

How to Robustly Combine Judgements from Crowd Assessors with AWARE [★]

Marco Ferrante¹, Nicola Ferro², and Maria Maistro²

¹ Department of Mathematics, University of Padua, Padua, Italy
`ferrante@math.unipd.it`

² Department of Information Engineering, University of Padua, Padua, Italy
`ferro@dei.unipd.it`, `maistro@dei.unipd.it`

Abstract. We propose the *Assessor-driven Weighted Averages for Retrieval Evaluation (AWARE)* probabilistic framework, a novel methodology for dealing with multiple crowd assessors, who may be contradictory and/or noisy. By modeling relevance judgements and crowd assessors as sources of uncertainty, AWARE directly combines the performance measures computed on the ground-truth generated by the crowd assessors instead of adopting some classification technique to merge the labels produced by them. We propose several unsupervised estimators that instantiate the AWARE framework and we compare them with *Majority Vote (MV)* and *Expectation Maximization (EM)* showing that AWARE approaches improve both in correctly ranking systems and predicting their actual performance scores.

Keywords: crowdsourcing, unsupervised estimators, AWARE

1 Introduction

Ground-truth is central to the data processing area, as in top-k ranking in databases, information retrieval, natural language processing, video and image processing, information extraction and many others. Although ground-truth enables the scoring and comparison of algorithms with respect to human judgements, creating a dataset and, in particular, gathering relevance assessments is an extremely demanding activity, therefore there is an increasing interest for more effective and affordable ways of gathering assessments [3].

Crowdsourcing [4] has emerged as a viable option for ground-truth creation since it allows to cheaply collect multiple assessments for each task. However, it raises many questions regarding the quality of the collected assessments. Therefore, in order to obtain a ground-truth good enough to be used for evaluation purposes, the possibility of discarding the low quality assessors and/or combining them with more or less sophisticated algorithms has been considered.

The problem of merging multiple crowd assessors has been addressed mostly from a classification point of view, with traditional approaches which focus

[★] Extended abstract of [2].

mainly on how to select assessors and/or discard low quality assessors and how to merge judgments from multiple assessors. We can consider this as a kind of “upstream” approach, because the aggregated ground-truth is created before systems are evaluated and performance scores are computed.

In this paper, we address the problem of ground-truth creation from a new angle, i.e. we investigate how to estimate performance measures in a way more robust to crowd assessors. In particular, we seek a better estimation of the true expected value of a performance measure, by leveraging its multiple observations, generated separately by the relevance judgements of each crowd assessor. We can consider this as a kind of “downstream” approach, since the aggregation happens after performance measures have been computed.

The main intuition behind our approach is based on the idea that the choice of the “best” relevance judgments, operated ahead at the pool level, may have a diverse impact on different systems and on various performance measures. Indeed, systems rank the same documents differently and therefore the same correctly labelled or mis-labelled documents impact the performances of different systems in different ways. Therefore, we propose the *Assessor-driven Weighted Averages for Retrieval Evaluation (AWARE)* probabilistic framework, which allows us to combine multiple versions of a performance measure, computed from the ground-truth created by each crowd assessor, into a single composite measure, referred as the AWARE version of it. The AWARE framework specifies how performance measures have to be merged on the basis of the estimated crowd assessor accuracies and we propose several unsupervised estimators of such accuracies. The experimentation shows that AWARE approaches improve in terms of capability of correctly ranking systems and predicting their actual performance scores.

The paper is organized as follows: Section 2 introduces the AWARE framework; Section 3 gives an intuitive overview of several unsupervised estimators for determining the assessors accuracies; Section 4 carry out the experimental evaluation using TREC collections; finally, Section 5 draws some conclusions.

2 The AWARE Framework

In [1] we introduced the following definitions: let D and T be a *set of documents* and a *set of topics*, respectively; let (REL, \preceq) be a totally ordered *set of relevance degrees*. For each pair $(t, d) \in T \times D$, the *ground-truth* GT is a map which assigns a relevance degree $rel \in REL$ to a document d with respect to a topic t .

In order to cope with and leverage crowd assessors, we assume that the relevance of a document is not deterministically known, but it is described by a probability distribution: instead of specifying a single value from REL as results of the relevance assessment, we model the uncertainty entailed in the assessment process as a whole distribution of possible values associated to each (t, d) pair. Furthermore, we assume that the ability of the crowd assessors is stochastically determined by a probability assigned to them, that we call their *accuracy*.

More precisely, we assume that there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which provides the source of randomness and encompasses the judgements done

by all the possible crowd assessors, on all the possible documents for any possible topic. Considering this space, we can extend the definition of the ground-truth as $GT : \Omega \times T \times D \rightarrow REL$. In this way, to any pair (t, d) we associate a random variable $GT(\cdot, t, d)$ with value on REL , whose distribution describes the relevance of the document d with respect to the topic t .

Let $\Lambda = \{W_1, \dots, W_l\}$ be a finite set of crowd assessors and let us assume that there exists a random variable, $W : \Omega \times T \rightarrow \Lambda$, whose distribution identifies the ability of a single crowd assessor with respect to any given topic. We call $a_k(t) = \mathbb{P}[T = t, W = W_k]$ the *accuracy* of crowd assessor W_k in assessing topic t and we assume that $a_k(t)$ is determined by the expected ability she/he demonstrates in assessing all the possible documents for that topic.

The easiest way to jointly cope with these random objects, i.e. ground-truth and crowd assessors, is to consider their expectations. The expected relevance of document d for topic t , by the law of total expectation, is given by

$$\mathbb{E}[GT(t, d)] = \mathbb{E}[\mathbb{E}[GT(t, d)|W]] = \sum_{k=1}^l \mathbb{E}[GT(t, d)|W = W_k] a_k(t) .$$

Then for a performance measure $m(\cdot)$, we can proceed in a similar way and define its AWARE version as its expectation with respect to \mathbb{P} :

$$\text{aware-}m(t, r_t) = \mathbb{E}[\mu(\hat{r}_t)] = \sum_{k=1}^l \mathbb{E}[\mu(\hat{r}_t)|W = W_k] a_k(t) ,$$

where μ is the scoring function associated to the performance measure $m(\cdot)$ [1], and \hat{r}_t is the judged run.

We estimate the first term by $\mu(\hat{r}_t^k)$, where \hat{r}_t^k represents the judged run under the assessments done by the crowd assessor W_k . However, the estimation of the accuracies $a_k(t) = \mathbb{P}[T = t, W = W_k]$ is somehow more problematic. We therefore take a *random assessor* as a comparison point. In the case of binary relevance, i.e. when $REL = \{0, 1\}$, an assessor W_k is a *random assessor of parameter* $p \in [0, 1]$, if for any pair (t, d) the conditional random variables $GT(t, d)|W = W_k \sim \text{Bin}(1, p)$, where $\text{Bin}(1, p)$ denotes a Binomial random variable with parameter p , and are mutually independent.

A random assessor, of any possible parameter p , is the prototype of a “bad” or at least a “shallow” assessor, since p is the same for any possible pair (t, d) . The basic idea that we will apply in the next section is that the farther a crowd assessor is from the random ones, the better she/he is and the higher her/his accuracy will be.

3 Estimating Crowd Assessor Accuracy

This sections aims at providing an intuitive overview of the proposed unsupervised estimators of the accuracy of a crowd assessor, more details can be found in [2]. Figure 1 shows the main steps (granularity, gap and weight) we use to



Measure	Gap G_k	Weight w_k		
		Minimal Dissimilarity	Minimal Squared Dissimilarity	Minimal Equi Dissimilarity
 M_h^p Random Assessors ρ_h^p	Measure Level - Frobenius Norm - RMSE	fro_md	fro_msd	fro_med
		rmse_md	rmse_msd	rmse_med
 M_k Crowd Assessor W_k	Distribution Level - KL Divergence	kld_md	kld_msd	kld_med
	Rankings Level - Kendall's Tau - AP Correlation	tau_md apc_md	tau_msd apc_msd	tau_med apc_med

Fig. 1: Approach to determine the accuracy of a crowd assessor W_k with respect to a random assessors ρ_h^p .

estimate the accuracy of a crowd assessor and the different estimators we can obtain by combining the various alternatives at each step. The idea is to compare the crowd assessor against a set of random assessors and how “different” this crowd assessor is from the random ones, i.e. how much better she/he is.

For each pool we generate, ρ_h^p , $h = 1, 2, \dots, H$, a set of H random assessors of level p , i.e. which randomly evaluate as relevant the p per cent of the documents in the pool. We consider three different classes of random assessors: *uniform random assessor* with $p = 0.5$, *underestimating random assessor* with $p = 0.05$, and *overestimating random assessor* with $p = 0.95$. Each of these random assessors gives origin to an assessor measure M_h^p for a given performance measure $m(\cdot)$.

Therefore, the intuitive idea described above boils down to determining some sort of “difference” between the measure M_k of a crowd assessor W_k and those M_h^p of the three random assessors ρ_h^p and turning this “difference” into an estimated accuracy a_t^k assigned to the crowd assessor W_k to compute the AWARE version of the performance measure $m(\cdot)$. This is achieved in two main steps:

- *gap G_k* : this quantifies what “different” means. We consider three alternatives:
 - *measure level*: this operates directly on the assessor measures by computing either the Frobenius norm of their difference (labelled **fro**) or their *Root Mean Square Error* (*RMSE*) (labelled **rmse**);
 - *distribution level*: this works on the performance distributions estimated from the assessor measures by using *Kernel Density Estimation* (*KDE*)

- and computes the *Kullback-Leibler Divergence (KLD)* between them (labelled **kld**);
- *rankings level*: this considers the system rankings induced by the assessor measures and compares them by using either the Kendall’s tau correlation (labelled **tau**) or the AP correlation (labelled **apc**);
- *weight w_t^k* : this turns the gap computed in the previous step into an estimated accuracy to be assigned to a crowd assessor. In particular, we reason in terms of *dissimilarity* from random assessors since, for a crowd assessor W_k , being close to a random one ρ_h^p can be considered as an indicator of her/his poor quality. We have three alternatives:
 - *minimal dissimilarity* (labelled **md**): this computes a weight which is proportional to the minimum gap from one of the random assessors class, i.e. the closer to one of the random assessors, the smaller the weight;
 - *minimal squared dissimilarity* (labelled **msd**): this is similar to the previous case but uses the minimum squared gap;
 - *minimal equi-dissimilarity* (labelled **med**): this computes a weight which is proportional to the crowd assessor being equally distant from all three families of random assessors.

For each of the three random assessor classes, we generate a set of H replicates to cope with the uncertainty of the random generation process and to obtain better estimates. Therefore, for each crowd assessor W_k , we obtain a set of H estimates and we need to aggregate them into a single one; we compute a mean gap \bar{G}_k , averaging over the set of H gaps computed with respect to each random assessor ρ_h^p .

Finally, the described procedure produces an estimated accuracy a_t^k to be assigned to a crowd assessor W_k for each topic $t \in T$; this is what we call *topic-by-topic score granularity*, labelled **tpc**. However, we are also interested in the case when a single accuracy score is assigned to a crowd assessor W_k , i.e. when the a_t^k are the same for all the topics; this is what we call *single score granularity*, labelled **sgl**.

4 Experimental Evaluation

4.1 Experimental Setup

We use the TREC 21, 2012, Crowdsourcing [6] data sets developed in the *Text Relevance Assessing Task (TRAT)*. The TRAT required participating groups to simulate the relevance assessing role of the NIST for 10 of the TREC 08, 1999, Ad-hoc topics [9]. Participating groups had to submit a binary relevance judgements for every document in the judging pools of the ten topics. Two TREC Adhoc tracks used these 10 topics over the years: the TREC 08, 1999, Ad-hoc track [9] (labeled T08), and the TREC 13, 2004, Robust track [8] (labeled T13).

When it comes to the measures for evaluating the effectiveness of the different approaches, we adopt two criteria used in the TREC 22, 2013, Crowdsourcing

track [7]: referred as *rank correlation* and *score accuracy*. We use *Average Precision (AP)* correlation [10] to compare the ranking of the systems produced for a given performance measure $m(\cdot)$, computed over the gold standard, with respect to the ranking produced for the same performance measure computed over the ground-truth, generated by one of the approaches under examination. In addition to correctly ranking systems, it is important that the performance scores are as accurate as possible. To this end, for a given performance measure $m(\cdot)$, we use the RMSE between the performance measure computed over the gold standard and the one computed over the ground-truth created by one of the approaches under examination.

When it comes to the assessor measures M_k and M_h^p , we consider *Average Precision (AP)*, *Normalized Discounted Cumulated Gain (nDCG)*, and *Expected Reciprocal Rank (ERR)*.

We consider three baselines, representing the state-of-the-art: the MV algorithm, labeled `mv`, and two variants of the EM algorithm: `emmv`, i.e. EM seeded by the pool generated by the MV algorithm, and `emneu`, i.e. EM initialized using the worker confusion matrix. Finally, we experiment also a fourth baseline labeled `uni`, representing AWARE in absence of any information, i.e. using uniform accuracies for all the merged crowd assessors.

4.2 Methodology

The goal of this section is to investigate how the AWARE approaches and the state-of-the-art baselines behave with respect to different factors, and to compare the AWARE approaches against those baselines. To this end, we adopt a *General Linear Mixed Model (GLMM)* model for the three-way *ANalysis Of VAriance (ANOVA)* with repeated measures [5]. We are interested in determining whether a factor effect is significant, i.e. its p -value is less than 0.05, as well as in which proportion of the variance is due to it.

AP Correlation The ANOVA table – not reported due to space limit [2] – shows that **Measure** is a large size effect and it explains the largest share of variance; **Systems** is a large size effect as well and it is the second largest main effect; finally, also **Approach** is a large size effect but about 2 times smaller than **Measure** effect and 1.25 times smaller than **Systems** effect. Overall, this supports the intuition that led to the development of the AWARE framework: performance **Measures** and **Systems** effects do matter a lot when merging assessors and they should be taken into the play.

The Tukey HSD multiple comparison analysis reported in Figure 2a highlights the top group (dashed blue line), the group of approaches not significantly different from the `uni` baseline (dashed bright red line), the group of approaches not significantly different from `mv` (dashed dark red line), and the group of approaches not significantly different from `emmv` and `emneu` (dashed orange line). We can note how the top group is separated from the others while the `uni` and `mv` groups partially overlaps. In particular, we can see that the approaches significantly better than all the others are `sgl_tau_msd` (the top one), `sgl_apc_msd`,

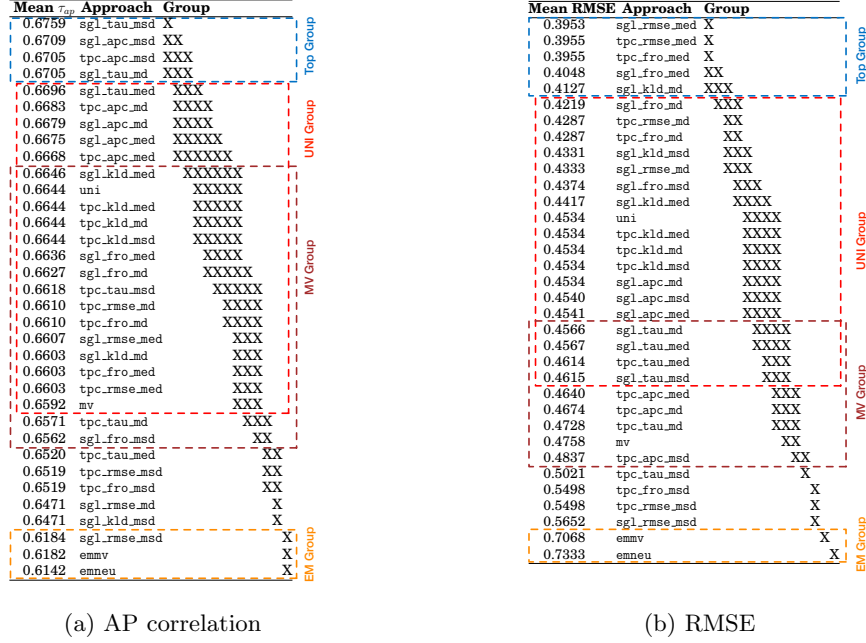


Fig. 2: Tukey HSD multiple comparison test for the Approach factor.

tpc_apc_msd, and **sgl_tau_md**, suggesting that the single score granularity is preferable to the topic-by-topic one and that the **tau** and **apc** gaps help to rank systems better. State-of-the-art approaches, namely **mv** (the best one in this group), **emmv**, and **emneu** are clearly separated from the top group. Finally, the AWARE **uni** baseline exhibits better performances than **mv**, even though it is not significantly different from it.

RMSE The ANOVA table – not reported due to space limit [2] – shows that the **Measure** factor is a large size effect with the greatest impact; **Approach** is a large size effect but, unlike the case of AP correlation, it is almost as important as **Measure**; finally, **Systems** is a large size effect but much smaller than the previous two. Overall, this further supports the intuition behind AWARE, but it also suggests that **Approaches** are much more prominent for the accurate estimation of the actual value of a performance measure, (assessed by the RMSE) than for ranking systems correctly (assessed by AP correlation).

The top group, reported in the Tukey HSD comparison of Figure 2b, consists of **sgl_rmse_msd**, **tpc_rmse_msd**, **tpc_fro_msd** (the top ones with extremely close performances), **sgl_fro_msd**, and **sgl_kld_md**; this suggests that there is more balance between single and topic-by-topic score granularities and that the gaps operating closer to the assessors measures (**fro**, **rmse**, **kld**) are more effective. State-of-the-art approaches are clearly distinct from the top group and, in this case, AWARE **uni** is significantly better than **mv** and the rest of them.

5 Conclusions and Future Work

In this paper, we presented the AWARE framework for robustly combining performance measures coming from multiple crowd assessors. The idea of AWARE stemmed from the observation of the potential impact of both performance measures and systems when it comes to correctly labeled/mis-labeled relevance judgements. Therefore, we proposed a probabilistic framework to take systems and performance measures into account during the estimation of the crowd assessors accuracies used to combine them. We then exemplified how to instantiate the proposed stochastic framework by introducing many unsupervised estimators of the accuracy of crowd assessors.

Finally, we conducted a thorough evaluation on TREC collections, comparing AWARE against state-of-the-art approaches and studying their influencing factors. The experimentation has provided multiple evidence supporting the intuition behind the AWARE framework. Moreover, it has shown that AWARE approaches perform better than state-of-the-art ones in terms of both ranking systems and correctly predicting their performance scores.

As future work we will investigate multi-feature estimators, i.e. estimators that take into account multiple performance measures at the same time to determine the accuracy of a crowd assessor, supervised estimators, i.e. estimators that leverage a gold standard instead of random assessors for determining the accuracy of a crowd assessor and extend the experiments to graded-relevance judgements.

References

1. Ferrante, M., Ferro, N., Maistro, M.: Towards a Formal Framework for Utility-oriented Measurements of Retrieval Effectiveness. In ICTIR, pp. 21–30, ACM, 2013.
2. Ferrante, M., Ferro, N., Maistro, M.: AWARE: Exploiting Evaluation Measures to Combine Multiple Assessors. In TOIS, 36(2), 20:1–20:38, 2017.
3. Halvey, M., Villa, R., Clough, P.: SIGIR 2014 Workshop on Gathering Efficient Assessments of Relevance (GEAR). In SIGIR, p. 1293, ACM, 2014.
4. Marcus, A., Parameswaran, A.: Crowdsourced Data Management: Industry and Academic Perspectives. In Foundations and Trends® in Databases, 6(1-2) pp. 1–16, Now Publishers, Inc, 2015.
5. Maxwell, S., Delaney, H.D.: Designing Experiments and Analyzing Data. A Model Comparison Perspective. Lawrence Erlbaum Associates, 2004.
6. Smucker, M.D., Kazai, G., Lease, M.: Overview of the TREC 2012 Crowdsourcing Track. In TREC, NIST, Special Publication 500-298, 2013.
7. Smucker, M.D., Kazai, G., Lease, M.: Overview of the TREC 2013 Crowdsourcing Track. In TREC, NIST, Special Publication 500-302, 2014.
8. Voorhees, E.M.: Overview of the TREC 2004 Robust Track. In TREC, NIST, Special Publication 500-261, 2004.
9. Voorhees, E.M., Harman, D.K.: Overview of the Eight Text REtrieval Conference (TREC-8). In TREC, pp. 1–24, NIST, Special Publication 500-246, 1999.
10. Yilmaz, E. and Aslam, J. A. and Robertson, S. E.: A New Rank Correlation Coefficient for Information Retrieval. In SIGIR, pp. 587–594, ACM, 2008.