On Judgments Obtained from a Commercial Search Engine

Emine Yilmaz, Gabriella Kazai
Microsoft Research Cambridge, UK
{eminey,v-gabkaz }@microsoft.com

Nick Craswell, S.M.M. Tahaghoghi
Microsoft, Bellevue, WA, USA
{nickcr,saied.tahaghoghi}@microsoft.com

ABSTRACT

In information retrieval, relevance judgments play an important role as they are used for evaluating the quality of retrieval systems. Numerous papers have been published using *judgments obtained from a commercial search engine* by researchers in industry. As typically no information is provided about the quality of these judgments, their reliability for evaluating retrieval systems remains questionable. In this paper, we analyze the reliability of such judgments for evaluating the quality of retrieval systems by comparing them to judgments by NIST judges at TREC.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3[Information Search and Retrieval]

General Terms

Experimentation, Human Factors, Measurement

Keywords

Crowdsourcing, Evaluation, Test Collection

1. INTRODUCTION

In information retrieval (IR), test collections are typically used to evaluate and optimize the performance of IR systems. The quality of a test collection can impact the conclusions of the evaluation, where the quality of the relevance judgments is a key factor. For example, evaluation outcomes are shown to be affected by using different judge populations [1] and different judging guidelines [3]. On the other hand, using different judges from the same population of NIST judges employed by TREC has been shown to lead to relatively stable conclusions as to which retrieval algorithm beats another [4].

In recent years, several papers using *judgments obtained from a commercial search engine* have been published [2]. Most of these papers use such judgments (which are typically not publicly available) to validate the superiority of their proposed methods over existing algorithms. Since judges employed by commercial search engine companies are likely to come from different populations than NIST judges and are likely to be subjected to different training and judging procedures, we may reason that judgments from a commercial search engine are likely to lead to different conclusions than judgments from NIST judges.

We analyze whether judgments obtained from a commercial search engine are reliable, in terms of leading to the same evaluation conclusions as when using NIST judgements.

2. EXPERIMENTAL RESULTS

We use the test collection from the TREC Web Track Adhoc tasks from 2009 and 2010. This dataset consists of nearly 50K NIST relevance labels, roughly 25K in each year, for 50 topics in each year. We took these 100 topics and using the topic titles as queries we scraped the top 10 search results from Google and Bing for each query. This gave us a total of 1603 unique query-URL pairs for the 100 topics.

We constructed three different collections by obtaining judgments from three judge groups : judges from (1) the TREC Web track ad-hoc task (**NIST**), (2) a commercial search engine (**ProWeb**), and (3) crowdsourcing (**Crowd**). ProWeb judges were experienced and highly trained judges, employed by the search engine company, while crowd workers, recruited via Clickworker, received no prior training on relevance assessing.

The NIST judgments differ across the two years. In 2009, three relevance levels were used (Highly Relevant, Relevant, and Not Relevant), while in 2010, five grades of relevance were used (Navigational, Key, Relevant, Non-relevant, and Junk). The ProWeb and Crowd judgments were obtained using a simple interface that asked judges to rate a search result's usefulness to a query using a five point scale that can be viewed as a variation of the 2010 Web Track scale (Ideal, Highly Relevant, Relevant, Somewhat Relevant, Non-relevant). Unlike the ProWeb and Crowd judges, the NIST judges were given descriptions (topic narrative) about what information need is associated with a particular query.

Using the different sets of judgments and the NDCG measure, we evaluate the effectiveness of the runs submitted to the TREC 2009 and 2010 Web Track ad-hoc task. Since we only have labels for a subset of the retrieved documents, we remove unjudged documents from the runs. To avoid variance due to having different documents labeled across the different judge groups, we only consider documents that were judged by all three groups.

To remove the inconsistency across different judge groups due to the different levels of relevance scales used, we converted all the judgments to the Web Track 2009 scale using the following mapping:*Navigational* (or *Ideal*) judgments to *Highly Relevant*, *Key* and *Relevant* (or *Highly Relevant* and

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

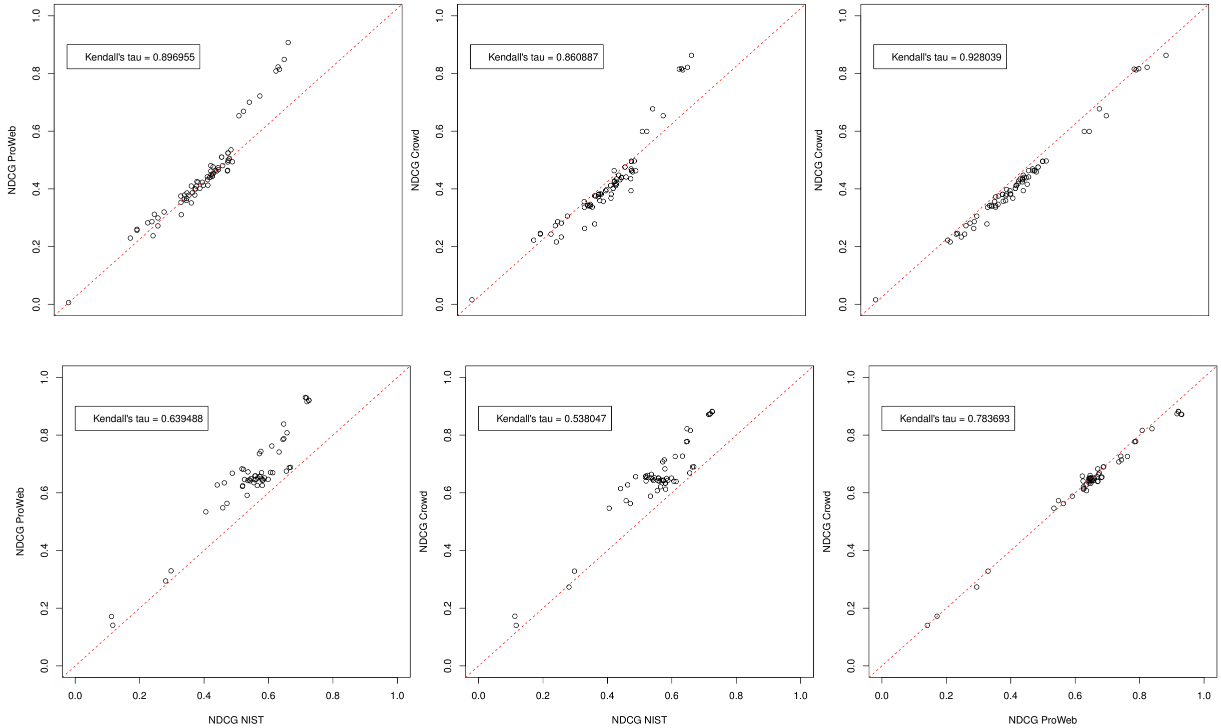

Figure 1: Comparisons of evaluation results for runs submitted to TREC 2009 (top) and TREC 2010 (bottom) using three different sets of judges and their judgements

Relevant) to *Relevant*, and *Non-relevant* and *Junk* (or *Somewhat Relevant* and *Non-relevant*) to *Not Relevant*. We have also experimented with various other mappings but all supported the same conclusions of this poster.

Figure 1 shows the obtained scatter plots: TREC 2009 results in the top row and TREC 2010 in the bottom row. For each year, we plot the results obtained from evaluating the runs using the NIST vs ProWeb (left plot), NIST vs. Crowd (middle plot) and Crowd vs. ProWeb (right plot) judgments. In each plot, we also include the Kendall's τ correlation between the resulting system rankings, obtained using the two respective sets of judgments.

We see that for TREC 2009, evaluations using the NIST, ProWeb and Crowd judgments mostly agree with each other (Kendall's τ of 0.86-0.93). This suggests that the judgments obtained from a commercial search engine are more or less consistent with NIST: using them will not cause major differences in the evaluation. Crowd judgments may be somewhat noisier, but still lead to stable evaluations. The differences in the three plots could be caused by the consistency in the number and description of relevance grades between the Crowd and ProWeb judges as compared to the NIST judges.

On the other hand, for TREC 2010, evaluations using the ProWeb and NIST judgments are quite different (Kendall's $\tau = 0.639$). At first glance, one might think that ProWeb judgments are not reliable at evaluating the systems and the high agreement on the TREC 2009 data is due to chance. However, if we compare the agreements between the evaluations using the Crowd and ProWeb judgments, we also see low agreements. This suggests that the low correlation is specific to this particular TREC. As the figure suggests, systems that were submitted to this particular TREC have very similar performance. Hence, the Kendall's τ statistic may be affected by these systems. Furthermore, when we considered the cases where NIST judgments highly disagree with the Crowd and ProWeb judgments, we found that there are quite a few documents that got the best rating by the Crowd and ProWeb judges but were labeled as non-relevant by the NIST judges. When we analyzed the disagreement cases, we realized that these differences could be caused by the specific topic description that was given to the NIST judges, which limited the possible intents associated with a query. Since the Crowd and ProWeb judges were not given such topic descriptions, they would have considered all possible intents for a query when assigning labels to the documents.

Overall, our conclusion is that even though judgments from a commercial search engine could lead to slightly different conclusions than NIST judges in some settings, evaluations using the two judge groups seem mostly consistent.

3. REFERENCES

[1] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P. de Vries, and Emine Yilmaz. Relevance assessment: are judges exchangeable and does it matter. In *Proc. of ACM SIGIR Conference*, pages 667–674. ACM, 2008.

[2] Christopher J. C. Burges, Robert Ragno, and Quoc Viet Le. Learning to rank with nonsmooth cost functions. In *NIPS*, pages 193–200. MIT Press, 2006.

[3] Charles L.A. Clarke, Nick Craswell, Ian Soboroff, and Azin Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *Proc. of ACM WSDM Conference*, pages 75–84. ACM, 2011.

[4] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proc. of ACM SIGIR Conference*, pages 315–323. ACM, 1998.