Transitivity, Time Consumption, and Quality of Preference Judgments in Crowdsourcing

Kai Hui[1,2(✉)] and Klaus Berberich[1,3]

[1] Max Planck Institute for Informatics, Saarbrücken, Germany
{khui,kberberi}@mpi-inf.mpg.de
[2] Saarbrücken Graduate School of Computer Science, Saarbrücken, Germany
[3] htw saar, Saarbrücken, Germany

Abstract. Preference judgments have been demonstrated as a better alternative to graded judgments to assess the relevance of documents relative to queries. Existing work has verified transitivity among preference judgments when collected from trained judges, which reduced the number of judgments dramatically. Moreover, strict preference judgments and weak preference judgments, where the latter additionally allow judges to state that two documents are equally relevant for a given query, are both widely used in literature. However, whether transitivity still holds when collected from crowdsourcing, i.e., whether the two kinds of preference judgments behave similarly remains unclear. In this work, we collect judgments from multiple judges using a crowdsourcing platform and aggregate them to compare the two kinds of preference judgments in terms of transitivity, time consumption, and quality. That is, we look into whether aggregated judgments are transitive, how long it takes judges to make them, and whether judges agree with each other and with judgments from TREC. Our key findings are that only strict preference judgments are transitive. Meanwhile, weak preference judgments behave differently in terms of transitivity, time consumption, as well as of the quality of judgment.

1 Introduction

Offline evaluation in information retrieval following the Cranfield [6] paradigm heavily relies on manual judgments to evaluate search results returned by competing systems. The traditional approach to judge the relevance of documents returned for a query, coined graded judgments, is to consider each document in isolation and assign a predefined grade (e.g., highly-relevant, relevant, or non-relevant) to it. More recently, preference judgments have been demonstrated [5,10,13] as a better alternative. Here, pairs of documents returned for a specific query are considered, and judges are asked to state their relative preference. Figure 1 illustrates these two approaches. Initiatives like TREC have typically relied on trained judges, who tend to provide high-quality judgments. Crowdsourcing platforms such as Amazon Mechanical Turk and CrowdFlower have emerged, providing a way to reach out to a large crowd of diverse workers for

J.M. Jose et al. (Eds.): ECIR 2017, LNCS 10193, pp. 239–251, 2017.
DOI: 10.1007/978-3-319-56608-5_19

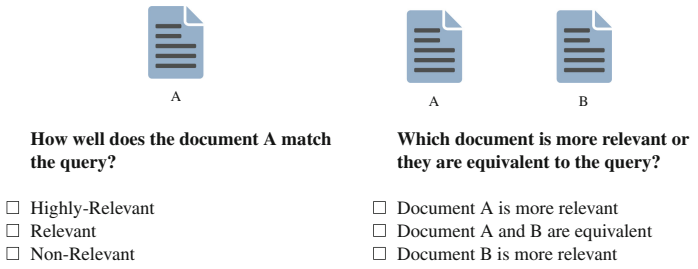

Fig. 1. Examples for graded (left) and preference judgments (right).

judgments. While inexpensive and scalable [1], judgments from those platforms are known to be of mixed quality [9,11,12].

Kazai et al. [10] demonstrated that preference judgments collected using crowdsourcing can be inexpensive yet high-quality. In their experiments preference judgments yielded better quality, getting close to the ones obtained from trained judges in terms of user satisfaction. Unfortunately, preference judgments are very expensive. To judge the relevance of n documents, $\mathcal{O}(n^2)$ preference judgments are needed, since pairs of documents have to be considered, whereas $\mathcal{O}(n)$ graded judgments suffice. Luckily, it has been shown that preference judgments are transitive [5,14] when collected from trained judges, which can be exploited to reduce their required number to $\mathcal{O}(n \log n)$. Whether transitivity still holds when preference judgments are collected using crowdsourcing is an open question as mentioned in [4]. In the aforementioned studies [5,14], trained judges stated their relative preference for all pairs of documents returned for a specific query. As a consequence, when considering a triple of documents, the same judge states relative preferences for all pairs of documents therein, making transitivity more of a matter of judges' self-consistency. When using crowdsourcing, in contrast, it is very unlikely that the same judge states relative preferences for all pairs of documents from a triple, given that workers typically only contribute a small fraction of work. Transitivity, if it exists, can thus only be a result of agreement among different judges. We examine whether transitivity holds when preference judgments are collected using crowdsourcing, when considering preference judgments aggregated from the stated preferences of multiple different judges.

Another difference between graded judgments and preference judgments, as reported by Carterette et al. [5], is that preference judgments tend to be less time consuming. Thus, in their experiments, trained judges took 40% less time to make individual preference judgments than to make individual graded judgments. We investigate whether this observation also holds when judgments are collected using crowdsourcing. If so, there is an opportunity to reduce cost by paying less for preference judgments.

Beyond that, previous works have considered different variants of preference judgments. When judges are asked to state strict preferences for two documents

d_1 and d_2, as done in [5,13,14], they can only indicate whether d_1 is preferred over d_2 ($d_1 \succ d_2$) or vice versa ($d_1 \prec d_2$). When asking for weak preferences, additional options are provided, allowing judges to state that the two documents are tied ($d_1 \sim d_2$) [10,15,16] or two documents are either equally relevant or equally non-relevant [4]. Allowing for ties is natural when judging search relevance, since it is unlikely that each of the possibly hundreds of returned documents has its own degree of relevance. We investigate whether weak preferences and strict preferences exhibit transitivity, and how they compare in terms of time consumption and quality.

Putting it together, we investigate the following research questions.

RQ1: *Do weak/strict preference judgments exhibit transitivity when collected using crowdsourcing?*

RQ2: *How do weak/strict preference judgments compare against graded judgments in terms of time consumption?*

RQ3: *Can weak/strict preference judgments collected using crowdsourcing replace judgments by trained judges?*

To answer these, we conduct an empirical study on CrowdFlower. Using topics and pooled documents from the TREC Web Track,[1] we collect graded judgments, weak preference judgments, and strict preference judgments. Akin to Carterette et al. [5], we examine transitivity by considering triples of documents. To analyze the time consumption for different kinds of judgments, our user interface is carefully instrumented to record the time that it takes judges to read documents and to make their judgment. We assess the inter-judge agreement for the different kinds of judgments and also examine to what extent they can replace judgments by trained judges from TREC.

We observe that transitivity holds over 90% for strict preference judgments collected using crowdsourcing; for weak preference judgments it only holds for about 75% of triples. In addition, we find that judges spend more time when asked for preference judgments than graded judgments in terms of total time consumption. Though time on making a single judgment is found to be lower for strict preference judgments. Finally, we see that preference judgments collected using crowdsourcing tend to show better agreement with TREC judges. Moreover, the agreement between strict preference judgments from crowdsourcing and judgments from TREC already match the agreement among trained judges reported from literature [5,10].

Organization. The rest of this paper is organized as follows. Section 2 recaps existing literature and puts our work in context. Following that, in Sect. 3, the setup of our empirical study is described. Section 4 describes its results and provides answers to the research questions stated above. Finally, in Sect. 5, we draw conclusions.

[1] http://trec.nist.gov/data/webmain.html.

2 Related Work

Preference judgments have been demonstrated as a better alternative to graded judgments, since there is no need to define graded levels [5], their higher inter-assessor agreement, and better quality [5,10,13]. Moreover, Carterette et al. [5] pointed out that preference judgments are less time-consuming than graded judgments.

Reduce the Number of Judgments in Preference Judgments. Assuming transitivity can dramatically bring down the number of judgments from $\mathcal{O}(n^2)$ to $\mathcal{O}(n \log n)$ [5]. To utilize transitivity, Rorvig [14] verified the transitivity among judgments from a group of undergraduates. Carterette et. al [5] tested transitivity among judgments from six trained judges, finding that the transitivity holds for 99% of document triples. Different from our settings, both works examined transitivity with trained judges, which is very different from the condition under crowdsourcing as indicated in Sect. 1. Moreover, both works applied strict preferences in their empirical studies. Meanwhile, follow-up works tend to extend this property to weak preferences [15]. Thus, in this work, we also examine the transitivity over weak preference judgments.

Weak Preferences Versus Strict Preferences. The choices between two kinds of preferences varied a lot among different works, even though some of them share similar motivations or research mythologies. Carterette et al. [5], Radinsky and Ailon [13] and Rorvig [14] employed strict preferences in their empirical studies for preference judgments. In the meantime, Kazai et al. [10] collected weak preference judgments from both trained judges and crowdsourcing workers to empirically explore the inter-assessors agreement and the agreement between the collected judgments and the real user satisfactions. Song et al. [15] introduced an option "same as" in the judging interface and assumed transitivity over the weak preferences in their QUICK-SORT-JUDGE method. Additionally, Zhu and Carterette [16] collected weak preferences through a "no preference" option in their research over the user preference for the layout of search results. It seems to us that the strict and weak preferences are regarded as interchangeable in existing works. However whether preference judgments with and without tie are the same in terms of judgment quality and judgment efforts remains unclear.

Crowdsourcing for Relevance Judgments. Existing works examined different ways to collect judgments from crowdsourcing [7] and provided a proper model to follow in collecting graded judgments from crowdsourcing [1]. Alonso and Mizzaro [2,3] demonstrated that it is possible to replace graded judgments from TREC using crowdsourcing. Additionally, Kazai et al. [10] compared graded and preference judgments from both trained judges and crowdsourcing, highlighting that preference judgments are especially recommended for crowdsourcing, where judgment quality can be close to the one from trained judges. Different from this work, Kaizai et al. [10] measured agreement based on individual judgments, instead of aggregated ones. As mentioned in [3], it is the aggregated judgments that can be used in practice. Moreover, the judgment quality is measured in terms of the agreement relative to user clicks, whereas in our work,

the measurement is based on judgments from TREC Web Track. Thereby, in the regards of empirical analysis over judgment quality, our work can be regarded as an extension to both [3,10].

3 Empirical Study on CrowdFlower

User Interface. We display queries together with their description from the TREC Web Track 2013 & 2014. Judges are instructed to consider both the query and its corresponding description as in Fig. 1. To help them understanding the topic, we also display a link to run the query against a commercial web search engine. When collecting preference judgments, we show the query and description together with two documents (A and B) and ask judges "Which document is more relevant to the query?". When collecting strict preferences, judges can choose between the options "Document A is more relevant" and "Document B is more relevant". A third option "Document A and B are equivalent" is added, when collecting weak preferences. When collecting graded judgments, the query and description are shown together with a single document. Judges are asked "How well does the document match the query?" and can click on one of the grades "Non-Relevant", "Relevant", and "Highly Relevant". In our instructions we include the same definitions of grades from TREC.

Quality Control. Unique tasks, in our case judgments, are referred as rows in CrowdFlower. Multiple rows are grouped into a page, which is the basic unit for payment and quality control. The major means to control quality are test questions, that is, rows with a known expected input from workers. Test questions can be used to run a qualification quiz, which workers have to complete upfront. By thresholding on their accuracy in the qualification quiz, unreliable workers can be filtered out. Moreover, test questions can be interspersed with rows to continuously control the quality of work. Workers can thus be banned once their accuracy on interspersed test questions drops below a threshold. The accuracy threshold is set as 0.7, following the default on CrowdFlower.

Job Settings. When collecting graded judgments a page consists of eleven judgments and a test question, and workers are paid $0.10 on successful completion. When collecting preference judgments, we pack eight document pairs and a test question into each page, and pay workers $0.15 on successful completion. The rationale behind the different pays is that workers receive the same amount of $0.0083 per document read. Each row is shown to workers until three trusted judgments have been collected.

Selection of Queries and Documents. Queries and documents are sampled from the TREC Web Track 2013 & 2014. From the 100 available queries, we sample a subset of twelve queries.[2] Among the sampled queries, one query is marked as ambiguous by TREC, five queries are marked as unambiguous (single), and six queries are faceted. The original relevance judgments contain up to six relevance

[2] Queries are available in http://trec.nist.gov/data/webmain.html.

levels: junk pages (*Junk*), non-relevant (*NRel*), relevant (*Rel*), highly relevant (*HRel*), key pages (*Key*), and navigational pages (*Nav*), corresponding to six graded levels, i.e., $-2, 0, 1, 2, 3, 4$. Different from other grades, *Nav* indicates a document can satisfy a navigational user intend, making the comparison relative to other documents depend on the information intent from the crowdsourcing judges. Hence, in our work, documents labeled *Nav* together with documents labeled *Junk* are removed. Due to the limit occurrences, documents labeled *Key* and *HRel* are both regarded as highly relevant. For each query we determine two sets of documents. Each set consists of twelve documents selected uniformly across graded levels, resulting in four documents per graded level. The first set is used to collect judgments; the second set serves to create test questions. When collecting graded judgments, the selected documents are directly used. To collect preference judgments, we generate for each query all 66 pairs of documents and randomly permute each document pair. Test questions are generated treating the judgments from TREC as ground truth. To ensure that workers on CrowdFlower see the same documents as trained judges from TREC, we host copies of ClueWeb12[3] documents on our own web server.

Time Consumption. To monitor the time consumed for reading documents and making judgments, we proceed as follows. We record the timestamp when judges start reading the shown document(s). To display available options for judging, workers have to click on a button "Click here to judge", and we record the instant when this happens. As a last timestamp, we record when the worker selects the submitted option. In recording timestamps, the order of clicks from judges are restricted by customized JavaScript, e.g., "Click here to judge" button is enabled only after document(s) is (are) read. We thus end up with three timestamps, allowing us to estimate the reading time, as the time passed between the first two timestamps, and the judgment time, as the time passed between the last two.

Judgment Aggregation. As mentioned, at least three trusted judgments are collected for each row. One straightforward option to aggregate them is to use majority voting as suggested by Alonso and Mizzaro [1]. However, in our setting, a simple majority vote may not break ties, given that there are more than two options to choose from. As a remedy we use workers' accuracies, as measured on test questions, in a weighted majority voting to break ties.

4 Results

We now report the results of our empirical study. After giving some general statistics about the collected judgments, we answer our three research questions, by comparing different groups of judgments over the same set of test queries employing statistical instruments like Student's t-test.

[3] http://lemurproject.org/clueweb12/index.php.

Table 1. General statistics about judgments collected using crowdsourcing.

	Graded Judgments	Strict Preferences	Weak Preferences
Total Cost	$9.36	$62.10	$76.80
#Judgments	919	2,760	2,931
#Judgments per Judge	28.80	55.00	20.10
Fleiss' κ	0.170	0.498	0.253

Distribution of Judgments					
"Highly-Relevant"	28%	$A \succ B$	51%	$A \succ B$	30%
"Relevant"	43%	$A \prec B$	49%	$A \prec B$	31%
"Non-Relevant"	29%	-		$A \sim B$	39%

4.1 General Statistics

Table 1 summarizes general statistics about the collected judgments. The collected judgments are publicly available.[4]

Inter-Judge Agreement. Similar to [3], Fleiss' κ is computed over each query and average Fleiss' κ among all queries is reported in Table 1. To put our results in context, we merge "Highly-Relevant" with "Relevant" and convert graded to binary judgments, ending up with Fleiss' $\kappa = 0.269$, which is close to 0.195 reported in [3]. In addition, Kazai et al. [10] reported Fleiss' $\kappa = 0.24$ (cf. Table 2 PC (e) therein) among weak preference judgments from crowdsourcing, which approximates 0.253 in our work. We further conduct two-tailed Student's t-test in between the three kinds of judgments over different queries. The p-value between strict preferences and graded judgments is smaller than 0.001; between weak preferences and graded judgments is 0.314; whereas it is 0.005 between the two kinds of preference judgments. It can be seen that the judges achieve better inter-agreement for strict preferences than for the others, meanwhile there is no significant difference between weak preferences and graded judgments. This aligns with the observations from [5], that strict preferences exhibit higher inter-judges agreement. The introduction of "ties" reduces the inter-judges agreement, which might due to more options are available.

4.2 RQ1: Transitivity

In this section, transitivity is examined over both strict and weak preference judgments. Different from in [5] and in [14], we investigate transitivity based on aggregated judgments. This is because the aggregated judgments are the ultimate outcome from crowdsourcing, and also because, as mentioned in Sect. 1, triples from a single judge are too few over individual queries to lead to any conclusions. The results per query are summarized in Table 2. It can be seen that over strict preferences, transitivity holds for 96% triples on average, and the number is between 91% and 100% over individual query. This number is close to the transitivity reported in [5], where average transitivity is 99% and

[4] http://people.mpi-inf.mpg.de/~khui/data/ecir17empirical.

Table 2. Transitivity over aggregated judgments. The ratio of transitive triples out of triples in different types is reported. The numbers in the bracket are the number of transitive triples divides the total number of triples.

Query	Strict Preferences	Weak Preferences			
	asymTran	*asymTran*	*s2aTran*	*s2sTran*	Overall
216	100% ($^{220}/_{220}$)	96% ($^{78}/_{81}$)	89% ($^{90}/_{101}$)	8% ($^{3}/_{38}$)	78% ($^{171}/_{220}$)
222	99% ($^{218}/_{220}$)	100% ($^{40}/_{40}$)	98% ($^{117}/_{120}$)	50% ($^{30}/_{60}$)	85% ($^{187}/_{220}$)
226	96% ($^{210}/_{220}$)	98% ($^{39}/_{40}$)	87% ($^{86}/_{99}$)	24% ($^{19}/_{81}$)	66% ($^{144}/_{220}$)
231	98% ($^{216}/_{220}$)	100% ($^{17}/_{17}$)	95% ($^{107}/_{113}$)	30% ($^{27}/_{90}$)	69% ($^{151}/_{220}$)
241	99% ($^{217}/_{220}$)	100% ($^{52}/_{52}$)	99% ($^{112}/_{113}$)	31% ($^{17}/_{55}$)	82% ($^{181}/_{220}$)
253	91% ($^{199}/_{220}$)	100% ($^{24}/_{24}$)	86% ($^{66}/_{77}$)	38% ($^{45}/_{119}$)	61% ($^{135}/_{220}$)
254	99% ($^{218}/_{220}$)	100% ($^{39}/_{39}$)	97% ($^{105}/_{108}$)	36% ($^{26}/_{73}$)	77% ($^{170}/_{220}$)
257	95% ($^{208}/_{220}$)	97% ($^{88}/_{91}$)	86% ($^{87}/_{101}$)	11% ($^{3}/_{28}$)	81% ($^{178}/_{220}$)
266	94% ($^{207}/_{220}$)	100% ($^{69}/_{69}$)	98% ($^{123}/_{125}$)	50% ($^{13}/_{26}$)	93% ($^{205}/_{220}$)
277	91% ($^{200}/_{220}$)	100% ($^{37}/_{37}$)	82% ($^{109}/_{133}$)	54% ($^{27}/_{50}$)	79% ($^{173}/_{220}$)
280	99% ($^{218}/_{220}$)	100% ($^{37}/_{37}$)	85% ($^{85}/_{100}$)	29% ($^{24}/_{83}$)	66% ($^{146}/_{220}$)
296	96% ($^{212}/_{220}$)	90% ($^{35}/_{39}$)	77% ($^{82}/_{106}$)	19% ($^{14}/_{75}$)	60% ($^{131}/_{220}$)
Avg.	96% ($^{212}/_{220}$)	98% ($^{46}/_{47}$)	90% ($^{98}/_{108}$)	32% ($^{21}/_{65}$)	75% ($^{164}/_{220}$)

at least 98% triples from a single judge are transitive. Meanwhile, for weak preferences, this number is only 75% on average, and the minimum percentage is 60% from query 296, indicating that transitivity does not hold in general. To explore the reasons, we further decompose transitivity according to different types of preferences within unique document triples. In particular, the "better than" and "worse than" options are referred as asymmetric relationships and the "tie" option is referred as symmetric relationship [8]. The transitivity can be categorized as: *asymTran*, which lies among asymmetric relationships (no tie judgment in a triple); *s2aTran*, which lies in between symmetric and asymmetric relationships (only one tie judgment in a triple) and *s2sTran*, which lies among symmetric relationships (at least two tie judgments in a triple). Over each query, the 220 triples are thereby categorized according to the three types on which transitive percentage is computed. From Table 2, we can see that *asymTran* holds even better than in strict preferences, meanwhile, *s2aTran* holds for 90% on average. However, over *s2sTran*, the transitivity does not hold anymore: the transitive percentage drops to 32% on average.

Answer to RQ1: We conclude that transitivity holds for over 90% aggregated strict preference judgments. For weak preference judgments, though, transitivity only holds among non-tie judgments (*asymTran*) and in between tie and non-tie judgments (*s2aTran*). Thus, given judgments $d_1 \sim d_2$ and $d_2 \sim d_3$, we can not infer $d_1 \sim d_3$. We can see that, in terms of transitivity, weak and strict preference judgments exhibit differently, and extra caution must be taken when assuming transitivity when collecting weak preferences via crowdsourcing.

4.3 RQ2: Time Consumption

We compare time consumption for different kinds of judgments looking both at total time, which includes the time for reading document(s) and judgment

Table 3. Average time consumption (in seconds) and quartiles over twelve queries.

Time consumption		Average	25^{th} percentile	Median	75^{th} percentile
Graded judgments	Judgment	2.60	1.37	1.52	1.82
	Total	24.24	11.73	19.55	28.88
Strict preferences	Judgment	1.79	1.24	1.37	1.58
	Total	34.17	17.84	25.28	40.98
Weak preferences	Judgment	2.07	1.40	1.57	1.91
	Total	32.43	15.77	24.57	39.10

time. The results are summarized in Table 3, based on aggregated statistics from twelve queries. For judgment time, it can be seen that judges spend least time with strict preferences. The p-values from two-tailed Student's t-tests between the three kinds of judgments are as follows. P-value equals 0.055 between strict preferences and graded judgments, equals 0.196, between weak preferences and graded judgments, and equals 0.100 between the two kinds of preference judgments. We can conclude that judges are slightly but noticeably faster in making judgments with strict preferences than in making the other two kinds of judgments, meanwhile the difference between the time consumption with weak preferences and with graded judgments is insignificant. As for total time, Table 3 demonstrates that judges are significantly faster in finishing single graded judgments after considering reading time, with p-value from two-tailed Student's t-test is less than 0.001 relative to both preference judgments. However, there is no significant difference for judges with weak and strict preferences – the corresponding p-value equals 0.168.

Answer to RQ2: Judges are faster in making strict preference judgments. When considering total time, judges need to read two documents in preference judgments, making total time consumption higher. Moreover, when comparing the two kinds of preference judgments, judges take significantly less time with strict preferences, meanwhile there is no difference in terms of total time consumption. Compared with [5,14], time consumption is measured among judges from crowdsourcing, who are with more diverse reading and judging ability and might be less skillful than trained judges. Actually, the web pages being judged require more than 20 s on average to read, making reading time dominate the total time consumption.

4.4 RQ3: Quality

We compare the quality of three kinds of judgment collected via crowdsourcing in terms of their agreement with judgments from TREC (qrel). We employ both percentage agreement, which counts the agreed judgments and divides it by the number of total judgments, and Cohen's κ as in [3], and use the latter for two-tailed Student's t-tests. When evaluating preference judgments from crowdsourcing, judgments from TREC are first converted to preference judgments, by

Table 4. Agreement between graded judgments from crowdsourcing (columns) and TREC (rows).

TREC	*Non-Relevant*	*Relevant*	*Highly-Relevant*	#Total
Non-Relevant	56.3%	39.6%	4.1%	48
Relevant	14.6%	54.2%	31.2%	48
Highly-Relevant	14.6%	37.5%	47.9%	48

Table 5. Agreement between preference judgments from crowdsourcing (columns) and the one inferred from TREC judgments (rows).

(a) strict preferences

TREC	$A \prec B$	$A \succ B$	#Total
$A \prec B$	83.0%	17.0%	282
$A \sim B$	46.8%	53.2%	216
$A \succ B$	20.4%	79.6%	294

(b) weak preferences

TREC	$A \prec B$	$A \sim B$	$A \succ B$	#Total
$A \prec B$	62.8%	30.9%	6.3%	285
$A \sim B$	17.6%	59.7%	22.7%	216
$A \succ B$	7.6%	32.0%	60.5%	291

comparing labels over two documents, resulting in "better than", "worse than" or "tie". The percentage agreement over three kinds of judgment relative to judgments from TREC are summarized in Tables 4 and 5, where the percentage is normalized per row. To put our results in context, we first measure agreement based on binary judgments, by merging the grades *Relevant* and *High-Relevant* in both TREC judgments and graded judgments from crowdsourcing. In [3], percentage agreement equals 77% and Cohen's $\kappa = 0.478$, relative to judgments from TREC-7 and TREC-8. Meanwhile we obtain 75.7% and Cohen's $\kappa = 0.43$ – slightly lower values. We argue that is due to the document collections in use: ClueWeb12, used in our work, consists of web pages which are more diverse and noisy, making it harder to judge; whereas disk 4 & 5 used in TREC-7 and TREC-8 consist of cleaner articles.[5] When using three grades, graded judgments from crowdsourcing achieve 52.8% and Cohen's $\kappa = 0.292$ relative to judgments from TREC. And the percentage agreement is 59.1% and Cohen's $\kappa = 0.358$ for strict preferences, whereas for weak preferences the numbers are 61% and 0.419 respectively. Compared with graded judgments from crowdsourcing, the corresponding p-values from paired sample t-tests over Cohen's κ among queries are 0.259 and 0.052, indicating weak preference judgments agree with TREC judgments better.

Note that, however, for documents with the same grade in TREC a tie is inferred, whereas strict preferences do not permit tie judgments. From Table 5(a), it can be seen that 216 document pairs are inferred as tied, where agreement is zero for strict preferences currently. To mitigate this mismatch, in line with [5], tie judgments in inferred preference judgments are redistributed as "A is better" or "B is better". In this redistribution, an agreement is assumed, coined as *aar*. In other words, the 216 document pairs that are inferred as tied in Table 5(a) are redistributed so that $216 \times aar$ random pairs are assigned with the same

[5] http://trec.nist.gov/data/docs_eng.html.

judgments as in collected strict preference judgments. The logic behind this is that the ground-truth strict preferences over these inferred ties are unknown and we need to assume an agreement over them to make strict preference judgments comparable. Thereby, two groups of agreement are reported for strict preference judgments at assumed agreement rates $aar = 50\%$ and 80%, respectively corresponding to random agreement and the average agreement under non-tie situations (average of 83% and 79.6% in Table 5(a)). Without influencing comparison results, graded judgments from crowdsourcing are also converted to preference judgments, making three kinds of judgments from crowdsourcing more comparable. In Table 6, it can be seen that Cohen's $\kappa = 0.530$ for strict preferences when assuming $aar = 50\%$, and the value for weak preferences is 0.419. Both preference judgments agree with TREC significantly better than graded judgments, with p-values from paired sample t-test equal 0.001 and 0.015 respectively. We further compare Cohen's κ from strict preferences ($aar = 50\%$) with the one from weak preferences, getting p-value from paired sample t-test equals 0.004, indicating strict preference judgments agree with judgments from TREC significantly better than weak preferences.

Answer to RQ3: From Table 6, it can be seen that agreement from strict preferences under $aar = 50\%$ and weak preferences are 88% and 49% higher than the collected graded judgments in terms of Cohen's κ. We further compare this agreement relative to TREC with the agreement among trained judges reported in literature, similar to [3]. Intuitively, if agreement between judgments from crowdsourcing and from TREC is comparable to the one among trained judges,

Table 6. Percentage agreement and Cohen's κ between inferred preference judgments from TREC and three kinds of judgments collected via crowdsourcing. For the column of strict preferences, tie judgments in the inferred judgments from TREC are redistributed by assuming different agreement rates. Results under $aar = 50\%$ and 80% are reported.

Query	Strict preferences				Weak preferences		Graded judgments	
	Break tie $aar = 50\%$		Break tie $aar = 80\%$					
	Percentage	Cohen's κ	Percentage	Cohen's κ	Percentage	Cohen's κ	Percentage	Cohen's κ
216	77%	0.594	85%	0.710	65%	0.466	53%	0.269
222	76%	0.569	83%	0.680	59%	0.391	65%	0.474
226	77%	0.589	79%	0.611	65%	0.473	62%	0.386
231	70%	0.494	83%	0.686	53%	0.310	65%	0.435
241	74%	0.557	83%	0.689	70%	0.543	59%	0.386
253	74%	0.533	77%	0.576	49%	0.248	36%	0.044
254	80%	0.649	91%	0.821	71%	0.573	65%	0.471
257	73%	0.529	83%	0.680	64%	0.445	61%	0.380
266	70%	0.459	73%	0.500	73%	0.588	38%	0.048
277	68%	0.397	70%	0.417	50%	0.261	38%	0.075
280	65%	0.389	74%	0.510	56%	0.345	44%	0.193
296	77%	0.601	85%	0.715	59%	0.386	50%	0.224
Avg	74%	0.530	81%	0.633	61%	0.419	53%	0.282

we can conclude that judgments from crowdsourcing are good enough to replace those from trained judges. Carterette et al. [5] reported agreement among six trained judges over preference judgments, and the percentage agreement is 74.5% (cf. Table 2(a) therein), whereas in our work agreement for strict preferences are 74% under $aar = 50\%$, and 81% under $aar = 80\%$. Kazai et al. [10] reported that Fleiss' κ among trained judges over preference judgments is 0.54 (cf. Table 2 PE (e) therein). Thus, we recompute the agreement between strict preference judgments and judgments from TREC in terms of Fleiss' κ, and get $\kappa = 0.504$ under $aar = 50\%$ and 0.637 under $aar = 80\%$. Note that strict preferences are collected in [5] and weak preferences are employed in [10]. Since the difference of these two kinds of preference judgments when collected from trained judges is unclear, we regard them the same. We can conclude that the agreement between strict preferences collected via crowdsourcing and TREC are comparable to the one among trained judges. Moreover, compared with strict preference judgments, we can conclude that judgment quality in crowdsourcing is significantly degraded when using weak preferences.

As reported in [2,3], we also observe judges from crowdsourcing can sometimes point out mistakes in TREC judgments. In total, we found around 20 such documents, especially via "test questions", by examining documents (or document pairs) that receive majority judgments opposing to the judgments from TREC. One example is clueweb12-0013wb-31-22050 and clueweb12-0806wb-32-26209 for query 280, "view my internet history". The former is labeled as "Highly-Relevant" and the latter is labeled as "Relevant" in qrel. However, none of them is relevant: the first page is a comprehensive list about history of internet & W3C, and the second page is a question on a forum about how to clean part of ones' browsing history.

4.5 Discussion

It has been demonstrated that weak and strict preferences are different in all three regards. To investigate the reasons, we reduce the number of options in weak preferences by merging "tie" with "A is better", merging "tie" with "B is better" or merging the two non-tie options, measuring the agreements among judges, getting Fleiss' $\kappa = 0.247, 0.266$, and 0.073 respectively. The corresponding p-values from two-tailed Student's t-tests relative to the one with three options are $0.913, 0.718$, and less than 0.001. It can be seen that judges tend to disagree more when making choices between ties and non-ties judgments. Put differently, the threshold to make a non-tie judgment is ambiguous and is varied among different judges. This implies that the tie option actually makes the judgments more complicated, namely, judges have to firstly determine whether the difference is large enough to be non-tied before judging the preferences.

5 Conclusion

In this work, we use crowdsourcing to collect graded judgments and two kinds of preference judgments. In terms of judgment quality, the three kinds of judgments

can be sorted as follows, graded judgments < weak < strict preference judgments. Moreover, our position for tie judgments is: it can be used but must be with more cautions when collected via crowdsourcing, especially when attempting to assume transitivity.

References

1. Alonso, O., Baeza-Yates, R.: Design and implementation of relevance assessments using crowdsourcing. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 153–164. Springer, Heidelberg (2011). doi:10.1007/978-3-642-20161-5_16
2. Alonso, O., Mizzaro, S.: Can we get rid of TREC assessors? Using mechanical turk for relevance assessment. In: SIGIR 2009 Workshop on the Future of IR Evaluation (2009)
3. Alonso, O., Mizzaro, S.: Using crowdsourcing for TREC relevance assessment. Inf. Process. Manag. **48**(6), 1053–1066 (2012)
4. Bashir, M., Anderton, J., Wu, J., Golbus, P.B., Pavlu, V., Aslam, J.A.: A document rating system for preference judgements. In: SIGIR 2013 (2013)
5. Carterette, B., Bennett, P.N., Chickering, D.M., Dumais, S.T.: Here or there: preference judgments for relevance. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 16–27. Springer, Heidelberg (2008). doi:10.1007/978-3-540-78646-7_5
6. Cleverdon, C.: The cranfield tests on index language devices. In: Aslib Proceedings, vol. 19 (1967)
7. Grady, C., Lease, M.: Crowdsourcing document relevance assessment with mechanical turk. In: NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (2010)
8. Hansson, S.O., Grne-Yanoff, T.: Preferences. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy (2012)
9. Kazai, G.: In search of quality in crowdsourcing for search engine evaluation. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 165–176. Springer, Heidelberg (2011). doi:10.1007/978-3-642-20161-5_17
10. Kazai, G., Yilmaz, E., Craswell, N., Tahaghoghi, S.M.: User intent and assessor disagreement in web search evaluation. In: CIKM 2013 (2013)
11. Moshfeghi, Y., Huertas-Rosero, A.F., Jose, J.M.: Identifying careless workers in crowdsourcing platforms: a game theory approach. In: SIGIR 2016 (2016)
12. Moshfeghi, Y., Rosero, A.F.H., Jose, J.M.: A game-theory approach for effective crowdsource-based relevance assessment. ACM Trans. Intell. Syst. Technol. **7**(4) (2016)
13. Radinsky, K., Ailon, N.: Ranking from pairs and triplets: information quality, evaluation methods and query complexity. In: WSDM 2011 (2011)
14. Rorvig, M.E.: The simple scalability of documents. J. Am. Soc. Inf. Sci. **41**(8), 590–598 (1990)
15. Song, R., Guo, Q., Zhang, R., Xin, G., Wen, J.R., Yu, Y., Hon, H.W.: Select-the-best-ones: a new way to judge relative relevance. Inf. Process. Manag. **47**(1), 37–52 (2011)
16. Zhu, D., Carterette, B.: An analysis of assessor behavior in crowdsourced preference judgments. In: SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (2010)