AWARE: Exploiting Evaluation Measures to Combine Multiple Assessors

MARCO FERRANTE, University of Padua
NICOLA FERRO, University of Padua
MARIA MAISTRO, University of Padua

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 takes the expectation of a generic performance measure, like *Average Precision (AP)*, composed with these random variables. In this way, it approaches the problem of aggregating different crowd assessors from a new perspective, i.e. directly combining 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 state-of-the-art approaches, i.e. *Majority Vote (MV)* and *Expectation Maximization (EM)*, on TREC collections. We found that AWARE approaches improve in terms of their capability of correctly ranking systems and predicting their actual performance scores.

CCS Concepts: • **Information systems** → **Relevance assessment**; **Retrieval effectiveness**; *Test collections*;

Additional Key Words and Phrases: crowdsourcing, performance measure, weighted average, unsupervised estimators, AWARE

ACM Reference format:
Marco Ferrante, Nicola Ferro, and Maria Maistro. 2017. AWARE: Exploiting Evaluation Measures to Combine Multiple Assessors. *ACM Transactions on Information Systems* 0, 0, Article 0 (June 2017), 38 pages.
DOI: 0000001.0000001

1 INTRODUCTION

Ground-truth is central to *Information Retrieval (IR)* evaluation since it enables the scoring and comparison of algorithms and systems with respect to human judgments, determining whether documents are relevant, or not, to user information needs.

Creating a dataset and, in particular, gathering relevance assessments is an extremely demanding activity: it involves sizable costs for hiring assessors and a fairly large amount of time to judge a pool of documents. Therefore, there is an increasing interest for more effective and affordable ways of gathering assessments [23], especially to face the ever increasing number of new search tasks that need an appropriate dataset to be evaluated.

Author's address:
M. Ferrante, Department of Mathematics, Via Trieste 63, 35121 Padova, Italy; email: ferrante@math.unipd.it
N. Ferro and M. Maistro, Department of Information Engineering, Via G. Gradenigo 6/B, 35131 Padova, Italy; email: {ferro,maistro}@dei.unipd.it.
 1046-8188/2017/6-ART0 $15.00
DOI: 0000001.0000001

Crowdsourcing [2, 39, 43, 47] has emerged as a viable option for ground-truth creation since it allows to cheaply collect multiple assessments for each document. 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, i.e. choosing among the set of possible judgements (labels) those best supported by the evidence provided by the crowd assessors. In detail, traditional approaches focus mainly on how to select assessors and/or discard low quality assessors, how to merge judgments from multiple assessors into a single assessor, and how to route tasks to assessors. They typically determine the "best" relevance judgements, combining those produced by multiple crowd assessors according to some criteria, and use them to compute a performance measure, like AP, and score systems. 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 in a crowdsourcing context from a new angle, i.e. we investigate how to estimate performance measures in a way more robust to crowd assessors. To the best of our knowledge, what happens when you aggregate the different performance scores directly computed on the judgements produced by multiple assessors is yet to be explored. 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 as a kind of "downstream" approach with respect to the classification ones, since the aggregation happens after performance measures have been computed.

The main intuition behind our approch 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. Moreover, performance measures embed different user models, weighting differently even the same system ranking; therefore, the same correctly labelled or mis-labelled documents have a different impact on different performance measures. As a consequence, even a small error over a whole pool of documents may affect systems and performance measures in quite different ways.

To make an intuitive yet extreme toy example, suppose that out of 10 relevant documents in a pool, just 1 document has been wrongly labelled as not relevant, thus there is a 10% error with respect to the whole pool. Now consider a run which retrieves that mis-labelled document, represented as a blue *R* in italics, somewhere in the ranks from 1 to 5 and it also retrieves a few other relevant documents in the ranks from 6 to 10, marked as a plain R.

Rank	1	2	3	4	5	6	7	8	9	10	P@5	AP
Run_1						R			R	R	0.0000	0.0765
Run_2					*R*	R			R	R	0.2000	0.1407
Run_3				*R*		R			R	R	0.2000	0.1463
Run_4			*R*			R			R	R	0.2000	0.1556
Run_5		*R*				R			R	R	0.2000	0.1741
Run_6	*R*					R			R	R	0.2000	0.2296

Run_1 represents the case where the mis-labelled document is not detected in any ranks from 1 to 5, while the other runs show what could have happened if it had been correctly labelled. You can see how for P@5, i.e. precision at 5 retrieved documents, wherever this document is in the

ranks from 1 to 5, it makes the difference between P@5 = 0 and P@5 = 20%, which represents a 100% error; for AP, it changes from AP = 7.65% to AP between 14.07% and 22.96%, i.e. an error ranging between 45.61% and 66.67%. In all these cases, the effect of a single mis-labelled document has a different impact on different runs and for different performance measures and, in the extreme example at hand, it is much greater than the error on the pool itself.

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, which we call 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. Intuitively, these unsupervised estimators compute some kind of "distance" between the selected performance measure computed on the ground-truth produced by the crowd assessor and the same performance measure computed on the ground-truth produced by different types of random assessors: the greater this "distance", the better the accuracy of the crowd assessor.

We conduct a thorough experimental evaluation, using the ground-truth created by the crowd assessors of the TREC 21, 2012, Crowdsourcing track [63] with respect to the systems submitted to the TREC 08, 1999, Ad-hoc [70] and the TREC 13, 2004, Robust [69] tracks. We experiment with the following performance measures: *Average Precision (AP)* [6], *Normalized Discounted Cumulated Gain (nDCG)* [28], and *Expected Reciprocal Rank (ERR)* [10]. 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 related works and provides a description of state-of-the-art algorithms for combining multiple assessors which will be used for comparison with the AWARE approach; Section 3 introduces the AWARE framework; Section 4 proposes several unsupervised estimators for determining the assessors accuracies to be used for combining AWARE measures; Section 5 describes the experimental setup; Section 6 and Section 7 carry out a thorough evaluation using TREC collections; finally, Section 8 draws some conclusions and presents an outlook for future work.

2 BACKGROUND

2.1 Crowdsourcing for Ground-truth Creation

One of the first investigated issues, assuming the quality of the assessors for granted, concerned the impact of the inter-assessor disagreement. What happens if we assign the same set of topics and documents to another assessor? Will the ranking of the systems remain stable? Several studies [7, 44, 67, 68] have shown that even a not negligible amount of inter-assessor disagreement does not severely impact the ability of ranking systems and, more recently, [74] has provided evidence that the rank of a document is a factor influencing the probability of disagreement among assessors. Other issues concern the expertise of the assessors on the domain of the topics they are judging: [3, 40] noted that this factor has some impact on the evaluation.

Moreover, regarding the comparison of different types of assessors, a lot of work was done to investigate the relation between domain experts and crowd assessors [11], authoritative and alternative assessors [75], primary and secondary assessors [72], NIST assessors and user studies participants [62], crowd assessors and university laboratory participants [61]. Finally, [58] studies the assessors' characteristics that lead to different relevance assessments and [60] investigates how to build test collections in order to optimize the assessor effort.

Research in crowdsourcing has focused on several different issues: aggregating labels from multiple assessors to improve the quality of the gathered assessments, by using unsupervised [4, 26], supervised [52, 53, 55], and hybrid [24] approaches; behavioural aspects [34]; proper and careful design of *Human Intelligence Tasks (HITs)* [1, 22, 27, 33], also using gamification to improve quality [14] and game theory to increase user engagement [49]; and, routing tasks to proper assessors [30, 42].

There is a growing concern about the quality of the gathered assessments [31, 35, 71], how assessor quality and errors impact evaluation [9, 32], how much tolerant evaluation measures are to these errors [45], and how crowd and editorial assessors agreement relates to user intent and click-based measures [36].

In recent years, several evaluation activities have focused on crowdsourcing for ground-truth creation, as witnessed by the TREC Crowdsourcing track series[1] from 2011 to 2013 [63, 64], the MediaEval Crowdsourcing tracks[2] in 2013 and 2014 [46, 77], or the CrowdScale 2013 Shared Task Challenge[3] [29]. There is also a growing interest and attention about how crowdsourcing affects the repeatability and reproducibility of IR experiments [5, 18, 19].

In this paper we are interested in aggregating labels from multiple assessors and, in the experimental part in Sections 6 and 7 we will compare our proposed approach, AWARE, with two state-of-the-art approaches for label aggregation, namely *Majority Vote (MV)* and *Expectation Maximization (EM)* [4, 26], which are briefly summarized in the following sections.

2.2 Majority Vote

In [17] 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*, i.e. they are defined on an ordinal scale [66], where we assume the existence of a minimum that we call the `non-relevant` relevance degree $\mathrm{nr} = \min(REL)$. In the following, and without any loss of generality, we consider $REL \subseteq \mathbb{R}_0^+$ with the constraint that $0 \in REL$ and the order relation $\preceq$ becomes the usual ordering $\leq$ on real numbers; the non-relevant degree is therefore given by $\min(REL) = 0$; in the following we restrict ourself to the case of binary relevance and we assume $REL = \{0, 1\}$.

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. This means that if the document d has relevance grade $g \in REL$, then $GT(t, d) = g$.

Moreover, let $\Lambda = \{W_1, \ldots, W_l\}$ be a finite set of assessors, we define as $GT_k(t, d)$ the discrete variable with values in $\{0, 1\}$, which represents the label given by the assessor k to the document d with respect to the topic t. Note that this is the only information that we are provided with, indeed we assume that the relevance judgments, $GT(t, d)$, are not known. We further suppose that each document receives at least one relevance label. Finally, let $\mathbb{1}_{\{GT_k(t,\cdot)=g\}}$ be a binary variable that is equal to 1 if the assessor k assigns the label g to the document d and zero otherwise.

The simplest way of estimating the true relevance labels is the *Majority Vote (MV)* algorithm, which views each worker as a voter. If the number of voters which consider a given document as relevant is greater than the number of voters that consider it as not relevant, that document will be classified as relevant. Hence, if $n_t[d, g] = \sum_{k=1}^{l} \mathbb{1}_{\{GT_k(t,d)=g\}}$ is the number of times that the document d is labeled as g for the topic t, we will assign to d the relevance g that maximizes $n_t[d, g]$, that is g such that $n_t[d, g] = argmax_g\{n_t[d, 0], n_t[d, 1]\}$. In the case of tie, i.e. $n_t[d, 0] = n_t[d, 1]$, a coin is tossed to determine whether the document is relevant or not.

[1] https://sites.google.com/site/treccrowd/
[2] http://www.multimediaeval.org/
[3] http://www.crowdscale.org/shared-task

2.3 Expectation Maximization

The *Expectation Maximization (EM)* algorithm is an alternative to MV for defining the relevance of the documents. We follow the same approach described in [26] to implement the EM algorithm.

Suppose that a latent confusion matrix, $\pi_t[\cdot,\cdot](k),\ k \in \{1,\dots,l\}$, is assigned to each assessor, this matrix has as many rows and columns as the number of relevance grades, i.e. two in the binary case. Each row represents the true relevance grade and each column the label given by the worker. We define $\pi_t[g,h](k) = \mathbb{P}\big[GT_k(t,\cdot) = h | GT(t,\cdot) = g\big]$, i.e. the probability that the assessor k assigns to a document the relevance grade h, given that the true relevance label of the document is g. For instance, $\pi_t[1,0](k)$ is the probability that the worker k labels a document as not relevant, given that this document is relevant. The matrix $\pi_t[g,h](k)$ could be estimated by:

$$\frac{\text{number of times the worker } k \text{ provides label } h \text{ while the true label is } g}{\text{number of labels provided by worker } k \text{ for documents of relevance } g}\ .$$

Note that, in the binary case:

$$\pi_t[g,0](k) + \pi_t[g,1](k) = 1 \quad \forall\ k \in \{1,\dots,l\} \text{ and } g \in \{0,1\}\ .$$

Moreover, we define $p_t[g] = \mathbb{P}\big[GT(t,\cdot) = g\big]$, the probability that a randomly chosen document has relevance grade g, i.e. $p_t[0]$ is the probability that a document drawn at random is not relevant and $p_t[1]$ is the probability that it is relevant.

The EM algorithm consists of five main steps that we will describe in the following, and we will indicate with the symbol $\tilde{}$ a possible estimate of the parameter or the variable under the $\tilde{}$.

Step 1: Initialization Firstly we initialize the parameters of our model, we adopt two different strategies that we will illustrate later in detail.

Step 2: Estimate the maximum likelihood Then we compute the maximum likelihood estimates of $\pi_t[\cdot,\cdot](\cdot)$ and $p_t[\cdot]$ as follows:

$$\tilde{\pi}_t[g,h](k) = \frac{\sum_{d=1}^{|D|} \mathbb{1}_{\{GT(t,d)=g\}} \mathbb{1}_{\{GT_k(t,d)=h\}}}{\sum_{h\in REL}\sum_{d=1}^{|D|} \mathbb{1}_{\{GT(t,d)=g\}} \mathbb{1}_{\{GT_k(t,d)=h\}}}\ ,$$

$$\tilde{p}_t[g] = \frac{\sum_{d=1}^{|D|} \mathbb{1}_{\{GT(t,d)=g\}}}{|D|}\ .$$

Step 3: Estimate the probability of relevance We compute the new estimate of the relevance judgments based on $\hat{\pi}_t[\cdot,\cdot](\cdot)$ and $\hat{p}_t[g]$:

$$\mathbb{P}\big[GT(t,d) = g | GT.(t,\cdot), \pi_t[\cdot,\cdot](\cdot)\big] = \frac{\tilde{p}_t[g]\prod_{k=1}^{l}\prod_{h\in REL}(\tilde{\pi}_t[g,h](k))^{\mathbb{1}_{\{GT_k(t,d)=h\}}}}{\sum_{g\in REL}\tilde{p}_t[g]\prod_{k=1}^{l}\prod_{h\in REL}(\tilde{\pi}_t[g,h](k))^{\mathbb{1}_{\{GT_k(t,d)=h\}}}}\ .$$

Step 4: Iterate We repeat the steps 2 and 3 until the results converge.

Step 5: Define the relevance labels Finally, for each document d, we assign the label g to the documents with the maximal probability of having relevance grade g; i.e. we compute $argmax_{g\in REL}\{\mathbb{P}\big[GT(t,d) = g | GT.(t,\cdot), \pi_t[\cdot,\cdot](\cdot)\big]\}$, then we set $GT(t,d) = g$. Notice that in the binary case all the documents with probability of relevance greater than 0.5 are considered as relevant, and documents with probability equal or lower than 0.5 are considered as not relevant.

The convergence of the EM algorithm strongly depends on many assumptions that, if not satisfied, could compromise the convergence of the algorithm [13, 76]. In particular, the starting point of the EM algorithm represents a criticality that has to be treated properly. Therefore, we define two

different instantiations of the EM algorithm, by interpreting the initialization step in two different ways:

EM-MV We use the algorithm of [26] and we set the initial relevance labels as the result of the MV algorithm, as done in [53, 55];

EM-NEU We initialize each worker confusion matrix and the probability p_t as done in [4]:

$$\tilde{\pi}_t[\cdot,\cdot](k) = \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix} \quad , \quad \tilde{p}_t = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \quad .$$

Hence, we make the hypothesis that each worker honestly assigns the relevance labels. Then, we initialize the relevance labels by computing the probability of relevance as in the third step of the EM algorithm.

3 THE AWARE FRAMEWORK

In the following we recall some further definitions introduced in [17]. Given a positive natural number n called the *length of the run*, we define the *set of retrieved documents* as $D(n) = \{(d_1, \ldots, d_n) : d_i \in D, d_i \neq d_j \text{ for any } i \neq j\}$, i.e. the ranked list of retrieved documents without duplicates, and the *universe set of retrieved documents* as $\mathcal{D} := \bigcup_{n=1}^{|D|} D(n)$. A *run* r_t, retrieving a ranked list of documents $D(n)$ in response to a topic $t \in T$, is a function from T into $\mathcal{D}$: $t \mapsto r_t = (d_1, \ldots, d_n)$. We denote by $r_t[j]$ the j-th element of the vector r_t, i.e. $r_t[j] = d_j$. We define the *universe set of judged documents* as $\mathcal{R} := \bigcup_{n=1}^{|D|} REL^n$. We call *judged run* the function $\hat{r}_t$ from $T \times \mathcal{D}$ into $\mathcal{R}$, which assigns a relevance degree to each retrieved document in the ranked list

$$(t, r_t) \mapsto \hat{r}_t = \big(GT(t, d_1), \ldots, GT(t, d_n)\big)$$

We denote by $\hat{r}_t[j]$ the j-th element of the vector $\hat{r}_t$, i.e. $\hat{r}_t[j] = GT(t, d_j)$.

A *performance measure*, like AP, is a function $\mathrm{m} : T \times \mathcal{D} \to \overline{\mathbb{R}}_0^+$ defined as $\mathrm{m} = \mu(\hat{r}_t)$, i.e. the composition of a judged run $\hat{r}_t$ with a *scoring function* $\mu : \mathcal{R} \to \overline{\mathbb{R}}_0^+$, which assigns to any sequence of judged documents a non negative number, representing the effectiveness of the run.

In order to cope with and leverage crowd assessors, we need to extend the definitions of [17] and frame them in a probabilistic context. In particular, 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 themselves 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 follows:

$$GT : \Omega \times T \times D \to REL$$

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. This distribution can be modeled by means of various parameters, for example, the expected relevance obtained by all the possible crowd assessors who judge that pair.

All the previous definitions (judged run, performance measure and so on) remain unchanged, provided that it is understood that all the objects are now random variables. For example, a *(random)*

judged run will be the random variable $\hat{r}_t$ from $\Omega \times T \times \mathcal{D}$ into $\mathcal{R}$, which assigns a (random) relevance degree to each retrieved document in the ranked list

$$(\omega, t, r_t) \mapsto \hat{r}_t = \big(GT(\omega, t, d_1), \ldots, GT(\omega, t, d_n)\big)$$

In the sequel, as it usually done in probabilistic frameworks, we omit to explicitly write the dependence of the random variables on ω.

Let $\Lambda = \{W_1, \ldots, W_l\}$ be a finite set of crowd assessors and let us assume that there exists a random variable, $W : \Omega \times T \to \Lambda$, whose distribution identifies the ability of a single crowd assessor with respect to any given topic. In practice, we can assume to be able, from the judgments of all the documents and with respect to a given topic t, to weight the average ability of any single crowd assessor with a positive number; the distribution on Λ can be then obtained from these numbers once normalized to 1. 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 ground-truth of a pair (t, d), i.e. the expected relevance of document d for topic t, by the law of total expectation, is given by

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

The conditional expectation $\mathbb{E}\big[GT(t,d)\big|W = W_k\big]$ in (1) represents the "best" possible approximation of $GT(t,d)$ given that the assessment has been provided by the crowd assessor W_k, where "best" refers to the minimal distance in mean square between them. This is, for example, the approach adopted by MV, under some strong assumptions: the crowd assessors W_k are *independent and identically distributed (i.i.d.)* and the accuracies $a_k(t)$ are uniformly distributed.

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

$$\text{aware-m}\big(t, r_t\big) = \mathbb{E}\Big[\mu\big(\hat{r}_t\big)\Big] = \sum_{k=1}^{l} \mathbb{E}\big[\mu\big(\hat{r}_t\big)\big|W = W_k\big]\; a_k(t) \tag{2}$$

To make this approach feasible, we need to have a simple but yet reasonable way to estimate $\mathbb{E}\big[\mu\big(\hat{r}_t\big)\big|W = W_k\big]$ and $a_k(t)$.

For the first term, we estimate $\mathbb{E}\big[\mu\big(\hat{r}_t\big)\big|W = W_k\big]$ by $\mu\big(\hat{r}_t^k\big)$, where $\hat{r}_t^k$ represents the judged run under the assessments done by the crowd assessor W_k. Indeed, we typically have available just one judgement for each (t, d) pair by each crowd assessor and therefore the expectation collapses into that single observation.

The estimation of the accuracies $a_k(t) = \mathbb{P}[T = t, W = W_k]$ is somehow more problematic. Indeed, the estimation of the probability $\mathbb{P}$ calls for multiple observations and this is addressed by state-of-the-art approaches like MV and EM by assuming that crowd assessors are somehow i.i.d.. However, this is quite a strong assumption since crowd assessors are very different from each other and even the same crowd assessor may have a quite different behavior across different topics.

Therefore, we remove the i.i.d. assumption about the crowd assessors and we look for something to compare our not-i.i.d. crowd assessors against, something that can be truly i.i.d. and allows us to perform inferential statistics. We therefore take a *random assessor* as a truly i.i.d. 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 $GT(t,d)|W = W_k \sim$

$Bin(1, p)$, where $Bin(1, p)$ 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). As the definition of the random assessor is purely theoretic, we can assume that we are able to produce a sample of i.i.d. random assessors with the same parameter p. This fact allows us to provide classical inferential constructions of the estimates of the accuracy $a_k(t)$, as will be described in detail in the next section. 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.

Thanks to these considerations, we define the estimated version of AWARE as follows

$$\widetilde{\text{aware-m}}\big(t, r_t\big) = \sum_{k=1}^{l} \mu\big(\hat{r}_t^k\big) a_t^k \tag{3}$$

where a_t^k represents an estimate of the unknown accuracies $a_k(t)$.

Let us discuss how equation (3) works and the potential benefits of the AWARE approach by means of a toy example. Let us consider AP as performance measure, a pool containing just 3 relevant documents, and a run of length 5 where the first and the third documents are relevant, while the second, fourth and fifth are not relevant:

$$\hat{r}_t = (1, 0, 1, 0, 0) \;\Rightarrow\; \text{AP}\big(\hat{r}_t\big) = 0.5556$$

Suppose that we have three crowd assessors, judging that documents as follows:

$$\hat{r}_t^1 = (1, 1, 0, 0, 0) \;\Rightarrow\; \text{AP}\big(\hat{r}_t^1\big) = 0.6667$$
$$\hat{r}_t^2 = (1, 1, 1, 0, 0) \;\Rightarrow\; \text{AP}\big(\hat{r}_t^2\big) = 1.0000$$
$$\hat{r}_t^3 = (0, 1, 1, 0, 1) \;\Rightarrow\; \text{AP}\big(\hat{r}_t^3\big) = 0.5889$$

By using the MV and EM approaches we can compute a merged ground-truth, which in this case is the same for both approaches, and thus we obtain:

$$\hat{r}_t^{\text{MV}} = \hat{r}_t^{\text{EM}} = (1, 1, 1, 0, 0) \;\Rightarrow\; \text{AP}\big(\hat{r}_t^{\text{MV}}\big) = \text{AP}\big(\hat{r}_t^{\text{EM}}\big) = 1.0000$$

which represents a 20% error in terms of relevance labels but an 80% error in terms of AP. If in equation (3) we take the simplest estimator possible of $a_k(t)$, i.e. a uniform distribution $a_t^k = \frac{1}{3},\ k = 1, 2, 3$, which basically is the same underlying uniform approach used by MV, we obtain

$$\widetilde{\text{aware-AP}}(\hat{r}_t) = 0.7518$$

which represents a 35% error in terms of AP.

4 ESTIMATING CROWD ASSESSOR ACCURACY

This sections aims at providing several unsupervised estimators of the accuracy a_t^k of a crowd assessor. We introduce some notation and an intuitive overview of the proposed estimators and then we go into their details.

4.1 Notation

Let S be the *set of systems* under experimentation and $s \in S$ be a generic system.

We call *assessor measure* the $|T| \times |S|$ matrix M_k containing the scores of each system for each topic, computed using a performance measure $\text{m}(\cdot)$, according to the ground-truth generated by the crowd assessor W_k.

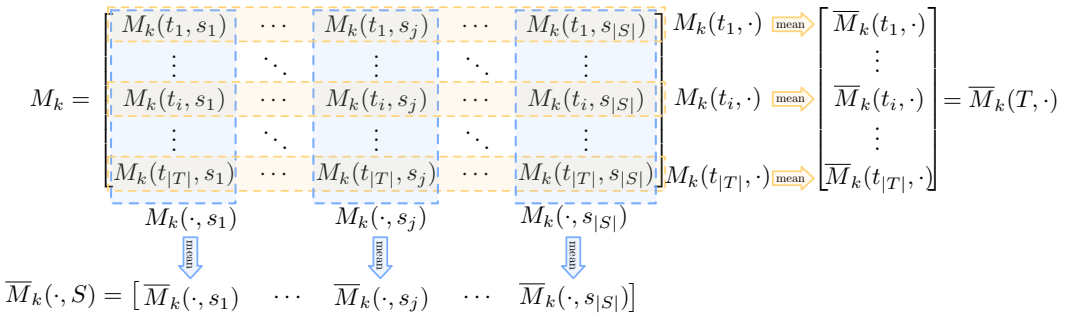


Fig. 1. Matrix notation for the assessor measure M_k.

The notation $M_k(\cdot, s)$ indicates a column vector containing all the performance figures for a given system s; the $\overline{M}_k(\cdot, s)$ indicates the average of the previous column vector; $\overline{M}_k(\cdot, S)$ indicates the average across the rows for all the systems; similarly, $M_k(t, \cdot)$, $\overline{M}_k(t, \cdot)$, and $\overline{M}_k(T, \cdot)$ indicate a row vector containing all the performance figures for a given topic t, its average, and the average across the columns for all the topics. Finally, the notation $M_k(:)$ indicates the linearization of the matrix, i.e. the row-wise concatenation of all its elements. A visualization of this matrix notation is reported in Figure 1.

For example, in the case of AP, each cell of AP_k contains the values of AP for system s on topic t according to assessor W_k; if we average over the topics $\overline{M}_k(\cdot, S)$, we obtain the *Mean Average Precision (MAP)* for all the systems $s \in S$ according to assessor W_k.

4.2 Intuitive Overview

Figure 2 shows the main steps (granularity, gap and weight) we use to estimate the accuracy of a crowd assessor and the different estimators we can obtain by combining the various alternatives at each step. The basic 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, $h = 1, 2, \ldots, H$, a set of random assessors of level p, i.e. which randomly evaluate as relevant the p per cent of the documents in the pool. As above, each of these random assessors gives origin to an assessor measure M_h^p for a given performance measure m(·). We consider three different classes of random assessors, each of which contains a set of H random replicates:

- *uniform random assessor* ρ_h^{uni}: this tosses a coin to judge a document, i.e. $p = 0.5$;
- *underestimating random assessor* ρ_h^{und}: this tends to judge documents as non relevant, e.g. $p = 0.05$;
- *overestimating random assessor* ρ_h^{ovr}: this tends to judge documents as relevant, e.g. $p = 0.95$.

Note that the idea of generating random assessors resembles [65] when they investigated the impact of random assessors compared to real assessors. However, to generate the random assessors [65] used a normal distribution with a proportion of relevant/not relevant documents derived by the same proportion in the case of real assessors. In our case, being a fully unsupervised approach, we do not have the real proportion of relevant documents available; when it comes to the distribution to be used, we chose the uniform distribution to avoid any assumption on assessor behavior, but a normal distribution or others could be an interesting future exploration.

Similarly, the approaches proposed by [9, 45] to simulate different types of assessors and different types of assessor errors cannot be applied in this unsupervised context, since they both start from a

Measure	Gap G_k	Weight w_k		
		Minimal Dissimilarity	Minimal Squared Dissimilarity	Minimal Equi Dissimilarity
	Measure Level - Frobenius Norm - RMSE	`fro_md` `rmse_md`	`fro_msd` `rmse_msd`	`fro_med` `rmse_med`
	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. 2. Approach to determine the accuracy of a crowd assessor W_k with respect to a random assessors ρ_h^p.

gold standard ground-truth and modify the assigned labels according to some desired distribution of truly/falsely relevant/not relevant documents. Even in [49] the authors presents a way of simulating assessors based on a probabilistic approach, however they are interested in simulating the time that each assessor spends in completing a task.

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(·). 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[4] of their difference (labelled `fro`, see Section 4.3.1) or their *Root Mean Square Error (RMSE)* (labelled `rmse`, see Section 4.3.2);
  - *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`, see Section 4.3.3);
  - *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`, see Section 4.3.4) or the AP correlation (labelled `apc`, see Section 4.3.5);
- *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

[4]We used the Frobenius norm because it is the Euclidean norm in the space $\mathbb{R}^{n \times m}$ and it has many desirable properties, such as invariance under rotations, which makes it robust for our purposes.

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`, see Section 4.4.2): this computes a weight which is proportional to the minimum gap from one of the random assessors (uniform, underestimating, and overestimating), i.e. the closer to one of the random assessors, the smaller the weight;
- *minimal squared dissimilarity* (labelled `msd`, see Section 4.4.3): this is similar to the previous case but uses the minimum squared gap;
- *minimal equi-dissimilarity* (labelled `med`, see Section 4.4.4): this computes a weight which is proportional to the crowd assessor being equally distant from all three random assessors (uniform, underestimating, and overestimating).

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`.

4.3 Gap

4.3.1 Frobenius Norm. Given an $m \times n$ matrix A, its Frobenius norm [21] is:

$$\left\|A\right\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\left|a_{ij}\right|^2} \tag{4}$$

which is also equal to the square root of the matrix trace $\left\|A\right\|_F = \sqrt{\mathrm{Tr}\left(AA^H\right)}$, where A^H is the transpose conjugate of A.

Single Score Granularity. This is given by the Frobenius norm of the matrices of the crowd and random assessor measures, as defined below:

$$G_k^p = \left\|M_k - M_h^p\right\|_F \tag{5}$$

Topic Score Granularity. For each topic $t \in T$, this is given by the Frobenius norm of row vectors of the crowd and random assessor measures for that topic, as defined below:

$$G_k^p(t) = \left\|M_k(t,\cdot) - M_h^p(t,\cdot)\right\|_F \tag{6}$$

4.3.2 Root Mean Square Error. Given two m elements vectors X and Y, their *Root Mean Square Error (RMSE)* [38] is:

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{m}\frac{\left(X_i - Y_i\right)^2}{m}} \tag{7}$$

Note that $\mathrm{RMSE} = \frac{1}{\sqrt{m}}\left\|X - Y\right\|_F$.

Single Score Granularity. This is given by the RMSE of the vectors of the crowd and random assessor measures averaged by topic, as defined below:

$$G_k^p = \text{RMSE}\big(\overline{M}_k(\cdot, S) - \overline{M}_h^p(\cdot, S)\big) \tag{8}$$

Topic Score Granularity. For each topic $t \in T$, this is given by the RMSE of row vectors of the crowd and random assessor measures for that topic, as defined below:

$$G_k^p(t) = \text{RMSE}\big(\overline{M}_k(t, \cdot) - \overline{M}_h^p(t, \cdot)\big) \tag{9}$$

4.3.3 KL Divergence. To compute the *Kullback-Leibler Divergence (KLD)* [41], we need the *Probability Density Function (PDF)* of the performance measures, which we estimate by using a *Kernel Density Estimation (KDE)* [73] approach.

Given a vector X of m elements, the KDE estimation of its PDF is given by

$$\hat{f}_X(x) = \frac{1}{mb} \sum_{i=1}^{m} K\left(\frac{x - X_i}{b}\right) \tag{10}$$

where b is a positive number called *bandwidth* or *window width*; $K(\cdot)$ is the *kernel* satisfying $\int_{-\infty}^{+\infty} K(x)dx = 1$.

Given two m elements vectors X and Y, the KLD between their PDFs is given by

$$D_{KL}\big(X \big\| Y\big) = \sum_x \ln\left(\frac{\hat{f}_X(x)}{\hat{f}_Y(x)}\right) \hat{f}_X(x) \tag{11}$$

$D_{KL} \in [0, +\infty)$ denotes the information lost when Y is used to approximate X [8]; therefore, 0 means that there is no loss of information and, in our settings, it will mean that two assessors are considered the same; $+\infty$ means that there is full loss of information and, in our settings, it will mean that two assessors are considered completely different. Note that D_{KL} is not symmetric and so, in general, $D_{KL}\big(X \big\| Y\big) \neq D_{KL}\big(Y \big\| X\big)$.

Single Score Granularity. This is given by the KLD of the vectors of the crowd and random assessor linearize measures, as defined below:

$$G_k^p = D_{KL}\big(M_k(:) \big\| M_h^p(:)\big) \tag{12}$$

Topic Score Granularity. For each topic $t \in T$, this is given by the KLD of row vectors of the crowd and random assessor measures for that topic, as defined below:

$$G_k^p(t) = D_{KL}\big(M_k(t, \cdot) \big\| M_h^p(t, \cdot)\big) \tag{13}$$

4.3.4 Kendall's Tau Correlation. Given two m elements vectors X and Y, their Kendall's τ correlation [37] is given by

$$\tau\big(X, Y\big) = \frac{C - D}{m(m-1)/2} \tag{14}$$

where C is the total number of concordant pairs (pairs that are ranked in the same order in both vectors) and D the total number of discordant pairs (pairs that are ranked in opposite order in the two vectors).

Single Score Granularity. This is given by the τ correlation of the vectors of the crowd and random assessor measures averaged by topic, as defined below:

$$G_k^p = \tau\big(\overline{M}_k(\cdot, S) - \overline{M}_h^p(\cdot, S)\big) \tag{15}$$

Topic Score Granularity. For each topic $t \in T$, this is given by the τ correlation of row vectors of the crowd and random assessor measures for that topic, as defined below:

$$G_k^p(t) = \tau\big(M_k(t,\cdot), M_h^p(t,\cdot)\big) \tag{16}$$

4.3.5 AP Correlation. AP correlation τ_{ap} [78] is a correlation coefficient inspired by the Kendall's τ correlation, but it puts more emphasis on the order of the top ranked systems.

Given two m elements vectors X and Y, their AP correlation is given by

$$\tau_{ap}(Y,X) = \frac{2}{m-1}\sum_{i=2}^{m}\frac{C(i)}{i-1} - 1 \tag{17}$$

where $C(i)$ is the number of items above rank i in X and correctly ranked with respect to the item at rank i in Y, which acts as a reference. Note that τ_{ap} is not symmetric and so, in general, $\tau_{ap}(Y,X) \neq \tau_{ap}(X,Y)$.

Note that τ_{ap} does not handle tied values in the two vectors, so we adopt the same approach suggested in the TREC 2013 Crowdsourcing track [64] where, in case of ties, they sample over possible orders and average the obtained τ_{ap} coefficients.

Single Score Granularity. This is given by the τ_{ap} correlation of the vectors of the crowd and random assessor measures averaged by topic, as defined below:

$$G_k^p = \tau_{ap}\big(\overline{M}_k(\cdot,S), \overline{M}_h^p(\cdot,S)\big) \tag{18}$$

Topic Score Granularity. For each topic $t \in T$, this is given by the τ_{ap} correlation of row vectors of the crowd and random assessor measures for that topic, as defined below:

$$G_k^p(t) = \tau_{ap}\big(M_k(t,\cdot), M_h^p(t,\cdot)\big) \tag{19}$$

4.4 Weight

As anticipated above, the basic idea is to understand how close a crowd assessor W_k is to a random one ρ_h^p and consider this as an indicator of being a poor quality assessor. Therefore, we are interested in reasoning in terms of dissimilarity from random assessors: the farther away from a random assessor the higher the accuracy assigned to a crowd assessor.

As shown in Figure 3, we can create a vector space whose base is given by the three random assessors ρ_h^p, represent each crowd assessor W_k in this space, and project the crowd assessor on the random assessors (indicated by W_k^{uni}, W_k^{ovr}, and W_k^{und} respectively); $\mathbf{b}$ is the bisector of the first quadrant. Note that the projections of the crowd assessor on the random assessors are given by the gaps described above and properly normalized as discussed in the following section.

4.4.1 Normalization. When you reason in terms of similarity between vectors, if two vectors $\mathbf{v}$ and $\mathbf{w}$ are equal, then the norm of $\mathbf{v} - \mathbf{w}$ will be equal to 0, i.e. 0 means equal. However, in the vector space of Figure 3, we reason in terms of dissimilarity between vectors: 0 means different from random assessor and 1 means equal to random assessor. Therefore, in the following section, first we normalize all the gaps to the range $[0,1]$; then, when needed, we also transform them, e.g. by reversing the $[0,1]$ range, to ensure that these normalized gaps have the expected meaning of 0 "different from random assessor" and 1 "equal to random assessor".

Frobenius Norm. The Frobenius norm is in the $\left[0, \sqrt{|T| \cdot |S|}\right]$ range, where 0 means equal to a random assessor. So we need to divide it by its maximum and reverse it so that 0 means different

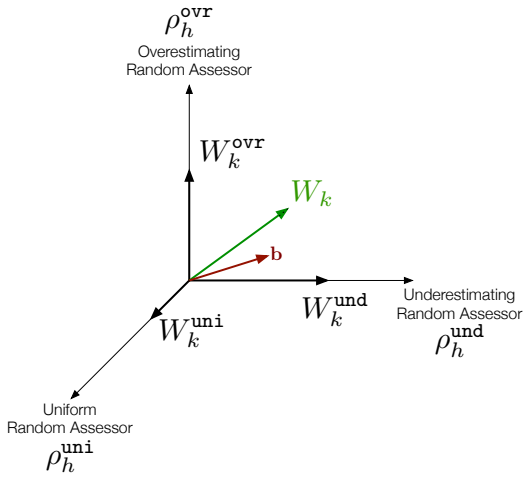


Fig. 3. Vector space representation of the crowd assessor W_k and the random assessors ρ_h^p.

from a random assessor:

$$G' = 1 - \frac{G}{\sqrt{|T| \cdot |S|}} \tag{20}$$

Note that when we consider the single score G_k, the equation holds as above; if we consider the topic score $G_k(t)$ we have to set $|T| = 1$ in the above equation.

Root Mean Square Error. The RMSE is in the $[0, 1]$ range, where 0 means equal to a random assessor. So we need to reverse it so that 0 means different from a random assessor:

$$G' = 1 - G \tag{21}$$

KL Divergence. The KLD is in the $[0, \infty)$ range, where 0 means equal to a random assessor. So we map it to the $(0, 1]$ range by the negative exponential so that 0 means different from a random assessor

$$G' = \mathrm{e}^{-\beta G} \tag{22}$$

where $\beta > 0$ is a positive real number.

Kendall's Tau Correlation. The Kendall's τ correlation is in the $[-1, 1]$ range, where 0 means different from a random assessor, 1 means equal to a random assessor and -1 completely opposite to a random assessor[5]. We consider -1 as 1:

$$G' = |G| \tag{23}$$

AP Correlation. The τ_{ap} correlation is in the $[-1, 1]$ range, where 0 means different from a random assessor, 1 means equal to a random assessor and -1 completely opposite to a random assessor[6]. We consider -1 as 1:

$$G' = |G| \tag{24}$$

[5]Consider an assessor that has correlation equal to -1 with one of the random assessors. This means that the assessor gives the exact opposite relevance judgement for each document. Therefore, this assessor can be considered a random assessor as well, and it is correct to give him a weight equal to 1.

[6]Same considerations as in the case of Kendall's τ hold here as well.

4.4.2 Minimal Dissimilarity. If we take the minimum between the dissimilarities of the assessor W_k from the random assessors, the assessor W_k cannot be closer to any of the random assessors more than this minimum. Therefore, we compute the minimum of the scalar products of the dissimilarity vector with the axes of the vector space shown in Figure 3:

$$w_k = \min\left(\left(G_k^{\mathsf{und}}\right)', \left(G_k^{\mathsf{uni}}\right)', \left(G_k^{\mathsf{ovr}}\right)'\right) \tag{25}$$

4.4.3 Minimal Squared Dissimilarity. We reason as in the previous case, but we consider the square of the gaps to have steeper behaviour:

$$w_k = \min\left(\left(\left(G_k^{\mathsf{und}}\right)'\right)^2, \left(\left(G_k^{\mathsf{uni}}\right)'\right)^2, \left(\left(G_k^{\mathsf{ovr}}\right)'\right)^2\right) \tag{26}$$

4.4.4 Minimal Equi-Dissimilarity. The bisector vector $\mathbf{b}$ represents the direction with the greatest equal dissimilarity from all the random assessors at the same time. Therefore, the closer the crowd assessor W_k is to the bisector $\mathbf{b}$, the farther away she/he is from all the random assessors at the same time. The scalar product between the crowd assessor vector and the bisector represents this quantity:

$$w_k = \left(G_k^{\mathsf{und}}\right)' + \left(G_k^{\mathsf{uni}}\right)' + \left(G_k^{\mathsf{ovr}}\right)' \tag{27}$$

4.5 Summary

Algorithm 1 shows the pseudo-code for computing the estimated accuracy of a crowd assessor W_k in the case of the single score granularity, while Algorithm 2 describes the case of the topic-by-topic score granularity. The inputs of the algorithms are the ground-truth produced by the crowd assessor W_k, i.e. the relevance judgments assigned by crowd assessor W_k, the ground-truths generated by each replicate of the random assessors with level p equals to 0.5 (`uni`), 0.05 (`und`) and 0.95 (`ovr`) and the performance measure to be computed. As output the algorithm will give the accuracy a_k for the crowd assessor W_k, which will be a single number for the single score granularity and a vector of length $|T|$ for the topic score granularity.

Firstly the performance measure is computed on the ground-truth provided by the crowd assessor W_k and by the H replicates of the three types of random assessors, obtaining respectively the $|T|\times|S|$ matrices M_k and M_h^p. Then the gap between the crowd assessor W_k and the random assessors is computed with respect to the strategies previously described:

- *Measure Level*:
  - Frobenius Norm: Equation (5) for single score granularity and Equation (6) for topic score granularity, Equation (20) to normalize the accuracy;
  - Root Mean Square Error: Equation (8) for single score granularity and Equation (9) for topic score granularity, Equation (21) to normalize the accuracy;
- *Distribution Level*:
  - KL Divergence: Equation (12) for single score granularity and Equation (13) for topic score granularity, Equation (22) to normalize the accuracy;
- *Ranking Level*:
  - Kendall's τ: Equation (15) for single score granularity and Equation (16) for topic score granularity, Equation (23) to normalize the accuracy;
  - AP Correlation: Equation (18) for single score granularity and Equation (19) for topic score granularity, Equation (24) to normalize the accuracy;

ALGORITHM 1: How to estimate assessor accuracy a_k for the single score granularity.

Data: r_t^k ground-truth generated by the k-th assessor; r_h^p ground-truth generated by the h-th random assessor of level p, where $h \in \{1, \ldots, H\}$ and $p \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$; $\mathrm{m}(\cdot)$ performance measure
Result: a_k single score granularity accuracy for the k-th assessor;

/* Compute the performance measure M_k for the k-th assessor and M_h^p for each random assessors */
$M_k \leftarrow$ compute $\mathrm{m}(\cdot)$ on r_t^k;
$M_h^p \leftarrow$ compute $\mathrm{m}(\cdot)$ on r_h^p, $\forall\, h \in \{1, \ldots, H\}$ and $\forall\, p \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$;

/* Compute the Gap $G_{k,h}^p$ with respect to each random assessor: $h \in \{1, \ldots, H\}$ and $p \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$ */
for $h \in \{1, \ldots, H\}$ **do**
  if *measure level* **then**
    if *frobenius norm* **then**
      $G_{k,h}^p = \left\| M_k - M_h^p \right\|_F \quad \forall\, p \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$;
      $(G_{k,h}^p)' = 1 - \frac{G_{k,h}^p}{\sqrt{|S|}} \quad \forall\, p \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$
    else if *RMSE* **then**
      $G_{k,h}^p = \mathrm{RMSE}\left(\overline{M}_k(\cdot, S) - \overline{M}_h^p(\cdot, S)\right) \quad \forall\, p \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$;
      $(G_{k,h}^p)' = 1 - G_{k,h}^p \quad \forall\, p \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$;
    end
  else if *distribution level* **then**
    $G_{k,h}^p = D_{KL}\left(M_k(:) \middle\| M_h^p(:)\right) \quad \forall\, p \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$;
    $(G_{k,h}^p)' = \mathrm{e}^{-\beta G_{k,h}^p} \quad \forall\, p \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$;
  else if *ranking level* **then**
    if *Kendall's Tau* **then**
      $G_{k,h}^p = \tau\left(\overline{M}_k(\cdot, S) - \overline{M}_h^p(\cdot, S)\right) \quad \forall\, p \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$;
      $(G_{k,h}^p)' = \left| G_{k,h}^p \right| \quad \forall\, p \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$;
    else if *AP Correlation* **then**
      $G_{k,h}^p = \tau_{ap}\left(\overline{M}_k(\cdot, S), \overline{M}_h^p(\cdot, S)\right) \quad \forall\, r \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$;
      $(G_{k,h}^p)' = \left| G_{k,h}^p \right| \quad \forall\, p \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$;
    end
  end
end

/* Aggregate the Gap with respect to the random assessor replicates */
$(G_k^p)' \leftarrow \mathrm{mean}\left((G_{k,h}^p)'\right) \quad \forall\, p \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$;

/* Compute the weight a_k */
if *minimal dissimilarity* **then**
  $w_k = \min\left((G_k^{\texttt{und}})', (G_k^{\texttt{uni}})', (G_k^{\texttt{ovr}})'\right)$;
else if *minimal squared dissimilarity* **then**
  $w_k = \min\left(\left((G_k^{\texttt{und}})'\right)^2, \left(G_k^{\texttt{uni}})'\right)^2, \left(G_k^{\texttt{ovr}})'\right)^2\right)$;
else if *minimal equi-dissimilarity* **then**
  $w_k = (G_k^{\texttt{und}})' + (G_k^{\texttt{uni}})' + (G_k^{\texttt{ovr}})'$;
end

ALGORITHM 2: How to estimate assessor accuracy a_k for the topic-by-topic score granularity.

Data: r_t^k ground-truth generated by the k-th assessor; r_h^p ground-truth generated by the h-th random assessor of level p, where $h \in \{1, \ldots, H\}$ and $p \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$; $\mathrm{m}(\cdot)$ performance measure
Result: a_k vector of length $|T|$ containing the topic score granularity accuracy for the k-th assessor;

/* Compute the performance measure M_k for the k-th assessor and M_h^p for each random assessors */
$M_k \leftarrow$ compute $\mathrm{m}(\cdot)$ on r_t^k;
$M_h^p \leftarrow$ compute $\mathrm{m}(\cdot)$ on r_h^p, $\forall\, h \in \{1, \ldots, H\}$ and $\forall\, p \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$;

/* Compute the Gap $G_{k,h}^p(t)$ with respect to each random assessor: $h \in \{1, \ldots, H\}$ and $p \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$ */
for $t \in \{1, \ldots, |T|\}$ **do**
  for $h \in \{1, \ldots, H\}$ **do**
    if *measure level* **then**
      if *frobenius norm* **then**
        $G_{k,h}^p(t) = \left\| M_k(t, \cdot) - M_h^p(t, \cdot) \right\|_F \quad \forall\, p \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$;
        $(G_{k,h}^p(t))' = 1 - \frac{G_{k,h}^p(t)}{\sqrt{|T| \cdot |S|}} \quad \forall\, p \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$
      else if *RMSE* **then**
        $G_{k,h}^p(t) = \mathrm{RMSE}\left(\overline{M}_k(t, \cdot) - \overline{M}_h^p(t, \cdot)\right) \quad \forall\, p \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$;
        $(G_{k,h}^p(t))' = 1 - G_{k,h}^p(t) \quad \forall\, r \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$;
      end
    else if *distribution level* **then**
      $G_{k,h}^p(t) = D_{KL}\left(M_k(t, \cdot) \middle\| M_h^p(t, \cdot)\right) \quad \forall\, p \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$;
      $(G_{k,h}^p(t))' = \mathrm{e}^{-\beta G_{k,h}^p(t)} \quad \forall\, p \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$;
    else if *ranking level* **then**
      if *Kendall's Tau* **then**
        $G_{k,h}^p(t) = \tau\left(M_k(t, \cdot), M_h^p(t, \cdot)\right) \quad \forall\, p \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$;
        $(G_{k,h}^p(t))' = \left| G_{k,h}^p(t) \right| \quad \forall\, p \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$;
      else if *AP Correlation* **then**
        $G_{k,h}^p(t) = \tau_{ap}\left(M_k(t, \cdot), M_h^p(t, \cdot)\right) \quad \forall\, r \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$;
        $(G_{k,h}^p(t))' = \left| G_{k,h}^p(t) \right| \quad \forall\, p \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$;
      end
    end
  end
end

/* Aggregate the Gap with respect to the random assessor replicates */
$(G_k^p(t))' \leftarrow \mathrm{mean}\left((G_{k,h}^p(t))'\right) \quad \forall\, p \in \{\texttt{uni}, \texttt{und}, \texttt{ovr}\}$ and $\forall\, t \in \{1, \ldots, |T|\}$;

/* Compute the weight w_k */
for $t \in \{1, \ldots, |T|\}$ **do**
  if *minimal dissimilarity* **then**
    $w_k(t) = \min\left((G_k^{\texttt{und}}(t))', (G_k^{\texttt{uni}}(t))', (G_k^{\texttt{ovr}}(t))' \right)$;
  else if *minimal squared dissimilarity* **then**
    $w_k(t) = \min\left(\left((G_k^{\texttt{und}}(t))'\right)^2, \left(G_k^{\texttt{uni}}(t))'\right)^2, \left(G_k^{\texttt{ovr}}(t))'\right)^2 \right)$;
  else if *minimal equi-dissimilarity* **then**
    $w_k(t) = (G_k^{\texttt{und}}(t))' + (G_k^{\texttt{uni}}(t))' + (G_k^{\texttt{ovr}}(t))'$;
  end
end

Finally, the normalized Gap iss averaged over the H replicates of each random assessors class and the weight of the crowd assessor W_k is computed with respect to one of the following methods:

- *Minimal Dissimilarity*: Equation (25);
- *Minimal Squared Dissimilarity*: Equation (26);
- *Minimal Equi-Dissimilarity*: Equation (27).

5 EXPERIMENTAL SETUP

5.1 Crowd Assessors Collection

We use the TREC 21, 2012, Crowdsourcing [63] 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 [70], using binary relevance. Participating groups had to submit a binary relevance judgment for every document in the judging pools of the ten topics. The 10 topics selected were: 411, 416, 417, 420, 427, 432, 438, 445, 446, and 447. In total 33 pools were submitted to TRAT; we excluded two of them (`INFLB2012` and `Orc2Stage`) because, for some topics, they did not assess any document as relevant; indeed, this prevents the computation of some evaluation measures because you lack the information about the recall base. Therefore, we actually used 31 out the 33 submitted pools for TRAT.

In TRAT, the majority vote of the submitted pools was compared to the NIST relevance judgments; when the majority vote differed from the NIST judgment, TRAT organizers adjudicated the final relevance judgment for a document. The TRAT adjudicated pool constitutes the gold standard for our experimentation.

5.2 Evaluation Measures

When it comes to measures for evaluating the effectiveness of the different approaches, we adopt two criteria used in the TREC 22, 2013, Crowdsourcing track [64]:

- *rank correlation*: we use AP correlation [78] to compare the ranking of the systems produced for a given performance measure $\mathrm{m}(\cdot)$ 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;
- *score accuracy*: 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 $\mathrm{m}(\cdot)$, 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.

Note that the above use of AP correlation and RMSE is not related to their use as gaps between assessors, explained in Section 4; here they are used as evaluation measures for comparing the different algorithms and methods under examination. Moreover, we do not adopt some of the evaluation measures used in the TREC Crowdsourcing tracks, such as the *Logistic Average Misclassification (LAM)* rate [12] and the *Area Under the ROC Curve (AUC)* [15], because these measures specifically deal with classification tasks and basically compare the assigned relevance labels, but this does not apply to our case because AWARE does not generate relevance labels.

5.3 Performance Measures

When it comes to the assessor measures M_k and M_h^p, we consider the following performance measures:

- *Average Precision (AP)* [6] represents the "gold standard" measure in IR, known to be stable and informative, with a natural top-heavy bias and an underlying theoretical basis as approximation of the area under the precision/recall curve [56];
- *Normalized Discounted Cumulated Gain (nDCG)* [28] is the normalized version of the widely-known DCG which discounts the gain provided by each relevant retrieved document proportionally to the rank at which it is retrieved. nDCG is defined for graded relevance judgments and we use nDCG@20, which is calculated up to rank position 20.
- *Expected Reciprocal Rank (ERR)* [10] is a measure defined for graded relevance judgments and it is particularly top-heavy since it highly penalizes systems placing not-relevant documents in high positions. We use ERR@20.

5.4 Systems

Two TREC Adhoc tracks used these 10 topics over the years: the TREC 08, 1999, Ad-hoc track [70] (labeled `T08`), which contains 129 runs and from which these topics were selected; and, the TREC 13, 2004, Robust track [69] (labeled `T13`), which contains 110 runs and whose goal was to specifically experiment against hard topics.

Both `T08` and `T13` adopt a corpus of about 528K news documents, i.e. disk 4 and 5 of the TIPSTER collection minus the Congressional Record.

5.5 Parameters Setup

For nDCG we use a log base $b = 2$ and gains 0 and 5 for not relevant and relevant documents, respectively. For ERR we use gains 0 and 5 for not relevant and relevant documents, respectively.

We generate $H = 1,000$ replicates of the random assessors in each class – uniform, underestimating and overestimating assessors.

Let $l = 31$ be the total number of available crowd assessors and $k < l$ the number of assessors we are merging using the AWARE framework or other approaches. For each of the above evaluation measures, we experimented all the $k = 2, 3, \ldots, 30$. For each value of k, there are $\binom{31}{k} = \frac{31!}{k!(31-k)!}$ possible ways of choosing the k assessors to be merged; we randomly sampled 1,000 k-tuples out of the $\binom{31}{k}$ possible ones. The evaluation measures we report – AP correlation and RMSE – are averaged over these 1,000 samples.

For the computation of AP correlation in the case of ties, we sample and average over 100 randomly generated orderings.

For the KDE of a performance measure in equation (10), we use 100 equally spaced values x in the range $[0, 1]$, a Gaussian kernel $K(\cdot)$, and a bandwidth $b = 0.015$.

For the normalization of the KLD in equation (22), we set $\beta = 1$.

For the EM algorithms we set a limit of 1,000 iterations and a tolerance of 10^{-3}.

All the experiments were developed using the *MATlab Toolkit for Evaluation of information Retrieval Systems (MATTERS)* library[7] and their source code is publicly available[8] to favour reproducibility.

5.6 Experiments

We experiment all the combinations of factors for the estimation of a crowd assessor accuracy, as described in Section 4:

- *granularity*: whether, for a crowd assessor, we compute a single accuracy (`sgl`) or a separate accuracy for each topic (`tpc`);

[7] http://matters.dei.unipd.it/
[8] https://bitbucket.org/frrncl/tois-aware

- *gap*: how we compute the "difference" between a crowd and a random assessor (`fro`, `rmse`, `kld`, `tau`, or `apc`);
- *weight*: how we turn a "difference" between a crowd and a random assessor into a final accuracy estimation (`md`, `msd`, or `med`).

The combination of these three factors gives raise to 30 different approaches for estimating a crowd assessor accuracy. We introduce the following notation to facilitate the comprehension of the main characteristics of an estimator from its name:

`<granularity>_<gap>_<weight>`

So, for example, the tag `sgl_apc_med` indicates a single crowd assessor accuracy a_k for all the topics using AP correlation as "difference" between crowd and random assessors and the minimal equi-dissimilarity weighting criterion.

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, as explained in Section 2.

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, as done in the toy example of Section 3.

We conduct the following experiments:

- a factorial analysis to isolate the contributions of different factors – k-tuple size, the performance measure under consideration, and the considered systems (Section 6). This analysis allows us to understand: (i) which approaches perform best across a wide range of influencing factors, net their effects; (ii) how these factors interact with each other;
- a break-down of the contribution of the different components of the AWARE estimators – namely granularity, gap, and weight (Section 7). This analysis allows us to dig into the AWARE estimators themselves and better understand how they work.

6 FACTORIAL ANALYSIS OF KTUPLE, APPROACH, MEASURE AND SYSTEM EFFECTS

6.1 Methodology

The goal of this section is to conduct a deep analysis to investigate how the AWARE approaches and the state-of-the-art baselines behave with respect to different factors, namely the k-tuple size, the performance measure under consideration, and the considered systems. To this end, we adopt the following *General Linear Mixed Model (GLMM)* model for the three-way *ANalysis Of VAriance (ANOVA)* with repeated measures [48, 57]:

$$Y_{ijkl} = \underbrace{\mu_{\dots} + \kappa_i + \alpha_j + \beta_k + \gamma_l}_{\text{Main Effects}} + \underbrace{\alpha\beta_{jk} + \alpha\gamma_{jl} + \beta\gamma_{kl}}_{\text{Interaction Effects}} + \underbrace{\varepsilon_{ijkl}}_{\text{Error}} \tag{28}$$

where: Y_{ijkl} is the score of the i-th subject in the j-th, k-th, and l-th factors; $\mu_{\dots}$ is the grand mean; κ_i is the effect of the i-th subject, i.e. the k-tuple size $k = 2, \dots, 30$; α_j is the effect of the j-th factor, i.e. both the AWARE and the state-of-the-art approaches; β_k is the effect of the k-th factor, i.e. the the performance measures under consideration, namely AP, nDCG@20, and ERR@20; and, γ_l is the effect of the l-th factor, i.e. the systems submitted to the `T08` and `T13` tracks. We consider also the interaction effects among approaches and performance measures ($\alpha\beta_{jk}$), approaches and systems ($\alpha\gamma_{jl}$), and performance measures and systems ($\beta\gamma_{kl}$). Finally, ε_{ijkl} is the error committed by the model in predicting the score of the i-th subject in the three factors j, k, l.

For each model, we report the ANOVA table which summarizes the outcomes of the ANOVA test on the above model indicating, for each factor, the *Sum of Squares (SS)*, the *Degrees of Freedom (DF)*, the *Mean Squares (MS)*, the F statistics, and the p-value of that factor. In the following, we consider a confidence level $\alpha = 0.05$ to determine if a factor is statistically significant.

We are not only interested in determining whether a factor effect is significant, i.e. its p-value in the ANOVA table is less than 0.05, but also which proportion of the variance is due to it. Therefore, we need to estimate its *effect-size measure* or *Strength of Association (SOA)*. The SOA is a "standardized index and estimates a parameter that is independent of sample size and quantifies the magnitude of the difference between populations or the relationship between explanatory and response variables" [51, 59]. We use the $\hat{\omega}^2_{\langle fact \rangle}$ SOA:

$$\hat{\omega}^2_{\langle fact \rangle} = \frac{df_{fact}(F_{fact} - 1)}{df_{fact}(F_{fact} - 1) + N} \tag{29}$$

which is an unbiased estimator of the variance components associated with the sources of variation in the design, where N is the total number of elements under analysis.

The common rule of thumb [50] when classifying $\hat{\omega}^2_{\langle fact \rangle}$ effect size is: 0.14 and above is a large effect, 0.06–0.14 is a medium effect, and 0.01–0.06 is a small effect. $\hat{\omega}^2_{\langle fact \rangle}$ values could happen to be negative and in such cases they are considered as zero.

In addition to the ANOVA table, we also show both the main effects and the interaction effects plots in order to get a better appreciation of the behaviour of the different levels of each factor. In particular, the main effects plot graphs the response mean for each factor level connected by a line. An interaction effects plot displays the levels of one factor on the X axis and has a separate line for the means of each level of the other factor on the Y axis; it allows us to understand whether the effect of one factor depends on the level of the other factor.

A *Type I* error occurs when a true null hypothesis is rejected and the significance level α is the probability of committing a Type I error. When performing multiple comparisons, the probability of committing a Type I error increases with the number of comparisons and we keep it controlled by applying the Tukey *Honestly Significant Difference (HSD)* test [25] with a significance level $\alpha = 0.05$. Tukey's method is used in ANOVA to create confidence intervals for all pairwise differences between factor levels, while controlling the family error rate. Two levels u and v of a factor are considered significantly different when

$$|t| = \frac{|\hat{\mu}_u - \hat{\mu}_v|}{\sqrt{MS_{error}\left(\frac{1}{n_u} + \frac{1}{n_v}\right)}} > \frac{1}{\sqrt{2}} q_{\alpha,k,N-k} \tag{30}$$

where $\hat{\mu}_u$ and $\hat{\mu}_v$ are the marginal means, i.e. the main effects, of the two factors; n_u and n_v are the sizes of levels u and v; $q_{\alpha,k,N-k}$ is the upper $100 * (1 - \alpha)$th percentile of the studentized range distribution with parameter k and $N - k$ degrees of freedom; k is the number of levels in the factor and N is the total number of observations.

In the following, we have a section dedicated to each evaluation measure, i.e. AP correlation and RMSE.

Note that when we analyse AP correlation, when can use the data as they are, since all the scores are in the same range $[0, 1]$ and they are measured in the same way. On the other hand, when we analyse RMSE, even if all the measures are in the range $[0, 1]$ and so also RMSE is, AP = 0.20 is not exactly the same as ERR@20 = 0.20 because of their different user models and they typically assume different values in the range $[0, 1]$. As a consequence, an RMSE 0.15 for AP is not directly comparable with an RMSE 0.15 for ERR@20. Therefore, we need to apply some kind of

Table 1. ANOVA table for AP Correlation considering the k-tuple size, approach, measure and systems effects.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$
K-tuple Size	3.5161	28	0.1256	580.7705	< 0.0001	
Approach	1.2264	33	0.0372	171.8716	< 0.0001	0.4880
Measure	13.0727	2	6.5364	30,230.1290	< 0.0001	0.9109
Systems	1.9857	1	1.9857	9,183.9134	< 0.0001	0.6082
Approach*Measure	2.0701	66	0.0314	145.0584	< 0.0001	0.6164
Approach*Systems	0.3008	33	0.0091	42.1620	< 0.0001	0.1867
Measure*Systems	5.3240	2	2.6620	12,311.4096	< 0.0001	0.8063
Error	1.2433	5,750	0.0002			
Total	28.7391	5,915				

normalization first to make the scores comparable and we normalize them by the maximum value achieved on the dataset, thus reasoning in term of ratios.

6.2 AP Correlation

Table 1 shows that all the main and interaction effects are statistically significant. As far as main effects are concerned, we can see 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, instead of optimizing upstream, as also illustrated in the toy example of Section 1.

When it comes to the interaction effects, `Approach*Measure` is a large size effect, about 1.27 times greater than the `Approach` effect alone, while `Approach*Systems` is a large size effect but less than half the `Approach` effect alone. These two facts further strengthen the intuition behind AWARE: not only do `Measures` and `Systems` effects play an important role alone, they also influence and interact a lot with the `Approaches` for merging assessors, where `Measures` have a greater impact on `Approaches` than `Systems`.

Finally, there is also a large size interaction effect between `Measure` and `Systems`, indicating that different measures score systems differently, but this is less interesting for the purposes of the present discussion because it is an intrinsic phenomenon of the relationship between performance measures and systems.

The main effects plot in Figure 4 shows the marginal mean contributions of each effect together with their confidence interval (shaded). Figure 4(a) shows the contributions of the different approaches across all the conditions and net of their effects, thus allowing us to appreciate the best and most stable approaches in many operational settings. We can see that the AWARE approaches lie in a somehow stable range of performances, with the only exception of `sgl_rmse_msd` which is the worst performing one but still better than `emmv` and `emneu`.

As expected, we can observe from in Figure 4(b) that increasing the number of merged assessors improves the performances; you can also note how the confidence interval slightly increases as the k-tuple size increases, denoting a higher variability due to the larger number of (potentially heterogenous) assessors merged.

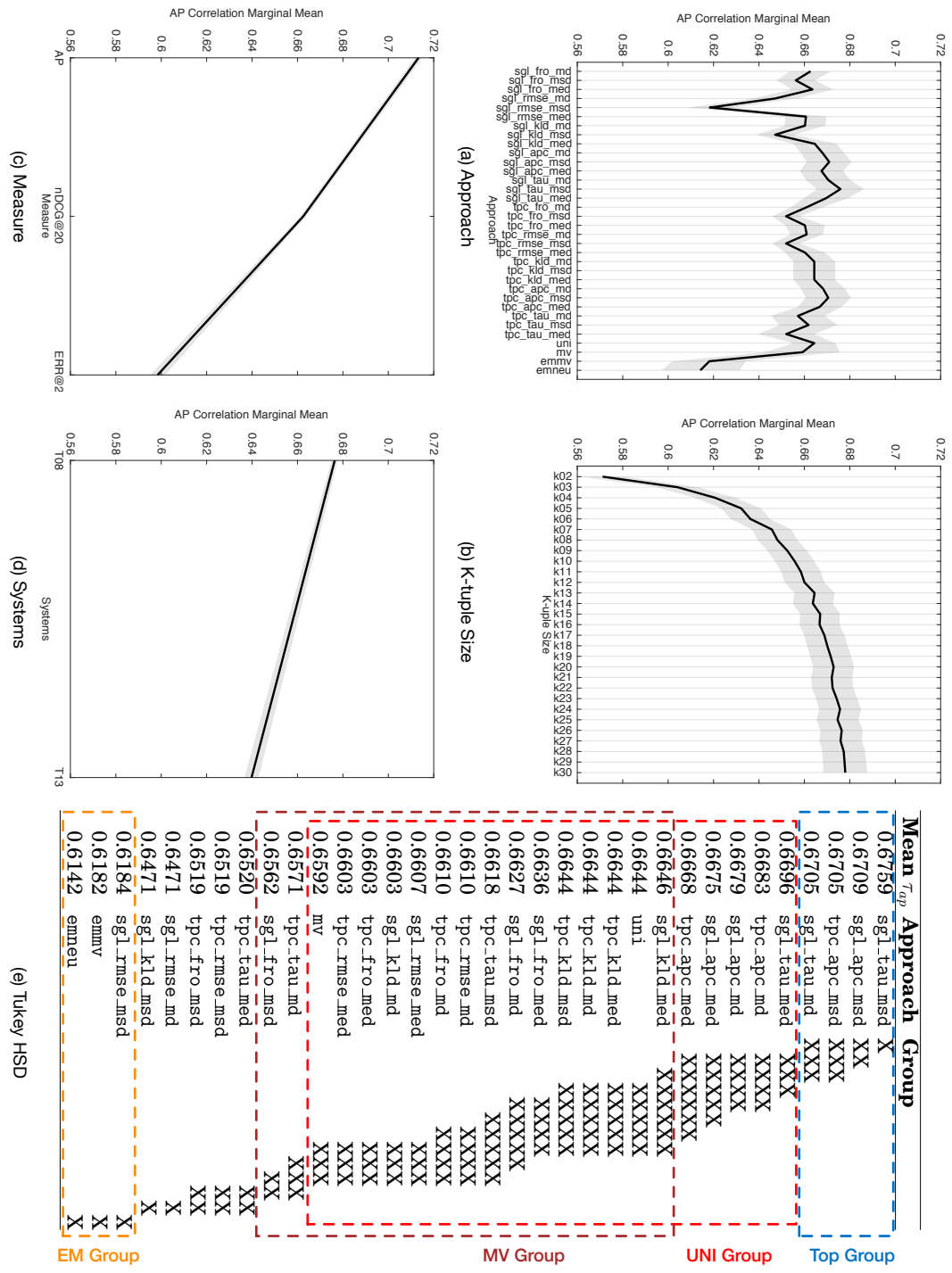


Fig. 4. AP correlation: main effects plots for `Approach` (a), `K-tuple Size` (b), `Measure` (c), `Systems` (d), and Tukey HSD multiple comparison test for the `Approach` factor (e).

Figure 4(c) shows how the different performance measures lead to quite different performances when it comes to merging assessors and, in particular, nDCG@20 and ERR@20 are more challenging than AP. Finally, Figure 4(d) highlights how the targeted systems affect the performances as well, with the T13 ones somehow being more difficult.

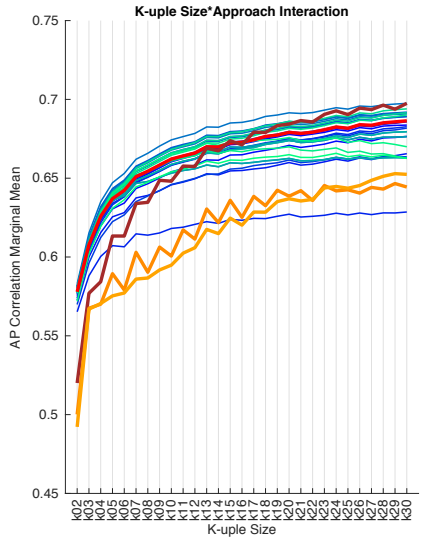
The Tukey HSD multiple comparison analysis reported in Figure 4(e) highlights the top group (dashed blue line), the group of approaches not significantly different from the uni baseline (dashed bright red line), the group of approaches not significantly different from mv (dashed dark red line), and the group of approaches not significanty 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, 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. As also shown in the toy example of Section 3, among the AWARE approaches, uni is the closest to mv, in that they both merge assessors attributing the same weight to all of them; yet performing this operation on the measures rather than on the relevance judgments proves to be slightly more effective.

Figure 5 shows the interaction plots. We used the following color convention: we selected cool colors for the proposed models, based on the AWARE framework, and warm colors for state-of-the-art models, i.e. mv, emmv, emneu and the AWARE uni baseline.

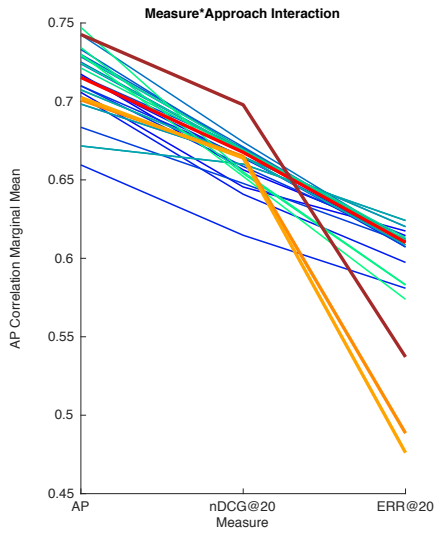
As shown in Figure 5(a), we can see that K-tuple Size has a positive effect for all the Approaches. Figure 5(a) also allows us to understand which approaches perform best for a given number of crowd assessors, i.e. for a given k-tuple size. AWARE approaches start higher for low k-tuple sizes while state-of-the-art ones grow faster as the k-tuple size increases. In particular, mv reaches uni at $k = 13$ merged assessors and surpasses it from $k = 17$ onwards, attaining an interaction level as positive as sgl_tau_msd just from $k = 25$ merged assessors. On the other hand, the emneu and emmv methods start to behave better at higher numbers of merged assessors and this is consistent with previous findings in the literature [54, 55].

Being effective already at low numbers of merged assessors is a clear advantage of the AWARE approaches, since this helps in containing the costs and effort for creating a pool. Moreover, when considering the increasingly better performances of the mv method with high numbers of merged assessors, we have also to remember how the gold standard has been created: TREC 2012 Crowdsourcing organizers took the majority vote of the submitted pools and then adjudicated it with respect to the NIST pool. Therefore, it is somehow natural that when you use almost all the crowd assessors, i.e. all the submitted pools, the performances of the majority vote tend to become the best ones, since you start converging towards what has been used as the gold standard.

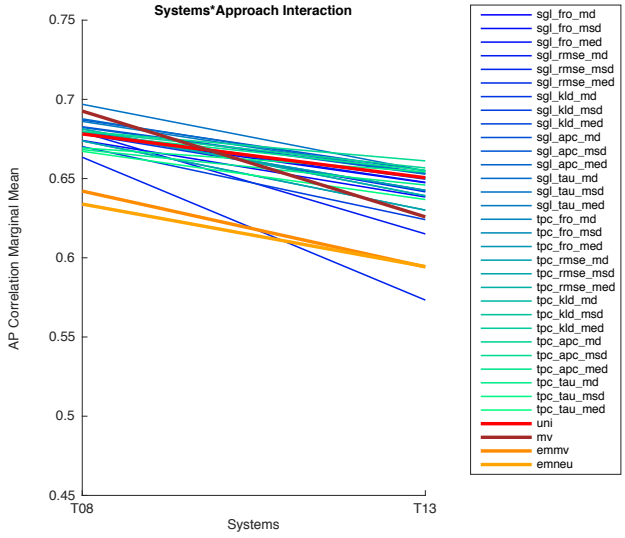
When it comes to the interaction between Measures and Approaches (Figure 5(b)), AWARE approaches react more proportionally to the increasing difficulty of the different performance measures; indeed, while mv is among the best interacting approaches for AP and the best one for nDCG@20, it suffers from a very consistent drop in the case of ERR@20 (and similarly for emmv and emneu). Finally, in the case of the interaction between Systems and Approaches (Figure 5(c)), AWARE approaches behave similarly while mv loses more when it comes to the T13 systems. Again, all of this supports the intuition behind the AWARE approaches about taking into account performance measures and systems in the merging process.



(a) K-tuple size and Approach



(b) Measure and Approach



(c) System and Approach

Fig. 5. AP correlation: interaction effects plots for K-tuple size (a), measure (b) and systems (c).

6.3 RMSE

Table 2 shows how all the main effects as well as all the interaction effects are statistically significant. 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

Table 2. ANOVA table for normalised RMSE considering the k-tuple size, approach, measure and systems effects.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$
K-tuple Size	20.6579	28	0.7378	272.9961	< 0.0001	
Approach	32.2530	33	0.9774	361.6465	< 0.0001	0.6680
Measure	56.7010	2	28.3505	10,490.3151	< 0.0001	0.7800
Systems	3.7700	1	3.7700	1,394.9723	< 0.0001	0.1907
Approach*Measure	45.4675	66	0.6889	254.9091	< 0.0001	0.7391
Approach*Systems	2.4886	33	0.0754	27.9039	< 0.0001	0.1305
Measure*System	0.6374	2	0.3187	117.9227	< 0.0001	0.0380
Error	15.5396	5,750	0.0027			
Total	177.5149	5,915				

behind AWARE, but it also suggests that Approaches are much more prominent for the accurate estimation of the actual value of a performance measure, (i.e. what is assessed by the RMSE) than for ranking systems correctly (i.e. what is assessed by AP correlation).

When it comes to the interaction effects, we can see that Approach*Measure and Approach*Systems are both large size effects and that the Approach*Measure is the second largest effect, a bit bigger than Approach alone; again, these two facts strengthen the motivations behind AWARE. Finally, the Measure*Systems factor is a small size effect but this is less relevant for our discussion, as explained in the previous section.

The main effects plots in Figure 6 show: (i) that increasing the number of merged assessors has the expected positive impact, with a greater variability when merging a higher number of (possibly heterogenous) assessors, see Figure 6(b); (ii) how the different performance measures influence the effectiveness, with AP being the most challenging one while nDCG@20 and ERR@20 display a somewhat similar behavior, see Figure 6(c); (iii) that the targeted systems affect the performances as well, with T08 being somehow more difficult, see Figure 6(d).

Figure 6(a) shows the main effects of the Approach factor: we can see that the AWARE approaches are quite good, but with a few more exceptions than in the case of AP correlation, namely sgl_rmse_msd, tpc_fro_msd, and tpc_rmse_msd. The top group, reported in Figure 6(e), consists of sgl_rmse_med, tpc_rmse_med, tpc_fro_med (the top ones with extremely close performances), sgl_fro_med, 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, see Figure 6(e).

If we look at the interaction effects plots in Figure 7, we can see that K-tuple size has a positive effect for all the Approaches, apart from emmv and emneu, see Figure 7(a). As in the case of AP correlation, AWARE approaches quickly gain at lower numbers of merged assessors, becoming more stable as the k-tuple size increases. Unlike the case of AP correlation, mv behaves like AWARE approaches up to $k = 16$ merged assessors whereas, afterwards, adding more assessors becomes even harmful.

When it comes to the Measure*Approach interaction effect in Figure 7(b), we can see that emmv and emneu react badly to it, while mv behaves similarly to the AWARE approaches, even though many of them benefit from the Measure effect more than mv, which is one of the worst interacting approach in the case of ERR@20. Finally, for the Systems*Approach interaction effect in Figure 7(c), emmv and emneu are almost insensitive to it and perform badly, while mv behaves better than most

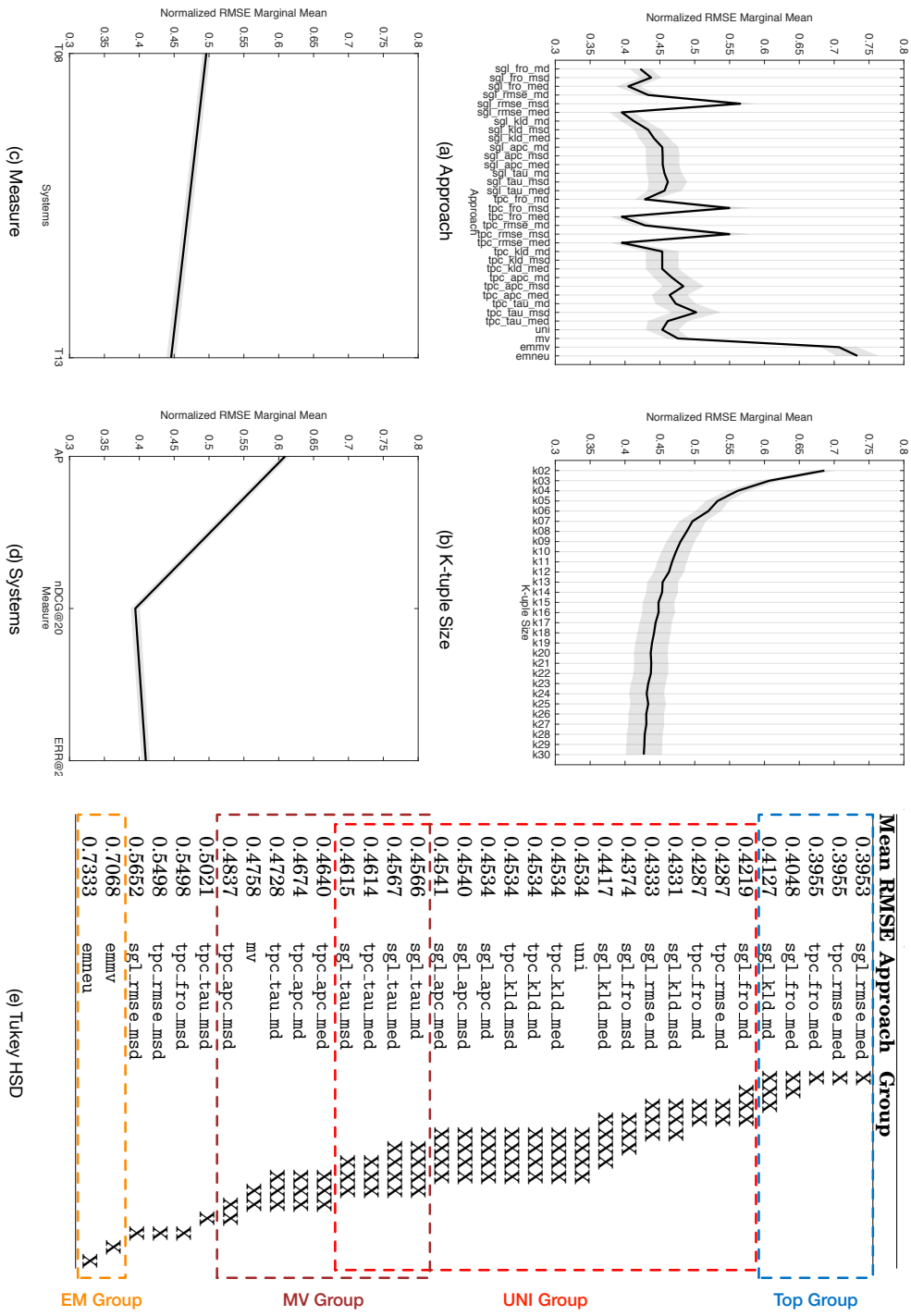
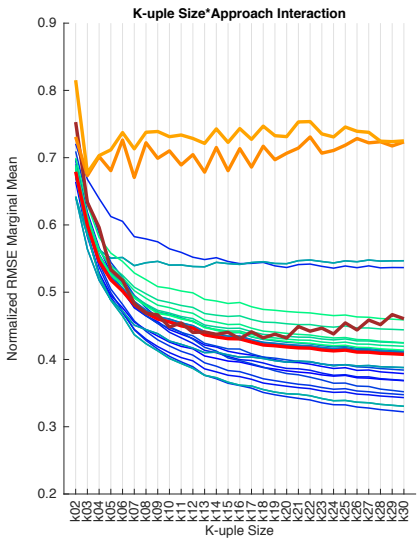
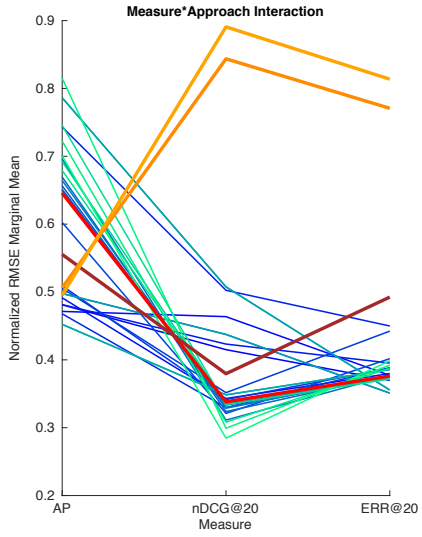


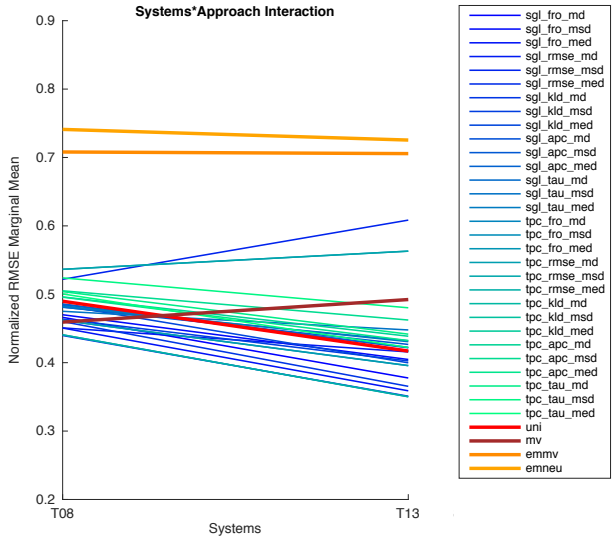
Fig. 6. RMSE: main effects plots for Approach (a), K-tuple Size (b), Measure (c), Systems (d), and Tukey HSD multiple comparison test for the Approach factor (e).



(a) K-tuple size and Approach



(b) Measure and Approach



(c) Systems and Approach

Fig. 7. AP correlation: interaction effects plots considering k-tuple size (a), measure (b) and systems (c).

of the AWARE approaches for T08, but worse than most of them in the case of T13. Overall, these facts are a further confirmation of the intuition which led to the development of AWARE.

7 FACTORIAL ANALYSIS OF AWARE COMPONENTS

7.1 Methodology

The goal of this section is to conduct a break-down analysis to investigate how the different components of the AWARE accuracy estimators, namely the granularity, gap, and weight, behave at the net of the other factors, namely the k-tuple size, the performance measure under consideration, and the considered systems. To this end, we adopt the following GLMM model for the three-way ANOVA with repeated measures:

$$Y_{ijklmn} = \underbrace{\mu..... + \kappa_i + \alpha_j + \beta_k + \gamma_l + \delta_m + \zeta_n}_{\text{Main Effects}} + \underbrace{\alpha\beta_{jk} + \alpha\gamma_{jl} + \beta\gamma kl}_{\text{Interaction Effects}} + \underbrace{\varepsilon_{ijklmn}}_{\text{Error}} \tag{31}$$

where: Y_{ijkl} is the score of the i-th subject in the j-th, k-th, l-th, m-th, and n-th factors; $\mu.....$ is the grand mean; κ_i is the effect of the i-th subject, i.e. the ktuple size $k = 2, \ldots, 30$; α_j is the effect of the j-th factor, i.e. the granularity either `sgl` or `tpc`; β_k is the effect of the k-th factor, i.e. the adopted gap, namely `fro`, `rmse`, `kld`, `apc`, or `tau`; γ_l is the effect of the k-th factor, i.e. the adopted weight, namely `md`, `msd`, or `med`; δ_m is the effect of the m-th factor, i.e. the the performance measures under consideration, namely AP, nDCG@20, and ERR@20; and, ζ_n is the effect of the n-th factor, i.e. the systems submitted to the `T08` and `T13` tracks. We consider also the interaction effects among granularity and gap ($\alpha\beta_{jk}$), granularity and weight ($\alpha\gamma_{jl}$), and gap and weight ($\beta\gamma_{kl}$). Finally, ε_{ijklmn} is the error committed by the model in predicting the score of the i-th subject in the five factors j, k, l, m, n.

As in the previous section, also in this case we normalize the RMSE score by their maximum value for each performance measure before proceeding with the analyses.

7.2 AP Correlation

Table 3 confirms that `K-tuple Size`, `Measure` and `Systems` are significant and large size factors that affect the performances as already observed in the previous section, with `Measure` and `Systems` being the most prominent effects. All the interaction effects are small size effects with `Granularity*Gap` and `Gap*Weight` quite similar in terms of size and `Granularity*Weight` about 6 times smaller.

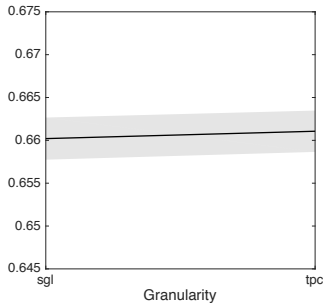
When it comes to the break-down of the AWARE components, we can observe that `Granularity` is not a significant factor. This can also be noted in: (i) the main effects plot in Figure 8(a), where `sgl` and `tpc` are connected by an almost straight line; (ii) the Tukey HSD multiple comparison analysis in Figure 8(d), which shows that `sgl` and `tpc` are not significantly different since their ranges overlap.

Both the `Gap` and the `Weight` factors are significant but small size effects, see Figures 8(b) and 8(c), even though `Gap` is about 5.8 times `Weight` in terms of explained variance. In particular, the top gaps are `apc` and `tau`, see Figure 8(e), while `med` and `md` are the top weights, see Figure 8(f). Overall, this suggests that, in terms of AP correlation, the key ingredient of the AWARE approaches is the `Gap` component and this is corroborated also by the top approaches emerging from Figure 4(e), i.e. `sgl_tau_msd` (the top one), `sgl_apc_msd`, `sgl_tau_md`, and `tpc_apc_msd`, which are a combination of the top `Gaps` and `Weights`.

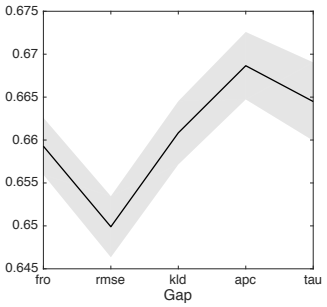
When it comes to the interaction between the different AWARE components in Figure 8(g), it turns out that the `apc` and `fro` gaps are almost insensitive to either the `sgl` or the `tpc` granularities and that the `tau` gap works better with the `sgl` granularity while the opposite is true for the `kld` and `rmse` gaps. Overall, this suggest that gaps closer to the assessor measures, i.e. `rmse` and `fro`,

Table 3. ANOVA table for AP correlation providing the break-down of AWARE components effects.

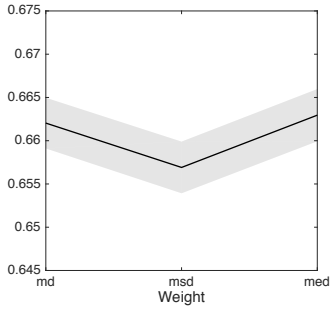
Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$
K-tuple Size	2.8154	28	0.1005	80.7404	< 0.0001	
Granularity	0.0009	1	0.0009	0.7746	0.3788	0
Gap	0.2049	4	0.0512	41.1420	< 0.0001	0.0298
Weight	0.0369	2	0.0185	14.8185	< 0.0001	0.0053
Measure	9.6331	2	4.8166	3,867.6402	< 0.0001	0.5970
Systems	1.6418	1	1.6418	1,318.3279	< 0.0001	0.2015
Granularity*Gap	0.1373	4	0.0343	27.5633	< 0.0001	0.0199
Granularity*Weight	0.0256	2	0.0128	10.3056	< 0.0001	0.0036
Gap*Weight	0.1263	8	0.0158	12.6764	< 0.0001	0.0176
Error	6.4347	5,167	0.0012			
Total	21.0570	5,219				



(a) Granularity



(b) Gap



(c) Weight

Mean τ_{ap}	Granularity	Group
0.6611	tpc	X
0.6602	sgl	X

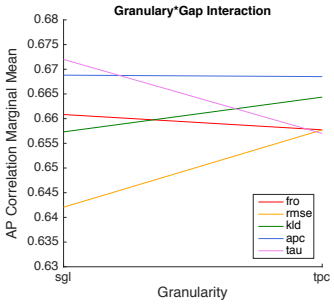
(d) Tukey HSD Granularity

Mean τ_{ap}	Gap	Group
0.6687	apc	X
0.6645	tau	XX
0.6608	kld	XX
0.6593	fro	X
0.6499	rmse	X

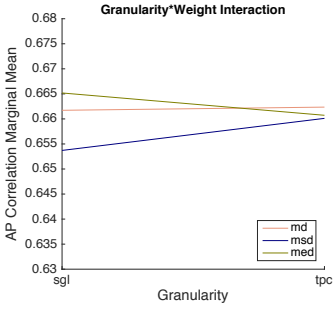
(e) Tukey HSD Gap

Mean τ_{ap}	Weight	Group
0.6630	med	X
0.6620	md	X
0.6569	msd	X

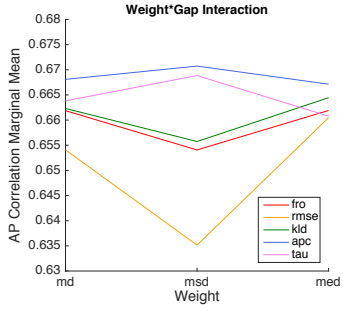
(f) Tukey HSD Weight



(g) Granularity and Gap



(h) Granularity and Weight



(i) Weight and Gap

Fig. 8. AP correlation: main effects plots (a), (b), (c), Tukey HSD multiple comparison tests (d), (e), (f), and interaction plots (g), (h), (i), for τ_{AP} considering granularity, gap, and weight.

benefit from a pinpoint granularity more than progressively less close ones, as the kld, tau, and apc gaps are.

As far as Granularity*Weight interaction is concerned in Figure 8(h), it is interesting to note the difference in behavior between the kinds of weighting schemes: the minimal (squared) dissimilarity ones, i.e. md and msd, benefit more from tpc than sgl (especially msd) while the opposite is true for the other weighting scheme, i.e. med.

Finally, the Weight*Gap interaction in Figure 8(i) reveals that all the Gaps are almost insensitive to the md and med weights while they either gain a lot (apc and tau) or lose a lot (kld, fro, rmse) with the msd weight. This suggests that the sharpness of the weighting scheme, i.e. minimal squared dissimilarity, affects the gaps more than the difference in the kind of weighting schemes, i.e. minimal dissimilarity vs minimal equi-dissimilarity, and this becomes more and more detrimental as you choose a gap closer and closer to the assessor measures.

7.3 RMSE

As in the case of AP correlation, Table 4 confirms that K-tuple Size, Measure and Systems are significant and large size factors, with the Measures and Systems being quite close in terms of size.

Unlike the case of AP correlation, for RMSE all the AWARE components factors are statistically significant and, while Granluarity and Gap are small size effects, Weight is a medium size effect. The interaction effects Granularity*Gap and Granularity*Weight are small size effects, while Gap*Weight is a medium size effect, greater than Weight alone.

Looking at the main effects and Tukey HSD multiple comparison analyses in Figure 9, we can see that: sgl granularity is the best, see Figure 9(d); the kld and fro gaps are the top ones, see Figure 9(e), suggesting that gaps moderately close to assessor measures are preferable to better predict a performance score; and the med weight is better than both md and msd, see Figure 9(f), indicating that its balanced distance from all the random assessors works best in predicting performance scores.

As suggested also by Table 4, the Weight*Gap interaction is the most prominent one: in Figure 9(i) the rmse and fro gaps lose most with the msd weight while they have a consistent gain with the med weight; the other gaps are almost insensitive to the weight, apart from a small drop with msd. This suggests that the closer the gap to the assessor measure, the stronger the interaction with the weights: a very negative one in the case of the msd weight, which is the sharpest one; a very positive one in the case of the med weight, which is the most balanced one.

When it comes to the Granularity*Gap interaction in Figure 9(g) gaps tend to improve passing from the tpc to the sgl granularity, especially fro, although rmse is an exception as slightly gains with the tpc granularity.

Finally, for the Granularity*Weight interaction in Figure 9(h) med and md are mostly insensitive to granularity, while msd improves using sgl.

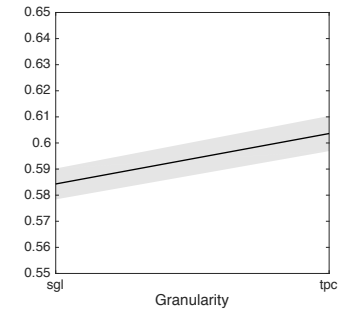
8 CONCLUSIONS AND FUTURE WORK

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/mislabeled 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.

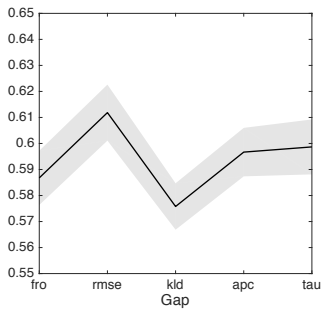
We then exemplified how to instantiate the proposed stochastic framework by introducing many unsupervised estimators of the accuracy of crowd assessors.

Table 4. ANOVA table for RMSE providing the break-down of AWARE components effects.

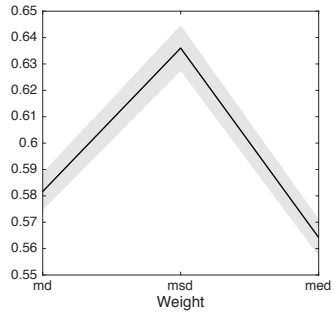
Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$
K-tuple Size	36.9657	28	1.3202	136.8070	< 0.0001	
Granularity	0.4871	1	0.4871	50.4805	< 0.0001	0.0094
Gap	0.7642	4	0.1910	19.7969	< 0.0001	0.0142
Weight	4.8828	2	2.4414	252.9934	< 0.0001	0.0880
Measure	25.0089	2	12.5044	1,295.7805	< 0.0001	0.3316
Systems	18.7670	1	18.7670	1,944.7464	< 0.0001	0.2713
Granularity*Gap	0.3381	4	0.0845	8.7584	< 0.0001	0.0059
Granularity*Weight	0.2805	2	0.1403	14.5350	< 0.0001	0.0052
Gap*Weight	5.8342	8	0.7293	75.5719	< 0.0001	0.1026
Error	49.8622	5,167	0.0097			
Total	143.1908	5,219				



(a) Granularity



(b) Gap



(c) Weight

Mean RMSE	Granularity	Group
0.5843	sgl	X
0.6036	tpc	X

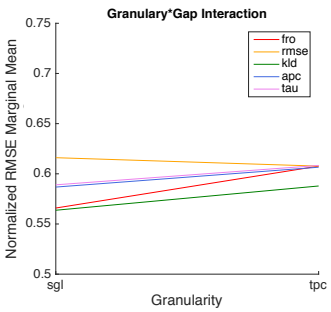
(d) Tukey HSD Granularity

Mean RMSE	Gap	Group
0.5758	kld	X
0.5868	fro	XX
0.5967	apc	XX
0.5987	tau	X
0.6119	rmse	X

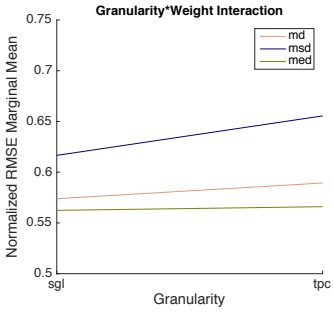
(e) Tukey HSD Gap

Mean RMSE	Weight	Group
0.5642	med	X
0.5816	md	X
0.6360	msd	X

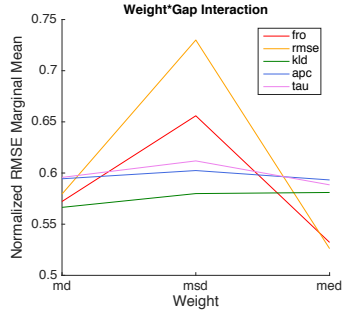
(f) Tukey HSD Weight



(g) Granularity and Gap



(h) Granularity and Weight



(i) Weight and Gap

Fig. 9. RMSE: main effects plots (a), (b), (c), Tukey HSD multiple comparison tests (d), (e), (f), and interaction plots (g), (h), (i), for τ_{AP} considering granularity, gap, and weight.

Finally, we conducted a thorough evaluation on TREC collections, comparing AWARE against state-of-the-art approaches and studying their influencing factors, namely performance measures and systems. We also investigated the contributions and interactions of the different components of the AWARE estimators.

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. Finally, it has provided insights about which estimators work best in which context.

Table 5 summarizes the top AWARE approaches, analyzed in detail in Section 6, as well as the best AWARE components, namely granularities, gaps and weights, analyzed in Section 7; the table shows these analyses for both AP correlation, i.e. as far as ranking systems is concerned, and RMSE, i.e. as far as predicting system performances is concerned.

`sgl_tau_msd` is the best approach in terms of AP correlation while `sgl_rmse_med` is the best approach for RMSE. In general, AWARE approaches outperform the state-of-the-art ones which are never part of the top group. Moreover, for both AP correlation and RMSE, we can observe that increasing the number of crowd assessors improve the performances – see Figures 4(b) and 6(b) – but the AWARE approaches are more effective than the state-of-the-art ones for low numbers of assessors, as shown in Figures 5(a) and 7(a). Therefore, besides better performance, AWARE provides the additional benefit of requiring less resources for ground-truth creation.

When it comes to components, in terms of AP correlation, the `sgl` and `tpc` granularities are not significantly different, even if the `sgl` granularity is predominant among top approaches. This is due to the interaction among components, analyzed in Figure 8, which boost the performances for some combinations of components, e.g. the `sgl` granularity performs best than all the others when it is combined with the `tau` gap, as shown in Figure 8(g). As far as gaps are concerned, the top group is represented by `apc` and `tau` while `med` and `md` are the weights in the top group. As before, the fact that top approaches mostly use the `msd` weight is due to the interaction between components; indeed, as shown in Figure 8(i), the performance of `msd` is boosted by the `apc` and `tau` gaps which, at the same time, lower the performance of `md` and `med`. The importance of the interaction effects is supported also by the effect sizes reported in Table 3, which shows that the Granularity*Gap and Gap*Weight interactions have size one order of magnitude greater than the Granularity*Weight interaction.

Table 5. Result Summary for AP Correlation and RMSE.

	AP Correlation	**RMSE**

Approach	`sgl_tau_msd`
	`sgl_apc_msd`
	`tpc_apc_msd`
	`sgl_tau_md`
		`sgl_rmse_med`
		`tpc_rmse_med`
		`tpc_fro_med`
		`sgl_fro_med`
		`sgl_kld_md`

Granularity	`sgl` `tpc`	`sgl`
Gap	`apc` `tau`	`kld` `fro`
Weight	`med` `md`	`med`

Respectively, for RMSE, the best granularity is `sgl` which is also the most frequent in the top group of approaches. The best gaps are `kld` and `fro` while `med` is the top weight. As discussed above, interaction plays an important role also in this case: indeed, the top approaches are `sgl_rmse_med` and `tpc_rmse_med` because of the strong positive interaction between `med` and `rmse`, shown in Figure 9(i) and supported by the medium effect size of the Gap*Weight interaction, which is two order of magnitude greater than all the other interactions effects, as reported in Table 4.

The proposed unsupervised estimators are, in a sense, mono-feature, since they operate on each performance measure separately. However, the experimentation has shown that the performance of the proposed estimators varies from measure to measure, e.g. ERR is more challenging than AP in terms of AP correlation. Therefore, as part of 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; in this way, we plan to exploit the differences among various evaluation measures to obtain more robust estimators.

Another direction for future work will concern the development of supervised estimators, i.e. estimators that leverage a gold standard instead of random assessors for determining the accuracy of a crowd assessor. Also in this case, we can envision both mono-feature and multi-feature estimators, in the sense explained above.

Finally, it would be interesting to experiment what happens in the case of graded-relevance judgments. Not only is this a natural setting for nDCG and ERR, it also opens up to other evaluation measures such as *Graded Average Precision (GAP)* and its extensions [16, 56] or effort-based measures such as Twist [20].

ACKNOWLEDGMENTS

The authors are grateful to Gianmaria Silvello for the fruitful discussions and the continuous feedback. The authors wish to warmly thank Matt Lease for having provided the TREC 21, 2012, Crowdsourcing dataset used in the paper as well as going into the details about it. We also sincerely thank the associate editor and the anonymous reviewers whose suggestions helped in improving this paper.

REFERENCES

[1] O. Alonso. 2013. Implementing Crowdsourcing-based Relevance Experimentation: an Industrial Perspective. *Information Retrieval* 16, 2 (April 2013), 101–120.

[2] O. Alonso and S. Mizzaro. 2012. Using Crowdsourcing for TREC Relevance Assessment. *Information Processing & Management* 48, 6 (November 2012), 1053–1066.

[3] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz. 2008. Relevance Assessment: Are Judges Exchangeable and Does it Matter?. In *Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, T.-S. Chua, M.-K. Leong, D. W. Oard, and F. Sebastiani (Eds.). ACM Press, New York, USA, 667–674.

[4] M. Bashir, J. Anderton, J. Wu, M. Ekstrand-Abueg, P. B. Golbus, V. Pavlu, and J. A. Aslam. 2013. Northeastern University Runs at the TREC12 Crowdsourcing Track. In *The Twenty-First Text REtrieval Conference Proceedings (TREC 2012)*, E. M. Voorhees and L. P. Buckland (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-298, Washington, USA.

[5] R. Blanco, H. Halpin, D. M. Herzig, P. Mika, J. Pound, and H. S. Thompson. 2011. Repeatable and Reliable Search System Evaluation using Crowdsourcing. In *Proc. 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011)*, W.-Y. Ma, J.-Y. Nie, R. Baeza-Yaetes, T.-S. Chua, and W. B. Croft (Eds.). ACM Press, New York, USA, 923–932.

[6] C. Buckley and E. M. Voorhees. 2005. Retrieval System Evaluation. In *TREC. Experiment and Evaluation in Information Retrieval*, D. K. Harman and E. M. Voorhees (Eds.). MIT Press, Cambridge (MA), USA, 53–78.

[7] R. Burgin. 1992. Variations in Relevance Judgments and the Evaluation of Retrieval Performance. *Information Processing & Management* 28, 5 (September-October 1992), 619–627.

[8] K. P. Burnham and D. R. Anderson. 2002. *Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach* (2nd ed.). Springer-Verlag, Heidelberg, Germany.

[9] B. A. Carterette and I. Soboroff. 2010. The Effect of Assessor Errors on IR System Evaluation. In *Proc. 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010)*, F. Crestani, S. Marchand-Maillet, E. N. Efthimiadis, and J. Savoy (Eds.). ACM Press, New York, USA, 539–546.

[10] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *Proc. 18th International Conference on Information and Knowledge Management (CIKM 2009)*, D. W.-L. Cheung, I.-Y. Song, W. W. Chu, X. Hu, and J. J. Lin (Eds.). ACM Press, New York, USA, 621–630.

[11] P. Clough, M. Sanderson, J. Tang, T. Gollins, and A. Warner. 2013. Examining the Limits of Crowdsourcing for Relevance Assessment. *IEEE Internet Computing* 17, 4 (July 2013), 32–38.

[12] G. Cormack and T. Lynam. 2005. TREC 2005 Spam Track Overview. In *The Fourteenth Text REtrieval Conference Proceedings (TREC 2005)*, E. M. Voorhees and L. P. Buckland (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-266, Washington, USA.

[13] A. P. Dawid and A. M. Skene. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28, 1 (March 1979), 20–28.

[14] C. G. Eickhoff, C. Harris and A. P. de Vries. 2012. Quality through Flow and Immersion: Gamifying Crowdsourced Relevance Assessments. In *Proc. 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2012)*, W. Hersh, J. Callan, Y. Maarek, and M. Sanderson (Eds.). ACM Press, New York, USA, 871–880.

[15] T. Fawcett. 2006. An Introduction to ROC Analysis. *Pattern Recognition Letters* 27, 8 (June 2006), 861–874.

[16] M. Ferrante, N. Ferro, and M. Maistro. 2014. Rethinking How to Extend Average Precision to Graded Relevance. In *Information Access Evaluation – Multilinguality, Multimodality, and Interaction. Proceedings of the Fifth International Conference of the CLEF Initiative (CLEF 2014)*, E. Kanoulas, M. Lupu, P. Clough, M. Sanderson, M. Hall, A. Hanbury, and E. Toms (Eds.). Lecture Notes in Computer Science (LNCS) 8685, Springer, Heidelberg, Germany, 19–30.

[17] M. Ferrante, N. Ferro, and M. Maistro. 2015. Towards a Formal Framework for Utility-oriented Measurements of Retrieval Effectiveness. In *Proc. 1st ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR 2015)*, J. Allan, W. B. Croft, A. P. de Vries, C. Zhai, N. Fuhr, and Y. Zhang (Eds.). ACM Press, New York, USA, 21–30.

[18] N. Ferro. 2017. Reproducibility Challenges in Information Retrieval Evaluation. *ACM Journal of Data and Information Quality (JDIQ)* 8, 2 (February 2017), 8:1–8:4.

[19] N. Ferro, N. Fuhr, K. Järvelin, N. Kando, M. Lippold, and J. Zobel. 2016. Increasing Reproducibility in IR: Findings from the Dagstuhl Seminar on "Reproducibility of Data-Oriented Experiments in e-Science". *SIGIR Forum* 50, 1 (June 2016), 68–82.

[20] N. Ferro, G. Silvello, H. Keskustalo, A. Pirkola, and K. Järvelin. 2016. The Twist Measure for IR Evaluation: Taking User's Effort Into Account. *Journal of the American Society for Information Science and Technology (JASIST)* 67, 3 (2016), 620–648.

[21] G. H. Golub and C. F. Van Loan. 2012. *Matrix Computations* (4th ed.). Johns Hopkins University Press, USA.

[22] C. Grady and M. Lease. 2010. Crowdsourcing Document Relevance Assessment with Mechanical Turk. In *Proc. NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, C. Callison-Burch and M. Dredze (Eds.). The Association for Computational Linguistics (ACL), USA, 172–179.

[23] M. Halvey, R. Villa, and P. Clough. 2014. SIGIR 2014 Workshop on Gathering Efficient Assessments of Relevance (GEAR). In *Proc. 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2014)*, S. Geva, A. Trotman, P. Bruza, C. L. A. Clarke, and K. Järvelin (Eds.). ACM Press, New York, USA, 1293.

[24] C. Harris and P. Srinivasan. 2013. Using Hybrid Methods for Relevance Assessment in TREC Crowd'12. In *The Twenty-First Text REtrieval Conference Proceedings (TREC 2012)*, E. M. Voorhees and L. P. Buckland (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-298, Washington, USA.

[25] Y. Hochberg and A. C. Tamhane. 1987. *Multiple Comparison Procedures.* John Wiley & Sons, USA.

[26] M. Hosseini, I. J. Cox, N. Milić-Frayling, G. Kazai, and V. Vinay. 2012. On Aggregating Labels from Multiple Crowd Workers to Infer Relevance of Documents. In *Advances in Information Retrieval. Proc. 32nd European Conference on IR Research (ECIR 2012)*, R. Baeza-Yaetes, A. P. de Vries, H. Zaragoza, B. B. Cambazoglu, V. Murdock, R. Lempel, and F. Silvestri (Eds.). Lecture Notes in Computer Science (LNCS) 7224, Springer, Heidelberg, Germany, 182–194.

[27] P. G. Ipeirotis and E. Gabrilovich. 2014. Quizz: Targeted Crowdsourcing with a Billion (Potential) Users. In *Proc. 23rd International Conference on World Wide Web (WWW 2014)*, C.-W. Chung, A. Broder, K. Shim, and T. Suel (Eds.). ACM Press, New York, USA, 143–154.

[28] K. Järvelin and J. Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (October 2002), 422–446.

[29] T. Josephy, M. Lease, P. Paritosh, M. Krause, M. Georgescu, M. Tjalve, and D. Braga. 2014. Workshops Held at the First AAAI Conference on Human Computation and Crowdsourcing: A Report. *AI Magazine* 35, 2 (2014), 75–78.

[30] H. J. Jung and M. Lease. 2015. A Discriminative Approach to Predicting Assessor Accuracy. In *Advances in Information Retrieval. Proc. 37th European Conference on IR Research (ECIR 2015)*, N. Fuhr, A. Rauber, G. Kazai, and A. Hanbury (Eds.). Lecture Notes in Computer Science (LNCS) 9022, Springer, Heidelberg, Germany.

[31] G. Kazai. 2011. In Search of Quality in Crowdsourcing for Search Engine Evaluation. In *Advances in Information Retrieval. Proc. 33rd European Conference on IR Research (ECIR 2011)*, P. Clough, C. Foley, C. Gurrin, G. J. F. Jones, W. Kraaij, H. Lee, and V. Mudoch (Eds.). Lecture Notes in Computer Science (LNCS) 6611, Springer, Heidelberg, Germany, 165–176.

[32] G. Kazai, N. Craswell, E. Yilmaz, and S. S. M. Tahaghoghi. 2012. An Analysis of Systematic Judging Errors in Information Retrieval. In *Proc. 21st International Conference on Information and Knowledge Management (CIKM 2012)*, X. Chen, G. Lebanon, H. Wang, and M. J. Zaki (Eds.). ACM Press, New York, USA, 105–114.

[33] G. Kazai, J. Kamps, M. Koolen, and N. Milić-Frayling. 2011. Crowdsourcing for Book Search Evaluation: Impact of HIT Design on Comparative System Ranking. In *Proc. 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011)*, W.-Y. Ma, J.-Y. Nie, R. Baeza-Yaetes, T.-S. Chua, and W. B. Croft (Eds.). ACM Press, New York, USA, 205–214.

[34] G. Kazai, J. Kamps, and N. Milić-Frayling. 2011. The Face of Quality in Crowdsourcing Relevance Labels: Demographics, Personality and Labeling Accuracy. In *Proc. 20th International Conference on Information and Knowledge Management (CIKM 2011)*, I. Ounis, I. Ruthven, B. Berendt, A. P. de Vries, and F. Wenfei (Eds.). ACM Press, New York, USA, 2583–2586.

[35] G. Kazai, J. Kamps, and N. Milic-Frayling. 2013. An Analysis of Human Factors and Label Accuracy in Crowdsourcing Relevance Judgments. *Information Retrieval* 16, 2 (April 2013), 138–178.

[36] G. Kazai, E. Yilmaz, N. Craswell, and S. S. M. Tahaghoghi. 2013. User Intent and Assessor Disagreement in Web Search Evaluation. In *Proc. 22h International Conference on Information and Knowledge Management (CIKM 2013)*, A. Iyengar, Q. He, J. Pei, R. Rastogi, and W. Nejdl (Eds.). ACM Press, New York, USA, 699–708.

[37] M. G. Kendall. 1948. *Rank Correlation Methods.* Griffin, Oxford, England.

[38] J. F. Kenney and E. S. Keeping. 1954. *Mathematics of Statistics – Part One* (3rd ed.). D. Van Nostrand Company, Princeton, USA.

[39] I. King, K.-T. Chen, O. Alonso, and M. Larson. 2016. Special Issue: Crowd in Intelligent Systems. *ACM Transactions on Intelligent Systems and Technology (TIST)* 7, 4 (May 2016).

[40] K. A. Kinney, S. B. Huffman, and J. Zhai. 2008. How Evaluator Domain Expertise Affects Search Result Relevance Judgments. In *Proc. 17th International Conference on Information and Knowledge Management (CIKM 2008)*, J. G. Shanahan, S. Amer-Yahia, I. Manolescu, Y. Zhang, D. A. Evans, A. Kolcz, K.-S. Choi, and A. Chowdhury (Eds.). ACM Press, New York, USA, 591–598.

[41] S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics* 22, 1 (March 1951), 79–86.

[42] E. Law, P. N. Bennett, and E. Horvitz. 2011. The Effects of Choice in Routing Relevance Judgments. In *Proc. 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011)*, W.-Y. Ma, J.-Y. Nie, R. Baeza-Yaetes, T.-S. Chua, and W. B. Croft (Eds.). ACM Press, New York, USA, 1127–1128.

[43] M. Lease and E. Yilmaz. 2013. Crowdsourcing for Information Retrieval: Introduction to the Special Issue. *Information Retrieval* 16, 2 (April 2013), 91–100.

[44] M. E. Lesk and G. Salton. 1968. Relevance Assessments and Retrieval System Evaluation. *Information Storage and Retrieval* 4, 4 (December 1968), 343–359.

[45] L. Li and M. D. Smucker. 2014. Tolerance of Effectiveness Measures to Relevance Judging Errors. In *Advances in Information Retrieval. Proc. 36th European Conference on IR Research (ECIR 2014)*, M. de Rijke, T. Kenter, A. P. de Vries, C. Zhai, F. de Jong, K. Radinsky, and K. Hofmann (Eds.). Lecture Notes in Computer Science (LNCS) 8416, Springer, Heidelberg, Germany, 148–159.

[46] B. Loni, M. Larson, A. Bozzon, and L. Gottlieb. 2013. Crowdsourcing for Social Multimedia at MediaEval 2013: Challenges, Data set, and Evaluation. In *Working Notes Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, M. Larson, X. Anguera, T. Reuter, G. J. F. Jones, B. Ionescu, M. Schedl, T. Piatrik, C. Hauff, and M. Soleymani (Eds.). CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-1043/.

[47] A. Marcus and A. Parameswaran. 2015. Crowdsourced Data Management: Industry and Academic Perspectives. *Foundations and Trends in Databases (FnTDB)* 6, 1–2 (December 2015), 1–161.

[48] S. Maxwell and H. D. Delaney. 2004. *Designing Experiments and Analyzing Data. A Model Comparison Perspective* (2nd ed.). Lawrence Erlbaum Associates, Mahwah (NJ), USA.

[49] Y. Moshfeghi, H. F. Huertas Rosero, and J. M. Jose. 2016. A Game-Theory Approach for Effective Crowdsource-Based Relevance Assessment. *ACM Transactions on Intelligent Systems and Technology (TIST)* 7, 4 (May 2016), 55:1–55:XXX.

[50] K. R. Murphy, B. Myors, and A. Wolach. 2014. *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests* (4th ed.). Routledge, Taylor & Francis Group, UK.
[51] S. Olejnik and J. Algina. 2003. Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs. *Psychological Methods* 8, 4 (December 2003), 434–447.
[52] I. Pillai, I. Fumera, and F. Roli. 2013. Multi-label Classification with a Reject Option. *Pattern Recognition* 46, 8 (August 2013), 2256–2266.
[53] V. C. Raykar and S. Yu. 2012. Eliminating Spammers and Ranking Annotators for Crowdsourced Labeling Tasks. *Journal of Machine Learning Research* 13 (February 2012), 491–518.
[54] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. Hermosillo Valadez, L. Bogoni, and L. Moy. 2009. Supervised Learning from Multiple Experts: Whom to Trust when Everyone Lies a Bit. In *Proc. 26th Annual International Conference on Machine Learning (ICML 2009)*, L. Bottou and M. Littman (Eds.). ACM Press, New York, USA, 889–896.
[55] V. C. Raykar, L. H. Zhao, G. Hermosillo Valadez, C. Florin, L. Bogoni, and L. Moy. 2010. Learning From Crowds. *Journal of Machine Learning Research* 11 (April 2010), 1297–1322.
[56] S. E. Robertson, E. Kanoulas, and E. Yilmaz. 2010. Extending Average Precision to Graded Relevance Judgments. In *Proc. 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010)*, F. Crestani, S. Marchand-Maillet, E. N. Efthimiadis, and J. Savoy (Eds.). ACM Press, New York, USA, 603–610.
[57] A. Rutherford. 2011. *ANOVA and ANCOVA. A GLM Approach* (2nd ed.). John Wiley & Sons, New York, USA.
[58] I. Ruthven. 2014. Relevance Behaviour in TREC. *Journal of Documentation* 70, 6 (2014), 1098–1117.
[59] T. Sakai. 2014. Statistical Reform in Information Retrieval? *SIGIR Forum* 48, 1 (June 2014), 3–12.
[60] M. Sanderson and J. Zobel. 2005. Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. In *Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, R. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, and J. Tait (Eds.). ACM Press, New York, USA, 162–169.
[61] M. D. Smucker and C. P. Jethani. 2011. The crowd vs. the Lab: A comparison of Crowd-Sourced and University Laboratory Participant Behavior. In *In Proceedings of the SIGIR 2011 Workshop on Crowdsourcing for Information Retrieval.*
[62] M. D. Smucker and C. P. Jethani. 2011. Measuring Assessor Accuracy: A Comparison of Nist Assessors and User Study Participants. In *Proc. 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011)*, W.-Y. Ma, J.-Y. Nie, R. Baeza-Yaetes, T.-S. Chua, and W. B. Croft (Eds.). ACM Press, New York, USA, 1231–1232.
[63] M. D. Smucker, G. Kazai, and M. Lease. 2013. Overview of the TREC 2012 Crowdsourcing Track. In *The Twenty-First Text REtrieval Conference Proceedings (TREC 2012)*, E. M. Voorhees and L. P. Buckland (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-298, Washington, USA.
[64] M. D. Smucker, G. Kazai, and M. Lease. 2014. Overview of the TREC 2013 Crowdsourcing Track. In *The Twenty-Second Text REtrieval Conference Proceedings (TREC 2013)*, E. M. Voorhees (Ed.). National Institute of Standards and Technology (NIST), Special Publication 500-302, Washington, USA.
[65] I. Soboroff, C. Nicholas, and P. Cahan. 2001. Ranking Retrieval Systems without Relevance Judgments. In *Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, D. H. Kraft, W. B. Croft, D. J. Harper, and J. Zobel (Eds.). ACM Press, New York, USA, 66–73.
[66] S. S. Stevens. 1946. On the Theory of Scales of Measurement. *Science, New Series* 103, 2684 (June 1946), 677–680.
[67] E. M. Voorhees. 1998. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998)*, W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel (Eds.). ACM Press, New York, USA, 315–323.
[68] E. M. Voorhees. 2000. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. *Information Processing & Management* 36, 5 (September 2000), 697–716.
[69] E. M. Voorhees. 2004. Overview of the TREC 2004 Robust Track. In *The Thirteenth Text REtrieval Conference Proceedings (TREC 2004)*, E. M. Voorhees and L. P. Buckland (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-261, Washington, USA.
[70] E. M. Voorhees and D. K. Harman. 1999. Overview of the Eight Text REtrieval Conference (TREC-8). In *The Eighth Text REtrieval Conference (TREC-8)*, E. M. Voorhees and D. K. Harman (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-246, Washington, USA, 1–24.
[71] J. B. P. Vuurens and A. P. de Vries. 2012. Obtaining High-Quality Relevance Judgments Using Crowdsourcing. *IEEE Internet Computing* 16, 5 (September-October 2012), 20–27.
[72] S. Wakeling, M. Halvey, R. Villa, and L. Hasler. 2016. A Comparison of Primary and Secondary Relevance Judgements for Real-Life Topics. In *Proc. 1st ACM on Conference on Human Information Interaction and Retrieval (CHIIR 2016)*, D. Kelly, R. Capra, N. Belkin, J. Teevan, and P. Vakkari (Eds.). ACM Press, New York, USA, 173–182.
[73] M. P. Wand and M. C. Jones. 1995. *Kernel Smoothing.* Chapman and Hall/CRC, USA.

[74] W. Webber, P. Chandar, and B. A. Carterette. 2012. Alternative Assessor Disagreement and Retrieval Depth. In *Proc. 21st International Conference on Information and Knowledge Management (CIKM 2012)*, X. Chen, G. Lebanon, H. Wang, and M. J. Zaki (Eds.). ACM Press, New York, USA, 125–134.

[75] W. Webber and J. Pickens. 2013. Assessor Disagreement and Text Classifier Accuracy. In *Proc. 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2013)*, G. J. F. Jones, P. Sheridan, D. Kelly, M. de Rijke, and T. Sakai (Eds.). ACM Press, New York, USA, 929–932.

[76] C. F. J. Wu. 1983. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics* 11, 1 (March 1983), 95–103.

[77] K. Yadati, P. S. N. Shakthinathan, C. Ayyanathan, and M. Larson. 2014. Crowdsorting Timed Comments about Music: Foundations for a New Crowdsourcing Task. In *Working Notes Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop*, M. Larson, B. Ionescu, X. Anguera, M. Eskevich, P. Korshunov, M. Schedl, M. Soleymani, P. Petkos, R. Sutcliffe, J. Choi, and G. J. F. Jones (Eds.). CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-1263/.

[78] E. Yilmaz, J. A. Aslam, and S. E. Robertson. 2008. A New Rank Correlation Coefficient for Information Retrieval. In *Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, T.-S. Chua, M.-K. Leong, D. W. Oard, and F. Sebastiani (Eds.). ACM Press, New York, USA, 587–594.

LIST OF ACRONYMS

ANOVA ANalysis Of VAriance
AP Average Precision
AUC Area Under the ROC Curve
AWARE Assessor-driven Weighted Averages for Retrieval Evaluation
DCG Discounted Cumulated Gain
DF Degrees of Freedom
EM Expectation Maximization
ERR Expected Reciprocal Rank
GAP Graded Average Precision
GLMM General Linear Mixed Model
HIT Human Intelligence Task
HSD Honestly Significant Difference
i.i.d. independent and identically distributed
IR Information Retrieval
KDE Kernel Density Estimation
KLD Kullback-Leibler Divergence
LAM Logistic Average Misclassification
MAP Mean Average Precision
MATTERS MATlab Toolkit for Evaluation of information Retrieval Systems
MS Mean Squares
MV Majority Vote
nDCG Normalized Discounted Cumulated Gain
PDF Probability Density Function
RMSE Root Mean Square Error
SOA Strength of Association
SS Sum of Squares
TRAT Text Relevance Assessing Task
TREC Text REtrieval Conference

Received 11 October 2016; revised 23 March 2017; revised 17 May 2017; accepted 14 June 2017