Considering Assessor Agreement in IR Evaluation

Eddy Maddalena
University of Southampton
Southampton, UK
e.maddalena@soton.ac.uk

Kevin Roitero
University of Udine
Udine, Italy
roitero.kevin@spes.uniud.it

Gianluca Demartini
University of Queensland
Brisbane, Australia
g.demartini@uq.edu.au

Stefano Mizzaro
University of Udine
Udine, Italy
mizzaro@uniud.it

ABSTRACT

The agreement between relevance assessors is an important but understudied topic in the Information Retrieval literature because of the limited data available about documents assessed by multiple judges. This issue has gained even more importance recently in light of crowdsourced relevance judgments, where it is customary to gather many relevance labels for each topic-document pair. In a crowdsourcing setting, agreement is often even used as a proxy for quality, although without any systematic verification of the conjecture that higher agreement corresponds to higher quality.

In this paper we address this issue and we study in particular: the effect of topic on assessor agreement; the relationship between assessor agreement and judgment quality; the effect of agreement on ranking systems according to their effectiveness; and the definition of an agreement-aware effectiveness metric that does not discard information about multiple judgments for the same document as it typically happens in a crowdsourcing setting.

CCS CONCEPTS

• **Information systems** → **Relevance assessment**;

KEYWORDS

TREC, evaluation, test collections, agreement, disagreement

ACM Reference format:
Eddy Maddalena, Kevin Roitero, Gianluca Demartini, and Stefano Mizzaro. 2017. Considering Assessor Agreement in IR Evaluation. In *Proceedings of ICTIR '17, Amsterdam, Netherlands, October 1–4, 2017,* 8 pages.
https://doi.org/10.1145/3121050.3121060

1 INTRODUCTION

Gathering relevance assessments is a crucial activity in Information Retrieval (IR) effectiveness evaluation. However, as it is well known, relevance is often uncertain: a query can be ambiguous, a term can have multiple meanings, a document can be considered from different viewpoints, an information need might be unclear, the "need behind a query" might be unknown, relevance assessors might work in non-ideal conditions, or be inadequate or even malicious, and so on. This is even more prominent when crowdsourcing is used to gather document relevance labels, as it has often been done in the last few years [1]. When crowdsourcing relevance assessments, usually each document is redundantly assessed by several crowd workers, who judge the relevance of the document to a specific topic. Generally speaking, all the relevance judgments received for a document by different workers are aggregated to compute the final relevance score: by doing so, only the final aggregated scores are subsequently considered to compute IR evaluation metrics, and all the information about the distributions of the judgments scores before the aggregation, and assessors's agreement, are lost. This could be valuable information, though: high disagreement between workers could suggest the presence of unreliable judgments, or a high intrinsic document or topic ambiguity.

Previous work has shown that while assessor agreement is typically low, the effect on the final IR system ranking generated using IR test collections is limited [7, 19]. While some work on measuring assessor disagreement and its effects on IR evaluation has been performed, it is somehow surprising that a comprehensive understanding of agreement in relevance assessment is still missing also given the recent rise in popularity of crowdsourcing.

In this paper we address this issue and propose novel evaluation metrics that preserve agreement information rather than just aggregating relevance labels collected from different human assessors for the same topic-document pair. More specifically, we focus on the following research questions, each one divided into sub-questions.

RQ1. **Agreement, topic, and relevance.** Is there a relationship between assessor agreement, topic and relevance level?
- **RQ1a**. Is agreement level different on different topics?
- **RQ1b**. Is agreement related to topic ease/difficulty?
- **RQ1c**. Is agreement different on different relevance levels?

RQ2. **Agreement and system evaluation.** Is there a relation between assessor agreement and effectiveness evaluation?
- **RQ2a**. What is the effect of agreement on measures of retrieval effectiveness?
- **RQ2b**. What is the effect of agreement on systems ranking? Are system ranks more affected when removing high agreement or low agreement topics?

RQ3. **Agreement-Aware metric**. Is it possible to take agreement into account and define an agreement-aware IR evaluation metric? What is the effect on evaluation and system rankings of such a metric?
- **RQ3a**. How to define an agreement-aware metric?

ICTIR '17, October 1–4, 2017, Amsterdam, Netherlands

ACM ISBN 978-1-4503-4490-6/17/10…$15.00
https://doi.org/10.1145/3121050.3121060

RQ3b. What is the effect of an agreement-aware metric on effectiveness evaluation and system rankings?

This paper is structured as follows. Section 2 surveys the related work in the area of assessor agreement. Section 3 presents the datasets used in our experiments, the agreement measures that we rely on, and highlights some issues. Each of Sections 4, 5, and 6 addresses one of the three research question RQ1, RQ2, and RQ3. Section 7 concludes the paper.

2 RELATED WORK

Inter-annotator agreement (IAA) is a measure of labeling quality used across several fields. For example, when creating linguistics collections IAA measures are used to validate the quality of the collection [10]. In IR evaluation the classic approach to create test collections is to rely on one human assessor to judge the relevance of a retrieved result with respect to the search topic. To preserve judgment consistency, in traditional IR evaluation one topic is entirely judged by one single assessor. Despite this common approach, previous work has looked at the expected level of agreement among assessors judging the same set of topic-document pairs.

Agreement in TREC Collections. Early research work has looked at agreement levels among different groups of human assessors (e.g., experts and non-experts) for TREC collections. For example, in [2, 15] authors looked at how the same evaluation collection created by different groups of assessors can lead to different evaluation results due to low IAA. An early measure used to look at disagreement was the overlap of relevant documents identified by different assessors [19]. In this and earlier work by Cleverdon [7], authors showed that, while IAA can be low, it has a limited impact on the evaluation results measured by the final IR system ranking correlation.

Causes of Disagreement. An analysis of what are the main causes for low IAA in relevance judgment tasks showed that document readability and length have a significant impact [6]. Another study which looked at the causes of judgment errors and disagreement [17] observed an existing 'assessor inertia' where a certain judgment is unconsciously affected by the previous ones done by the human assessor. Another cause of disagreement is the more or less detailed assessment guidelines [20]. Related to this is a recent study on how the order of documents presented to assessors impacts IAA levels, which showed that ordering documents by decreasing levels of relevance leads to higher levels of disagreement [8].

Crowdsourced Relevance Judgments. With the rising of the use of crowdsourcing as a means to collect relevance judgments for search results, a new approach to create IR evaluation collections is to collect *multiple* relevance labels for the same topic-document pair and to then aggregate such labels into a final relevance judgment [1]. This is typically done to improve the quality of the judgments by removing possible noise introduced by randomly assigned labels or adversaries. Previous work on crowdsourced relevance judgments has shown how collections built with such an approach lead to reliable results, and are repeatable if created again [3].

Several ways to aggregate labels collected from the crowd have been proposed, starting with the simple majority vote [12] up to complex weighted aggregation models that combine labels together by looking at common patterns at the crowd level [18]. However, performing such operation of label aggregation leads to a loss of information about the level of agreement among the different human assessors who looked at the same topic-document pair. On the other hand, we claim it is important to preserve such information in the IR evaluation process and to incorporate assessor agreement levels into the evaluation by, for example, giving less importance to judgments where lower agreement was observed.

Agreement-aware IR evaluation. Related to our work is the model presented in [9] where the hypothesis used is that non-relevant results judged with high disagreement should be retrieved higher than those with high agreement. Authors proposed an evaluation framework that integrates information about disagreement. Based on such previous work, in our paper we present a comprehensive study of assessor agreement and of the effects that high/low agreement rates have on IR system evaluation: We first perform an extensive analysis of assessor disagreement across different test collections and measure the effects on IR system evaluation. We then propose IR evaluation measures that integrate assessor agreement levels by looking at the distribution of relevance labels and compare such novel agreement-aware metrics with traditional ones in terms of the generated IR system ranking.

3 EXPERIMENTAL METHODOLOGY

3.1 Datasets

To study agreement, we need datasets featuring several relevance assessments for the same ⟨topic, document⟩ pair. We use two such datasets in our experiments, as detailed below.

3.1.1 RF: Relevance Feedback. The first dataset, denoted in the following by *RF* for Relevance Feedback, is described by Demeester et al. [9]. For the Relevance Feedback track 2010 [4] some of the ClueWeb documents used in TREC 2009 Million Query Track [5] have been re-assessed by up to 11 assessors each. Since we are interested in assessor agreement, we select the documents that were judged by at least 5 assessors; when more than 5 assessments are available, we select the first five only. We thus obtain around 15,000 usable re-assessments.

We also transform the original scale (that included −2 for "broken link" and −1 for "unknown (no gold label)") into 0, 1, 2 (by collapsing −1 and −2 into 0). For all these documents the original NIST assessments are available; we refer to them as *Gold* assessments.

The RF dataset contains re-assessments on 100 topics, however in some cases we cannot use all of them. Of the 100 topics, 11 do not have re-judged documents for all the three Gold relevance levels. In other terms, it is only for 89 topic that we can find re-assessed documents for all Gold relevance levels of 0, 1, and 2. Since in some cases we are interested to measure, and compare, the agreement of the documents at a given relevance level, sometimes we use those 89 topics only. Furthermore, some of the original summary files in the TREC 2009 Million Query Track data do not contain the NDCG values. This is not a problem when measuring agreement; however, when comparing agreement and effectiveness or ease (as we will sometimes do in the following) we need to filter some data. This leaves us with 81 topics. Finally, when combining all the above filters, we are left with 71 topics to work with.

3.1.2 ME: Magnitude Estimation. The second dataset, denoted *ME* for Magnitude Estimation, is the datasets used for the experiments by Maddalena et al. [14]. They re-assessed 18 TREC-8 topics

by crowdsourcing. For each topic, the top ten retrieved documents by the systems that participated in TREC-8 were pooled and re-assessed, each one by 10 Crowdflower workers. Several quality checks were used to ensure that the collected data were reliable. One particularity of this dataset is that relevance scores were not on a graded, or category, scale as it is usually done; instead, magnitude estimation was used. With this psychophysical technique workers could express the relevance of a document to a topic using any number in the $]0, +\infty[$ range (note that 0 was not admissible).

For those documents Gold from NIST assessors (the original TREC-8 assessors) are available.

3.2 Measures: Effectiveness, Ease, Agreement

In this paper we use mainly NDCG as a measure of system effectiveness. For each topic we define *topic ease* as the average effectiveness on that topic of the systems participating in the evaluation exercise (for RF we use the Million Query 2009 runs; for ME we use the TREC-8 runs). On terminology, we notice that we prefer to use "topic ease" instead of the probably most common "topic difficulty" for symmetry with system effectiveness: higher e would mean both higher topic ease and higher average system effectiveness.

Several agreement measures have been defined in the past: variance, standard deviation, entropy, ICC, Cohen's kappa, Fleiss's kappa, etc. Throughout this paper we use Krippendorff's α [13], a standardized measure of agreement that adapts to items having different numbers of evaluators, different scales, missing values. α assumes values ranging from -1 (complete disagreement) through 0 random (agreement obtained by random evaluations) to 1 (complete agreement). We use two versions of α:

1. α interval/ordinal for RF data. The scale is the usual graded scale 0, 1, 2 and using the interval version means that we weight more the difference between 0 and 2 than the difference between 0 and 1. Results are equivalent to using α ordinal.
2. α interval for ME data. Normalized ME scores are log-normally distributed [14]. We take the logarithm the scores, to obtain a normal distribution and avoid precision issues. Since ME scores are on a ratio scale [14], by taking the logarithm we obtain an interval scale; therefore we adopt the interval version of α.

α (and other agreement measures) works on a set of assessments and computes the overall agreement for that set. Using it to compute the agreement on a single ⟨topic, document⟩ pair would be absolutely not standard and open to criticisms. Therefore our experiments are topic-based. As we will see shortly, it is indeed the case that there is a topic-related notion of agreement, and it makes sense to speak of *agreement for a topic*: agreement is somehow intrinsic in a topic, and there are high- and low-agreement topics.

Table 1 summarizes the measures for the two datasets. For each topic we have its ease e, its agreement computed using all the assessed documents (α^A, A stands for "All"), and its agreement computed using a subset of the documents having a specific relevance value in the Gold. For RF we have Gold on a three level scale (0, 1, 2), so we can compute α^0, α^1, and α^2; for ME we have Gold on a binary scale (0, 1) and therefore only α^0 and α^1 can be computed. We will also refer to α^{01}, α^{02}, and α^{12} as the α values computed on the basis of documents having two Gold relevance values.

Table 1: Datasets. The α superscripts indicate the set of documents used to compute α, i.e., A=All, or having Gold = 0, Gold = 1, or Gold = 2 (the latter only for the RF dataset).

Topic	Ease	Agreement			
	(e)	(α^A)	(α^0)	(α^1)	(α^2)
t_1	e_1	α_1^A	α_1^0	α_1^1	α_1^2
t_2	e_2	α_2^A	α_2^0	α_2^1	α_2^2
…					
t_n	e_n	α_n^A	α_n^0	α_n^1	α_n^2

3.3 Justifications and Issues

We provide a justification for the choice of those two datasets, and highlight some issues for each of them. Of course we need datasets featuring several relevance assessments for the same ⟨topic, document⟩ pair. In this respect, both datasets are adequate.

For some of our experiments on RF data we had to "interpolate" some effectiveness values by filling NDCG values for specific ⟨topic, system⟩ pairs. 26% of such values we missing. We created those values by taking the mean of all the other available NDCG values for that topic and that system. Intuitively this interpolation makes sense. In practice it does not affect the results significantly: when evaluating system effectiveness or topic ease with the "interpolated" values or with the original values, the results is in practice unchanged (we obtain correlations around 0.98).

RF has three levels relevance judgments. When the relevance scale used is bounded, as it is in this case, some subtle issues arise with agreement metrics. When focusing on one end of the scale (maximum relevance or complete irrelevance), a decrease in agreement implies a decrease in judgment quality (since agreement can be decreased only by "going away in one direction" from the true value). This is not true for the central part of the scale, where, however, another phenomenon arises: for the central value of the scale, maximum disagreement means also maximum quality of the aggregated judgment (since the maximum disagreement is obtained "going away in two symmetrical directions" from the true value). In an attempt to overcome these problems, we also use ME data. ME does not have the same limitations, since it is based on an unbounded scale $]0, +\infty[$; however a somehow non-standard scale is used, and therefore another dataset is needed to find confirmations and avoid results depending on effects of the used scale only.

We additionally report results obtained by grouping together topics with similar agreement levels to smooth noise in the data: after ordering topics by their α score, we create groups of topics of the same size (referred to as topic binning).

4 RQ1: AGREEMENT, TOPIC, RELEVANCE

RQ1 is aimed at understanding if there is a relationship between assessor agreement, topic, and relevance level. Each sub-question is addressed in the following subsections.

4.1 RQ1a: Agreement Over Topics

To answer RQ1a, we analyze the variation of Krippendorff's α over topics. Figure 1 shows that there is quite some variation of α values over the 100 topics of the RF dataset: the range is approximately $[-0.1, 0.3]$. On the ME dataset the variation is more limited, approximately in the $[0.25, 0.45]$ range (this can be seen from the red top series in Figure 3, bottom chart—the figure will be described in

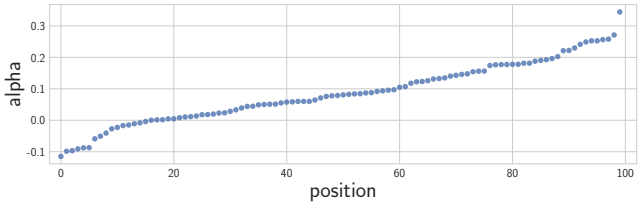
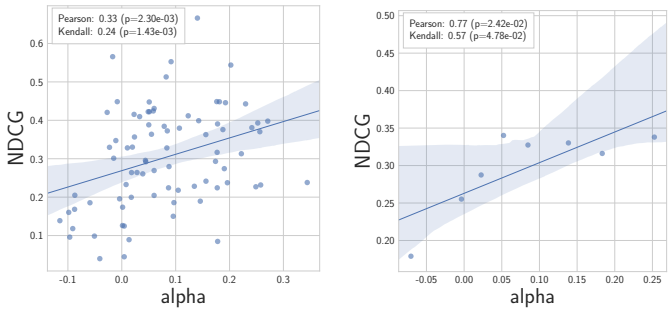


Figure 1: Agreement over topics



(a) (b)

Figure 2: Agreement and topic ease: individual topics (a) and topics binned according to their agreement (b)

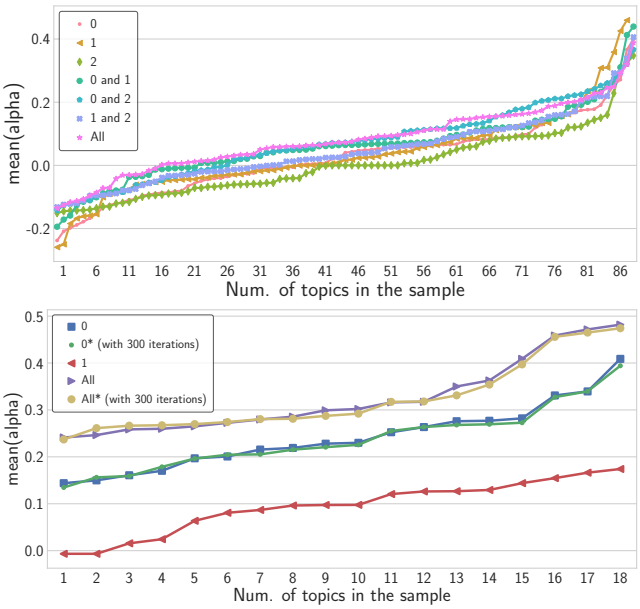


Figure 3: Agreement over topics: breakdown on relevance levels and combinations, RF above and ME below.

Section 4.3), but the topics are only 18. After this first analysis, there seems to be a significant effect of topic on agreement, so the answer to RQ1a is positive. Also note that this result should be quite stable since it is obtained using topics with dozens of documents each.

4.2 RQ1b: Agreement and Topic Ease/Difficulty

As anticipated in Section 3.2 we speak of topic ease. To understand if agreement is related to topic ease (RQ1b) we plot for each topic its agreement value (α) against its ease value (e, NDCG in our case). Figure 2(a) shows the scatter plot for RF data (using 81 topics, as detailed in Section 3.1.1): there is a mild, though significant, correlation between agreement and topic ease on the RF dataset. In other terms, topics with a higher agreement among the five assessors also tend to also be easier topics, i.e., systems tend to obtain a higher NDCG on them. The correlation is mild but it becomes stronger (and still statistically significant) when binning the topics, as shown in Figure 2(b). In this figure each dot is a bin of topics of similar agreement; the values on the axes are the averages of α and NDCG for the topics in that bin.

However, when turning to the ME datasets, we do not find any significant correlation between α and NDCG. We do not report the data for space limits, but correlations are around zero and not significant. We can formulate a conjecture for understanding this difference. One important difference on the two datasets is that the topics in ME are TREC-8 topics, with the usual Title plus Description plus Narrative rich structure; whereas RF topics are Million Query 2009 queries, selected from a query log and consisting of simply a few terms. When the latter contain ambiguous terms, those are ambiguous for both human assessors (thus probably leading to low agreement) and systems (thus probably leading to lower effectiveness as well). TREC-8 topics, instead, are probably less ambiguous for human assessors, that can exploit topic Description and Narrative to build a context and disambiguate meaning; of course, systems can rely on the query terms only, and therefore "human" and "system" ambiguity become more independent.

As an answer to RQ1b we can state that there seems to be some relation between topic ease and agreement, although a positive and significant correlation has been observed in the RF dataset only.

4.3 RQ1c: Agreement Over Relevance Levels

As discussed in Section 3.1, the individual documents of each topic have a Gold relevance assessment. Then, we can repeat the previous analyses to understand if the same results hold when a breakdown of the documents on the different relevance levels is done. Also, we can study what are the relations among the different relevance levels. Therefore, to answer RQ1c we proceed with a breakdown on relevance levels according to the Gold, keeping in mind that RF has three levels (0, 1, 2), whereas ME has two (0, 1). Since there seems to be an effect of the topics on agreement, we still work on the individual topics (instead of putting all the topics together).

Figure 3 shows, for the two datasets, the agreement values considering the documents at a given relevance level, across the various topics: α^A, α^0, α^1, and (for RF only) α^2, as well as α^{01}, α^{02}, and α^{12} values for each topic. On the x axis, topics are ranked by increasing agreement (within each series). As anticipated in Section 3.1.1, to be able to compare agreement over different relevance levels, we use 89 topics for RF. The figure shows that the relevance level causes some variation, on both datasets. The curves do not overlap (as they would do if there were no effect of the relevance level).

The charts show another result: agreement is lower for relevant documents: when focusing on documents having Gold = 2 in RF or Gold = 1 in ME, the α values are lower. However, in RF the

Table 2: Pearson's ρ correlations of agreement computed over different topic subsets on the basis of relevance levels for the RF dataset (* means $p < .005$).

	α^0	α^1	α^2	α^{01}	α^{02}	α^{12}	α^A
α^0		.13	.06	.78*	.80*	.19	.71*
α^1			.06	.51*	.16	.74*	.47*
α^2				.07	.29*	.53*	.31*
α^{01}					.7*	.46*	.87*
α^{02}						.45*	.88*
α^{12}							.68*

intermediate relevance (Gold = 1) does not have lower α values than irrelevant (Gold = 0). We verified that in our data lower α values do not depend on the number of documents involved in each α calculation. For the ME dataset we recomputed α on a subset of all the documents, and on a subset of the irrelevant documents, having a cardinality equal to the set of relevant documents for that topic. We also repeated the process 300 times and took the average α. The resulting α values are plotted on the bottom chart of Figure 3 as the two more pale curves. Those are in practice overlapping with the curves obtained using the full data. If the amount of data used had an effect, those two curves would be much more similar to the red curve. It has to be noted however that, the metric α decreases when judgments are distributed across more values (which makes sense because intuitively "it is easier to disagree when there is more space / possibility to disagree").

4.4 RQ1 Again: Agreement Over Topics and Relevance, Revisited

To answer RQ1 in a more complete way, we report on some further analysis. Since there seems to be an effect of *both* topics and relevance levels, we study the two combined.

From the results presented in Table 2 it is clear that there is no correlation between α^i, α^j, and α^k, neither between α^{ij} and α^k for $i, j, k \in \{0, 1, 2\}$. Note that the higher correlation of α^i and α^{ij} with α^A (last column) is to be expected since for the latter a superset of the same data is used. The same results hold for the ME dataset (although of course in that case we can speak only of α^0, α^1, and α^A). In other terms, even when focusing on a topic, the agreement level on documents having a certain relevance value according to the gold is not a good predictor of the agreement level on documents with a different relevance value.

As a general conclusion about RQ1, we can state that agreement does not seem related to topic only. One needs to take into account the relevance levels too.

5 RQ2: EFFECT OF AGREEMENT ON RETRIEVAL EVALUATION

With RQ2 we are interested not only in the agreement per se, but more specifically in its effect on effectiveness evaluation.

5.1 RQ2a: Effect on Effectiveness Measures

We begin by analyzing how NDCG values vary when varying the topic subset on the basis of agreement, and also taking into account the relevance level in the Gold (RQ2a).

Table 3: α - NDCG Pearson's ρ correlation over relevance levels combinations, for RF data (* means $p < .005$).

num. of bins	α^0	α^1	α^2	α^{01}	α^{02}	α^{12}	α^A
8	.76	.78	.84*	.71	.80	.66	.77
16	.67*	.51	.65*	.69*	.67*	.62	.64*
32	.48*	.29	.58*	.58*	.61*	.46*	.62*
89 = no bins	.42*	.24	.40*	.43*	.41*	.39*	.48*

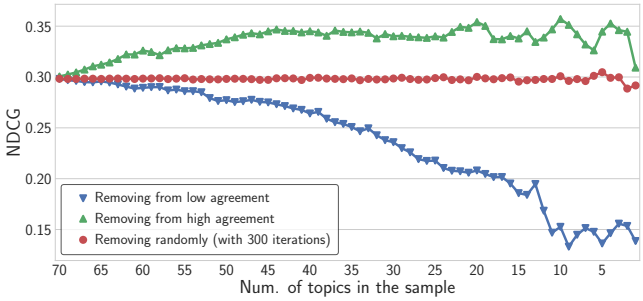


Figure 4: NDCG variation when removing topics (with high agreement, low agreement and randomly).

Table 3 shows correlation values between α and NDCG values, on RF data. This is a similar table to Figure 2, with only fewer data items used because we selected the 71 RF topics having at least one document for each relevance level and for which we found an NDCG value from the summary files (see Section 3.1.1). The results show that, although no correlation is found between α^i, α^j, and α^k, neither between α^{ij} and α^k for $i, j, k \in \{0, 1, 2\}$ (as we have seen in Table 2), the correlation between agreement (α^A) and ease shown by Figure 2 does persist when agreement is computed on a subset of the topics only.

To better understand how NDCG is affected by agreement levels we observe how NDCG changes by computing it only on high/low agreement topics and removing the others. Figure 4 provides some further insight. In the chart, the x axis represents the number of topics used to compute NDCG: at the extreme left all 71 topics are used, and topics are removed while going to the right of the chart. We use three strategies to remove the topics: randomly (with 300 random repetitions), removing the high agreement topics first, and removing the low agreement topic first. The figure (consistently with the previous results) shows that when computing NDCG after removing high agreement topics NDCG decreases, as expected since those are the easier topics. Conversely, when removing low agreement topics, NDCG increases. As a baseline, when removing topics randomly, NDCG remains constant (with some normal fluctuations). Similar patterns are observed when binning topics.

5.2 RQ2b: Effect on System Ranks

Having understood that high and low agreement topics impact differently on NDCG values, it is natural to ask how system ranks are likewise affected (RQ2b). We address this issue by trying to estimate system ranks by using a subset of topics. In other terms, we compute the Kendall's τ correlation between (i) NDCG computed on the whole topic set and (ii) NDCG computed on a subset formed by selecting the high agreement topics, or the low agreement ones, or random ones (with the usual 300 sampling repetitions). This is

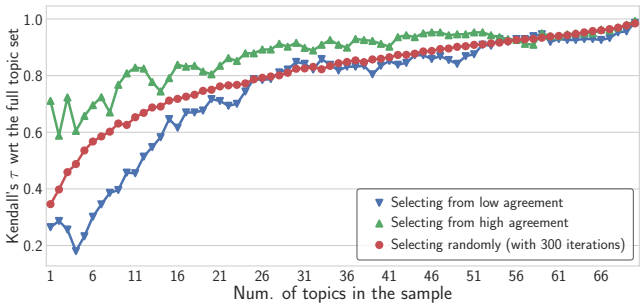


Figure 5: Effect of removing topics on system ranks. The Y axis shows the Kendall's τ correlation between the rankings obtained computing NDCG using the full 80 topics or a subset of the size shown on the x axis. Subsets are formed by low agreement, high agreement, and random topics.

a similar methodology to that used by Guiver et al. [11]. Figure 5 shows the result. The x axis represents the number of topics in the topic subset (differently from previous figures here cardinality increases while moving right; we can use 71 topics from RF). The three series show the τ values for the three topic subset selection strategies. The figure shows that when using the topics with high agreement one can predict better the system ranks than when using low agreement topics. In other terms, system ranks are more affected by high agreement topics.

6 RQ3: AN AGREEMENT AWARE METRIC

Having shown that agreement has an effect on evaluation, we now turn to RQ3. Whereas in the previous section we have been focusing on RF data only, in this section we use the ME dataset as it provides relevance judgments on a continuous scale thus displaying a score distributions that can be considered in the proposed metrics. We first define an agreement aware IR evaluation metrics and provide some intuition. We then re-evaluate systems with the new effectiveness and topic ease metrics on ME data, comparing the outcome with the official evaluation results.

6.1 RQ3a: Definition of the Agreement-Aware NDCG Metric

To answer RQ3a we provide a generalization of NDCG@k.

6.1.1 Generalizing DCG. Classic[1] DCG@k is defined as:

$$DCG@k = g_1 + \sum_{i=2}^{k} \frac{g_i}{\log(i)}.$$

This assumes that it is possible to define a gain value g_i, which in turn assumes that it can be determined with certainty. A first generalization incorporating the notion of assessor agreement could be to consider gain intervals in place of individual gains g_i. Interval DCG@k (IDCG@k) could be defined as:

$$IDCG@k = (g_1 \pm \delta) + \sum_{i=2}^{k} \frac{(g_i \pm \delta)}{\log(i)}.$$

[1]We follow the well known original formulation where the first retrieved document is considered individually. We use $\log_2$ in this paper without loss of generality.

IDCG@k is an interval, not a number. We might use generic, i.e., non symmetric, intervals, but we can generalize it even more by considering in place of individual gains g_i not simply intervals, but *gain distributions* G_i. The individual gain g_i is the expected value of G_i. We therefore define Uncertain DCG at k (UDCG@k) as:

$$UDCG@k = G_1 + \sum_{i=2}^{k} \frac{G_i}{\log(i)}.$$

Thus UDCG@k is neither a number, nor an interval but a (discounted) gain distribution. In other terms, we are not cumulating the gain values g_i but we are cumulating (discounted) gain distributions G_i. This is a generalization of IDCG@k, that can be obtained from UDCG@k by using uniform distributions over an interval.

6.1.2 Generalizing NDCG. However, DCG@k needs to be normalized to take into account different "quantities of relevance" for different topics. Classic normalized DCG@k ($NDCG@k$) is obtained by dividing the obtained DCG@k by the ideal DCG@k, i.e., the DCG@k that would be obtained with the ideal ranking of documents. Ideal DCG@k is defined as:

$$iDCG@k = ig_1 + \sum_{i=2}^{k} \frac{ig_1}{\log(i)},$$

where i stand for ideal and ig for the gain of the ideal document. Then NDCG@k is defined as: $NDCG@k = \frac{DCG@k}{iDCG@k}$. We can do almost the same to normalize UDCG@k, using mixture distributions [16], that are weighted sums of probability distributions. They are almost what we need; mixture distributions have the requirement that the sum of the weights is 1, which is not true in our case (our weights are the inverse of the logarithms of the ranks). But we can divide our sums in UDCG@k by the sum of the weights. We can define the constant value:

$$C@k = 1 + \sum_{i=2}^{k} \frac{1}{\log(i)}.$$

We can therefore build a mixture distribution for both the actual rank of documents retrieved by a system and the ideal rank, provided that we divide the results by $C@k$

$$DCG'@k = \frac{DCG@k}{C@k} \text{ and } iDCG'@k = \frac{iDCG@k}{C@k}.$$

We now normalize $DCG'@k$ by subtracting the expected value of $iDCG'@k$ (we use the mean μ): $NDCG@k = DCG'@k - \mu$. This has the effect of translating the distribution "towards zero". Note that μ of the ideal case will alway be greater than or equal to the actual case. The alternative of dividing by μ is also sensible and perhaps more similar to what is done in original NDCG@k; however we feel that the effect of subtracting μ is more intuitive and therefore we leave the alternative as future work.

With this process we can associate a gain distribution to each ⟨system, topic⟩ pair. The gain distributions can provide insight on both effectiveness / ease and agreement. Furthermore, from each gain distribution we can derive its mean (that can be used as single value of effectiveness, or ease) and its standard deviation (that can be used to measure assessor agreement). The mean over all topics is an effectiveness / ease measure, that we call AANDCG@k (Agreement Aware NDCG@k).

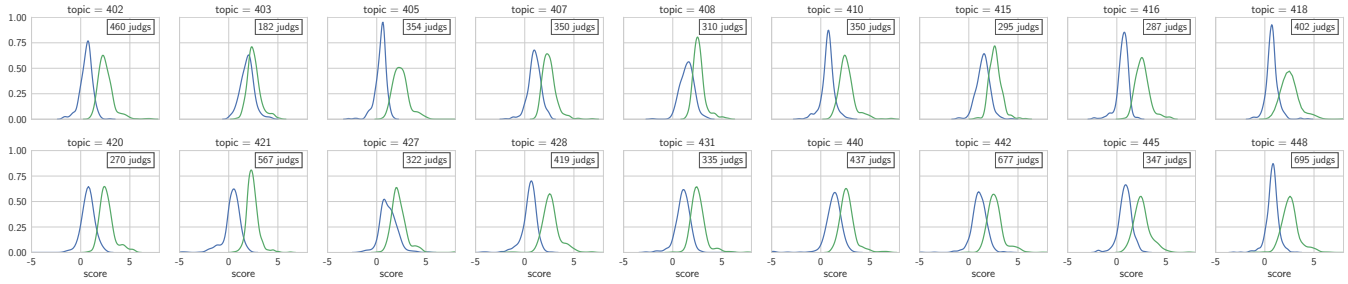


Figure 6: Distribution of the relevance scores for the 18 ME topics. Relevant documents are represented by the green distributions, not relevant documents by the blue distributions.

6.1.3 Relevance / Gain Distributions. We attempt to make more concrete the idea that each document can have its own relevance / gain distribution, in place of the standard single value. In the ME dataset, for each of the 18 topics, there are two documents, named H_K and N_K in the original paper [14], that received many relevance assessment. The number of relevance assessment varied across topic, from 182 to 695. Those documents, one highly relevant (H_K) and one very irrelevant (N_K) were used in the original work as quality checks and to normalize the scales, but by using those assessment we can have an idea of what relevance distribution to expect, at least for those relevance levels. Figure 6 shows the distributions of the (logarithm of) the relevance scores. The figure shows that, as expected, relevant documents (green distributions) have higher values than irrelevant (blue ones), but it overall confirms the idea of associating to each document its own relevance distribution rather than a single relevance value as a measure of assessor agreement. Then, in place of associating a gain value g_i to each relevance value as it is done in DCG, it is natural to think of gain distributions G_i as done in Section 6.1.2.

6.2 RQ3b: AANDCG at Work

To analyze the effect of agreement on system rankings (RQ3b) we provide two kinds of results. The first is in graphic form. Figure 7 represents the gain distributions for a given ⟨system, topic⟩ pair, and shows the subset of 100 intervals having positive values. The figure shows that there is indeed some variation across both systems and topics. Some of the gain distributions are unimodal, some are bimodal; some have a lower standard deviation; the means are in general different. Both system effectiveness and topic ease are affected by agreement.

The second result is numeric. When computing AANDCG@k (i.e., the means as described at the end of Section 6.1.2) over systems and topics, and comparing them to the original evaluation of system effectiveness and topic ease we obtain the two scatter plots, with their correlation values, in Figure 8. It is clear that the two measurements are related, but different: assessor agreement has an effect on system ranks as well as on topic ease.

7 CONCLUSIONS AND FUTURE WORK

IR evaluation results are heavily influenced by relevance judgments done by human assessors. The use of crowdsourcing to scale the collection of relevance judgments has introduced the novel aspect of judgments for the same topic-document pair being collected from multiple assessors and their labels being then aggregated. Such approach has the advantage of increasing judgment quality, but, at the same time, it disregards the information about agreement levels among assessors judging the same document.

In this paper we performed an in-depth study of the effects of assessor agreement on IR evaluation results across different datasets looking at how topics with different agreement levels affect the evaluation. We additionally proposed agreement-aware IR evaluation metrics (AANDCG) that preserve and leverage information about assessor agreement. We have observed how using topics with high-agreement judgments lead to more robust evaluation results. Our experimental results also show the benefits of defining relevance as a distribution (and thus incorporating agreement information) rather than as an absolute value attached to a retrieved document with respect to the search topic.

This paper leaves many indications for future work. We worked on a topic-wise definition of agreement; this allowed to use the standard agreement measure by Krippendorff but of course it leaves room to different approaches. Also, variants of AANDCG@k (like dividing by the mean in place of subtracting) need to be studied. We took particular care to verify our results on two independent datasets, but of course further confirmation on other data is needed.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 732328 and was partially supported by the UK EPSRC grant number EP/N011589/1. We want to thank the Erasmus+ traineeships program for facilitating collaborations, and the European Science Foundation for funding the Science Meeting SM 5917 under the ELIAS Research Networking Programme. We thank the crowd workers who participated in our experiments.

REFERENCES

[1] Omar Alonso and Stefano Mizzaro. 2012. Using Crowdsourcing for TREC Relevance Assessment. *Inf. Process. Manage.* 48, 6 (Nov. 2012), 1053–1066.
[2] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P. de Vries, and Emine Yilmaz. 2008. Relevance Assessment: Are Judges Exchangeable and Does It Matter. In *Proceedings of the 31st ACM SIGIR*. 667–674.
[3] Roi Blanco, Harry Halpin, Daniel M. Herzig, Peter Mika, Jeffrey Pound, Henry S. Thompson, and Thanh Tran Duc. 2011. Repeatable and Reliable Search System Evaluation Using Crowdsourcing. In *Proceedings of the 34th ACM SIGIR*. 923–932.
[4] Chris Buckley, Matthew Lease, and Mark D Smucker. 2010. Overview of the TREC 2010 Relevance Feedback Track. In *19th Text Retrieval Conference*.
[5] Ben Carterette, Virgiliu Pavlu, Hui Fang, and Evangelos Kanoulas. 2009. Million Query Track 2009 Overview. In *TREC*.

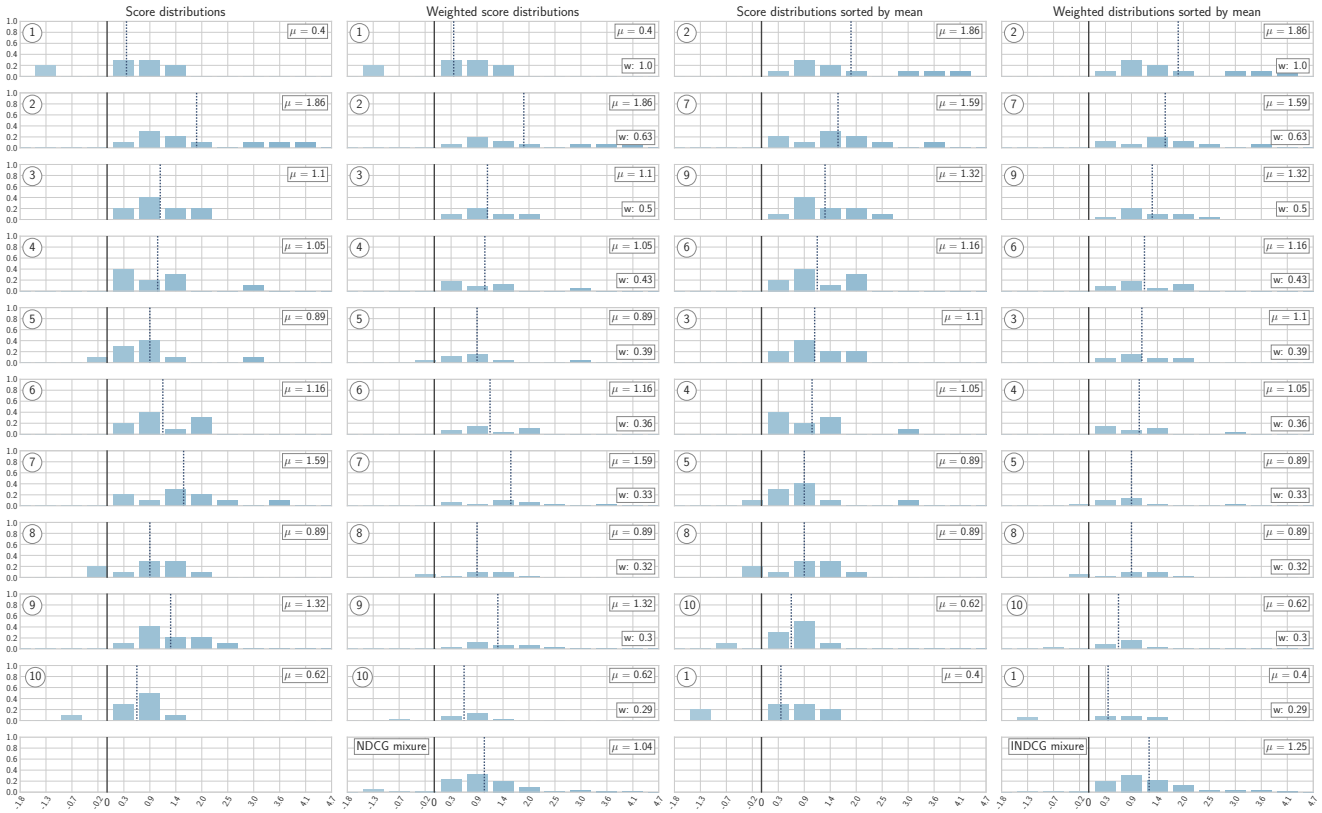


Figure 7: Gain distributions for the ME datasets. Each plot shows the binned distribution of the document relevance scores, obtained considering crowd judgments. The first column shows the first 10 documents as retrieved by an IRS; the second column shows the same distribution discounted by the log of the rank; the third column shows the top 10 documents in the ideal rank (i.e., the most relevant first; the mean value is shown, and graphically represented as a dotted vertical line); the fourth column shows the ideal rank discounted by the log of the rank (i.e., as col.2). In the bottom row we see 2 mixture distributions: on the left, the mixture distribution obtained cumulating the distributions produced by the IRS; on the right, the distribution obtained cumulating the distributions according to the ideal order.

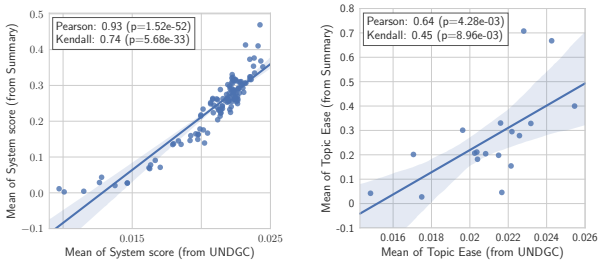


Figure 8: AANDCG vs MAP, for systems (left) and topics (right) on the ME dataset.

[6] Praveen Chandar, William Webber, and Ben Carterette. 2013. Document Features Predicting Assessor Disagreement. In *36th ACM SIGIR*. 745–748.

[7] Cyril W Cleverdon. 1970. *The effect of variations in relevance assessments in comparative experimental tests of index languages.* Technical Report. Cranfield University; Aslib.

[8] Tadele T. Damessie, Falk Scholer, Kalvero Järvelin, and J. Shane Culpepper. 2016. The Effect of Document Order and Topic Difficulty on Assessor Agreement. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval (ICTIR '16)*. 73–76.

[9] Thomas Demeester, Robin Aly, Djoerd Hiemstra, Dong Nguyen, Dolf Trieschnigg, and Chris Develder. 2014. Exploiting User Disagreement for Web Search Evaluation: An Experimental Approach. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM '14)*. 33–42.

[10] Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. Broad Twitter Corpus: A Diverse Named Entity Recognition Resource. In *COLING 2016, 26th International Conference on Computational Linguistics*. 1169–1179.

[11] J. Guiver, S. Mizzaro, and S. Robertson. 2009. A few good topics: Experiments in topic set reduction for retrieval evaluation. *ACM TOIS* 27, 4 (2009).

[12] Gabriella Kazai. 2011. In Search of Quality in Crowdsourcing for Search Engine Evaluation. In *33rd ECIR Conference*. 165–176.

[13] Klaus Krippendorff. 2007. *Computing Krippendorff's alpha reliability.* Technical Report. University of Pennsylvania. 43 pages.

[14] Eddy Maddalena, Stefano Mizzaro, Falk Scholer, and Andrew Turpin. 2017. On Crowdsourcing Relevance Magnitudes for Information Retrieval Evaluation. *ACM Trans. Inf. Syst.* 35, 3, Article 19 (Jan. 2017), 32 pages.

[15] Joao Palotti, Guido Zuccon, Johannes Bernhardt, Allan Hanbury, and Lorraine Goeuriot. 2016. *Assessors Agreement: A Case Study Across Assessor Type, Payment Levels, Query Variations and Relevance Dimensions.* 40–53.

[16] Surajit Ray and Bruce G Lindsay. 2005. The topography of multivariate normal mixtures. *Annals of Statistics* (2005), 2042–2065.

[17] Falk Scholer, Andrew Turpin, and Mark Sanderson. 2011. Quantifying Test Collection Quality Based on the Consistency of Relevance Judgements. In *Proceedings of the 34th ACM SIGIR*. 1063–1072.

[18] Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. 2014. Community-based Bayesian Aggregation Models for Crowdsourcing. In *23rd International Conference on World Wide Web (WWW '14)*. 155–164.

[19] Ellen M. Voorhees. 1998. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In *Proceedings of the 21st ACM SIGIR*. 315–323.

[20] William Webber, Bryan Toth, and Marjorie Desamito. 2012. Effect of Written Instructions on Assessor Agreement. In *35th ACM SIGIR*. 1053–1054.