On Real-time Ad-hoc Retrieval Evaluation

Stephen E. Robertson
Microsoft Research
7 JJ Thomson Avenue
Cambridge CB3 0FB, UK
stephenerobertson@hotmail.co.uk

Evangelos Kanoulas [*]
Information School
University of Sheffield
Sheffield, UK
ekanoulas@gmail.com

ABSTRACT

Lab-based evaluations typically assess the quality of a retrieval system with respect to its ability to retrieve documents that are relevant to the information need of an end user. In a real-time search task however users not only wish to retrieve the most relevant items but the most recent as well. The current evaluation framework is not adequate to assess the ability of a system to retrieve both recent and relevant items, and the one proposed in the recent TREC Microblog Track has certain flaws that quickly became apparent to the organizers. In this poster, we redefine the experiment for a real-time ad-hoc search task, by setting new submission requirements for the submitted systems/runs, proposing metrics to be used in evaluating the submissions, and suggesting a pooling strategy to be used to gather relevance judgments towards the computation of the described metrics. The proposed task can indeed assess the quality of a retrieval system with regard to retrieving both relevant and timely information.

Categories and Subject Descriptors

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

General Terms

Experimentation, Measurement

Keywords

evaluation, measures, pooling, real-time retrieval, microblog

1. INTRODUCTION

Lab-based evaluations typically assess the quality of a retrieval system with respect to its ability to retrieve documents that are relevant to the information need of an end user. There are retrieval tasks however for which relevance is only one of the dimensions on which we want to evaluate retrieval systems. In particular, we consider a real-time ad-hoc retrieval task, where the user wishes to retrieve information that is not only relevant, but also recent – the focus of the recent TREC Microblog Track [1]. The corpus for the task was composed of about 16 million tweets with time-stamps. Information needs were represented by queries at specific time-stamps. Systems were expected to respond to each query with the most recent relevant tweets, prior to the query time.

Evaluating a system over a real-time ad-hoc task clearly requires a measure that accounts both for relevance and recency. Instead of inventing a new measure the TREC Microblog Track chose to use traditional set-based evaluation measures on a time-ordered ranked list of the submitted by participants ranked lists of tweets/runs. Each participant was asked to submit a set of up to 1000 tweets, submitted tweets were then ordered by reverse time, the top 30 were pooled, handed to NIST assessors to be assessed with respect to relevance ("interestingness"), and each system was then evaluated by precision@30 over these 30 time-ordered tweets.

What became immediately apparent was that such an experiment could not properly measure the effectiveness of a system to provide both timely and relevant tweets. A system could achieve optimal performance by returning just 30 relevant tweets, ignoring the time dimension. Hence, the optimal strategy for each run was to return a set of the top-30 tweets ordered by a probability (score) of relevance, even if these were not the most recent relevant tweets. An intuitive explanation of the flaw in this evaluation experiment is that runs were not evaluated against a single global target set of tweets (the most relevant, recent tweets) and thus the evaluation measure was only aware of the relevant and non-relevant tweets returned by each individual system separately. The aim of this poster is to correct this flaw by redefining a real-time ad-hoc task experiment, in a form suitable for the Microblog track, in which systems are evaluated for their ability to return both relevant and timely information.

2. REAL-TIME AD-HOC TASK

In what follows we define the following elements of the task:

- general description of the task
- submission requirements
- metric(s)
- pooling/assessment strategy

As in the Microblog Track [1] we assume that there is an effective timeline for items, and queries are time-specific.

[*] We gratefully acknowledge the support provided by the European Commission grant FP7-PEOPLE-2009-IIF-254562.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

To simplify the evaluation of runs we assume that judgments are binary, i.e. items are either relevant or not relevant to a query. We briefly discuss graded relevance in Section 3 that follows.

General description of the task: The aim for each system is to retrieve the 30 most recent relevant items.

Submission requirements: Participants will need to submit (a) a set of not more than 30 items; (b) a ranked list of 1000 top previous items by relevance only. Evaluation will be based on the set only (see Metrics below), but the ranked list will be used in constructing the assessment pools (see pooling/assessment strategy below).

Metric(s): For each topic, we define a single Target Set, consisting of the 30 most recent relevant items known. Basic metrics will be recall and precision in relation to the target set. Further metrics might include e.g. F1. Defining a single target set informs the metrics of the optimal behaviour of a retrieval system, which is to retrieve the most recent 30 relevant items, allows the observation of suboptimal behaviour of submitted systems when the items they return in the submitted set are either irrelevant or too old, and makes scores comparable across all submitted runs.

Pooling/assessment strategy: The pooling process is iterative rather than one-off – involving further additions to pools after initial items have been judged. The goal of the pooling process is to identify the items in the target set, that is the 30 relevant recent items. The iterative process follows:

1. Pool all sets of the 30 items submitted, together with the top 1 ranked item from all the ranked lists. Judging all sets avoids having holes in the sets of the 30 items over which runs are evaluated and thus enforces fairness across all submissions.

2. Obtain judgments for the current pool.

3. Identify the 30 most recent relevant items so far in the pool and set a time window from the time of the 30th most recent to the time of the query. This time window allows pruning of the space of items to be judged. Any item with a time stamp out of this time frame, i.e. older than the 30th most recent relevant item, does not need to be judged for relevance.

4. Take the next item from each result list which falls within the window, add it to the pool, judge unjudged documents. If additional relevant documents are found, iterate.

As mentioned this method guarantees to evaluate all submissions fairly according to the metric, and also provides good material for participants to do diagnostic work on their mixture of relevance and time-like criteria for retrieval. For future use, it identifies what is likely to be a good approximation to the true target set, and also labels for relevance a reasonable number of older items, again helpful for diagnostic work.

3. DISCUSSION

So far we have only considered binary relevance assessments; often multiple grades are used. In the Microblog Track tweets were judged as non-relevant, relevant, or vital (highly relevant) [1]. A simple way to account for graded relevance in the proposed framework is to define the target set to contain not only the 30 most recent relevant items but also all previous vital items known. This is a rather crude balance between the importance of relevance and time but it allows the measurement process to remain quite simple.

Note that the current set up does not prohibit a future definition of a measure that combines relevance and time through some gain function for instance. Under such a development, the submission requirements, as described here, should remain similar with participants submitting an auxiliary list of items ranked by relevance only to allow the construction of a target set/ranking. Further, pruning of the judgment space during pooling should also be possible in a way similar to the one described in the proposal above.

Further it should be also possible to consider giving extra weight to the retrieval of all vital items, but that complicates the metrics somewhat and a discussion on this is out of the scope of the current poster. An extension of set measures to graded relevance can be found in Robertson et al. [2].

Another point regarding the proposed experiment is that one could continue adding items to the pool even if no relevant documents are found for a small number of iterations, so that any recent relevant items that appear low in the ranked list by all systems gets a chance to be pooled. The number of iterations that would be required is a point of further investigation on the implementation of the proposed experiment and out of the scope of this poster.

Last, note also that set retrieval typically requires ranking and thresholding; thresholding is known to be hard. However, for the task as defined, participants will be able to get away with thresholding at fixed rank (20 or 25 or 30) without disastrously damaging effectiveness. There will be room for more sophisticated thresholding methods to be used.

4. CONCLUSION

In this poster we proposed a experiment to test the effectiveness of retrieval systems towards returning both relevant and timely information. We discussed a rather simple case, where relevance is binary and the only measures considered are set measures. Extensions of the proposed experiment for graded relevance, ranking measures are possible and follow the general principles of the proposed framework with somewhat more complex details on the implementation side. The proposed framework resolves the flaw of the current evaluation framework used by the TREC Microblog Track.

5. REFERENCES

[1] Craig Macdonald, Iadh Ounis, Jimmy Lin, Abdur Choudhury, and Ian Soboroff. TREC Microblog Track. `https://sites.google.com/site/microblogtrack/`, 2011.

[2] Stephen E. Robertson, Evangelos Kanoulas, and Emine Yilmaz. Extending average precision to graded relevance judgments. In *SIGIR '10: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 603–610, New York, NY, USA, 2010. ACM.