Overview of the TREC 2013 Crowdsourcing Track

Mark D. Smucker[1], Gabriella Kazai[2], and Matthew Lease[3]

[1]Department of Management Sciences, University of Waterloo
[2]Microsoft Research, Cambridge, UK
[3]School of Information, University of Texas at Austin

February 11, 2014

Abstract

In 2013, the Crowdsourcing track partnered with the TREC Web Track and had a single task to crowdsource relevance judgments for a set of Web pages and search topics shared by the Web Track. This track overview describes the track and provides analysis of the track's results.

1 Introduction

Now in its third year, the overall goals of the TREC Crowdsourcing track remained to build awareness and expertise with crowdsourcing in the Information Retrieval (IR) community, to develop and evaluate new methodologies for crowdsourced search evaluation on a shared task and data set, and to create reusable resources to benefit future IR community experimentation.

While the first year of the track was explicitly focused on crowdsourcing, last year we decided to loosen the crowdsourcing requirements and instead focus on a goal of obtaining document annotations by any means. The advantage of this change was that it gave groups freedom in the creation of their solutions towards methods that combined contributions from humans and computer algorithms. This year, we followed in the same vein and further emphasized the combined human-computer aspect.

As in previous years, the track set as its challenge the task of 'crowdsourcing' quality relevance judgments. Unlike last year, when we had a textual document and an image relevance judging task, this year the track consisted of a single task that required collecting relevance judgments for Web pages and search topics taken from the TREC Web Track. Participants thus had *almost* the same task as NIST assessors: judging the relevance of Web pages retrieved by teams who partook in the Web Track challenge. Whereas NIST judges were required to assess all sub-topics of each topic, track participants were only required to judge the first sub-topic. This kept the scale of the task similar to last year, with around 20k documents needing to be judged. To ease participation, we also offered a reduced scale task, with only around 3.5k documents to be judged.

Four groups participated in the track, submitting 11 runs in total. We next describe details of the task, the data set used, the evaluation methods, and finally the results.

2 Task Overview

The task required collecting relevance judgments for Web pages and search topics, taken from the TREC Web Track. The Web pages to be judged for relevance were drawn from the recently released ClueWeb12[1] collection. The search topics were created by NIST judges. While the Web Track participants were only provided with the topic titles, the Crowdsourcing Track participants were given the full topic descriptions, matching the setup for the NIST judges. After the Web Track participants submitted their retrieval runs, NIST identified a subset of the submitted documents to be judged for each topic. In parallel with NIST assessment, the Crowdsourcing Track participants were given the same topic and document pairs to label as the NIST judges.

In this 'crowdsourcing' track, participants were free to use or not use crowdsourcing techniques however they wished. For example, judgments could be obtained via a fully-automated system, or using traditional relevance assessment practices,

[1]http://www.lemurproject.org/clueweb12/

or a mix of these. Participants could use a purely crowdsourcing-based approach, or employ a hybrid approach combining automated systems, crowds, trained judges or other resources and techniques. Crowdsourcing could be paid or non-paid. It was left entirely up to each team to innovate the best way to obtain accurate judgments, in a reliable and scalable manner, while minimizing the time, cost, and effort required.

The track offered two entry levels for participation. Participants could choose to enter at either or both levels:

- Basic: approx. 3.5k documents (subset of NIST pool, 10 topics)
- Standard: approx. 20k documents (entire NIST pool, 50 topics)

The task in both cases was to obtain relevance labels for the documents and search topics included in the entry level set.

Judgments needed to be collected on a six-point scale:

- 4=Nav This page represents a home page of an entity directly named by the query; the user may be searching for this specific page or site.
- 3 = Key This page or site is dedicated to the topic; authoritative and comprehensive, it is worthy of being a top result in a web search engine.
- 2 = HRel The content of this page provides substantial information on the topic.
- 1 = Rel The content of this page provides some information on the topic, which may be minimal; the relevant information must be on that page, not just promising-looking anchor text pointing to a possibly useful page.
- 0 = Non The content of this page does not provide useful information on the topic, but may provide useful information on other topics, including other interpretations of the same query.
- -2 = Junk This page does not appear to be useful for any reasonable purpose; it may be spam or junk.

We informed participants that non-English documents would be judged non-relevant by NIST assessors, even if the assessor understands the language of the document and the document would be relevant in that language. If the location of the user matters, the assessor will assume that the user is located in Gaithersburg, Maryland. In addition, the NIST assessor guidelines were also made available to participants.

3 Data Set

Participants were provided with details of the search topics and a list of (topic-ID, document-ID) pairs to be judged in both entry level sets. The document-IDs identified the documents to be judged from the ClueWeb12 collection.

Thanks to Jamie Callan at CMU, and Gaurav Baruah at Waterloo, it was possible to participate in the track without purchasing ClueWeb12 and to download a self-contained corpus of the Web pages included in the entry level sets.

All topics were expressed in English. The standard entry level set contained all 50 queries from the Web Track ad hoc task; the same set that was sent to NIST judges for assessment. Of the 50 search topics, 25 were multi-faceted, i.e., they contained several sub-topics, representing different possible user intents. For example, the topic below is one of the multi-faceted topics:

```
<topic type="faceted" number="206">
  <query>wind power</query>
  <description>
    What are the pros and cons of using
    wind power.
  </description>
  <subtopic type="inf" number="1">
    What are the pros and cons of using
    wind power.
  </subtopic>
  <subtopic type="inf" number="2">
    Find information on wind power in
    the USA.
  </subtopic>
  <subtopic type="inf" number="3">
    Find information on wind power
    companies.
  </subtopic>
  <subtopic type="inf" number="4">
    Find information on residential
    (home) wind power.
  </subtopic>
  <subtopic type="inf" number="5">
    Find information on how wind
    turbines work.
  </subtopic>
```

Team	Basic	Standard
Hrbust	1	-
NEUIR	1	-
PRIS	1	-
udel	-	8

Table 1: Submitted runs to the basic and standard entry levels.

```
  <subtopic type="inf" number="6">
    Find pictures of wind turbines used
    for wind power.
  </subtopic>
  <subtopic type="inf" number="7">
    Find pictures of a wind farm.
  </subtopic>
</topic>
```

In the case of such topics, track participants were instructed to only consider the description field and ignore the various subtopics. Note that the description is always repeated as subtopic number 1. This was done in the interest of keeping the scale of the task to a defined limit, e.g., 20k topic-document pairs. NIST judges were required to assess the relevance of all documents retrieved for a given topic to each of the sub-topics.

The basic set contained 10 topics, which were randomly selected from the 50 TREC Web Track ad-hoc task topics: 202, 214, 216, 221, 227, 230, 234, 243, 246, 250. These were mostly single faceted topics with the exception of 202, 216 and 243.

4 Submitted Runs

Four groups submitted a total of 11 runs, see Table 1. Only one group (udel) submitted runs for the standard entry level set, the other three groups crowdsourced labels for the basic set of 3.5k topic-document pairs only.

The four groups followed very different approaches:

- The Hrbust team proposed a solution that leveraged social networking sites and multiple different crowds. They differentiated between three groups of crowds: Expert Group, Trustee Group and Volunteer Group. The expert group judged all 3.5k topic-document pairs, and asked their friends to contribute further judgments (trustee group). Additional judgments were collected for topic-document pairs by posting them on various social networking platforms (volunteer group). A consensus method was then used to obtain the final labels.

- The run by PRIS collected explicit rankings of documents from workers on Amazon's Mechanical Turk[2]. The result implemented a quality control mechanism at the task level based on a simple reading comprehension test.

- NEUIR's approach was based on using preference judgments made by a single judge (a graduate student), where the number of comparisons required for preference judgments was reduced following a Quick-sort like method, which partitioned the documents into n pivot documents and $n+1$ groups of documents that were compared to each other.

- The udel team submitted 8 fully automated runs that made use of three search engines and no human judges. The rating of a document was derived based on its position in each of the systems' rankings, where the different runs had slightly different rules.

5 Evaluation Measures

Using the judgments obtained from the trusted and highly trained NIST judges as gold standard, we measured the quality of the teams' submitted judgments. Participants may report in their own write-ups on the time, cost, and effort required to crowdsource the judgments.

We report results for the following three metrics:

- Rank Correlation: The Web Track participants' ad-hoc IR systems are scored based on NIST judgments according to the primary Web Track metric, ERR@20, inducing a ranking of IR systems. A similar ranking of IR systems is then induced from each Crowdsourcing Track participant's submitted judgments. Rank correlation is then calculated, indicating how accurately crowd judgments can be used to predict the NIST ranking of IR systems. The measure we use for rank correlation is Yilmaz et al.'s AP Correlation (APCorr) [4], which improves upon Kendall's Tau as a measure of rank correlation by emphasizing the order of the top ranked systems. To the best of our knowledge, the original version of APCorr

[2] www.mturk.com

does not handle ties; we handle ties by sampling over possible orders.

- Score Accuracy: In addition to correctly ranking systems, it is important that the evaluation scores be as accurate as possible. We use root mean square error (RMSE) for this measure.

- Label Quality: Direct comparison of each participant's submitted judgments against the NIST judgments (no evaluation of Web track IR systems). Label quality provides the simplest evaluation metric and can be correlated with the other measures predicting performance of IR systems. In previous years we reported logistic average misclassification rate (LAM), developed for the Spam Track [1], and Area Under Curve (AUC). However, LAM does not give any preference to ordering, and only works with binary classification. We introduced AUC to address ordering, but again, AUC deals with binary classification. Average precision (AP) is an alternative to AUC, but AP is also based on binary classification. Hence this year, we use graded average precision (GAP) [2]. The GAP is computed by ordering the documents as per the score assigned to the document and then using the qrels provided by NIST.

We provide our implementations of all 3 measures online in the active participants section of the TREC website[3].

6 Results

We report two sets of performance measurements. The first set of measurements is based on computing the mean ERR@20 for the 34 ad-hoc web track runs. Given a set of qrels submitted by a Crowdsourcing track participant, we compute the ERR@20 for each of the 10 randomly selected topics and then a mean ERR@20 across these 10 topics. Using the mean ERR@20 scores, we rank the web track runs and compare using APCorr the NIST qrels induced ranking to the Crowdsourcing track participant qrels induced ranking. Likewise, we compute RMSE based on the mean ERR@20. We take the mean ERR@20 produced using the NIST qrels as truth and measure the RMSE for the mean ERR@20 values produced by a participant's submitted qrels. These results are shown in Table 2. We note the number of documents used for the evaluation to highlight that NIST did not judge as much of the pool as anticipated, and thus there are fewer documents used for the evaluation than were actually judged in the runs submitted by Crowdsourcing Track participants.

[3] http://trec.nist.gov/act_part/tracks13.html

Team	#Docs	APCorr	RMSE
Hrbust	2758	0.480	0.135
NEUIR	2758	0.461	0.085
PRIS	2758	0.362	0.234
udel	2758	-0.172	0.155

Table 2: Evaluation results for the four primary runs based on the basic set of topics. The APCorr and RMSE values are computed based on the mean ERR@20 for the 34 ad-hoc web track runs.

The second set of measurements is based on averaging the per-topic APCorr, RMSE, and GAP scores. In this case, for each participant we have 10 scores and report the average. As can be seen in Table 3, the average per-topic APCorr and RMSE are worse than the APCorr and RMSE scores reported in Table 2. Ranking systems based on the mean ERR@20 should and does perform better than ranking systems based on ERR@20 for single topics for all groups except udel; the same result holds for RMSE.

When we look at the overall results in Table 2, we see that for APCorr, the highest score was obtained by the Hrbust team. APCorr reflects the agreement with the system rankings obtained using the ERR@20 measure and NIST judgments and a team's submitted judgments. When we look at the per-topic average APCorr in Table 3, we see that NEUIR has the best APCorr. While Hrbust obtained the highest APCorr score overall, and NEUIR has the highest average APCorr, the difference in average APCorr between Hrbust and NEUIR is not statistically significant by a paired, two-sided Student's t test (p-value = 0.16), nor are most differences between the groups with their per-topic APCorr values. It appears that 10 topics is not enough to distinguish performance differences in APCorr.

The best performance in terms of smallest obtained error between the obtained system scores, when using ERR@20 and the NIST qrels vs. the submitted judgments, is achieved by the NEUIR team. The average per-topic RMSE of NEUIR is also better than the next best system, Hrbust, but the difference is not statistically significant (p=0.06).

The highest quality of when looking at label quality directly is obtained by the NEUIR team both

Team	#Topics	Mean APCorr	Mean RMSE	Mean GAP
Hrbust	10	0.251	0.241	0.392
NEUIR	10	0.375	0.171	0.584
PRIS	10	0.203	0.311	0.481
udel	10	0.051	0.250	0.356

Table 3: Evaluation results for the four primary runs based on the basic set of topics. Here the results are the average of the per-topic scores.

overall and for the average per-topic GAP, and NEUIR's GAP performance is statistically significant compared to the next best by PRIS (p=0.006).

Both the NEUIR and the Hrbust teams relied on human judges, where the former collected judgments from a single (likely trusted) judge, while the latter relied on different groups of crowds as well as high redundancy (minimum 15 labels per topic-document pair) and a consensus method. In contrast the run by the udel team was fully automatic with no human input and did not match the performance achieved by other teams. This result seems to be consistent with earlier findings of Soboroff et al. in which blind evaluation by inducing qrels from rank fusion techniques is unable to distinguish whether outlier rankings are weak or strong [3]. Consequently, some degree of human labeling appears to remain necessary.

The difference between the NEUIR, Hrbust and the PRIS runs suggests the need for additional quality control methods on anonymous crowdsourcing platforms such as Amazon's Mechanical Turk.

7 Conclusions

The track this year tackled the issue of crowdsourcing relevance labels for web pages that were retrieved by the participants of the Web Track. This is the same task that faces NIST judges. However, in the interest of keeping the workload stable, we ended up not exactly reproducing NIST judging, in that participants did not have to judge sub-topics. Thus, it is still left for future work to try to reproduce NIST's full workload. Going even further, we planned, but did not run an additional "total recall" task with the goal to gather relevance labels for the 50 search topics over the complete ClueWeb12 corpus. This remains an exciting challenge for the future.

The overall performance scores obtained this year highlight the need for further research into methods to improve the quality of crowdsourced relevance judgments for IR evaluation. However, this challenge may be better addressed by collaborative efforts that combine expertise from multiple fields, beyond IR, such as incentive engineering, HCI aspects or game design. We take the low level of participation this year as further evidence to this. Thus, the crowdsourcing track will not run again next year at TREC. Research will, however, continue to be facilitated by similar initiatives, such as the CrowdScale 2013 Shared Task[4] that was launched recently with a broader set of annotation types and larger-scale real-world data sets. The MediaEval benchmarking initiative has also launched a crowdsourcing track[5] this year.

8 Acknowledgments

Special thanks to Gaurav Baruah for his on-going and extensive assistance with the preparation of the corpus and aspects of the evaluation. We are grateful to Ellen Voorhees and Ian Soboroff for their help and support with running the track. Thanks to the organizers of the Web Track, Kevyn Collins-Thompson, Paul N. Bennett, Fernando Diaz and Charles Clarke, for their help and collaboration. We thank Amazon and Crowd Computing Systems for sponsoring the track and providing credits or discounted prices to track participants.

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), the facilities of SHARCNET, the University of Waterloo, and by National Science Foundation Grant No. 1253413. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

References

[1] G. Cormack and T. Lynam. Trec 2005 spam track overview. In *Proceedings of TREC 2005*. NIST, 2005.

[4] www.crowdscale.org/shared-task

[5] www.multimediaeval.org/mediaeval2013/crowd2013

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

[3] I. Soboroff, C. Nicholas, and P. Cahan. Ranking retrieval systems without relevance judgments. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 66–73. ACM, 2001.

[4] E. Yilmaz, J. A. Aslam, and S. Robertson. A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 587–594, New York, NY, USA, 2008. ACM.