appear. The number of samples N was specified using the absolute ϵ-approximation, which approximates the measures with probability at least $1 - \delta$, where $\epsilon = 0.1$, and $\delta = 0.9$. An initial analysis with the INEX'08 runs (considering a passage as relevant when it contained at least one relevant character) showed that this probability is exponentially decaying with increasing rank.

3.3 Stopping Conditions

Our method runs in iterations where in each iteration, k documents from each non-stable system are chosen for assessment. After each iteration, partial and approximated iP are computed, and we check if we can stop based on two stopping conditions:

- **Test Statistics:** We statistically compare partial iP to approximated iP, where we test the hypothesis if they are approximately similar to each other. $H_0 : \mathbf{AiP} = \mathbf{E}[AiP]$, and $H_1 : \mathbf{AiP} \neq \mathbf{E}[AiP]$. It is expected that the null hypothesis is rejected in most cases due to the difference in partial and approximated measures.
- **Pairwise Measure Comparison:** We compare if partial and approximated iP agree pairwise on the ranking of a system compared to all other systems. If that is the case, that system is marked as stable. We can stop if all systems are marked as stable, i.e., if the ranking computed from the partial measure agrees with the ranking computed from the approximated measure.

4. RESULTS & DISCUSSIONS

We evaluated our method with data from the Focused Task of the INEX 2009 AdHoc track [3], which consisted of ***68*** topics. The comparison performance of systems on topic level was measured based on ***AiP***. The rank probability distribution was constructed using INEX'08 data. In total only 4080 relevance judgments, with approx.60 per topic, were needed until the stopping condition fired after 3 iterations (we stopped when more than half of the systems were marked as stable). We compared the produced ranking to the ranking with full INEX assessments (restricted to a standard pool of size 500 built for the considered runs) using Kendall's τ correlation. Figure 2 shows the correlation for all recall points, and Table 1 shows the correlation at four selected recall points after each iteration, and for PoolSize-55 (which uses 60 assessments per topic). It is evident that our method reaches an agreement level of 0.9. It is also more effective than standard INEX pooling with 50 documents per topic, which achieves a correlation slightly below 0.9 (Figure 2).

iteration	**iP[0.00]**	**iP[0.01]**	**iP[0.05]**	**iP[0.10]**
iter-1	87.66	88.87	87.54	91.17
iter-2	92.38	90.8	89.59	92.62
iter-3	93.71	**91.65**	**90.44**	**93.35**
PoolSize-55	**93.80**	90.64	88.31	90.64

Table 1: τ-coefficient on official recall points.

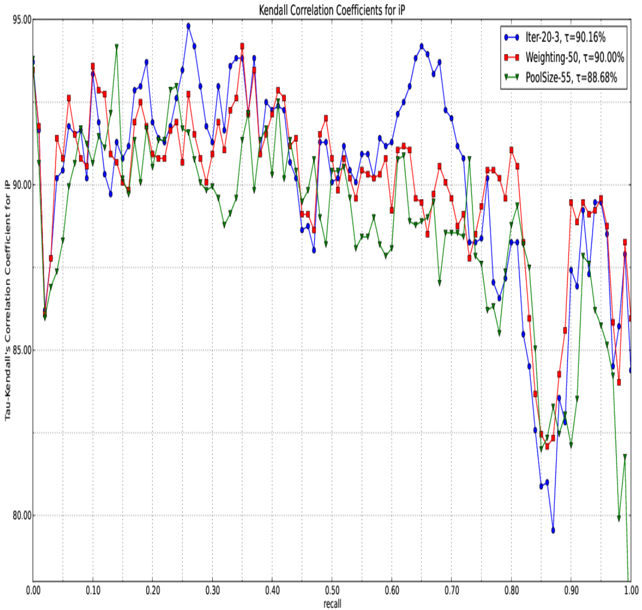

Figure 2: τ-coefficient on all 101 recall points.

5. CONCLUSIONS

The main goal of our was to rank multiple systems with minimal number of relevance judgments for focused retrieval. We proposed a new algorithm for selecting passages for assessment under recall-based evaluation metrics such as interpolated precision. We iteratively pick new passages to assess until all systems achieved a stable ranking, based on two stopping conditions.

6. REFERENCES

[1] Omar Alonso, Ralf Schenkel, and Martin Theobald. Crowdsourcing assessments for xml ranked retrieval. In *ECIR*, pages 602–606, 2010.

[2] Ben Carterette, James Allan, and Ramesh K. Sitaraman. Minimal test collections for retrieval evaluation. In *SIGIR*, pages 268–275, 2006.

[3] Shlomo Geva, Jaap Kamps, Miro Lehtonen, Ralf Schenkel, James A. Thom, and Andrew Trotman. Overview of the inex 2009 ad hoc track. In *INEX*, pages 4–25, 2009.

[4] Jaap Kamps, Jovan Pehcevski, Gabriella Kazai, Mounia Lalmas, and Stephen Robertson. Inex 2007 evaluation measures. In *INEX*, pages 24–33, 2007.

[5] Gabriella Kazai, Jaap Kamps, Marijn Koolen, and Natasa Milic-Frayling. Crowdsourcing for book search evaluation: impact of hit design on comparative system ranking. In *SIGIR*, pages 205–214, 2011.

[6] Sukomal Pal, Mandar Mitra, and Jaap Kamps. Evaluation effort, reliability and reusability in xml retrieval. *JASIST*, 62(2):375–394, 2011.