TREC 2011 Crowdsourcing Track

Version: 18 August 2011

Questions or comments regarding the track may be addressed to the coordinators, or more generally to the track mailing list for questions potentially of interest to other track participants.

Participation

To participate in the track, you first need to submit a formal application directly to NIST.

Participating teams will be expected to:

- Perform crowdsourcing experiments and submit their results, which will then be shared with other participants;
- Write a research paper describing their approach for inclusion in the TREC conference notebook (due Oct 15, 2011);
- Attend the TREC conference (Nov 15-18, 2011, at NIST's campus in Gaithersburg, MD); Per NIST policy, only those actively participating in the track can attend the conference; A limited number of speaking slots (typically 2-3) are typically awarded per track;
- Submit revised paper for inclusion in the final conference proceedings (due Feb 1, 2012).

We encourage collaboration between teams via discussions on the track mailing list and sharing of knowledge and resources. We will share information about participants and maintain a central repository of any resources (e.g., guidelines, code, templates) made available by participants.

Background: Crowdsourcing and Search Engine Evaluation

In Information Retrieval (IR), relevance judgments (aka relevance assessments or labels) are often collected from human judges (aka assessors or annotators) to evaluate search engine accuracy in retrieving relevant “documents” (e.g., web pages). Given the ever-growing size of digital collections searched today, such evaluation has become increasingly problematic for the IR field at large due to the cost, time, effort, and bias of traditional methods for collecting relevance judgments from in-house judges.

Crowdsourcing has emerged in recent years as a new avenue for obtaining labeled data from annotators online, be it through a Games with a purpose approach or via vendors such as Amazon Mechanical Turk (MTurk) and CrowdFlower.

While crowdsourcing offers tremendous promise in this regard, existing work suggests that achieving such benefits in practice requires a careful balance of various design issues such as: human factors (user interface, interaction mechanisms, incentive structures, questionnaire design), worker relations (recruiting, communications, retention), and quality control mechanisms (e.g. spammer detection, consensus algorithms). The importance of studying and better understanding such issues is clear in order to facilitate continuing evaluation of search engines at ever larger scales using human judgments.

Sponsorship

Teams are free to perform crowdsourcing using any platform or home-grown strategy they wish. Both pay and non-pay based approaches are possible.

In the event that teams choose to use (either or both) Amazon Mechanical Turk and CrowdFlower, we negotiated with the two vendors the following offers of support:

- **CrowdFlower** will provide a $100 credit towards the cost of a track participant's experiments performed on CrowdFlower. They can also assist teams with an introduction to the CrowdFlower self-service platform. See their 4/26/11 blog post. Please email Joseph Childress (joseph at crowdflower dot com), with cc to the track organizers, using the subject "TREC 2011 Crowdsourcing Track credit request" to receive your credit and inquire regarding technical support.
- We have *__tentative__* sponsorship from Amazon Mechanical Turk: Amazon would provide teams with a credit of approximately $300 to reimburse expense of using MTurk. We received verbal commitment, but are still waiting on funds.

Track Goals

The 2011 TREC Crowdsourcing Track, run at Text REtrieval Conference (TREC), is being organized to address the problem of *studying the use of crowdsourcing for search engine evaluation*.

The goals for this first year are to investigate:

1. **How to obtain high-quality relevance judgments from individual crowd workers;**
2. **How to effectively compute consensus[1] judgments from individual judgments;**
3. **Interaction between these (i.e., worker accuracy vs. subsequent consensus accuracy).**

Some of the research questions that participants may want to explore include:

- What are the most effective crowdsourcing strategies to obtain high quality topical relevance labels for a web page?
- What incentives and reward structures work best to attract and retain reliable workers?
- What does the accuracy vs. time/cost space look like as we collect additional redundant labels from multiple workers?
- What effort should be invested in redundancy vs. quality of workers for quality control?
- What factors influence the behaviour of workers?
- How can bias in well-meaning worker assessment be detected and corrected?
- How can sloppy workers, random clickers, adversarial behaviour be detected, corrected or filtered?
- What are possible sources of error, e.g., in task design, language, worker treatment?
- What is the impact of noise in the gathered relevance data for test collections and search engine evaluation?

Track Overview and Setup

[1] Computing consensus from multiple labels for the same given document (here a Webpage) is often referred to as "plurality", "multi-labeling", or "redundant labeling" and is used to compensate for "noise" / error in labels obtained from each individual annotator.

To investigate the issues outlined in the track goals, the track adopts a two-task task format, where participants can take part in either or both tasks:

Task 1: assessment. To investigate the effectiveness of crowdsourcing methods to gather relevance labels, in this task, participating teams will collect topical relevance judgments from multiple crowd workers for a subset of the ClueWeb09 collection.

Task 2: consensus. In the second task, teams will compute consensus (aka "label aggregation") over a set of individual worker labels provided by the organizers.

Task 1: Assessment

This task aims to evaluate the effectiveness of crowdsourcing methods (strategies and HIT designs[2]) to effectively collect relevance judgements from members of the crowd. Effectiveness is defined in terms of the quality of the workers attracted to the task measured based on the quality of the labels they contributed, the time taken to gather the labels and the incurred costs.

Participating teams will need to *collect topical relevance judgments from crowd workers for sets of topic-document pairs*, where a *set* comprises 5 documents for a given topic. The documents are web pages from a subset of the ClueWeb09 collection (see Data Provided by Organizers section).

Relevance judgements may be submitted in one or both of the following two forms:

- *Classification* judgments of each document's relevance to the given topic, expressed as a probability of relevance in the range [0,1]. Binary judgments indicate certainty of relevance (1) or non-relevance (0), while probabilities can be used to indicate non-certainty. "na" indicates no classification was performed.
- *Ranking* judgments over the set of 5 documents. Ranks are expressed by integers in the range [1,5], with 1 indicating the highest ranked document (judged to be most relevant). "na" indicates no ranking was performed.

We will separately evaluate both forms of judgments, so teams may choose to provide either or both forms of judgments. (While one might induce one form of the label from the other, the organizers will not do so and will rely on the explicit labels for each provided by each team).

Rules

Teams will need to submit the collected labels organised into runs, where the following rules apply:

- A run should be produced using a single HIT design (or template) and platform, i.e., the HIT design and platform cannot change within a run.
- A run should include at least one label per topic-document pair in the test set.
- A challenge of the task is to have workers judge all 5 documents in each set they judge. A worker who judges less than the full set of 5 documents in a given set will not be included in the evaluation.
- Judgments for evaluation must reflect the first time each given worker viewed the topic-document pairs judged.
- There is no limit on the number of labels teams can collect for a given topic-document pair.

[2] "HIT" stands for Human Intelligence Task, a term coined by MTurk. We use the term loosely in this document to refer to a user interface / task design presented to a human judge using any form of crowdsourcing.

- Teams need to submit **ALL** collected judgments on the test data from each crowd worker used. This includes any rejected work.

Submissions will be evaluated in terms of *average worker quality* obtained in a given run based on ranking and classification metrics (see Evaluation section) against NIST judgements and consensus label across teams. Secondary evaluation will be based on self-reported time and cost.

Task 2: Consensus

Teams will be provided a set of individual worker labels generated by the organizers for sets of topic-document pairs, for which teams need to compute consensus labels. Ground truth NIST judgments (i.e., labels) will be provided for a subset of the examples. This data is intended to facilitate development and self-evaluation of supervised, semi-supervised, or fully-unsupervised consensus methods.

As in Task 1, submitted consensus judgments will be evaluated using ranking and classification metrics (see Evaluation section). Accuracy will be computed with respect to the same two versions of “ground truth”: prior NIST judgments and majority vote for each example across teams.

Submissions

Runs may be submitted to NIST at:

http://ir.nist.gov/trecsubmit/crowd.html (this site is not yet live)

Up to 3 runs may be submitted for each of the two tasks: one run designated as primary for main evaluation, and up to two additional runs for secondary evaluation. Teams are responsible for controlling against worker-training effects, e.g. the same worker judging the same (topic,document) pair in different runs.

Task 1

The naming of a run file should include the team’s name (NIST ID, that was used to register to participate in the track), the task number (i.e., “Task1”), the crowdsourcing platform used, and an informative name (as far as possible) describing the HIT design, e.g., “RedEaglesTeam.Task1.MTurk.trainingGame.txt”. For homegrown systems, please use short but meaningful names; for Amazon’s Mechanical Turk use “MTurk” and for CrowdFlower use “CF”.

The run file should be a tab-separated file, where each **line** should contain judgements for a given topic-document pair by a given worker, sorted by the <worker-id>:

```
<team-id> <worker-id> <set-id> <topic-id> <doc-id> <rank-label> <class-label>
<assignment-id> <worker-time> <label-cost> <label-info>
```

Fields are defined as follows:

<team-id>: the team number that was assigned to your team when you confirmed participation during the training phase. See Appendix.

<worker-id>: the unmodified ID of the worker as provided by the crowdsourcing platform (or home-grown system) used. Use “na” if worker IDs are not available in the system. Note however that in that case you can only submit a single label per topic-document pair as per rules ‘one worker’ cannot judge the same pair more than once. The organizers will anonymize these IDs in any distributed data, but still allowing to inspect internally the degree to which the same worker has done judging for multiple teams (i.e. worker training effects).

<set-id>: the ID number of a set of 5 topic-document pairs in the test data where all 5 pairs in a set need to be judged by a given worker. Use “na” for non-test data the worker judged for training or quality control, etc.

<topic-id>: the ID number associated with the topic in the test data set (or in the training set), see Data and Software Provided by Organizers section.

<doc-id>: the ClueWeb document number, identifying the Webpage in the ClueWeb collection, see Data Provided by Organizers section.

<rank-label>: the rank value (1 to 5) assigned by the worker, starting from 1 for the topmost rank. If no ranking labels were collected, use “na” in this column.

<class-label>: probability of relevance in range [0,1]. Binary judgments indicate certainty of relevance (1) or non-relevance (0), while probabilities allow for non-certainty. If no classification labels were collected, use “na” in this column.

<assignment-id>: a unique ID number, identifying a specific HIT instance that the given worker completed (use “na” if this is not applicable to your crowdsourcing design/platform). On Amazon Mechanical Turk, this is reported in the AssignmentID column. This information will tell us if a worker judged all documents in a set for a given topic within a single HIT or across multiple HITs. For example, if a HIT contains multiple topic-document pairs, the assignment-id will be the same for all these topic-document pairs judged by the given worker (different workers will be associated with different assignment IDs).

<worker-time>: the total time in seconds that the given worker spent on the task divided by the number of topic-document pairs included in the task. For example, if Amazon’s Mechanical Turk was used, where the HIT collected two relevance judgements, then the WorkTimeInSeconds, as reported by Amazon, should be divided by 2 when generating information for the run submission. This should be done even if only 1 of the 2 topic-document pairs was actually labeled by the worker in the HIT.

<label-cost>: the cost per label that was paid to the worker plus any commission charged by the platform used. So, if a HIT on Amazon Mechanical Turk included 6 documents and paid $0.30, then a label cost $0.05 plus the 10% commission by Amazon. For rejected work that did not incur a cost, use $0. Same applies if a non-pay method was used to gather labels, e.g., a game.

<label-info>: a simple enumeration indicating additional metadata about the label:

0. default
1. “rejected” label: where you would have filtered this label out before subsequent use
2. automated label: label was produced by automation (“artificial artificial artificial intelligence”)
3. training / quality-control label: used in training/evaluating worker, not for labeling test data

In addition to the above, participants are encouraged to submit any additional data relating to the HITs and workers in a separate file, named the same as the run but ending with “.extra” (e.g., “RedEaglesTeam.Task1.MTurk.trainingGame.extra”). This may contain basic information, such as the document file format used (e.g., PDF, image, html), the overall time needed to collect all the labels, whether the assignment restricted worker participation to workers who passed a qualification test or workers from a specific country, etc. The file may be a simple text file, or a comma separated file, listing the attributes of the HITs or assignments.

Task 2

As in Task 1**,** the naming of a run file should include the team's name (NIST ID, that was used to register to participate in the track) and the Task number, e.g., "RedEaglesTeam.Task2". The run file should similarly be a tab-separated file, where each **line** should contain consenus judgements for each given topic-document pair. As in Task 1, teams can focus on classification and/or ranking metrics. Submission format is a subset of the fields used in Task 1 - see above for descriptions of the fields.

```
<topic-id> <doc-id> <rank-label> <class-label>
```

Training and Test Phases

Both the assessment and consensus Tasks distinguish between development (aka training) and test phases, with their own separate data sets (see Data Provided by Organizers section). During training, teams may experiment with different techniques and platforms using the training data, evaluating these using the available ground-truth data.

Any experiments and obtained labels on the test data must be submitted for evaluation.

Data Provided by Organizers

Rendered Subset of the ClueWeb Collection

The track uses a subset of the ClueWeb09 collection (see also: ClueWeb09 wiki). Teams will **not** need to purchase the ClueWeb09 collection to participate. To gain access to the collection, please follow the instructions on https://sites.google.com/site/treccrowd2011/ (in section "Download Crowdsourcing Track subset of ClueWeb09").

To address the problem of possibly malicious code in web pages, participants are provided with "rendered" versions of all Webpages in the ClueWeb subset: images, PDF, and plain text. Teams can use one or more of these files and report which formats were used.

Additional issues of note:

- Some Webpages may contain objectionable content;
- Differences in Web browsers or screen resolutions may impact relevance judgments;
- ClueWeb09 images were crawled several months after the initial pages and so some may be missing due changes in the intervening period.

Alternative Collection Formats

The originally-crawled .html and images are available at request for any teams wishing to have assessors judge this data directly, to re-render it into one of the formats above, or to render it into a different format. However, we explicitly want to avoid confounding effects of better accuracy being achieved by any team as the result of the different rendering quality or use of a unique format being judged, so we will make the following stipulations:

- Any team performing judging based on the originally-crawled data must publicize in advance how they will protect workers from possibly malicious code, and if any resource or infrastructure is used to do so, provide that infrastructure to other participating teams;
- Similarly, any team that re-renders Webpages should either: re-render and distribute all Webpages used by the track, or distribute the software (or its name if 3rd party) and instructions for other teams to re-render their pages using this software;
- Any team that re-crawls ClueWeb09 pages to correct errors from the initial crawls must similarly distribute all Webpages used in the track or the software (or its name if 3rd party) and instructions for its use.

Whatever format is used, it will be left to participants to host individual files online for crowd workers to judge.

Topics

The track uses 270 topics in total. Most of these are used in the training phase, leaving 28 topics for the test phase. Topics are identified by an ID number and have several fields (any of which can be used in the crowdsourcing experiments). An example topic is shown below.

```
<field name="number"> 20001 </field>
<field name="query">obama family tree</field>
<field name="description">Find information on President Barack Obama's family history, including genealogy, national origins, places and dates of birth, etc. </field>
<field name="narrative">Relevant documents will give information on Barack Obama's family history, both immediate and past. A list of names of Obama's family members is not relevant without further information.</field>
<field name="category">1</field>
```

Topics are available from https://sites.google.com/site/treccrowd2011/ (under the “Task 1” heading).

Topic-document pairs

A topic-document pair identifies a topic by its ID number and the document by its ClueWeb ID number, e.g., 20374 clueweb09-en0000-14-22146.

Training data

For Task 1 training, each team who has confirmed participation in the Task was assigned a team number (currently between 0 and 14), with data partitions assigned to each team number. Each team is assigned topic-document pairs that they can get judgements for from crowd workers and thus test their crowdsourcing methods and designs. There are several files with different sets of topic-document pairs that teams can use however they wished:

- The “assigned” files contain topic-document pairs, where teams should only use those that are assigned to their team number.
- The “shared” topic-document pairs can be used by all teams.
- For both “assigned” and “shared”, some of the topic-document pairs have prior NIST judgments (“judged” sets), and some do not (“unjudged” sets).
- The “judged” sets are divided into “balanced” and “imbalanced” sets. In the balanced set, there are sets of 5 topic-document pairs which all have prior NIST judgments, and those judgments are

balanced such that there is a (randomized) mix of relevant and non-relevant documents in the set of 5 documents per topic. For example, for team 2, for the topic 20374, the following 5 documents are included in the balanced set (the last column shows the relevance labels provided by NIST):

```
2  20374  clueweb09-en0000-00-00000  0
2  20374  clueweb09-en0000-14-22146  2
2  20374  clueweb09-en0002-75-21862  0
2  20374  clueweb09-en0003-48-01649  1
2  20374  clueweb09-en0003-48-03560  1
```

The “imbalanced” data provides further judgments for the topics in the balanced set assigned to teams, but this data is imbalanced in one or two senses: insufficient documents (less than 5 documents for a topic) and/or not diverse (all documents being relevant or irrelevant).

- In addition to the above, where unique topic-document pairs are assigned to teams, the ".topics" files assign teams exclusive right to certain topics and all documents (judged and unjudged) associated with those topics.

For the training phase of Task 2, the organizers will provide prior examples of worker judgments produced outside of the track. The number and quality of worker labels from the training data may differ from that distributed in the testing phase based on Task 1 (assessment) submissions by teams.

The topic-document pairs with/without prior NIST judgements for the development phase for both Task 1 (assessment) and Task 2 (consensus) are available from https://sites.google.com/site/treccrowd2011/ (under the Task 1 and Task 2 headings).

Test data

Task 1:
The test data set for Task 1 consists of previously unseen topic-document pairs. The test set comprises sets of 5 topic-document pairs, some with prior NIST labels. There are two sub-sets:

- The “assigned” topic-document pairs, where teams should only collect labels for those that are assigned to their team number. Each team has 5 topics, where the topics vary from team to team. The format of the file is as follows, with columns <team-id> <set-id> <topic-id> <doc-id> <NIST-label-or-minus1-if-no-NIST-label-is-given>:

```
5  823  20424  clueweb09-en0001-90-18599  1
5  823  20424  clueweb09-en0000-56-04197  2
5  823  20424  clueweb09-en0001-94-08915  0
5  823  20424  clueweb09-en0004-90-07845  0
5  823  20424  clueweb09-en0011-54-04607  1
```

- The “shared” topic-document pairs that should be judged by all teams. Each team will work with the same set of 15 topics (different from those in the assigned set). The file is in the form with columns <set-id> <topic-id> <doc-id> <NIST-label-or-minus1-if-no-NIST-label-is-given>:

```
403  20542  clueweb09-en0000-00-29790  -1
403  20542  clueweb09-en0000-21-11312  -1
403  20542  clueweb09-en0000-35-23724  -1
403  20542  clueweb09-en0000-36-01164  -1
```

```
403  20542  clueweb09-en0000-94-06426  -1
```

In total, teams will need to collect relevance labels (ranking or binary labels) for about 2200 topic-document pairs across 30 topics (20 topics per team). Both assigned and shared sets include a small number of topic-document pairs with NIST judgements. Participants may use these labelled sets for training workers or as honey-pots, but keeping to the same rules as described in Task 1 Task section.

Task 2:
Teams will compute consensus for provide prior examples of worker judgments produced outside of the track.

Evaluation

Primary runs will serve as the primary basis for evaluating teams; optional runs provide secondary evaluation information.

Classification labels will be used with classification metrics, and similarly ranking labels for ranking metrics.

Classification Metrics

Short-hand used: t = true, f = false, p=positive (relevant), n=negative (non-relevant)
Precision = tp / (tp + fp)
Recall = tp / (tp + fn)
Accuracy = tp + tn / (tp + fp + tn + fn) = # correct / # of examples   (classes may be imbalanced)
NOTE: counts above may be fractional rather than integer but metrics remain well defined (e.g. submitting label "0.5" for a single relevant example would yield precision 0.5/1 = 0.5).

Task 1: Evaluation will only consider judgments for each set from workers who judged the entire set, and average per-worker metric performance across those workers.

Task 2 submissions will generate a single consensus label per example, for which we will compute the same metrics as with Task 1.

Ranking Metrics

We will report the following ranking metrics; others may be added based on suggestions from participants and available time.

- Normalized Discounted Cumulative Gain (NDCG)
- Mean Average Precision (MAP)

Ground Truth

Ground Truth will be computed with respect to two different benchmark sets:

- Human judgements by NIST (for a subset of Webpages for which such judgments exist)
- Generated judgments by computing consensus across teams

Note that prior NIST judgements are ternary (non-relevant (0), relevant (1), and highly relevant (2)

) whereas consensus will yield binary labels. NDCG is defined for both, while for MAP and accuracy, relevant and highly-relevant NIST labels will be conflated to reduce the ternary judgments to binary.

Generated judgments. In the absence of gold-standard judgments, we will evaluate teams on the basis of generated judgments produced by taking consensus across teams.

With Task 1 class labels from workers, we can produce a single class label for each example for each team by averaging across that team's worker labels for that example. To create ground truth, we then average labels across teams as well. Similarly, Task 2 consensus labels from each team can be averaged across teams. This produces a probability of relevance as the ground truth for classification-based evaluation. The primary evaluation will simply round these probabilities to {0,1} to produce binary labels for evaluation using established metrics. Secondary evaluation can instead preserve this uncertainty in the ground truth and compute one or more divergence measures between submitted judgments and this probabilistic ground truth.

With rank labels, we must be able to arrive at a consensus ranking given multiple input rankings (e.g. coming from multiple workers and/or multiple teams). For simplicity, transparency, and clarity, we adopt the following trivial method for this rank fusion. Assume documents A-E are ranked from most relevant (A) to least relevant (E). We will assign documents "points" according to a simple linear discount, with A getting 5 points to E getting one point. We will then sum these points for each ranking across all rankings, then generate a consensus ranking of documents from most to least points. We recognize the lack of sophistication of this method of rank fusion, but again it is simple, transparent, and unbiased, and it provides a starting point for comparison against more sophisticated methods.

Secondary evaluation

In addition to the accuracy of the crowdsourced solutions, we are interested in time and cost. We ask teams to self-report these as part of the submission runs, restricting to the direct crowdsourcing costs (i.e., payments to workers) and the direct time that a worker took to generate a label. Additional time and effort information should be reported in the research papers, e.g., time taken to set up experiments, any training or manual tuning.

Resulting data

Participants will gain access to all collected data, with any information removed that may identify individual workers. A signed form may need to be submitted to NIST to obtain some of this data (details forthcoming).

Schedule

August 8: Release of Task 1 test data
ASAP: Release of Task 2 test data
September 15: Task 1 & 2 submissions due
October 1: Preliminary results announced to participants
October 15: Team papers due to NIST for inclusion in TREC conference notebook
Nov 15-18: Text REtrieval Conference (TREC 2011) in Gaithersburg, MD
February 1, 2012: Final team papers due to NIST for inclusion in TREC conference proceedings

Track Coordinators

- Gabriella Kazai, Microsoft Research
- Matthew Lease, University of Texas at Austin

Appendix - Team id numbers for the Task 1 Assessment task

If your registered with NIST to participate and your team is not listed below and you would like to participate in the Task 1 task, then please contact the track coordinators with your NIST ID.

0. USCUD
1. uogTr
2. NLPLDLUT
3. BUPT-WILDCAT
4. uc3m
5. RMIT
6. GeAnn
7. Microsheffware
8. wis_tudelft
9. L3S
11. MSRC
12. TUD_DMIR
13. UWaterlooMDS
14. UCSC