

University of Glasgow (qirdcsuog) at TREC Crowdsourcing 2011: TurkRank - Network-based Worker Ranking in Crowdsourcing

Stewart Whiting, Jesus Rodriguez Perez, Guido Zuccon, Teerapong Leelanupab and
Joemon M. Jose

{stewh, jesus, guido, kimm, jj}@dcs.gla.ac.uk

School of Computing Science, University of Glasgow,
Scotland, G12 8QQ, UK.

Abstract. For TREC Crowdsourcing 2011 (Stage 2) we propose a network-based approach for assigning an indicative measure of worker trustworthiness in crowdsourced labelling tasks. Workers, the gold standard and worker/gold standard agreements are modelled as a network. For the purpose of worker trustworthiness assignment, a variant of the PageRank algorithm, named TurkRank, is used to adaptively combine evidence that suggests worker trustworthiness, i.e., agreement with other trustworthy co-workers and agreement with the gold standard. A single parameter controls the importance of co-worker agreement versus gold standard agreement. The TurkRank score calculated for each worker is incorporated with a worker-weighted mean label aggregation.

1 Introduction and Related Work

In the TREC 2011 Crowdsourcing track we worked on the Stage 2 Task to extract single relevance labels from an aggregated set of crowdsourced document relevance labels. In this report we describe our approach to the problem using a network-based algorithm to assign worker trustworthiness.

Crowdsourcing is a broad term used to describe any technique of obtaining data from multiple people on a large-scale. Whilst the concept of crowdsourcing has undoubtedly existed for a long time, web-based platforms such as Amazon Mechanical Turk¹ (AMT) or CrowdFlower² have become a popular tool to facilitate markets in which geographically-dispersed workers complete the tasks of requesters for financial reward. Such platforms allow requesters to package their work as relatively small tasks (commonly known as Human Intelligence Tasks, or, HITs) and instantly advertise them to the pre-recruited worker population, with satisfactory completion attracting a reward and often bonus payment.

However, whilst the use of crowdsourcing may seem very favourable for repetitive labour-intensive tasks such as labelling document relevance to a query, the reliability and quality of the collected labels has been strongly questioned. With the incentive

¹ <http://www.mturk.com>

² <http://www.crowdflower.com>

of maximising earnings, some workers carelessly random-click in the hope of work acceptance. Likewise, automated systems (i.e. bots) setup to repeatedly submit random data have also become increasingly common with the increased financial incentives. To combat these issues, a number of approaches to improve data quality have been proposed and researched in the context of labelling tasks.

Aside from pre-screening workers through qualification requirements (e.g. previous work acceptance rates) or tests, intelligent HIT design can provide a first-line defence against casual random-clickers and bots. Validation can be through CAPTCHAs, whereby workers are asked to answer a simple question (such as $6 + 5 = ?$) or repeat a string of characters shown in a distorted image (to avoid optical character recognition). Similarly, many have proposed task-specific validation questions, the answers for which can be found within the content of the task itself [2]. Alternatively, rigidly controlling the workflow with submission control timers and strict input validation can also be used.

Multiple redundant labels can be sourced for each judgement, and following data collection, aggregation can determine the most likely label. Majority voting is the simplest method, however, obtaining multiple labels for each HIT can be expensive, and even then, not necessarily yield the correct aggregated label if there are multiple low-quality labels. Additional judgements can be crowdsourced (or low-quality ones removed) to improve certainty by using statistics such as Fleiss' or Cohen's Kappa coefficient to measure the inter-worker agreement, and the chance agreement occurred randomly. There has been a significant amount of work to apply machine learning (ML) strategies to detecting poor quality workers [3]. Meanwhile, some have suggested heuristics, such as work time or reward level, approximately correlate with work quality, whereas others have reported no such connection.

Similar to previous work [1], our approach provides a score of worker quality, or *trustworthiness*. We model workers, the gold standard and worker/gold standard agreements as a network. Using a variant of the PageRank algorithm, named TurkRank, we combine evidence that suggests worker trustworthiness, i.e., agreement with other trustworthy co-workers as well as agreement with the gold standard. The TurkRank assigned to each worker is then used to weight the importance of the worker's label during aggregation.

2 Approach

We consider worker trustworthiness to be a measure of the likelihood that a worker is performing valuable work: providing correct labels. Evidence for worker trustworthiness can be incrementally accumulated through co-worker label agreement and gold standard label agreement; with the availability of either source of evidence determining its importance. The more a worker agrees with other worker labels, the more trustworthy they are likely to be. However, particularly important is the trustworthiness of the agreeing worker (whether it is another crowdsourced worker, or a gold standard NIST assessor). A transitive relationship where workers agree with other workers and so on suggests an implicit network of trust propagation. Extensively studied in many applications is the PageRank (PR) [4] family of algorithms to measure relative importance

within a network (e.g. web page authority in a link graph). We make use of the extended PageRank with Priors (PRwP) [6] algorithm to variably combine the evidence provided by both co-worker and gold standard agreement, naming our approach TurkRank.

TurkRank models crowdsourced workers, the gold standard and label agreements as a graph. Vertices are crowdsourced workers or the gold standard worker (i.e. NIST assessor). Edges are undirected positive label agreements between workers (binary, with $weight = 1$). An example network can be seen in Figure 1.

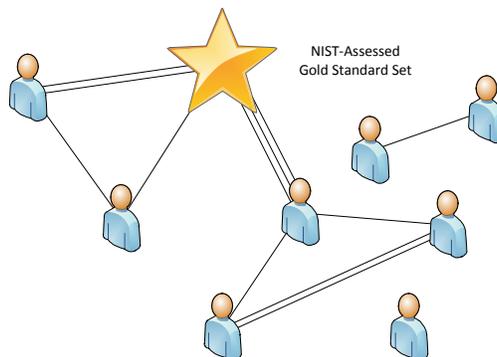
The PR algorithm is effective for discovering nodes with a high relative importance in a network, in this scenario, workers who have strong work agreement with other workers. To combine both worker and gold standard agreement in this estimation of relative importance, PRwP extends traditional PR by including vertex priors. The PageRank π (i.e. measure of trust) for a worker vertex v at iteration i is therefore calculated as:

$$\pi(v)^{(i+1)} = (1 - \beta) \left(\sum_{u=1}^{d_{in}(v)} p(v|u) \pi^{(i)}(u) \right) + \beta p_v \quad (1)$$

Priors influence the likelihood of the random walker jumping to a given vertex when teleporting, if the probability of teleporting is > 0 . PRwP has a single parameter regulating teleport probability, β .

Rather than teleporting to any vertex with equal probability (as in traditional PR), the gold standard vertex is assigned a 100% prior probability. Therefore, when $\beta = 1$, the random walker will always teleport to the gold standard worker, and so only it will accumulate trust. Conversely, when $\beta = 0$, the random walker will never teleport and so will move using edge-based co-worker agreements only, excluding the gold standard. With $0 < \beta < 1$, worker agreement and gold standard agreement will be combined. With a higher β , greater trust will be given to those workers who agree with the gold standard, whereas, a lower β will give more trust to those workers who agree more with similarly agreeing workers. The sum of the TurkRanks assigned to all workers is always $= 1$.

Fig. 1. Example network model, showing label agreements (edges) between co-workers and the gold standard NIST assessor.



2.1 Worker Label Aggregation

A weighted mean is used to incorporate the worker TurkRank in label aggregation, thus emphasising label contributions from more trustworthy workers. t is the set of worker trust ranks and l is the set of binary worker relevance labels (0/1, non-relevant/relevant).

$$\bar{l} = \frac{\sum_{i=1}^n t_i l_i}{\sum_{i=1}^n t_i} \quad (2)$$

\bar{l} is rounded to the nearest integer, such that ≤ 0.49 is considered non-relevant and ≥ 0.5 relevant.

3 Experiment and Results

3.1 Methodology

Evaluation was performed using the TREC 2011 Crowdsourcing Stage 2 dataset. The dataset contains 19,033 topic-document pairs, with judgement labels made by 762 workers, who produced a total of 89,624 binary relevance judgments. 2,275 of the topic-document judgements have prior “gold” relevance judgements by trusted NIST assessors (1,000 non-relevant, 1,275 relevant).

To investigate the effect of different β settings (i.e. importance of the gold standard agreement versus co-worker agreement) we vary $0 \geq \beta \leq 0.9$ at 0.1 intervals. For simplicity we report $\beta = 0, 0.2, 0.4, 0.6, 0.8$ only in this paper. Our primary TREC submission run for the track (as team “qirdcsuog”) was with $\beta = 0.4$. We submitted both binary relevance labels and a relevance ranking order, based on the descending order of weighted means.

3.2 Results and Discussion

Figure 2 shows the quantity of relevant labels produced after weighted mean aggregation for each β setting. A linear relationship is indicated with decreasing relevant judgements as the importance of agreement with the gold standard is increased in TurkRank assignment.

We are not certain of the cause of this relationship. One speculation posed, based on the design of the HIT form in the data collection study [5] is that poor workers may simply click the first available relevance selection (i.e. “Very Relevant” in this case) before submission. As these poor workers are filtered out with an increasing β for more gold standard reliance, the value of their previously majority judgements are reduced by the weighted mean.

Figure 3 shows the logarithmic distribution of worker TurkRanks for each β setting. Noticeable is the increasing decline in TurkRank distribution towards the lower-ranked workers, particularly at higher β levels (e.g. 0.8).

Figure 4 presents the raw weighted means calculated for each topic-document judgement, before binary rounding. There are clearly substantially more definitely relevant topic-document judgements than definitely non-relevant judgements. A large quantity of the weighted means are marginally above or below the cut-off 0.5 threshold. For these, additional labels should be sought to make a more reliable binary judgement.

Fig. 2. Effect of varying β on the number of “relevant” labels produced after aggregation.

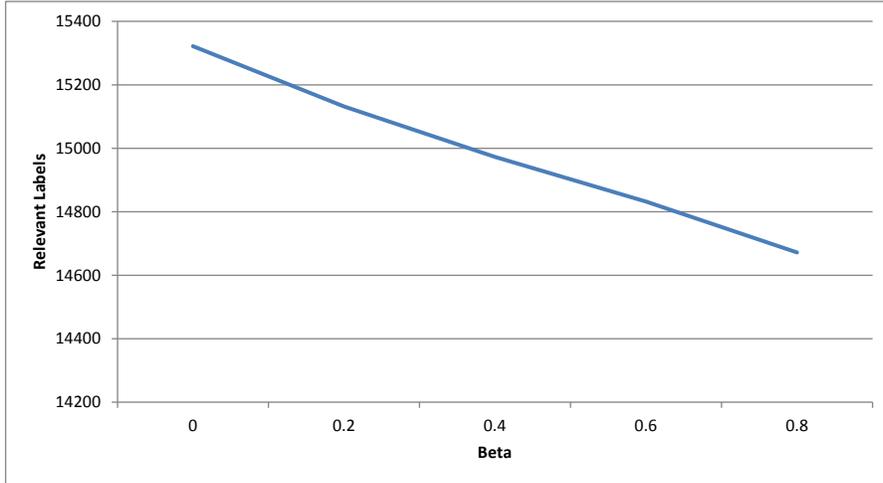


Fig. 3. Logarithmic distribution of worker TurkRank for all β settings.

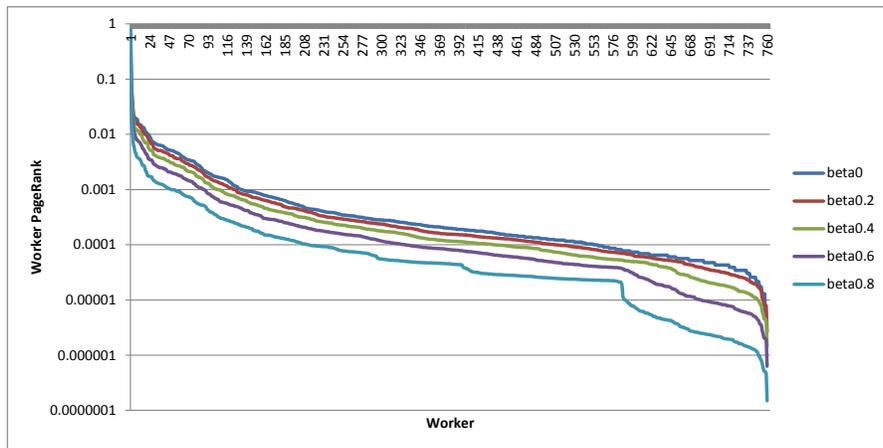


Fig. 4. Distribution of TurkRank-weighted mean for all topic-document judgements (before binary label rounding) for all β settings.

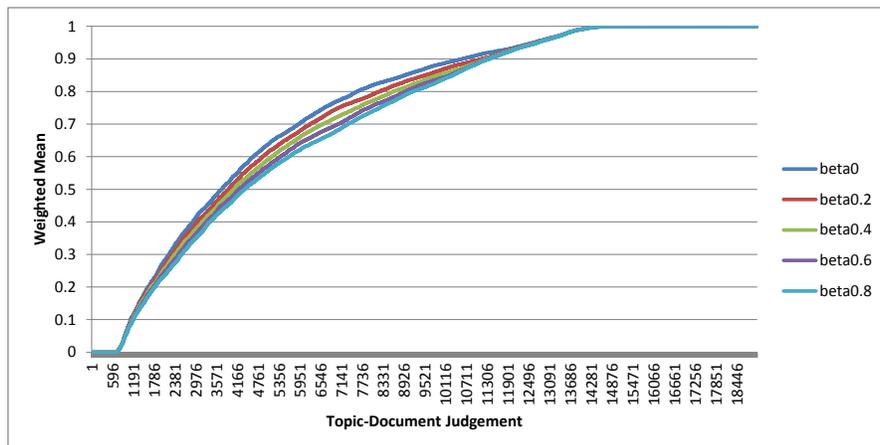


Figure 5 presents the effect of the β parameter setting on recall and precision (against the *Gold 1000* ground truth). There is relatively little variance in precision, however, recall does increase as β is reduced.

Table 1. Preliminary results for our primary submitted run, with $\beta = 0.4$.

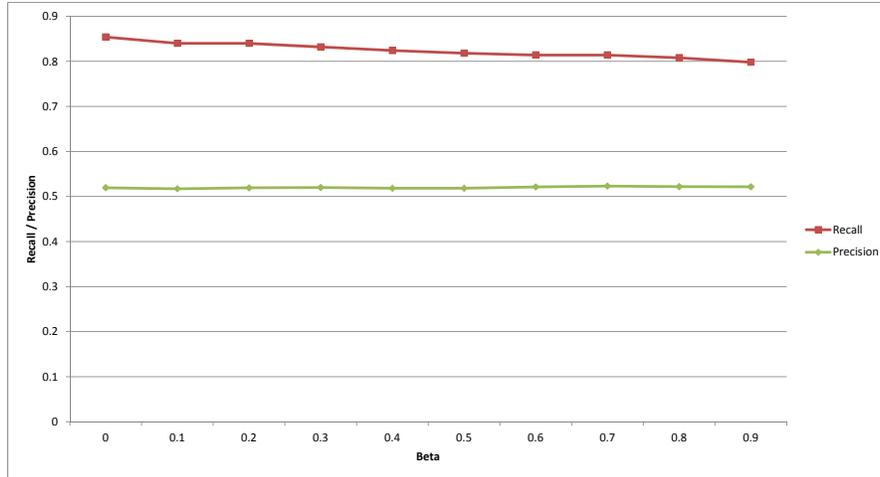
	Run	Accuracy	Recall	Precision	Specificity	Log Loss	KL-Div.	RMSE	MAP	NDCG
<i>Consensus 16785</i>										
	Consensus	76.9%	80.0%	83.0%	71.6%	4630.0	8384.0	0.0%	99.9%	99.9%
	qirdcsuog	72.0%	89.8%	72.6%	41.3%	43168.9	50317.7	40.4%	76.3%	91.8%
<i>Consensus 1000</i>										
	Consensus	77.0%	80.1%	83.3%	71.5%	275.6	496.3	0.0%	19.1%	47.3%
	qirdcsuog	70.3%	88.9%	71.6%	37.2%	2735.5	3050.2	41.2%	10.6%	37.7%
<i>Gold 1000</i>										
	Consensus	61.7%	73.3%	59.5%	50.2%	647.1	647.1	47.0%	16.3%	41.2%
	qirdcsuog	52.9%	82.4%	51.8%	23.4%	4338.1	4338.2	68.6%	8.6%	32.6%

Table 1 shows the preliminary evaluation results for our primary submitted run, with $\beta = 0.4$. The baseline is provided by the consensus for all examples: 16785 judgements, 1000 judgements and the 1000 gold set judgements. With the exception of recall, TurkRank at $\beta = 0.4$ reduces performance compared to the baseline. Recall is enhanced, however at the cost of precision.

4 Conclusion

Preliminary evaluation results suggest that while TurkRank may be ranking workers based on their trust, optimal effectiveness (and possible baseline improvement) is very

Fig. 5. Relevance label recall and precision at $0 \leq \beta < 1$ using *Gold 1000* ground truth labels.



dependent on the β parameter setting. An arbitrary β of 0.4 is clearly not adequate in the evaluation scenario. A deeper analysis is required to achieve optimisation, in particular on the behaviour of β at different levels, given different scenarios with varying levels of gold standard and redundant worker labelling evidence. Similarly, the weighted mean approach to aggregating worker TurkRank and label evidence may not be the best approach to incorporating a worker trustworthiness measure.

References

1. P. G. Ipeirotis, F. Provost, and J. Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation, HCOMP '10*, pages 64–67, New York, NY, USA, 2010. ACM.
2. G. Kazai, J. Kamps, M. Koolen, and N. Milic-Frayling. Crowdsourcing for book search evaluation: impact of hit design on comparative system ranking. In *SIGIR*, pages 205–214, 2011.
3. M. Lease. On Quality Control and Machine Learning in Crowdsourcing. In *Proceedings of the 3rd Human Computation Workshop (HCOMP) at AAAI*, pages 97–102, 2011.
4. L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.
5. W. Tang and M. Lease. Semi-supervised consensus labeling for crowdsourcing. In *ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR)*, 2011.
6. S. White and P. Smyth. Algorithms for estimating relative importance in networks. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '03*, pages 266–275, New York, NY, USA, 2003. ACM.