



UvA-DARE (Digital Academic Repository)

Studying Topical Relevance with Evidence-based Crowdsourcing

Inel, O.; Haralabopoulos, G.; Li, D.; Van Gysel, C.; Szlávik, Z.; Simperl, E.; Kanoulas, E.; Aroyo, L.

DOI

[10.1145/3269206.3271779](https://doi.org/10.1145/3269206.3271779)

Publication date

2018

Document Version

Final published version

Published in

CIKM '18

License

Article 25fa Dutch Copyright Act (<https://www.openaccess.nl/en/policies/open-access-in-dutch-copyright-law-taverne-amendment>)

[Link to publication](#)

Citation for published version (APA):

Inel, O., Haralabopoulos, G., Li, D., Van Gysel, C., Szlávik, Z., Simperl, E., Kanoulas, E., & Aroyo, L. (2018). Studying Topical Relevance with Evidence-based Crowdsourcing. In *CIKM '18: proceedings of the 2018 ACM International Conference on Information and Knowledge Management : October 22-26, 2018, Torino, Italy* (pp. 1253-1262). The Association for Computing Machinery. <https://doi.org/10.1145/3269206.3271779>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (<https://dare.uva.nl>)

Studying Topical Relevance with Evidence-based Crowdsourcing

Oana Inel*
VU Amsterdam
oana.inel@vu.nl

Giannis
Haralabopoulos†
University of Nottingham
Ghara@nottingham.ac.uk

Dan Li
University of Amsterdam
d.li@uva.nl

Christophe Van Gysel‡
Apple Inc.
cvangysel@apple.com

Zoltán Szilávik
IBM CAS Benelux
zoltan.szilavik@nl.ibm.com

Elena Simperl
University of Southampton
E.Simperl@soton.ac.uk

Evangelos Kanoulas
University of Amsterdam
E.Kanoulas@uva.nl

Lora Aroyo
VU Amsterdam
l.m.aroyo@gmail.com

ABSTRACT

Information Retrieval systems rely on large test collections to measure their effectiveness in retrieving relevant documents. While the demand is high, the task of creating such test collections is laborious due to the large amounts of data that need to be annotated, and due to the intrinsic subjectivity of the task itself. In this paper we study the topical relevance from a user perspective by addressing the problems of subjectivity and ambiguity. We compare our approach and results with the established TREC annotation guidelines and results. The comparison is based on a series of crowdsourcing pilots experimenting with variables, such as relevance scale, document granularity, annotation template and the number of workers. Our results show correlation between relevance assessment accuracy and smaller document granularity, *i.e.*, aggregation of relevance on paragraph level results in a better relevance accuracy, compared to assessment done at the level of the full document. As expected, our results also show that collecting binary relevance judgments results in a higher accuracy compared to the ternary scale used in the TREC annotation guidelines. Finally, the crowdsourced annotation tasks provided a more accurate document relevance ranking than a single assessor relevance label. This work resulted in a reliable test collection around the TREC Common Core track.

KEYWORDS

IR evaluation, crowdsourcing, TREC Common Core track

ACM Reference Format:

Oana Inel, Giannis Haralabopoulos, Dan Li, Christophe Van Gysel, Zoltán Szilávik, Elena Simperl, Evangelos Kanoulas, and Lora Aroyo. 2018. Studying Topical Relevance, with Evidence-based Crowdsourcing. In *The 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*, October 22–26, 2018, Torino, Italy. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3269206.3271779>

*Contact author.

†Work performed while at the University of Southampton.

‡Work performed while at the University of Amsterdam.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM '18, October 22–26, 2018, Torino, Italy

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-6014-2/18/10...\$15.00

<https://doi.org/10.1145/3269206.3271779>

1 INTRODUCTION

Information retrieval (IR) systems depend on test collections to evaluate their performance on retrieving relevant documents [25]. The effectiveness of these systems is typically measured with regard to topical relevance, which indicates whether a document is relevant to a topic or a search query. Due to the massive and continuously growing amount of information and document availability, the need for large volumes of manual topical relevance annotations is increasing. Smaller scale test collections are also needed and can be reused across tasks or systems [12]. Nonetheless, the creation of a topical relevance dataset requires expert annotations and the demand of expert annotators is rising.

Research indicates that the relevance assessment task is highly prone to disagreement between judges [31, 32] and very expensive [24]. The Text Retrieval Conference (TREC)¹, a well-established workshop series that provides large test collections for IR communities, is constantly in need of such judges. The test collections of these workshops are judged by the National Institute of Standards and Technology² (NIST) assessors. However, NIST typically employs one or, very rarely, two assessors per topic [19]. Thus, the task accommodates a single viewpoint. Through crowdsourcing, we have access to a high number of annotators [4, 5, 19, 22], but little measures have been taken to deal with the inherent ambiguity of the topical relevance task. Alonso and Mizzaro [6] found that the crowd relevance labels contradict the labels given by experts. However, individual worker judgments are still combined, without accounting for disagreement and ambiguity, by using majority vote or expectation maximization algorithms [22].

In this paper we focus on understanding how we can improve the accuracy and the reliability of topical relevance assessment, by studying the elements of the annotation process: (i) relevance scale, (ii) annotation guidelines, (iii) document granularity in comparative settings between NIST assessors and crowd annotators. In other words, our guiding research questions are:

RQ1 Can we improve the accuracy of topical relevance assessment by adapting annotation guidelines and document granularity?

RQ2 Can we improve the reliability of the topical relevance assessment by ensuring consistency of annotator input, with optimal relevance annotation scale?

We tackle these research questions from a user-centric perspective, by focusing on how people assign relevance to a document

¹<http://trec.nist.gov>

²<https://www.nist.gov>

with regard to a given topic. We adapt existing annotation practices of NIST assessors, and execute the adapted annotation task with crowd workers on FigureEight³ (formerly known as CrowdFlower) in a variety of settings guided by four hypotheses (see Section 4.1). Our methodology addresses the problem of ambiguity by capturing the diverse opinions of annotators and uses the disagreement among them to define novel quality measures.

The main contributions⁴ of this paper are:

- *a methodology for topical relevance assessment* consisting of empirically derived annotation guidelines and templates;
- *relevance metrics* that harness the diverse opinions and disagreement among the annotators to produce a relevance ranking for topic-document pairs;
- *annotated test collection for topical relevance* consisting of 23,554 English topic-document pairs that cover 250 topics annotated and ranked based on their topical relevance that is aligned with the NIST annotation guidelines.

2 RELATED WORK

In this section we review annotation practices for topical relevance via experts (Section 2.1) and crowds (Section 2.2). The most common way of annotating topical relevance is by using various ordinal scales: binary relevance scale [33], ternary relevance scale [14, 32], 4-point relevance scale [22, 27] or 6-point relevance scale [14], among others [29]. Even though there is a multitude of scales to choose from, Tang et al. [29] showed that there is no universal scale to fit all cases. Maddalena et al. [21] assign topical relevance to a set of documents using a magnitude estimation technique: the annotators assign relevance numbers that depict their perceived relevance over the documents. While the approach is intuitive, the annotator precision in choosing the ratios can be highly influenced by individual subjectivity. In this paper we experiment with various annotation practices at the level of the whole document, but also at the level of document paragraphs. We use a binary relevance scale to annotate each paragraph of a document, which we later aggregate to compute a ranked list of documents and assign a binary and a ternary relevance label at the document level.

2.1 Experts-based Topical Relevance

In IR, test collections for topical relevance are typically created by NIST employed assessors [19, 24, 31]. The assessors are given a set of annotation guidelines to ensure a uniform understanding of the relevance annotation task. Bailey et al. [9], Damessie et al. [15] identify three types of relevance judgments: (i) gold, from topic originators and subject experts, (ii) silver, from subject experts that are not topic originators, and (iii) bronze, from neither topic originator nor subject experts. Their experimental results suggest that bronze judges are still under-performing when compared to gold and silver TREC assessors. However, according to Dumitrache et al. [16, 17], bronze annotators can perform at least as well as subject experts on a variety of tasks and domains. In our study, we consider that gold or silver annotations can exhibit subjective insights. We believe that people who are familiar with the topic, or the subject, can introduce bias, while people that may not necessarily have prior experience or knowledge about them have the ability to accurately assess relevance. Disagreement in such cases can be a result of

an ambiguous topic, an ambiguous document, or an ambiguous annotation task design. Unlike NIST [19, 24, 31], we obtain topical relevance judgments from non-expert crowd workers.

2.2 Crowdsourcing-based Topical Relevance

Extensive studies have been published, comparing the relevance annotations of NIST assessors to non experts such as volunteers or crowd workers [1, 6, 13, 22]. As we mention in Section 5, the main finding of these papers is on par with our own finding, *i.e.*, the relevance annotations of TREC assessors are prone to inconsistencies and they do not always capture the most suitable relevance label. Furthermore, Al-Maskari et al. [1] show that the level of agreement among annotators is highly influenced by the ambiguity of a topic.

In addition, current research in topical relevance assessment addresses the role of ambiguity. Barrón-Cedeño et al. [10] investigate ambiguity in the context of ranking comments based on their relevance. They note that relevance can not be expressed as a boolean variable: it is not either relevant or not relevant, but there are also ambiguous instances. On a binary relevance scale, Alonso and Mizzaro [6] show that there is a higher disagreement at the level of relevant topic-document pairs, than at the level of not relevant topic-document pairs and argue that the number of judgments requested per topic-document pair influences the performance. The latter proved to be true in a variety of natural language processing tasks [26], as well. Furthermore, due to the "liberal" notion of relevance in TREC, Sormunen [27] argues the suitability of binary relevance scales and proposes the study of graded relevance scales. In other domains (*e.g.*, the medical domain), restricting the annotation guidelines and making them very precise and over-specified, showed an increase in inter-annotator disagreement [7].

While the majority of research in this field aims at replicating the task of NIST assessors with crowd workers, Alonso and Mizzaro [5], McDonnell et al. [22], Trotman et al. [30] marginally address the topical relevance at the level of document paragraphs instead of full documents. Previously, Callan [11] and White et al. [34] showed that with the increase of documents' length, it becomes natural to retrieve relevant documents by looking at their passages and respectively at their sentences, but no crowdsourcing topical relevance study confirmed their hypotheses. McDonnell et al. [22] report on a set of three pilot crowdsourcing tasks that also investigate the role of rationale in assessing the topical relevance. Although similar in nature with our methodology, their approach does not focus on understanding the performance impact of the relevance scale. Our main relevance assessment experiment emerges from the observations shaped during the pilot studies, while the main experiment reported in [22] does not follow the pilots.

3 DATA AND HUMAN ANNOTATORS

In our study we used a subset of the NYTimes Corpus⁵ that covers the 250 TREC Robust Track [2] topics. The topic-document pairs were selected from the documents of the participating teams in the TREC 2017 Core Track [3]. In total, we annotated 23,554 topic-document pairs from 250 topics (see *main* experiment in Table 1), covering short documents. We express the document length as bins - documents in bin1 have between 0 and 500 words and documents in bin2 have between 501 and 1,000 words. Only a fraction of the data was annotated by NIST, *i.e.*, 5,946 documents from 50 topics

³<https://www.figure-eight.com>

⁴The crowdsourced topical relevance annotations and the analysis are available at <https://github.com/CrowdTruth/NYT-Crowdsourcing-Topical-Relevance>

⁵<https://catalog.ldc.upenn.edu/ldc2008t19>

Table 1: Dataset Overview of Pilot and Main Crowdsourcing Annotation Experiments

Exp. Type	#Topics	#Doc	#Doc. per Topic	Document Length		NIST Assessors' Document Relevance Distribution			Reviewers' Document Relevance Distribution		
				Bin1	Bin2	Highly	Relevant	Not	Highly	Relevant	Not
						Relevant		Relevant	Relevant		
Pilot	10	120	12	90	30	25	30	65	27	40	53
Main	250	23,554	≈94	10,979	12,575	929	1,421	3,596	-	-	-

(929 highly relevant, 1,421 relevant and 3,596 not relevant), while the entire dataset was annotated by crowd workers, see Table 2.

For the *small scale crowdsourcing annotation pilots* we used a subset of 120 topic-document pairs annotated by NIST (see *pilot* experiment in Table 1). The 120 topic-document pairs were selected as follows. We randomly selected 10 topics and for each topic we sampled 12 short documents (bin1 and bin2) such that for each topic there was at least one document relevant or highly relevant and at least one document not relevant. Thus, the documents' relevance distribution is: 25 highly relevant, 30 relevant and 65 not relevant.

Table 2: Human annotators used in our annotation experiments: NIST assessors, crowd workers and reviewers.

Human Annotators	Exp. Type	Topics	Topic-Doc Pairs	Annotators / Topic-Doc Pair
Crowd	Pilot	10	120	15
	Main	250	23,554	7
NIST	Pilot	10	120	1 or 2
	Main	50	5,948	1 or 2
Reviewer	Pilot	10	120	3
	Main	-	-	-

We consider three types of human annotators: crowd workers, NIST assessors and quality reviewers (see Table 2). Each topic-document pair was annotated by one or two NIST assessors. Our empirical evaluation of the annotated corpus exposed some ambiguous relevance judgments of the NIST assessors (see Section 5). Therefore, to verify the reliability of our results, 3 authors of the paper acted as independent annotators (reviewers) and annotated the topic-document pairs used in the pilot studies using the ternary relevance scale. The reviewers were familiar with the NIST annotation guidelines but they did not know the relevance value provided by the NIST assessors. The reviewers annotated the topic-document pairs independently, using their own judgment and given the definition of each relevance value. We employed the reviewers in order to properly understand the reliability of the crowd on judging topical relevance and the degree of ambiguity inherent in this task.

4 RELEVANCE ASSESSMENT METHODOLOGY

Our topical relevance assessment methodology is empirically derived through a series of crowdsourcing experiments aiming at an optimal combination of annotation template, setup and quality metrics to gather topical relevance annotations. Our goal is to define the crowdsourcing template that provides the best crowd annotations to efficiently gather human relevance annotations at scale. The methodology follows the steps below:

- *crowdsourcing data collection*: we performed eight small scale crowdsourcing experiments on FigureEight with different crowdsourcing templates (Section 4.1);
- *crowdsourcing data analysis*: we adapted the CrowdTruth metrics [8, 18] to evaluate and compare the crowd annotations with the NIST annotations (Section 4.2);
- *evaluate results of NIST and crowd*: we performed a manual evaluation of the experiments in order to understand how well each type of annotators performs (Section 3, Section 5.1);
- *rank documents according to their topical relevance*: we aggregated the crowd annotations to rank and also derive optimal thresholds for labeling the relevance of a document to a topic (Section 4.3).

4.1 Crowdsourcing Experiments

To answer the two research questions introduced in Section 1, we performed eight pilot studies (Section 4.1.1) and one main experiment (Section 4.1.2) guided by the following hypotheses:

H1.1 (accuracy) Using text highlight to motivate the relevance choice increases the *accuracy* of the results.

H1.2 (accuracy) Assigning topical relevance at the level of document paragraphs instead of full documents increases the *accuracy* of the results.

H2.1 (reliability) The 2-point relevance annotation scale assigns more *reliable* relevance values of documents to topics.

H2.2 (annotation settings) Large amounts of crowd workers *perform as well as or better* than NIST assessors.

4.1.1 Crowdsourcing Pilot Studies. We performed eight crowdsourcing pilots on FigureEight in which we experimented with various annotation templates⁶ and settings, as shown in Table 3. In the annotation template of each pilot we tested different combinations of the following variables: (1) *relevance scale* (3-point and 2-point scales); (2) *annotation guidelines* (relevance value, text highlight to motivate the relevance value); (3) *document granularity* (full document, ordered and randomized document paragraphs); (4) *number of crowd annotators per topic-document pair* (from 3 to 15 annotators). In each pilot we used 120 topic-document pairs (TDP) covering short documents (bin1 and bin2), and requested 15 annotations for each TDP. The workers were level 2 contributors (experienced, higher accuracy contributors) according to FigureEight and were located in English-speaking countries (US, UK, CAN, AUS).

The annotation guidelines for the crowd did not provide any definition of the relevance values. Since relevance is usually internalized differently by every person, our aim was to allow for meaningful disagreement and not bias crowd workers. The crowd was given a topic description and either the full content of a document or the list of paragraphs in the document. The crowd was

⁶ Available at: <https://github.com/CrowdTruth/NYT-Crowdsourcing-Topical-Relevance>

Table 3: Overview of crowdsourcing experiments to derive optimal annotation settings and template

Type	Experiment	Input Data		Crowdsourcing Annotation Template			
		Topic-Doc Pairs	Doc Length	Relevance Annotation Values	Document Granularity	Document Paragraph Order	Annotation
Pilot	<i>3P-Doc-NoHigh</i>	120	Bin1 Bin2	3-point scale (Highly Relevant, Relevant, Not Relevant)	Full Document	-	Relevance Value
	<i>3P-Doc-High</i>	120	Bin1 Bin2	3-point scale (Highly Relevant, Relevant, Not Relevant)	Full Document	-	Relevance Value + Text Highlight
	<i>2P-Doc-NoHigh</i>	120	Bin1 Bin2	2-point scale (Relevant, Not Relevant)	Full Document	-	Relevance Value
	<i>2P-Doc-High</i>	120	Bin1 Bin2	2-point scale (Relevant, Not Relevant)	Full Document	-	Relevance Value + Text Highlight
	<i>2P-OrdPar-NoHigh</i>	116	Bin1 Bin2	2-point scale (Relevant, Not Relevant)	Document Paragraphs	Ordered	Relevance Value
	<i>2P-OrdPar-High</i>	116	Bin1 Bin2	2-point scale (Relevant, Not Relevant)	Document Paragraphs	Ordered	Relevance Value + Text Highlight
	<i>2P-RndPar-NoHigh</i>	116	Bin1 Bin2	2-point scale (Relevant, Not Relevant)	Document Paragraphs	Random	Relevance Value
	<i>2P-RndPar-High</i>	116	Bin1 Bin2	2-point scale (Relevant, Not Relevant)	Document Paragraphs	Random	Relevance Value + Text Highlight
Main	<i>2P-RndPar-High</i>	23,554	Bin1 Bin2	2-point scale (Relevant, Not Relevant)	Document Paragraphs	Random	Relevance Value + Text Highlight

asked to either choose the most appropriate relevance value for the document or for each paragraph, or choose the most appropriate relevance value for the document or for each paragraph and motivate their choice by highlighting relevant word phrases. The crowd annotations for each pilot were compared with the NIST and the reviewers' annotations to allow us to choose the best performing annotation setup and template, with which we performed the main annotation experiment on the main dataset introduced in Section 3.

4.1.2 Crowdsourcing Main Experiment. The main experiment replicates the *2P-RndPar-High* (Table 3, last row) annotation template. Since this is a large scale experiment, to keep the cost within reasonable margin, we only asked for 7 annotations per TDP.

4.2 Crowdsourcing Data Analysis

We evaluated the outcome of the crowdsourcing tasks, *i.e.*, the quality of each worker, the clarity of each TDP and the frequency of each relevance value, by applying the disagreement-aware methodology CrowdTruth [8, 18]. CrowdTruth models the inter-annotator disagreement using the triangle of reference [20], where the corners are represented by workers, TDP and relevance annotation values. The underlying assumption is that ambiguity in any of the corners disseminates and influences the other corners of the triangle. For example, an unclear TDP or an ambiguous relevance value can cause more disagreement between workers [18], and thus, both need to be accounted for when measuring the quality of the workers. Following, we introduce CrowdTruth main concepts and metrics:

- $WorkerVect(i, u)$: models the assessment of worker i on the TDP u as a binary vector. The length of the vector is equal to the total number of possible annotations (*i.e.*, relevance values) shown to the worker. If the worker selects an annotation value a , its corresponding element would be marked with 1, and 0 otherwise.
- $TDPVect(u) = \sum_{i \in workers(u)} WorkerVect(i, u)$: computed for every TDP as the sum of all $WorkerVect$ for that TDP.

- $TDP-RelVal$ score: expresses the likelihood of each relevance value (RelVal) a to be expressed by the TDP u . The $TDP-RelVal$ score is computed as the ratio of workers that picked the relevance value a over all the workers that annotated the TDP u , weighted by the worker quality.

The worker quality is the product of two metrics, *i.e.*, the *Worker-TDP agreement* and the *Worker-Worker agreement*. The *Worker-TDP agreement* indicates how much a worker i agrees with the rest of the workers; it is computed as the average cosine distance between the $WorkerVect(i, u)$ and the sum of all the other workers j $WorkerVect(j, u)$ that annotated TDP u . The *Worker-Worker agreement* is a pair-wise comparison between every two workers that participated in the task. The metric indicates how close a worker performs compared to the rest of the workers and is computed as the average cosine distance between the annotations of a worker i and all other workers that have worked on the same TDP as worker i . Therefore, the annotations with lower quality (*i.e.*, submitted by workers with lower quality score) have less of an impact on the final results.

4.3 Quality of Crowd Annotations

Following the preliminaries from Section 4.2 we compared the crowd performance against the NIST assessments and the reviewers' annotations. For each TDP, *i.e.* topic-document pair, we gathered 15 crowd annotations. Each crowd annotation on a TDP is stored as a vector of relevance values, $WorkerVect$, which has a variable number of elements based on the document granularity and the relevance annotation values used in the annotation template:

- For the 3-point scale relevance on full documents: the $WorkerVect$ has three elements - highly relevant, relevant and not relevant;
- For the 2-point scale relevance on full documents: the $WorkerVect$ has two elements - relevant and not relevant;
- For the 2-point scale relevance on document paragraphs: the $WorkerVect$ has the number of elements equal to the number of

Algorithm 1 Algorithm to Define Relevance Thresholds

```

1:  $t_1 \leftarrow \min(S)$ 
2:  $t_n \leftarrow \max(S)$ 
3:  $f_1(T_o) \leftarrow 0$ 
4:  $i \leftarrow 1$ 
5: while  $i \leq n$  do
6:   Calculate  $f_1(t_i)$ :
7:   if  $f_1(t_i) \geq f_1(T_o)$  then
8:      $f_1(T_o) \leftarrow f_1(t_i)$ 
9:      $T_o \leftarrow t_i$ 
10:   $i = i + 1$ 

```

paragraphs in the document + the value ["none"] (e.g., a document that is split in 10 paragraphs requires 11 elements).

All elements of a *WorkerVect* are initially set to 0. The elements of a *WorkerVect* that are picked by the worker are assigned a value of 1. As defined in Section 4.2, we use the sum of these vectors to compute the *TDP-RelVal* score. For the pilots that require the annotation of the full document, the *TDP-RelVal* score indicates the likelihood of each possible relevance annotation value, or scale. Therefore, we compute a *TDP-RelVal* for each possible relevance annotation value.

For the pilots that require the annotation of document paragraphs, the *TDP-RelVal* score indicates the relevance likelihood of each paragraph. To compute the relevance score of the full document to the topic we need a second aggregation step. We experimented with multiple aggregation methods for defining such relevance score using the *TDP-RelVal* score of each paragraph. The variations were based on permutations of the following: (1) Max, mean, median calculations, (2) All paragraphs, top n paragraphs, first n paragraphs, (3) Exclude, include zero values. The [max, mean, median] calculations correspond to the selection of the *TDP-RelVal* score for each *TDP*. The [all, top, first] paragraphs denote the selection of all paragraphs, the most relevant paragraphs based on their *TDP-RelVal* score, the first paragraphs in order of appearance in the document. The inclusion or exclusion of zeroes refers to the elements of the vector with *TDP-RelVal* score = 0, and whether they were kept or dropped before computing the final *relevance* score.

To summarize, in the pilots that annotate the relevance of document paragraphs we obtain a *TDP-RelVal* score between [0,1], i.e., each *TDP* is ranked based on its topical relevance. To align with the NIST assessors who assign a relevance value (highly relevant, relevant, not relevant) to each *TDP*, we experimentally determine optimal thresholds for these relevance scores to categorize the given *TDP* according to NIST. We obtained the appropriate thresholds using a heuristic method based on the final relevance score *TDP-RelVal*. The thresholds are evaluated via the F1-score that factors the selected (predicted) relevance values and the reviewers (true) relevance values for each *TDP*. The heuristic method calculates the F1-scores for all possible sets of thresholds and selects the best performing set based on the highest F1-score, i.e., the highest F1-score of identifying relevant and not relevant *TDP*.

The algorithmic formulation is as follows. For each aggregation method, let C be the number of relevance values, $S = (s_1, s_2, \dots, s_n)$ the set of relevance scores for each *TDP* in the corpus, $T = (t_1, t_2, \dots, t_n)$ an ordered set of thresholds and $f_1(T)$ the F1-score (micro- and macro-F1 score where the classes are not relevant, (highly) relevant) for thresholds T . We define a set of optimal thresholds T_o with

cardinality $C - 1$ that fulfills: $f_1(T_o) \geq f_1(T), \forall T$. Further, we use Algorithm 1 to obtain the optimal thresholds T_o . The aggregation that proved to provide the best relevance score for each *TDP* is equal to the $\max(\text{TDP-RelVal})$ of all the document paragraphs. We report on the exact performance in Section 5.

5 RESULTS ON CROWDSOURCING PILOTS

In the crowdsourcing annotation pilots we gathered 14,160 judgments from 221 unique crowd workers. The cost of these experiments was \$340. We analyzed and compared the crowd outcome using the methodology described in Section 4.2 and Section 4.3. We report on the performance of the crowd and NIST assessors in each pilot using binary and ternary relevance values. Thus, we evaluate both the crowd and NIST annotations against the reviewers annotations. For each pilot we identify the best *TDP-RelVal* score threshold and the F1-score at this threshold. For the pilots annotating relevance at the paragraphs level, we assign relevance values (highly relevant, relevant, not relevant) based on various thresholds of the *TDP* relevance score. Each relevance label has a lower bound threshold and an upper bound threshold. Naturally, for highly relevant documents the upper bound approaches 1.0 and for not relevant documents the lower bound approaches 0.0. The exact thresholds are computed using Algorithm 1, as explained in Section 4.3. Table 4 and Table 5 report on these values that are further analyzed in the remainder of the section.

Our main findings, in line with our four hypotheses, are:

- NIST assessors do not always capture the most suitable relevance label for a *TDP*, but the crowd is able to identify it when sufficient number of annotations are gathered (**H2.2**);
- The crowd workers are more accurate when they are asked to provide a reason for their relevance choice (**H1.1**);
- The assessment of topical relevance at the level of document paragraphs increases the accuracy of the results (**H1.2**);
- The ternary relevance scale is more ambiguous than the binary relevance scale and the relevance rank, i.e., when the *TDP* is assigned a relevance score instead of a relevance label (**H2.1**);

Table 4: Crowd and NIST results given the highest F1 *TDP-RelVal* score threshold (@T) for binary relevance. The values in bold show the best performance and the cases where the crowd performs significantly better (according to McNemar's test) than the NIST assessors for the given setup.

	Relevant		Not Relevant	
	@T	F1	@T	F1
NIST Assessors	-	0.80	-	0.79
3P-Doc-NoHigh-merged	≥ 0.79	0.85	<0.79	0.81
3P-Doc-High-merged	≥ 0.68	0.91	<0.68	0.90
2P-Doc-NoHigh	≥ 0.82	0.90	<0.82	0.88
2P-Doc-High	≥ 0.90	0.85	<0.90	0.84
2P-OrdPar-NoHigh	≥ 0.78	0.92	<0.78	0.91
2P-OrdPar-High	≥ 0.62	0.93	<0.62	0.90
2P-RndPar-NoHigh	≥ 0.65	0.90	<0.74	0.87
2P-RndPar-High	≥ 0.47	0.95	<0.47	0.94

5.1 NIST Assessors vs. Reviewers

First, the three reviewers assessed independently the relevance of each *TDP* and then they discussed the *TDP* cases on which they

Table 5: Crowd and NIST results given the highest F1 TDP-RelVal score threshold (@T) for ternary relevance. The values in bold show the best performance and the cases where the crowd performs significantly better (according to McNemar's test) than the NIST assessors for the given setup.

	Highly Relevant		Relevant		Not Relevant	
	@T	F1	@T	F1	@T	F1
<i>NIST Assessors</i>	-	0.57	-	0.45	-	0.79
<i>3P-Doc-NoHigh</i>	0.40	0.57	0.30	0.63	0.30	0.86
<i>3P-Doc-High</i>	0.30	0.69	0.40	0.64	0.40	0.92
<i>2P-OrdPar-NoHigh</i>	≥0.94	0.58	0.78-0.94	0.22	≤0.78	0.91
<i>2P-OrdPar-High</i>	≥0.84	0.56	0.62-0.84	0.40	≤0.62	0.90
<i>2P-RndPar-NoHigh</i>	≥0.91	0.55	0.69-0.91	0.35	≤0.69	0.87
<i>2P-RndPar-High</i>	≥0.63	0.66	0.47-0.63	0.56	≤0.47	0.94

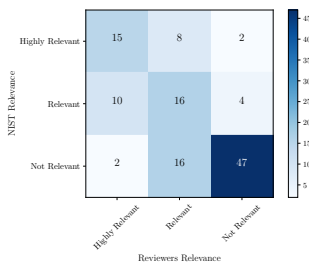


Figure 1: Comparison between NIST assessors' document relevance and reviewers' document relevance

highly disagreed, in order to reach a consensus. We measured the inter-rater reliability for the three reviewers using Fleiss' κ^7 , before and after reaching a consensus. Before the discussion the three reviewers had a moderate agreement with $\kappa = 0.48$ for ternary relevance and a substantial agreement with $\kappa = 0.67$ for binary relevance. After the discussions, their overall agreement improved: $\kappa = 0.67$ for ternary relevance and $\kappa = 0.79$ for binary relevance. Although the final agreement among the three reviewers is substantial for both binary and ternary relevance, we acknowledge that topical relevance is difficult, ambiguous, prone to inconsistencies and that each annotator interprets relevance in a subjective way. The last two columns of Table 1 and Figure 1 show the comparison between NIST and reviewers annotations, after reaching a consensus. The inter-rater agreement between the NIST assessors and the combined annotations of the reviewers using Cohen's κ^8 was equal to 0.37 for ternary relevance and 0.58 for binary relevance before the reviewers' discussion and 0.44 for ternary relevance and 0.60 for binary relevance after the reviewers' discussion. Therefore, the agreement between NIST and reviewers is moderate.

The main diagonal in Figure 1 shows that there are 78 TDP on which the reviewers agree with NIST, and 42 TDP on which they do not agree with NIST. There are 6 TDP that NIST labeled as highly relevant or relevant and the reviewers as not relevant. On 18 TDP

⁷https://en.wikipedia.org/wiki/Fleiss%27_kappa

⁸https://en.wikipedia.org/wiki/Cohen%27s_kappa

they disagree whether a document is highly relevant or just relevant to a topic and on 18 TDP they disagree whether a document is relevant or not relevant at all to a topic. We emphasize that a single annotator, as in the case of NIST assessors, can not capture the entire truth. Our observation is also confirmed in various other topical relevance studies [6, 22]. This, together with the F1 performance of the crowd (Table 4 and Table 5) shows that the crowd is able to identify the ambiguous cases where a single NIST assessor is not sufficient and to correctly label the relevance of these TDP. This observation aligns with our H2.2 hypothesis. Overall, we observe that the crowd agrees more with the relevance labels provided by the reviewers than the ones provided by the NIST assessor.

5.2 Annotation Guidelines

The pilot experiments are variations of each other in terms of the annotation guidelines used in the crowdsourcing template: relevance value *-NoHigh* or relevance value and text highlight to motivate the relevance choice *-High*. Figures 2, 3, 4, 6, 7, 8 and 9 show the F1-score calculated at each TDP-RelVal score threshold for every pair of pilots. We evaluate the performance of the crowd in terms of F1-score against the reviewers annotations. The best performing thresholds (@T) and the F1-score (F1) at these thresholds are given in Table 4 and Table 5. In all the aforementioned plots we also show the F1-score of NIST assessors (depicted by horizontal lines). In these plots we observe that *the crowd performs better mostly in the second setup, i.e., in the pilots where they had to motivate their relevance choice through text highlight*, and thus, confirming our H1.1 hypothesis. Even though we do not evaluate the text highlighted by the crowd workers similar to the approach published by McDonnell et al. [22], we see that the crowd provides better topical relevance annotations when motivating their choice.

For each experiment performed, we computed McNemar's test [23] at the best performing threshold reported in Table 4 and Table 5 (the cases in which the crowd performs significantly better than the NIST assessors are shown in bold in the two tables). When representing the TDP relevance on a binary scale, the crowd performs significantly better than NIST assessors on relevant documents, especially in the pilots that request both relevance value and text highlight (with p-values lower and much lower than 0.05), with the exception of pilot *2P-Doc-NoHigh* where the crowd seems to perform better in the setup that does not request text highlight as motivation. However, when we look at the F1-score of the crowd in these two paired pilots, in Figure 4, we observe that the best F1-score for pilot *2P-Doc-NoHigh* is only a peak, while for pilot *2P-Doc-High* we have a more uniform distribution.

Overall, pilot *2P-RndPar-High* gives the best results in identifying relevant TDP and performs significantly better than NIST assessors according to McNemar's test, on both binary and ternary relevance.

5.3 Number of Annotators

As part of our H2.2 we analyzed whether the number of crowd workers that annotate a TDP influences the overall performance. Therefore, for each number of workers, between 3 and 15 workers, we averaged the crowd performance in F1-score. The averaging was performed at the best performing TDP-RelVal score threshold, for 100 runs by randomly generating sets of i workers, where $i \in [3,15]$ for pilot *2P-RndPar-High* (Figure 5b) on binary and ternary relevance. For both binary and ternary relevance we observe that the crowd performance increases as the number of workers increases.

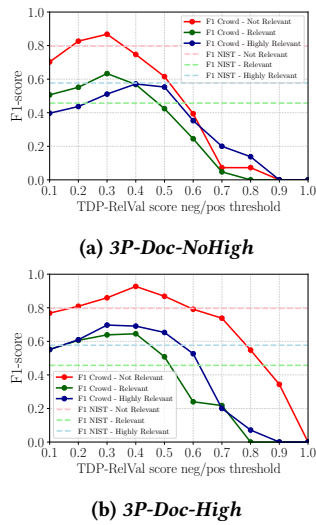


Figure 2: Annotation quality F1 per neg/pos crowd TDP-RelVal score threshold on ternary relevance for pilots 3P-Doc-NoHigh and 3P-Doc-High.

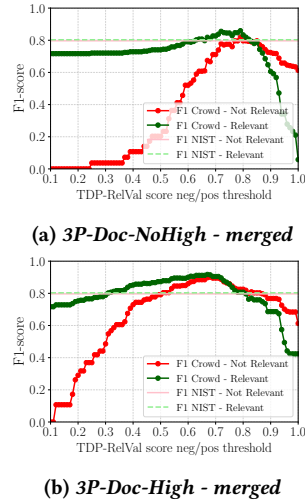


Figure 3: Annotation quality F1 per neg/pos crowd TDP-RelVal score threshold on binary relevance for pilots 3P-Doc-NoHigh and 3P-Doc-High, when merging Highly Relevant&Relevant.

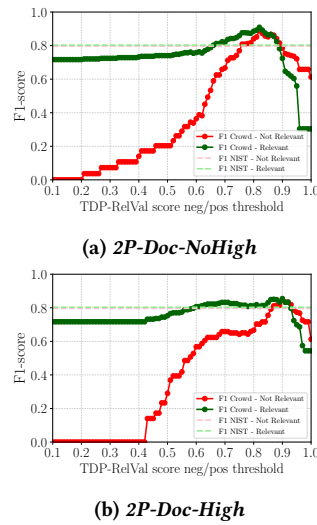


Figure 4: Annotation quality F1 per neg/pos crowd TDP-RelVal score threshold on binary relevance for pilots 2P-Doc-NoHigh and 2P-Doc-High.

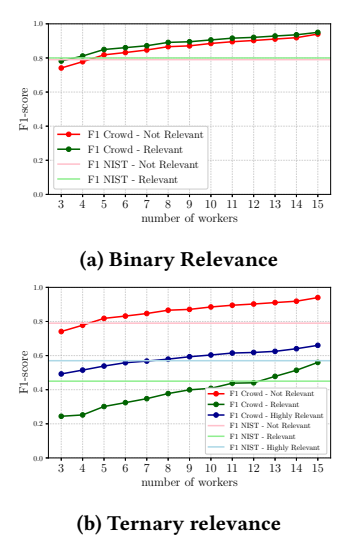


Figure 5: Crowd F1 performance at the best TDP-RelVal score thresholds for various number of workers.

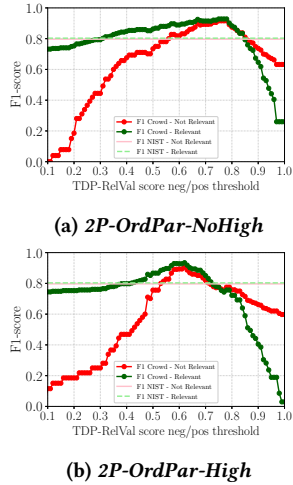


Figure 6: Annotation quality F1 per neg/pos crowd TDP-RelVal score threshold on binary relevance for pilots 2P-OrdPar-NoHigh and 2P-OrdPar-High.

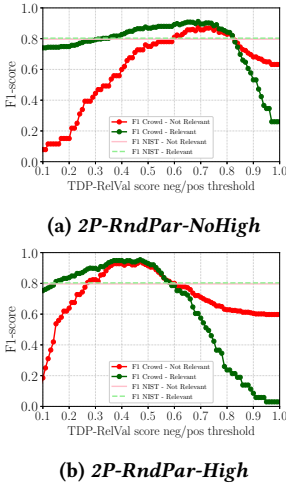


Figure 7: Annotation quality F1 per neg/pos crowd TDP-RelVal score threshold on binary relevance for pilots 2P-RndPar-NoHigh and 2P-RndPar-High.

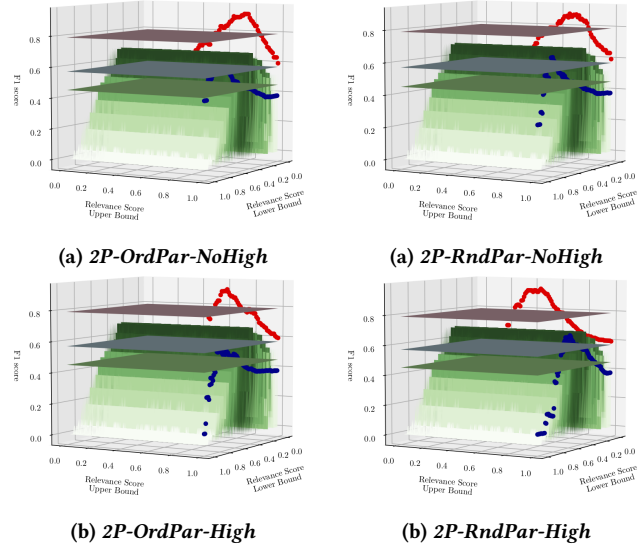


Figure 8: Annotation quality F1 per neg/pos crowd TDP-RelVal score threshold on ternary relevance for pilots 2P-OrdPar-NoHigh and 2P-OrdPar-High.

Figure 9: Annotation quality F1 per neg/pos crowd TDP-RelVal score threshold on ternary relevance for pilots 2P-RndPar-NoHigh and 2P-RndPar-High.

However, while for binary relevance we observe that 5 workers are sufficient to perform comparable with the NIST assessors, for ternary relevance the performance of the crowd is not as stable. At around 7 to 8 workers, the crowd performs at least as well as the NIST assessors only on highly relevant and not relevant documents.

5.4 Annotation Relevance Scale

Overall, we observe that the crowd performs the best at identifying the relevance of TDP on a binary scale. The options relevant and

highly relevant seem to be more ambiguous, e.g., they are often confused or used inconsistently. However, the highly relevant documents seem to be better identified by the crowd, as shown in Table 5. When examining the NIST assessors' performance compared to the reviewers, we observe a similar trend (see Figure 1). There is little ambiguity on not relevant documents, but relevant and highly relevant documents are more often confused between each other.

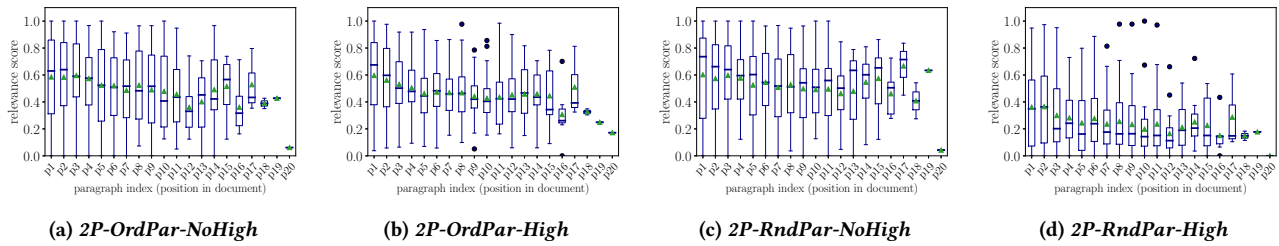


Figure 10: Crowd Document Paragraphs' Relevance Score Distribution across Relevant and Highly Relevant Documents

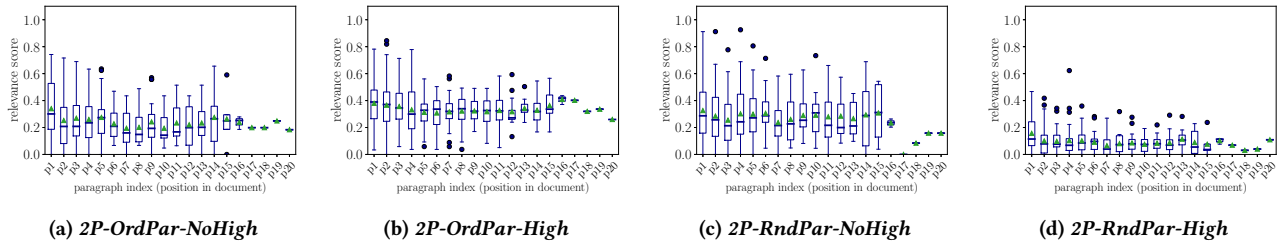


Figure 11: Crowd Document Paragraphs' Relevance Score Distribution across Not Relevant Documents

To show that the crowd and NIST assessors performance is influenced by the subjective nature of relevance scales, we plot in Figure 3 the F1-score at each TDP-RelVal score threshold, when merging the highly relevant and relevant options for all types of annotators, *i.e.*, crowd, reviewers and NIST assessors, for pilots *3P-Doc-NoHigh* and *3P-Doc-High*. In Figures 6a, 6b, 7a, 7b we plot the same analysis on the 4 pilots at the level of document paragraphs. The increased performance in F1-score of both NIST assessors and crowd workers shows that merging the relevant and highly relevant values significantly increases the performance (the statistically significant increase of performance is shown with bold font in the tables). The merged value approach supports the belief that a ternary relevance scale is more ambiguous and yields worse results, due to people internal inconsistencies and subjectivity on interpreting the relevance scales. We correlate this analysis and observation with our **H2.1** hypothesis. For the binary relevance scale the crowd performs the best in the *2P-RndPar-High* pilot, and performs significantly better than the NIST assessors, with p -values < 0.05 .

5.5 Document Granularity

The topical relevance experiments in literature [5, 22] are typically performed on the full document. However, in a crowdsourcing environment we can not assume that a crowd worker will read a long document. Our results in Table 4 indicate that people can better grasp the relevance of smaller document granularities, such as document paragraphs. The aggregated relevance of each document paragraph gives a better interpretation of the overall document relevance, as specified in our **H1.2** hypothesis. When interpreting the relevance on a ternary scale (Table 5) the crowd exhibits similar or even better performance than the NIST assessors.

In pilots *3P-Doc-NoHigh* and *3P-Doc-High* the crowd efficiently identifies highly relevant and relevant documents. However, the best performing thresholds that we identified do not ensure a single relevance value for a TDP. More precisely, the same TDP can have a likelihood of being highly relevant higher than 0.4, and a likelihood

of being relevant higher than 0.3 (*i.e.*, the document is a true positive in both cases). Therefore, we consider that (1) ranking documents based on their relevance is a more reliable solution than providing likelihood relevance values and (2) having a method that can clearly differentiate between relevant and not relevant documents results in a more reliable topical relevance test collection.

5.6 Paragraphs Order

The order in which the document paragraphs are shown to the crowd workers seems to also influence the accuracy of the results. When the paragraphs are shown in order (Figures 10a, 10b, 11a and 11b), crowd workers tend to choose more paragraphs as being relevant, which results in a decrease of the overall F1-score (Table 4). Conversely, random order of paragraphs increases the accuracy of the results. We believe that when people acknowledge the fact that the current paragraph is related to the previous paragraph, they tend to pick the same relevance label for both. However, when following paragraphs are not necessarily related, the crowd can more objectively assess whether they are relevant.

A crowdsourcing task in which the crowd workers are asked to mark the relevant part of every paragraph in a long document can become lengthy and cumbersome. Although in this work we focused only on short documents, we plan to run the same task on longer documents. In Figure 10 and Figure 11 we plotted for all relevant and respectively for all not relevant documents the TDP-RelVal score distribution per paragraph. The paragraph index, p_i , refers to the position of the paragraph in the document, where $i \in [1, \text{maximum number of document paragraphs}]$ (*e.g.*, p_1 refers to the first paragraph in the document). For relevant documents we observe that the paragraphs at the beginning of the document tend to have higher relevance scores. When looking at the scores per pilot, we notice that there is a tendency for paragraphs, from both relevant and not relevant documents, to have much higher scores when crowd workers are not asked to highlight the relevant text. However, as shown in Tables 4 and 5 these pilots usually

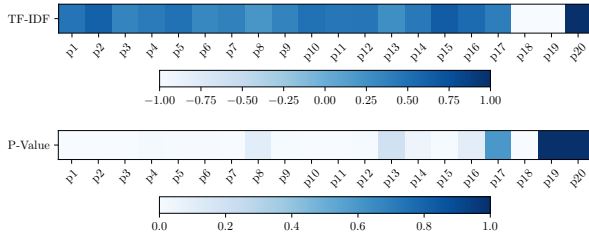


Figure 12: Spearman correlation between the crowd paragraph relevance score and the semantic similarity between the topic and the paragraph using TF-IDF

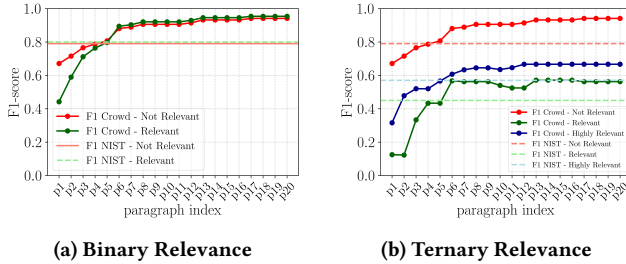


Figure 13: Crowd F1 performance at the best TDP-RelVal score thresholds when assessing the relevance of first i paragraphs in a document using the $2P\text{-}RndPar\text{-}High$ annotation template.

underperform. The high F1-score exhibited by pilot $2P\text{-}RndPar\text{-}High$ is also correlated with the paragraphs distribution scores. In Figure 11d we observe that all paragraphs from not relevant documents have relevance scores lower than 0.47, with only a few outliers.

We also looked at the semantic similarity between the topic and each paragraph of a document. We computed the semantic similarity using TF-IDF [28] to see whether higher crowd paragraphs relevance scores are correlated with higher similarity between topic and paragraph. In Figure 12 we plotted the Spearman’s correlation (ρ coefficient and p-values) between the crowd paragraph relevance score in pilot $2P\text{-}RndPar\text{-}High$ and the TF-IDF similarity score. Based on the p-values, the correlation seems to be statistically significant for the paragraphs in the beginning of the document. However, the ρ coefficient, equal to 0.522, is only marginally positive.

We further investigate the possibility to optimize such crowdsourcing task by only showing a subset of the paragraphs in the document. We plot in Figure 13 the crowd performance in F1-score at the best performing threshold for pilot $2\text{-}RndPar\text{-}High$ when the crowd is asked to inspect only the first i paragraphs in each document. We observe that for binary relevance values (Figure 13a), it is sufficient -for the crowd- to inspect only the first 6 paragraphs, to perform as well as the NIST assessors. However, on a ternary relevance score (Figure 13b) the F1-score of the crowd fluctuates, especially on relevant documents, and does not seem to stabilize when showing only 6 paragraphs.

6 RESULTS ON MAIN EXPERIMENT

In Section 5 we presented results from experimenting with independent variables such as: annotation guidelines (H1.1), document granularity (H1.2), relevance annotation scale (H2.1) and number of workers (H2.2), to improve the accuracy (RQ1) and reliability

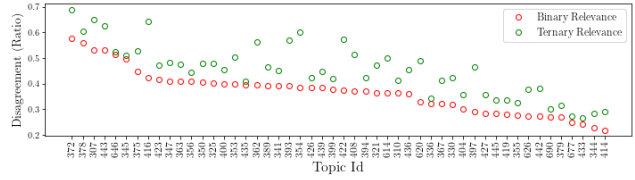


Figure 14: Relevance annotation disagreement between crowd and NIST assessors. The disagreement is computed per topic and per relevance scale (binary and ternary) as the ratio of documents in which the two types of annotators disagree.

(RQ2) of topical relevance annotations. The results showed that the $2P\text{-}RndPar\text{-}High$ annotation pilot performs better than the rest.

In the main experiment we annotated 23,554 TDP following the settings of pilot $2P\text{-}RndPar\text{-}High$. In total, we gathered 164,878 annotations from a total of 463 unique workers. We paid \$4,000 to gather these crowd annotations. Each TDP is assigned a relevance score. To align our relevance scores with the NIST annotations, we use Algorithm 1 to identify the best performing thresholds for the crowd, when compared to the NIST assessors. We use these thresholds to differentiate among highly relevant, relevant, and not relevant documents. On a binary scale, the crowd and NIST assessors agree on 63% of the TDP and on a ternary scale on 54% of the TDP. Due to the scale of this main experiment, it was not feasible to also employ quality reviewers.

Table 6: Correlation of systems’ ranking using NIST relevance assessments and crowd relevance assessments for binary and ternary relevance

Measure	Binary		Ternary	
	τ	τ_{AP}	τ	τ_{AP}
nDGC	0.6317	0.5485	0.5650	0.4879
MAP	0.5679	0.4576	0.4637	0.3728
R-Prec	0.5703	0.4849	0.4658	0.3751

In Figure 14 we plotted the disagreement between the crowd and the NIST assessors relevance annotations per topic, on both binary and ternary scales. The disagreement is computed as the ratio of documents on which the two annotators disagree. We manually inspected three topics with the most disagreement: 372 - "Native American casino", 378 - "Euro opposition" and 307 - "New Hydroelectric Projects". On all three topics the NIST assessors seem to be very restrictive and choose a limited number of documents as being relevant. On the one hand, many documents that discuss various hydroelectric projects in Niagara, or issues around casinos in Native American areas where more people are getting into trouble, were labeled as not relevant by NIST assessors and relevant by the crowd. Given that the hydroelectric projects are extensively described in such documents, we are not sure why the NIST assessors did not label them as relevant. On the other hand, in the case of topic 378, we observe that the crowd annotates as relevant documents that talk in general about the European currency and possible adherence to it, and not only documents that explicitly discuss the opposition.

In Table 6 we report on the Kendall’s τ and Yilmaz et al. [35] τ_{AP} rank correlation coefficient between the official TREC [3] systems’ ranking and the systems’ ranking using the crowdsourced

test collection. We report the correlation on three measures: normalized discounted cumulative gain, mean average precision and R-Precision on both binary and ternary relevance. Overall, the correlation between the two rankings is moderate, while it drops as we focus more on the top-ranked systems. This is an indication that crowd workers and NIST professionals agree to a significant level, but yet they often also contribute different aspects of relevance to the final dataset. We leave as future work the possible expansion of expert data with crowd data for the purpose of capturing diversity.

7 CONCLUSION AND FUTURE WORK

In this paper we proposed a crowdsourcing methodology for annotating and creating a reliable topical relevance test collection. It is empirically derived from various crowdsourcing annotation pilots based on established scientific literature. In these pilots we experiment with multiple independent variables such as the relevance scale, the document granularity, the annotation guidelines and the number of annotators. The research is guided by two question that focus on improving the accuracy and the reliability of topical relevance annotations. On the one hand, we show that we can improve the accuracy of topical relevance assessment annotations by adapting the annotation guidelines and the document granularity used to assess the relevance. On the other hand, we show that we can improve the reliability of topical relevance assessment annotations by asking many crowd workers to annotate each topic-document pair, and by using an objective binary relevance scale.

Our results show that (1) relevance aggregation on document paragraphs level results in a more accurate relevance, compared to assessing the full document, and (2) collecting relevance judgments with a binary relevance scale results in a higher relevance assessment reliability, compared to the ternary scale used in the TREC annotation guidelines. Based on these evidences, we apply the best performing crowdsourcing annotation template and setting at scale, in order to create a topical relevance test collection.

In future work we plan to extend this test collection with longer documents in the NYT Corpus. We will consider some of the optimization techniques mentioned in this paper, that are related to the number of paragraphs the crowd is asked to assess at once. We also plan to look into different approaches to determine the optimal thresholds for categorizing documents as relevant (highly relevant and relevant) and not relevant, without the need to have expert reviewers or NIST assessors' annotations. Furthermore, our purpose is to use the final topical relevance test collection for testing and evaluating various IR systems and to better understand how crowd-sourced topical relevance influences the ranking of the systems. We also plan to perform this analysis by looking at different topics.

8 ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their valuable comments. The authors would also like to thank the anonymous crowd workers that participated in the crowdsourcing tasks. This work was supported by the IBM PhD Fellowship program, as well as by "QROWD - Because Big Data Integration is Humanly Possible" (grant 732194), "STARS4ALL - A Collective Awareness Platform for Promoting Dark Skies in Europe" projects (grant 688135), the Bloomberg Research Grant program, and the Google Faculty Research Awards program. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

- [1] A. Al-Maskari, M. Sanderson, and P. Clough. Relevance judgments between trec and non-trec assessors. In *SIGIR*, pages 683–684. ACM, 2008.
- [2] J. Allan. Overview of the trec 2004 robust retrieval track. In *TREC*, volume 13, 2004.
- [3] J. Allan, E. K. Donna Harman, D. Li, C. V. Gysel, and E. Voorhees. Trec 2017 common core track overview. In *TREC*, 2017.
- [4] O. Alonso and R. Baeza-Yates. Design and implementation of relevance assessments using crowdsourcing. In *ECIR*, pages 153–164. Springer, 2011.
- [5] O. Alonso and S. Mizzaro. Can we get rid of trec assessors? using mechanical turk for relevance assessment. In *SIGIR Workshop on the Future of IR Evaluation*, volume 15, page 16, 2009.
- [6] O. Alonso and S. Mizzaro. Using crowdsourcing for trec relevance assessment. *IPM*, 48(6):1053–1066, 2012.
- [7] L. Aroyo and C. Welty. Crowd Truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *ACM Web Science*, 2013.
- [8] L. Aroyo and C. Welty. The Three Sides of CrowdTruth. *Journal of Human Computation*, 1:31–34, 2014.
- [9] 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 *SIGIR*, pages 667–674. ACM, 2008.
- [10] A. Barrón-Cedeño, G. Da San Martino, S. Filice, and A. Moschitti. On the use of an intermediate class in boolean crowdsourced relevance annotations for learning to rank comments. In *SIGIR*, pages 1209–1212. ACM, 2017.
- [11] J. P. Callan. Passage-level evidence in document retrieval. In *SIGIR*, pages 302–310. Springer-Verlag New York, Inc., 1994.
- [12] B. Carterette. Robust test collections for retrieval evaluation. In *SIGIR*, pages 55–62. ACM, 2007.
- [13] P. Clough, M. Sanderson, J. Tang, T. Gollins, and A. Warner. Examining the limits of crowdsourcing for relevance assessment. *IEEE Internet Computing*, 17(4):32–38, 2013.
- [14] K. Collins-Thompson, C. Macdonald, P. Bennett, F. Diaz, and E. M. Voorhees. Trec 2014 web track overview. Techn. report, University of Michigan, Ann Arbor, 2015.
- [15] T. T. Damessie, T. P. Nghiem, F. Scholer, and J. S. Culpepper. Gauging the quality of relevance assessments using inter-rater agreement. In *SIGIR*, pages 1089–1092. ACM, 2017.
- [16] A. Dumitrache, L. Aroyo, and C. Welty. Crowdsourcing ground truth for medical relation extraction. *ACM TIS*, 2017.
- [17] A. Dumitrache, O. Inel, B. Timmermans, C. Ortiz, R.-J. Sips, and L. Aroyo. Empirical methodology for crowdsourcing ground truth. *Semantic Web Journal, Special Issue on Human Computation and Crowdsourcing (in review)*, 2017.
- [18] A. Dumitrache, O. Inel, L. Aroyo, B. Timmermans, and C. Welty. Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement. 2018.
- [19] C. Grady and M. Lease. Crowdsourcing document relevance assessment with mechanical turk. In *NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 172–179, 2010.
- [20] J. Knowlton. On the definition of "picture". *AV Communication Review*, 14(2): 157–183, 1966.
- [21] E. Maddalena, S. Mizzaro, F. Scholer, and A. Turpin. On crowdsourcing relevance magnitudes for information retrieval evaluation. *TOIS*, 35(3):19, 2017.
- [22] T. McDonnell, M. Lease, T. Elsayad, and M. Kutlu. Why is that relevant? collecting annotator rationales for relevance judgments. In *HCOMP*, page 10, 2016.
- [23] Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.
- [24] S. I. Moghadasi, S. D. Ravana, and S. N. Raman. Low-cost evaluation techniques for information retrieval systems: A review. *Journal of Informetrics*, 7(2):301–312, 2013.
- [25] M. Sanderson et al. Test collection based evaluation of information retrieval systems. *FNTIR*, 4(4):247–375, 2010.
- [26] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *EMNLP*, pages 254–263, 2008.
- [27] E. Sormunen. Liberal relevance criteria of trec-: Counting on negligible documents? In *SIGIR*, pages 324–330. ACM, 2002.
- [28] K. Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.
- [29] R. Tang, W. M. Shaw Jr, and J. L. Vevea. Towards the identification of the optimal number of relevance categories. *JASIST*, 50(3):254, 1999.
- [30] A. Trotman, N. Pharo, and D. Jenkinson. Can we at least agree on something? In *SIGIR Workshop on Focused Retrieval*, pages 49–56, 2007.
- [31] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *IPM*, 36(5):697–716, 2000.
- [32] E. M. Voorhees. The philosophy of information retrieval evaluation. In *CLEF*, volume 1, pages 355–370. Springer, 2001.
- [33] E. M. Voorhees and D. Harman. Overview of trec 2001. In *TREC*, 2001.
- [34] R. M. White, I. Ruthven, and J. M. Jose. Finding relevant documents using top ranking sentences: an evaluation of two alternative schemes. In *SIGIR*, pages 57–64. ACM, 2002.
- [35] E. Yilmaz, J. A. Aslam, and S. Robertson. A new rank correlation coefficient for information retrieval. In *SIGIR*, pages 587–594. ACM, 2008.