Pairwise Crowd Judgments: Preference, Absolute, and Ratio

Ziying Yang
The University of Melbourne
Melbourne, Australia

Alistair Moffat
The University of Melbourne
Melbourne, Australia

Andrew Turpin
The University of Melbourne
Melbourne, Australia

ABSTRACT

Relevance judgments are conventionally formed by small numbers of experts using ordinal relevance scales defined by two or more relevance categories. Such judgments often contain many ties: documents in the same category that cannot be separated by relevance. Here we explore the use of crowd-sourcing and combined three-way relevance assessments using pairwise preference, absolute relevance, and relevance ratio, with forced choice testing and embedded quality control processes, seeking to reduce assessment ties, and to increase judgment consistency. In particular, the crowd-sourced judgments from these three approaches were normalized into numeric relevance scores, and compared against judgments arising via three previous techniques: NIST binary; Sormunen; and magnitude estimation. The relationship between generated judgment reliability and number of document pairs assessed was also explored, as was the effect that factors such as document length, topic difficulty, number of documents judged, and assessment time, have on assessment reliability. Lastly, we investigate the extent to which the methodology used to collect judgments affects the ability of an experiment to discriminate between IR systems.

KEYWORDS

Pairwise preference; relevance assessment; crowd-sourcing; test collection

ACM Reference Format:
Ziying Yang, Alistair Moffat, and Andrew Turpin. 2018. Pairwise Crowd Judgments: Preference, Absolute, and Ratio. In *23rd Australasian Document Computing Symposium (ADCS '18), December 11–12, 2018, Dunedin, New Zealand*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3291992.3291995

1 INTRODUCTION

Information retrieval (IR) system performance is often assessed by batch evaluation techniques. For a set of given information needs (topics), each system that is to be evaluated computes a *similarity score* between topics and documents, and returns a ranked list, or *run*, containing documents in decreasing similarity score order for each topic. An evaluation *metric* such as Average Precision (AP) or Rank-Biased Precision (RBP) [12] is then employed, to compute a *score* for each run based on a set of *relevance judgments*. The judgments are typically constructed using ordinal scales with either two (binary) levels, or (often) four levels of relevance [8]. They are recorded as *qrels*, with each row specifying a topic, a document identifier, and a *relevance grade* for that topic-document pair.

Conventionally, relevance judgments were assessed by small numbers of trained experts. In recent years researchers have also explored the use of techniques such as *pairwise preferences* (PP) [6]; *magnitude estimation* (ME) [19]; and *fine-grained scales* (S100) [13]; collecting relevance judgments via *crowd-sourcing* platforms [11]. The goal, is to construct a qrels set that is of good quality, but at moderate cost, where "quality" is a complex notion.

To form a set of qrels using the PP or ME methods, assessors answer questions about relevance of documents relative to one or more other documents. These preferences or ratios then have to be distilled into a score for each document. For S100 and binary judging, the documents are assessed against an absolute scale, without forcing the assessor to consider documents relative to one another. Typically these absolute values are used as the relevance grade in the resultant qrels file.

Our work here builds on these previous approaches. To better understand user perceptions of relevance, we investigate the variation of relevance judgments collected using three different relevance scales: pairwise preference; absolute relevance; and relevance ratio, all using a standard crowd-sourcing platform and non-specialist assessors. We take advantage of previously proposed methods and combine them in our experimentation, including forced choice answers and embedded quality control processes. As a result, we are able to consider the following research questions:

RQ1: Does the combination of relevance judging techniques collected by via crowd-sourcing give similar relevance scores to previous methods?
RQ2: Is it be possible to collect those judgments at lower cost than previous collection schemes?
RQ3: What factors might affect the quality of relevance assessments?
RQ4: Do our crowd-sourced judgments alter IR system evaluation compared to previous judgment schemes?

2 BACKGROUND

Relevance judgment variation. Relevance is usually subjectively judged by individuals (often experts), usually working independently. In early work, Katter [10] stated that the reliability of such relevance judgments was low. Even when relevance was assessed by groups, as was argued by Voorhees [20], there were still substantial disparity between assessors' opinions [2]. In the relevance assessment of TREC-4, up to 200 relevant documents plus 200 randomly selected irrelevant documents judged by the primary assessor (the person who created the topic) became a pool that was then passed over to another two secondary assessors [20]. Although these assessors had a similar background to the initial ones, and had been through the same TREC/NIST training, their assessment

ADCS '18, December 11–12, 2018, Dunedin, New Zealand

ACM ISBN 978-1-4503-6549-9/18/12…$15.00
https://doi.org/10.1145/3291992.3291995

agreements were under 50% [20]. Turpin and Scholer [18] carry out similar tests (using another data set, TREC GOV2) and concluded that assessors did not agree well with each other at both task and topic levels.

There are several factors that may affect relevance assessments [10, 18, 20]. First, documents are usually assessed by limited numbers of experts in this area, and viewpoints of relevance may be different from assessor to assessor due to individual factors such as gender, age, background and region [2]. Second, because the relevance scales used are selected by experiment designers rather than the assessors, the distinctions of relevance levels expected by assessors may be diverse [14, 15], with their perceptions of relevance hard to incorporate into a single relevance scale [19]. The third issue is that consistency may be lost in varying degrees as more documents are judged [16]. Finally, if two documents receive the same relevance grades (that is, are classified into the same relevance category), it is no longer possible to tell which document is preferred over the other, even if the judge held an opinion in that regard at the time the documents were examined.

Absolute relevance. Prior to the late 1990s, a binary relevance grades (in effect, "0" for non-relevant, "1" for relevant) was the dominant scale used. Subsequently, relevance scales with more granularity have allowed users to measure the pertinence of document and topic to more finely distinguished degrees. The relative usefulness of documents in regard to topics can be interpreted from qrels that use *ordinal relevance scales*. For example in TREC-2005 [8], irrelevant documents were graded as "0"; relevant documents as "1"; and documents assessed as being highly relevant as "2".

However the criteria that assessors use to determine the distinctions of relevance levels is diverse. Some experiment results have shown that *generous* users (defined as assessing more than 50% of level "0" documents as relevant) are more likely to judge documents as relevant [15], but *parsimonious* users (defined as assigning under 50% of level "1" documents as relevant) usually only judged the level "2" documents as being relevant [14].

The relevance scale used for assessment is usually selected by experiment creators, and not the assessors. But the distinction between relevance levels expected by assessors may not match that of the creators. Assessors perception of relevance is hard to incorporate into a single, discrete relevance scale [19].

Pairwise Preference. Carterette and Petkova [4] appear to be the first to make explicit use of *pairwise preferences* in IR, using them to determine a ranked list. Later, Carterette et al. [5] propose the use of *preference judgments* that record which document is preferred over another in a document pair to determine qrels. Judges then only need to make a preference choice for each pair, perhaps cognitively easier than assigning absolute relevance scores for each document, which is what is required in the conventional approach.

Compared to ordinal relevance scales, pairwise preferences help to reduce the complexity and increase the consistency of relevance assessing [7], reducing the cost of collecting valid relevance judgments with a smaller number of tied relevance scores. The preference between documents in pairs can then be used to generate a list containing all judged documents for each topic with the goal of having fewer relevance *ties* (documents with the same assessed relevance score). For example, when a binary relevance scale is used, the scope for documents having the same tied relevance score is high, whereas if relevance is measured on via Sormunen [17] categories (a four-level scale), there will most likely be (but, of course, not guaranteed to be) fewer ties. Note that ties in the relevance judgments is a factor that contributes to ties in run scores [22].

Crowdsourcing. In conventional relevance assessments, the number of assessors is relatively small, and each is trained in a way that makes carrying out the assessments a professional "work" task. If researchers wish to add to the judgment pool – or to study other aspects of retrieved documents such as novelty and freshness – that degree of expert professionalism may be hard to achieve in a defensible and independent manner; and it may also require non-trivial expenditure.

Crowdsourcing platforms, such as Amazon Mechanical Turk and Figure8 (previously CrowdFlower), provide access to human participants drawn from a large community offering a broad range of mostly non-expert skills. Workers are paid modest amounts of money to complete human intelligence tasks (HITs) published by task creators [1], and it is natural to consider making use of these services for relevance judgments [11]. The low per-unit cost means that assessment inconsistencies can be reduced by adding quality checks such as test questions and *gold standard* activities. Each crowd-worker may complete a large number of distinct small tasks including quality control tasks; and by assigning each such task to multiple individuals at the same time, a representative score of that population of individuals can be distilled.

Although crowd-sourcing workers tend to be "bronze standard" judges according to the classification of Bailey et al. [2], the addition of quality control processes, and collection of judgments from multiple workers, should reduce assessment noise. The Figure8 interface also provides an option of choosing participants in different quality levels, based on their performance on HITs provided in connection with other tasks. The number of *high quality* workers is relatively small, and (naturally) they can be selective in terms of the tasks they take on. The ability to set a minimum assessor threshold introduces a three-way tradeoff between quality, cost, and the duration it requires to get HITs assigned and completed.

3 EXPERIMENTAL METHODOLOGY

To measure the extent to which crowd-worker assessments can provide quality evaluations, we designed a collection activity based on a three-question interface that allowed internal consistency checking and mandated a forced choice between document pairs. In this section we first describe the context used for that data collection, and the collection process itself. Section 4 then explores the data that was collected, and considers each of the research questions posed in Section 1.

Dataset. We worked with the nine topics from the TREC-8 [21] Ad Hoc collection as already used in previous experimentation [17, 19], and isolated the top-10 documents returned by each of the contributing TREC systems, using them to form a pool of documents to be judged for each topic. We denote that pool size by *NumDocs*, a quantity that varies across topics.

For each topic, the *NumDocs* pooled documents were randomly and equally partitioned into *groups* a total of X times, with each

group containing *DocsPerGroup* documents, including two special ones – a HR (relevant) one, and a NR (not relevant) one (as judged by Sormunen [17]). Presented together, those two form a *gold standard*, used for assessment quality control. The other *DocsPerGroup* − 2 documents in each group were randomly selected from the pool; if *NumDocs* was not an integer multiple of *DocsPerGroup* for any particular topic, sufficient further documents from outside the pool were also included to make it so. The repetition multiplier X was chosen for each topic so that each document was paired with approximately 15% of the other documents in that topic's pool. The total number of groups across all partitionings for each topic is thus:

$$NumGroups = NumDocs/DocsPerGroup \cdot X\,. \tag{1}$$

Then, for each such group, a list of document pairs is generated in which each document in the group is coupled with K other documents from the group. Thus, the number of pairs formed for each group is:

$$PairsPerGroup = DocsPerGroup \times K/2\,. \tag{2}$$

To further reduce the assessment complexity, the sequence of pairs was constructed such that only one document varied from one pair to the next. For example, if A, B, C, and D are documents, and $K = 2$, then one sequence of pairs might be $\langle (A, D), (C, D), (C, B), (A, B)\rangle$. In other words, once they have absorbed the first pair in each group, participants only need to read one new document as each subsequent pair is presented.

Parameters for the nine selected topics are listed in Table 1. In our experiment, we grouped eight documents (including a HR and a NR, that is, *DocsPerGroup* = 8), and randomly paired each with $K = 3$ other documents in the same group, and hence *PairsPerGroup* = 12. Each assessment task was deemed to be finished when it had received valid judgments from $Y = 3$ different workers. For example, there are 216 documents in the top-10 pool formed for Topic 405 and, as already noted, we used *DocsPerGroup* = 8 throughout our trials. Hence, $27 = \lceil 216/8 \rceil$ HR documents were required. These were determined according to the Sormunen judgments, together with another 27 NR documents. As we chose *DocsPerGroup* = 8, the remaining 162 documents are randomly divided into 27 groups.

The HR documents were primarily selected from the "H" and "R" grades established by Sormunen. For Topics 402, 405, 407 and 416, there were too few "H" and "R" documents, and some "M" documents were also used. For Topics 402, 403, 407 and 440, even the "M" documents were insufficient, and resampling was employed to make up the required number. One HR and one NR document were then selected randomly and added to each group, to make $6 + 2 = 8$ documents. As already noted, we took $K = 3$, so that each document in the group appeared in three different pairs, and each unit contained a total of 12 pairs. For Topic 405 the partitioning process was repeated $X = 11$ times making $11 \times 27 = 297$ groups in total, with each document paired with $K \times X = 3 \times 11 = 33$ other documents, or 15.3% of the topic's pool. Finally, each group was assigned to a task unit to considered by $Y = 3$ different workers. That is, in total, each document for Topic 405 was presented to workers $K \times X \times Y = 99$ times for evaluation.

User Interface. The interface employed to assess each pair is shown in Figure 1. Workers were presented with two documents

Topic ID	*NumDocs*	*X*	*NumGroups*	Date
402	278	14	532	2017.12
403	120	6	96	2017.07
405	216	11	297	2018.05
407	225	11	330	2017.11
408	192	10	240	2018.05
415	184	9	207	2017.12
416	176	9	198	2017.12
431	208	10	260	2017.12
440	270	13	468	2017.12

Table 1: Data and parameters for different topics in the experiment on Figure8 as used in Equation 1. *NumDocs* includes any additional documents required to allow the integer division in Equation 1. The "Date" column shows the month in which the data was collected.

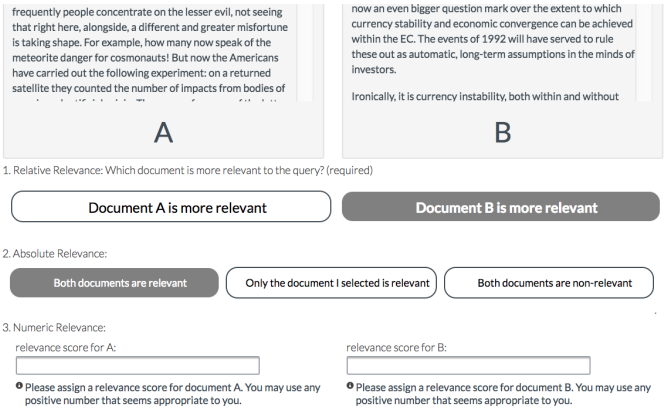

Figure 1: Screenshot of a single document pair assessment on Figure8. The two documents are displayed side-by-side in a scrolling box, with the three relevance questions below. Workers answered Question1, Question2, and Question3 for each displayed pair. Each HIT consisted of the twelve pairs associated with a single group, with each of eight documents presented three times as part of the pairs associated with the group.

in side-by-side panes, and then asked to address three questions. For Question1 they needed to make a preference choice between the two documents by clicking a single button. If both documents in the pair are considered as irrelevant and cannot be distinguished, workers were instructed to "choose one of them as best as you can". Next, Question2 asks the worker to give an absolute relevance assessment (relevant or not) for the two documents via a choice amongst three further buttons (see the screenshot). In Question3 the worker was asked to assign a numeric score (any positive number for a relevant document, and zero for irrelevant documents) to each document to indicate its relevance in relation to the paired one. This third question was intended to capture the *relevance ratio* between the paired documents. The exact instructions are available at *URL removed for blind review*.

Quality control. At the beginning of the task, workers were asked to read the task specification and the topic description carefully. We

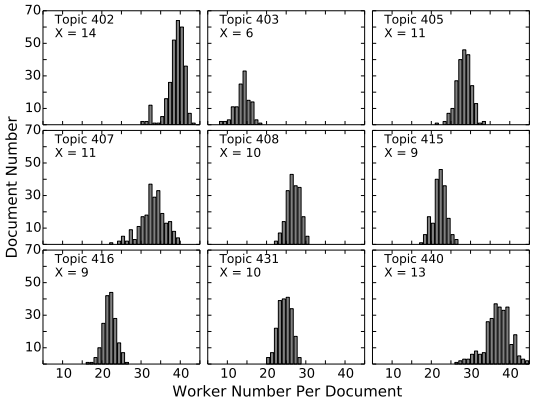

Figure 2: Distribution of crowd-workers across topics and documents. Each sub-graph is for one topic, with the topic number and the partition count of that topic shown in the upper left corner. The horizontal axis shows the number of distinct workers who judged the same document and the vertical axis shows the number of such documents. Nine sub-graphs share the same scales.

then employed Figure8's *quiz mode* to filter out low-quality workers, with *PairsPerGroup* known-answer test questions presented, and workers required to achieve a minimum accuracy requirement of 84%. The test questions contained *pseudo documents*, hand-crafted short summaries relative to the topic, with known and distinct relevance levels. These were constructed by the authors. We also employed forced choice testing and embedded quality control processes in the *work mode* HITs, to reduce criteria drift. For example, answers for Question3 were required to be in accordance with the preferences expressed in Question1.

Overall outcomes. Workers were paid USD$0.12 when they successfully completed the twelve assessments associated with the pairs in one group. In total a set of 806 different workers completed the 2628 pair groups (HITs) associated with the nine selected topics, at a total cost of USD$1,339. The most prolific of the workers completed 32 HITs; and there were 112 workers who only completed a single HIT. Most documents received judgments from multiple different crowd-workers, and the relevance judgments generated by the crowd-sourced answers are unlikely to be worker-biased. Figure 2 makes this point, by showing, for each of the topics, the number of workers involved across the set of documents, and the number of documents judged by them.

4 AGGREGATED JUDGMENTS

Frequencies and scores. To get some measure of veracity of jointly collecting pairwise preference (Question1), absolute relevance (Question2), and relevance ratio (Question3) judgments with our interface, we first compare overall outcomes against the existing NIST binary qrels and the Sormunen qrels [17].

Figure 3 shows the relevance scores of all pooled documents across the nine tested topics, collected using answers of each question. Each document-topic judgment, represented as a colored circle, is compared with Sormunen (in the first five columns) and NIST binary (in the right-hand three columns). The horizontal axis shows the categories that documents were classified in Sormunen ("H", "M", "R", "N") and binary ("1", "0") judgments. Documents which were not assessed by Sormunen or NIST are labeled as "U" in each case.

To obtain the scores plotted on the vertical axis in each of the graphs in Figure 3, the crowd-worker outputs are aggregated and processed according to the three questions. The number of times a document is the one preferred by a crowd-worker in Question1 is its *preference frequency*; its normalized *preference frequency* is then a value in the range $[0, 1]$ calculated as the preference frequency divided by the upper bound of preference frequency for each topic: $K \times X \times Y$ (recall that X varies by topic). The *normalized relevance frequencies* (obtained from Question2) of documents are computed similarly and shown in the middle pane of Figure 3.

To compute scores from the Question3 assessments, the following steps are used, on a per-topic basis. The key idea is to iteratively calculate a complete $NumDocs \times NumDocs$ pairwise matrix M, such that $M_{i,j}$ indicates the relative scores of the ith and the jth of the documents, based on the partial (and possibly inconsistent) evidence provided by the crowd-workers.

(1) First, M is initialized to contain the known values generated from the groups assessed by the workers, with cell $M_{i,j}$ assigned the geometric mean of the collected relevance ratios between documents i and j, if any such evaluations were collected. The ratio for each worker is calculated as $(s_i + \epsilon)/(s_j + \epsilon)$ where s_i and s_j are the non-negative (but possibly zero) numerical answers given by the assessor in Question3, and $\epsilon = 1$.
(2) As the set of document pairs is not complete, at this stage M is only partially assigned. We now infer the missing values by calculating the geometric mean of each column i, denoted as S_i, as regarding it as an estimate of the relevance score for document i.
(3) For any two documents which were not paired in the experiment (say, documents u and v), the value of $M_{u,v}$ is then (re-)estimated as S_u/S_v.
(4) Steps 2 and 3 are then iterated until a fixed point is reached and S is stable.
(5) Finally, the S_i s are normalized by dividing through by $S_{\max}$.

This process for completing the matrix M is the *Logarithmic Least Squares Method*, and was shown to lead to a unique optimal solution by Bozóki et al. [3]. The normalized scores of all document-topic combinations are plotted in the right graph of Figure 3.

The patterns in Figure 3 show that the documents considered as relevant using the two ordinal scales (Sormunen and binary) on average also receive high normalized scores as a result of our three questions, with overall medians and scores in alignment. At the top end of the scale, the relevance distance between grades "H" and "M" in the Sormunen scale are close in our normalized distributions, perhaps reflecting that it is difficult to distinguish these. Note that the normalization process applied in regard to Question1 and Question2 gives different distributions, and even lowly-ranked documents can get Question1 scores of around 0.4 as a consequence of the workers being required to preference one or the other of the two documents in a pair, even if both were non-relevant.

The top half of Table 2 shows the fractions of concords and discords on a "pairs of documents" basis across the nine topics,

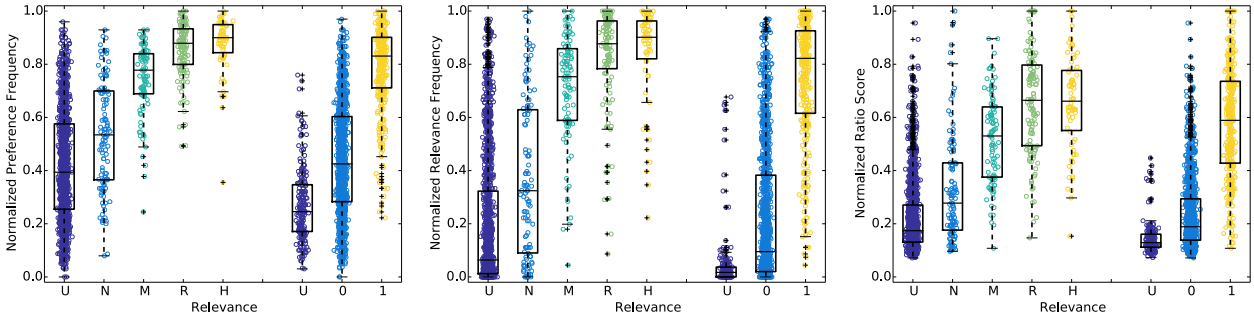

Figure 3: Normalized crowd-sourced scores (all nine topics combined) collected using three methods: pairwise preference (left), absolute relevance (middle), and relevance ratio (right), categorized by the relevance labels assigned by Sormunen (left five columns in each pane) and NIST (right three columns). Each document-topic combination is represented as a point, and has a normalized score between zero and one. Documents having no judgments in the Sormunen or NIST qrels respectively are shown in the "U" (unjudged) columns.

Judgments	Binary			Sormunen			ME			Pref			Rele			Ratio		
	A	D	U	A	D	U	A	D	U	A	D	U	A	D	U	A	D	U
Binary		–		.334	.005	.661	.301	.030	.669	.300	.028	.672	.307	.022	.671	.306	.024	.670
Sormunen	.938	.062	.000		–		.522	.181	.297	.579	.112	.309	.561	.129	.309	.551	.152	.297
ME	.898	.102	.000	.873	.127	.000		–		.810	.177	.013	.796	.152	.052	.821	.179	.000
Pref	.874	.126	.000	.858	.142	.000	.945	.055	.000		–		.824	.109	.067	.867	.119	.013
Rele	.886	.114	.000	.858	.142	.000	.951	.049	.000	.955	.045	.000		–		.858	.087	.055
Ratio	.887	.113	.000	.863	.137	.000	.942	.058	.000	.956	.044	.000	.970	.030	.000		–	

Table 2: Judgment agreement using relative document order (upper half, above the lead diagonal) and relative system order (lower half), aggregated over nine topics. In the upper half of the table, the three numbers for each pair of assessment methodologies indicate the fraction of document pairs that are clear concords (label "A", agreements), clear discords ("D", disagreements), and where either or both of the judgment sets results in a tie ("U", unknown). In the lower half of the table the 123 TREC-8 systems are compared in pairs using RBP ($\phi = 0.9$) based on the two corresponding qrels sets, and categorized similarly as concords or discords based on average score over the nine topics.

quantifying the relationships illustrated in Figure 3. When compared to binary, the other assessment approaches all result in large numbers of "U" outcomes, because in the binary qrels file many document pairs have both documents judged non-relevant. Agreement between the other regimes is higher. In particular, the methodologies associated with Question1, Question2, and Question3 all agree with each other well, and also with judgments generated by the magnitude estimation experiments of Turpin et al. [19].

System rankings. The lower half of Table 2 then shows the effect of using judgments on system score orderings (averaged over nine topics), comparing the $123 \times 122/2 = 7503$ system pairs, and again categorizing each pair as either a concord ("A") or a discord ("D"), and with Kendall's τ computed as the latter subtracted from the former. The same patterns of behavior can be seen, with the Question1, Question2, and Question3 outcomes being (unsurprisingly) strongly correlated, and the NIST binary, Sormunen, and ME system scores also yielding high correlation levels.

Worker consistency. We also explored the consistency of the crowd-workers when judging Question1 (preference) and Question2 (relevance). Worker consistency in Question1 was measured by the transitivity of the preferences made: a list of eight documents in a group would be generated to maximize agreement with the preference judgments given by the worker, and then the proportion of document discords counted. The top graph in Figure 4 illustrates the Question1 inconsistency of worker across the nine topics; the lower graph is the fraction of times a worker gave a different Question2 outcome when they saw a document repeated within a group, or as part of a different partition. There seems to be no relationship between the number of documents judged by a worker and the level of inconsistency in their judgments ($\tau = 0.029, p = 0.188$).

Worker accuracy. As each group contains a gold standard pair (constructed from an HR and an NR document), we can report the proportion that the worker chose HR rather than NR in Question1 of these gold standard pairs. This is shown in the final column of Table 3. As can be seen, there were topic differences, but mean accuracies are generally high.

Worker preference. After completing each HIT the workers were asked which of the three questions they liked. Table 3 shows the average worker choice, macro averaged across workers (so that each worker's aggregate opinion across the groups and topics they completed was valued equally) broken down by topic. The workers liked the preference evaluation (Question1) and absolute relevance assessment (Question2) rather more than the relevance ratio decision

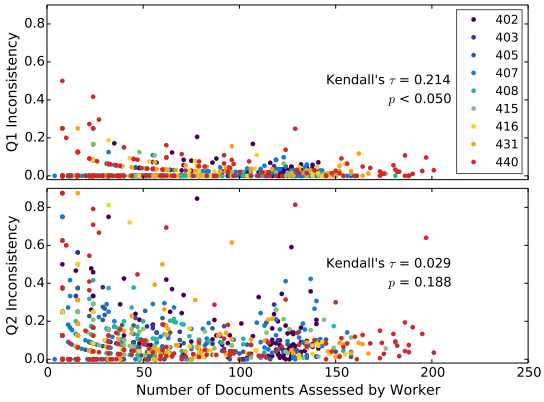

Figure 4: Fraction of documents affected by non-transitive preference judgments in Question1 (top); and receiving at least one inconsistent judgment from assessors in Question2 (bottom). Each dot in the graph represents an assessor, plotted according to the total number of documents they judged (horizontal scale), and with the different colors representing the nine topics. On average, each worker judged 60.2 documents, with an inconsistency score of 0.02 in Question1, and of 0.11 in Question2 (paired t-test, $p < 0.0001$).

Topic	Q1	Q2	Q3	Workers	AP	Accuracy
402	.397	**.635**	.217	152	0.204	87% (20%)
403	**.716**	.423	.245	47	0.708	89% (27%)
405	.493	**.584**	.210	126	0.182	94% (11%)
407	.530	**.573**	.324	103	0.279	83% (21%)
408	**.569**	.531	.188	125	0.198	91% (20%)
415	.456	**.529**	.223	72	0.400	89% (18%)
416	**.482**	.466	.294	74	0.295	88% (22%)
431	.513	**.549**	.240	79	0.329	81% (19%)
440	.408	**.642**	.328	133	0.154	76% (28%)

Table 3: Worker preferences in regard to the three questions. Columns 2 to 4 show the average rate at which workers liked that method (more than one could be selected); the largest in each row is highlighted. The number of workers who contributed to the assessment of the topic is shown in the "Workers" column. The "AP" column includes the average of the TREC-8 systems' mean Average Precision scores using the binary judgments, which may be indicative of the difficulty of the topic. The final column shows mean and standard deviation for gold standard accuracy, see text for details.

(Question3), perhaps the latter requires more thought and time to complete. The choice between Question1 and Question2 was significantly influenced by the number of groups that the worker assessed ($p = 0.013$ using ANOVA on a generalized linear mixed model). Workers who completed more groups within a topic increasingly preferred Question2 to Question1. Moreover, Question1 tended to be preferred for easier topics, where "easier" was taken to be indicated by the mean (across across TREC-8 systems) Average Precision score [9] ($p = 0.002$ using ANOVA based on a generalized linear mixed model).

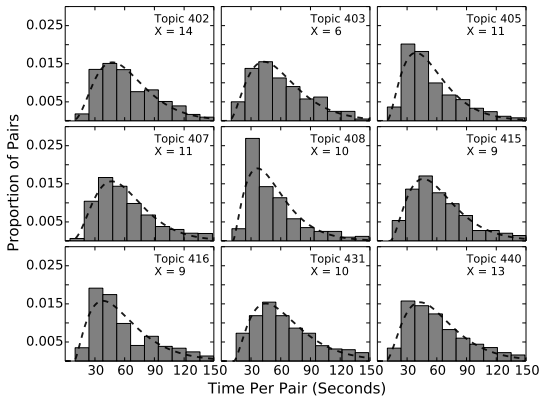

Figure 5: The distribution of judging time per document pair for each topic, with three questions to be answered per pair. In each sub-graph, the sum of bar areas is 1.0. A best-fit gamma distribution is also shown for each subgraph.

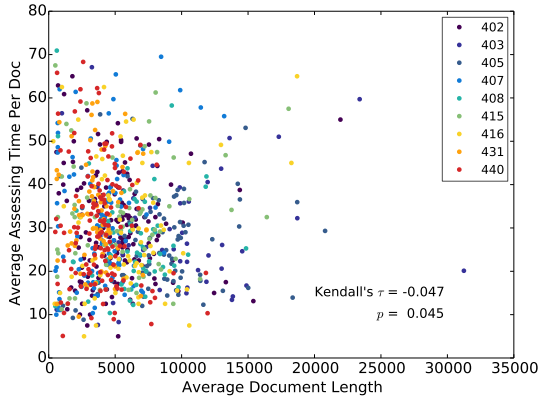

Figure 6: Average judging time per document as a function of average document length. Each dot in the graph represents a worker; colors indicate topics.

Judging time. We also measured the time that Figure8 workers spent judging document pairs. Figure 5 plots, for each of the nine topics, the distribution of worker per-pair reading and decision times. All of the topics show a similar distribution of time-spent per assessment pair, with reading of the two documents, and answering of the three questions, taking on average around 60 seconds.

Figure 6 shows the correlation between the average length of documents assessed by each worker and the average judging time that the worker spent on each document to answer all of Question1, Question2 and Question3. For each worker, the average assessed document length is plotted as the horizontal axis, and the average time the worker spent per document is plotted on the vertical axis. The Kendall's τ of all topic-worker combinations (dots) shown in Figure 6 is -0.047, and indicates that the time taken to judge documents is not related to the document length.

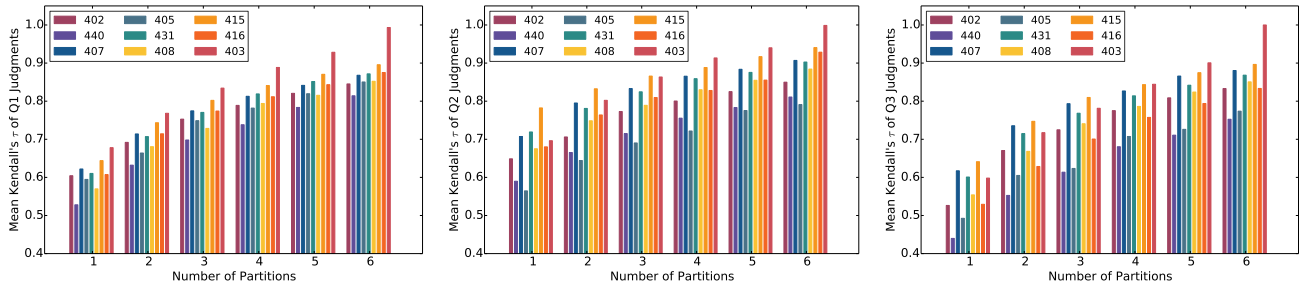

Figure 7: Mean Kendall's τ of document orderings given by judgments using different number of randomly selected partitions and the full judgments built using all partitions (as the reference), for three collection methods. In each graph and for each count of partitions used, the bars for the nine tested topics are sorted in decreasing pool size order, and shown as distinct colors. For each topic, Kendall's τ is measured between the document relevance ordering given by the full judgments, and an incrementally growing subset of the topic's partitions. Each bar represents the average of ten τ values computed over ten randomly generated permutations of the available partitions for that topic, with each such permutation dictating the incremental order that partitions are added in.

Agreement and consistency of judgments. Figure 7 shows the convergence of document score orderings that arises when partitions are incrementally added. Each group of bars represents some number of partitions being used, with partitions added incrementally towards the full set. The bar height is the Kendall's τ correlation between the document ordering induced by the scores computed from that subset of the partitions, and the document ordering induced by the scores generated when all partitions are used. Each bar is the average τ taken over ten randomly generated permutations of the topic's partitions (the column labeled X in Table 1 gives the total number of partitions used for each topic). Within each group of nine bars, the topics are sorted by the number of pooled documents from largest to smallest; that is, by decreasing number of available partitions.

By construction, these τ scores are increasing, since more and more of the judging information is being incorporated; what is of interest is the relative heights of the bars for small numbers of partitions, and the rate at which they increase in height in the left-half of each of the three graphs. The three different questions yield very similar convergence, but the τ values for relevance ratio assessments is a little lower than for the other two questions.

qrels	System Pair			
	TP	FP	FN	TN
Sormunen	2734	490	492	3787
Q1, Q2 and Q3	2773	828	453	3449
Q1 and Q3	2783	851	443	3426
Q2 and Q3	2767	822	459	3455
Q1 and Q2	2743	843	483	3434
Q1	2744	906	482	3371
Q2	2735	826	491	3451
Q3	2763	824	463	3453
Random	206	245	3020	4032

Table 4: Applying a two-tailed paired t-test (significance level 0.05) to compare all 123 TREC-8 deeply-judged systems in pairs (7503 pairs in total) using the NIST binary judgments as the reference point, and counting the number of significance differences when comparing systems using the Sormunen judgments and the judgments generated via Question1, Question2, and Question3.

Consistent discrimination. The end goal of collecting judgments is usually to determine if one system is better than another. To examine this ability, we took all pairs of systems from the 123 TREC-8 systems and evaluated them using mean Ranked-Biased Precision (RBP) with parameter $\phi = 0.9$ over the 9 topics using a variety of qrels. For each pair, if the paired t-test between the nine mean RBP scores had $p < 0.05$, the pair was *distinguished*, otherwise it was *tied*. Then, using NIST binary judgments as the reference, the system discriminations for a different set of qrels Q can be characterized as either:

- TP – both binary and Q distinguished the pair;
- TN – both binary and Q tied the pair;
- FP – binary tied the pair, but Q distinguished the pair; and
- FN – binary distinguished the pair, but Q tied the pair.

The number of each of these events is given in Table 4 for each of the 7503 pairs of systems for a variety of qrels. Those qrels also include *combined* relevance scores for documents by computing the geometric means of the three different scores arising from Question1, Question2, and Question3, to see if the whole was in some way superior to any of the parts.

For example, in the first row of Table 4, there are 2734 pairs of systems indicated as significantly different (distinguished) by both the Sormunen and the binary judgments, and 3787 pairs of systems that could not be distinguished by either of the two sets of qrels. Another 490 system pairs were tied according to the binary judgments, but distinguished by the Sormunen judgments (FP). However, the Sormunen judgments failed to separate another 492 pairs of systems which the binary judgments did distinguish.

As Table 4 shows, the preference judgments (Question1) alone distinguish the greatest number of system pairs (TP plus FP, totaling 3650), while the binary judgments distinguish 3226 system pairs and the Sormunen qrels distinguish 3224 (significantly different with $p < 0.05$ using a χ^2 test). Using any one of Question1, Question2,

Question3 and all possible combinations built judgments that distinguish more systems than NIST and Sormunen. The judgments of Sormunen have the most similar system comparing results (TP plus TN of 6521, or 86.9%) with judgments of binary. The combined judgments of Question1 and Question3 have the highest agreement of identifying different systems with binary (TP equal to 2783).

5 CONCLUSIONS

Instead of judging the relevance of documents using an ordinal scale, we divided documents into groups and paired each document the same number of times with other documents to collect pairwise relevance judgments by asking: their preference, an absolute binary judgment and the ratio of their relevance. We aggregated and normalized the pairwise judgments of each method and computed a numeric relevance score for each topic-document combination. That is, we stepped away from using broad-grained ordinal judgment categories to more fine-grained numeric ones, in the same way as has recently been suggested for S100 [13].

We proposed four research questions in conjunction with this approach. To address **RQ1** we compared the judgments generated from the answers we received to our three questions with the judgment sets (qrels) provided by NIST binary, Sormunen, and ME processes. Figure 3 and Table 2 shows the agreements between judgments in aspects of score, ordering of documents in pairs and ordering systems. We conclude that our judgments are not dissimilar to the previous schemes.

The second question, **RQ2**, is harder to resolve. We do not have costs for NIST binary process, but a reasonable estimate might be that an expert would be paid USD$50 an hour or more, and might judge one document per minute. If so, the total cost to judge the pool of 12,856 documents that we constructed would be around $11,000, not counting training time. Our total Figure8 cost was approximately one tenth of that, primarily because of the much lower pay rate involved; and even with multiple evaluations carried out on each document, the overall cost was markedly lower.

In regard to **RQ3**, we were interested in factors that might affect relevance assessment quality. Perhaps the greatest – and most obvious – factor is volume of opinion. The more crowd-workers that provide an evaluation of a document, the more consistent the overall assessment of that document is likely to be. Task difficulty, worker volume and experience, and document length may also play a part, although our analysis of these effects over the crowd-sourced data was inconclusive.

We also queried whether crowd-sourced judgments gave similar system comparison outcomes to conventional relevance methods (**RQ4**). The evidence provided in Figure 3 and Table 2 suggests that they do; and if anything, the crowd-sourced qrels allow finer system distinctions to be identified (Table 4).

In addition to the objective quantifications above, we also surveyed workers about which judgment collection method(s) they preferred. The analysis in Table 3 shows that workers did not like assigning a relevance ratio for document pairs, and would rather make a pairwise preference choice or assign a binary score. Their choices between Question1 and Question2 were significantly related to topic difficulty and workload. Because of the complexity of these factors and their inter-relationships, choosing an assessment method to suit users in general might not be straightforward ("pleasing some of the people, some of the time").

Overall, the results presented here indicate that collecting judgments using pairwise preferences is cheap, effective, and preferred by assessors. Moreover, the resulting qrels appear to be as reliable, or more reliable, than those collected via other approaches. We are currently running further Figure8 HITs in which only one question at a time is asked of the workers, to allow comparison of judgment quality and time taken. One hypothesis behind the initial experimental design was that when asked three similar questions, workers would take greater care, knowing that the answers to the questions might be checked against each other. With the additional data being collected we will seek to confirm that conjecture.

REFERENCES

[1] O. Alonso, D. E. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15, 2008.

[2] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz. Relevance assessment: Are judges exchangeable and does it matter? In *Proc. SIGIR*, pages 667–674, 2008.

[3] S. Bozóki, J. Fülöp, and L. Rónyai. On optimal completion of incomplete pairwise comparison matrices. *Math. and Comp. Modelling*, 52(1):318 – 333, 2010.

[4] B. Carterette and D. Petkova. Learning a ranking from pairwise preferences. In *Proc. SIGIR*, pages 629–630, 2006.

[5] B. Carterette, P. N. Bennett, D. M. Chickering, and S. T. Dumais. Here or there: Preference judgments for relevance. In *Proc. ECIR*, pages 16–27, 2008.

[6] P. Chandar and B. Carterette. Using preference judgments for novel document retrieval. In *Proc. SIGIR*, pages 861–870, 2012.

[7] X. Chen, P. N. Bennett, K. Collins-Thompson, and E. Horvitz. Pairwise ranking aggregation in a crowdsourced setting. In *Proc. WSDM*, pages 193–202, 2013.

[8] C. L. A. Clarke, F. Scholer, and I. Soboroff. The TREC 2005 terabyte track. In *Proc. TREC*, 2005.

[9] T. Damessie, F. Scholer, and J. S. Culpepper. The influence of topic difficulty, relevance level, and document ordering on relevance judging. In *Proc. ADCS*, 2016.

[10] R. V. Katter. The influence of scale form on relevance judgments. *Inf. Str. & Retri.*, 4(1):1–11, 1968.

[11] M. Lease and E. Yilmaz. Crowdsourcing for information retrieval: Introduction to the special issue. *Inf. Retr.*, 16(2):91–100, 2013.

[12] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Sys.*, 27(1):2.1–2.27, 2008.

[13] K. Roitero, E. Maddalena, G. Demartini, and S. Mizzaro. On fine-grained relevance scales. In *Proc. SIGIR*, pages 675–684, 2018.

[14] F. Scholer and A. Turpin. Metric and relevance mismatch in retrieval evaluation. In *Proc. AIRS*, pages 50–62, 2009.

[15] F. Scholer, A. Turpin, and M. Wu. Measuring user relevance criteria. In *Proc. EVIA*, pages 50–62, 2008.

[16] F. Scholer, A. Turpin, and M. Sanderson. Quantifying test collection quality based on the consistency of relevance judgements. In *Proc. SIGIR*, pages 1063–1072, 2011.

[17] E. Sormunen. Liberal relevance criteria of TREC: Counting on negligible documents? In *Proc. SIGIR*, pages 324–330, 2002.

[18] A. Turpin and F. Scholer. Modelling disagreement between judges for information retrieval system evaluation. In *Proc. ADCS*, page 51, 2009.

[19] A. Turpin, F. Scholer, S. Mizzaro, and E. Maddalena. The benefits of magnitude estimation relevance assessments for information retrieval evaluation. In *Proc. SIGIR*, pages 565–574, 2015.

[20] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Inf. Proc. & Man.*, 36(5):697–716, 2000.

[21] E. M. Voorhees and D. Harman. Overview of the eighth Text REtrieval Conference. In *Proc. TREC*, 1999.

[22] Z. Yang, A. Moffat, and A. Turpin. How precise does document scoring need to be? In *Proc. AIRS*, pages 279–291, 2016.