Retrieval System Evaluation: Automatic Evaluation versus Incomplete Judgments

Claudia Hauff
University of Twente
Enschede, The Netherlands
c.hauff@ewi.utwente.nl

Franciska de Jong
University of Twente
Enschede, The Netherlands
f.m.g.dejong@ewi.utwente.nl

ABSTRACT

In information retrieval (IR), research aiming to reduce the cost of retrieval system evaluations has been conducted along two lines: (i) the evaluation of IR systems with reduced (i.e. incomplete) amounts of manual relevance assessments, and (ii) the fully automatic evaluation of IR systems, thus foregoing the need for manual assessments altogether. The proposed methods in both areas are commonly evaluated by comparing their performance estimates for a set of systems to a ground truth (provided for instance by evaluating the set of systems according to mean average precision). In contrast, in this poster we compare an automatic system evaluation approach directly to two evaluations based on incomplete manual relevance assessments. For the particular case of TREC's Million Query track, we show that the automatic evaluation leads to results which are highly correlated to those achieved by approaches relying on incomplete manual judgments.

Categories and Subject Descriptors: H.3.3 Information Storage and Retrieval: Information Search and Retrieval
General Terms: Experimentation, Performance
Keywords: Automatic System Evaluation

1. INTRODUCTION

In information retrieval (IR), research aiming to reduce the cost of retrieval system evaluations has been conducted along two lines: (i) the evaluation of IR systems with reduced amounts of manual relevance assessments, and (ii) the fully automatic evaluation of IR systems, thus foregoing the need for manual assessments altogether. The two most important approaches in the first category are the determination of good documents to assess (the *MTC* approach) [6] and the proposal of alternative pooling methods (the *statAP* approach) [4]. Both, *MTC* and *statAP*, are now accepted system evaluation metrics at TREC[1]. They stand in contrast to the depth pooling methodology which has until recently been employed at TREC; due to the ever increasing size of test collections and query sets though, pooling the top 100 documents of each retrieval run participating in a benchmark and assessing those documents manually for their relevance, has become infeasible. The earliest method for a fully automatic evaluation was proposed by Soboroff et al. (the *RS* approach) [7]. It relies on drawing random samples from the pool of top retrieved documents.

The quality of *statAP*, *MTC* and *RS* is usually evaluated by comparing the performances of a set of retrieval runs for which sufficient relevance judgments are available according to a standard effectiveness metric (mean average precision) with the estimated system performances. Generally missing though is a direct comparison between *statAP*/*MTC* and an automatic method such as *RS*.

In recent work [3], we found the commonly reported problem of automatic evaluation approaches (the severe misranking of the very best retrieval runs [5]) not to be inherent to automatic system evaluation methods. The extent of this problem is strongly related to the degree of human intervention in the best retrieval runs: the larger the amount of human intervention, the less able automatic approaches are to identify the best runs correctly.

In this poster, we turn to investigating how closely the automatic evaluation of retrieval runs approximates the evaluation with incomplete manual relevance assessments. We perform this analysis in a setting which favors automatic evaluation: TREC's Million Query tracks of 2007 (MQ-2007) [2] and 2008 (MQ-2008) [1]. Due to the size of the query sets, creating retrieval runs with a great amount of human intervention is virtually impossible. We thus expect the *RS* approach to lead to similar estimates of system performances as *statAP* and *MTC* respectively. If this would indeed be the case, it would bring into question the need for manual assessments in this type of setting.

2. EXPERIMENTS

For our experiments, we relied on the twenty-nine retrieval runs submitted to MQ-2007 and the twenty-four[2] runs submitted to MQ-2008. Both sets of retrieval runs as well as their retrieval effectiveness scores according to *statAP* and *MTC* are available from the TREC website. Specifically, for MQ-2007, TREC provides the *statAP* measures[3], while for MQ-2008 both, *MTC* and *statAP*, are provided. Of the 10000 queries that were released for each year, 1153 (MQ-2007) and 564 (MQ-2008) queries respectively have valid *statAP* measurements; 784 (MQ-2008) queries have valid *MTC* measurements. These are the queries we also rely on in the *RS* approach.

[1] `http://trec.nist.gov/`

[2] In total, twenty-five runs exist, though one is not accessible from the TREC website and thus had to be ignored.

[3] The MTC measures are not accessible from the TREC website.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

Pool Depth p	Avg. Sampled Documents	Kendall's Tau
MQ-2007 *statAP*		
10	4.6	**0.803**
50	22.2	0.783
100	43.0	0.754
250	102.6	0.719
MQ-2008 *statAP*		
10	6.0	0.768
50	28.8	0.812
100	55.7	0.833
250	132.4	**0.841**
MQ-2008 *MTC*		
10	6.1	0.722
50	29.4	0.759
100	56.8	0.773
250	135.0	**0.780**

Table 1: Kendall's Tau rank correlation coefficient between the automatic *RS* approach and *statAP*/*MTC* respectively. All correlations are significant ($p < 0.01$). Column 2 contains the average number of sampled documents from the pool.

For the automatic evaluation, we implemented the random sampling approach [7]: first, the top p retrieved documents of all retrieval runs for a particular query are pooled together such that a document that is retrieved by x runs, appears x times in the pool. Then, a number m of documents are drawn at random from the pool; those are now considered to be the *pseudo relevant* documents. This process is performed for each query and the subsequent evaluation of each system is performed with *pseudo relevance judgments* instead of relevance judgments. Due to the randomness of the sampling, we performed 20 trials per query and averaged the pseudo relevance based system performance. We fixed the number m of documents to sample 5% of the number of unique documents in the pool and evaluated pool depths of $p = \{10, 50, 100, 250\}$.

In Table 1 (column 3) we report the rank correlation coefficient Kendall's Tau (τ) between the performance scores estimated by the automatic *RS* approach and the performance scores estimated by *statAP*/*MTC* which exploit manual relevance assessments. In the ideal case, $\tau = 1.0$, that is, *RS* leads to the same rank estimate of system performances as *statAP*/*MTC*. It is apparent, that although the correlations are not perfect, the correlation coefficients are consistently high; in the worst instance the correlation reaches $\tau = 0.72$ for MQ-2007 *statAP* and a pool depth of $p = 250$; at best the correlation reaches $\tau = 0.84$ for MQ-2008 *statAP* and $p = 250$.

Figures 1 and 2 show scatter plots of MQ-2007 *statAP* scores versus *RS* scores and of MQ-2008 *MTC* scores versus *RS* scores respectively. It is evident that the best retrieval runs as identified by *statAP*/*MTC* are also identified correctly by the automatic *RS* approach.

3. DISCUSSION AND CONCLUSION

In this poster, we investigated the ability of an automatic system evaluation approach (*RS* [7]) to approximate the system performance estimates as derived by two evaluation methods that rely on manually derived incomplete relevance judgments: *statAP* and *MTC*. Experiments on TREC's Million Query tracks showed that *RS* is highly correlated to *statAP* and *MTC*, an outcome which implies that retrieval runs, which are automatic in nature, can be evaluated by an automatic approach such as *RS* which requires no manual assessments at all.

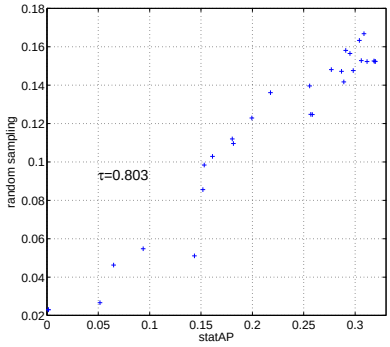

Figure 1: MQ-2007 *statAP* scores (x-axis) versus *RS* scores (y-axis) for a pool depth of $p = 10$.

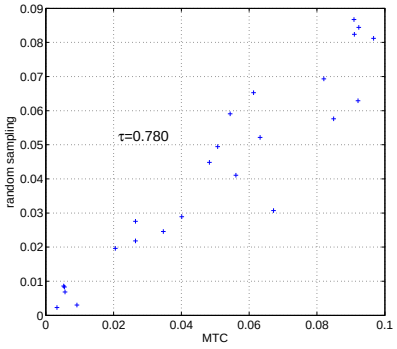

Figure 2: MQ-2008 *MTC* scores (x-axis) versus *RS* scores (y-axis) for a pool depth of $p = 250$.

One direction of future work will be the adaptation of *RS* to further improve the method's correlation with *statAP* and *MTC* by for instance taking advantage of the relationship between queries of a query set (as is possible for larger sets of queries) in contrast to the current approach where each query is viewed in isolation.

4. REFERENCES

[1] J. Allan, J. A. Aslam, V. Pavlu, E. Kanoulas, and B. Carterette. Million Query Track 2008 Overview. In *TREC 2008*, 2008.

[2] J. Allan, B. Carterette, B. Dachev, J. A. Aslam, V. Pavlu, and E. Kanoulas. Million Query Track 2007 Overview. In *TREC 2007*, 2007.

[3] C. Hauff, D. Hiemstra, L. Azzopardi, and F. de Jong. A Case for Automatic System Evaluation. In *ECIR '10*, pages 153–165, 2010.

[4] J. Aslam, V. Pavlu, and E. Yilmaz. A statistical method for system evaluation using incomplete judgments. In *SIGIR '06*, pages 541–548, 2006.

[5] J. A. Aslam and R. Savell. On the effectiveness of evaluating retrieval systems in the absence of relevance judgments. In *SIGIR '03*, pages 361–362, 2003.

[6] B. Carterette, J. Allan, and R. Sitaraman. Minimal test collections for retrieval evaluation. In *SIGIR '06*, pages 268–275, 2006.

[7] I. Soboroff, C. Nicholas, and P. Cahan. Ranking retrieval systems without relevance judgments. In *SIGIR '01*, pages 66–73, 2001.