Gauging the Quality of Relevance Assessments using Inter-Rater Agreement

Tadele T. Damessie
RMIT University
Melbourne, Australia

Thao P. Nghiem
RMIT University
Melbourne, Australia

Falk Scholer
RMIT University
Melbourne, Australia

J. Shane Culpepper
RMIT University
Melbourne, Australia

ABSTRACT

In recent years, gathering relevance judgments through non-topic originators has become an increasingly important problem in Information Retrieval. Relevance judgments can be used to measure the effectiveness of a system, and are often needed to build supervised learning models in learning-to-rank retrieval systems. The two most popular approaches to gathering *bronze* level judgments – where the judge is not the originator of the information need for which relevance is being assessed, and is not a topic expert – is through a controlled user study, or through crowdsourcing. However, judging comes at a cost (in time, and usually money) and the quality of the judgments can vary widely. In this work, we directly compare the reliability of judgments using three different types of bronze assessor groups. Our first group is a controlled *Lab* group; the second and third are two different crowdsourcing groups, *CF-Document* where assessors were free to judge any number of documents for a topic, and *CF-Topic* where judges were required to judge all of the documents from a single topic, in a manner similar to the *Lab* group. Our study shows that *Lab* assessors exhibit a higher level of agreement with a set of ground truth judgments than *CF-Topic* and *CF-Document* assessors. Inter-rater agreement rates show analogous trends. These finding suggests that in the absence of ground truth data, agreement between assessors can be used to reliably gauge the quality of relevance judgments gathered from secondary assessors, and that controlled user studies are more likely to produce reliable judgments despite being more costly.

1 INTRODUCTION

Gathering relevance judgments using humans is a key component in building Information Retrieval test collections. However, human interpretation of "relevance" is an inherently subjective process [11]. According to Tang and Solomon [16], judging relevance is a dynamic, multidimensional process likely to vary between assessors, and sometimes even with a single assessor at different stages of the process. For example, Scholer et al. [13] found that 19% of duplicate document pairings were judged inconsistently in the TREC-7 and TREC-8 test collections. Understanding the factors that lead to such variation in relevance assessments is crucial to reliable test collection development.

To address this issue, Bailey et al. [3] proposed three classes of judges – gold, silver and bronze – based on the expertise of the assessor. *Gold* judges are topic originators as well as subject experts; whereas *silver* judges are subject experts but not topic originators. *Bronze* judges are neither topic originators nor subject experts. But are all judges in a single class really the same? Secondary assessors who are neither topic creators nor experts are all bronze assessors, but there are in fact many different types of assessors who fall into this class. As assessment at the bronze level is now becoming a common practice in IR, in particular with the growing popularity of crowdsourcing, we set up an experiment to investigate the homogeneity of assessment quality using three different variants of bronze judges. The classes used in this study are:

- *Lab*: This group of assessors carried out a relevance assessment task in a monitored lab environment, with a requirement to assess a pre-determined number of 30 documents in relation to a single search topic.
- *CF-Topic*: This group of assessors are an exact replica of the *Lab* group task except that the task was administered using the CrowdFlower crowdsourcing platform.
- *CF-Document*: This group of assessors performed the task using CrowdFlower just as the *CF-Topic* group, but unlike the other two groups, each participant could judge as few (minimum 1) or as many (maximum 30) documents as they liked for a topic.

Our main research question can formally be stated as:

Research Question: *Are there differences in the quality of relevance judgments gathered from different sub-classes of bronze-level judges?*

2 RELATED WORK

The subjective nature of relevance is likely to result in disagreement between judges [11, 15]. Voorhees [18] was among the first to study this phenomenon, and quantified agreement in relevance assessment using overlap between primary TREC assessors and two secondary assessors on 48 topics. A total of 30% of the documents judged relevant by the primary assessor were judged non-relevant, and less than 3% of the documents initially judged as non-relevant by the primary assessor were judged relevant by the secondary assessors.

SIGIR'17, August 7–11, 2017, Shinjuku, Tokyo, Japan

978-1-4503-5022-8/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3077136.3080729

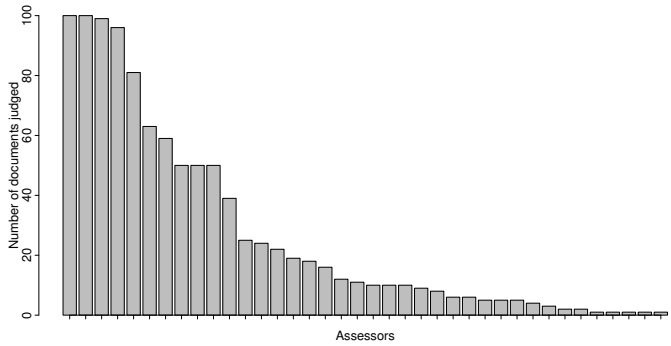

Figure 1: Distribution of number of documents judged per assessor by the *CF-Document* group.

Sormunen [14] also compared judgments from a group of 9 master's students using a 4-point ordinal relevance scale with TREC binary assessments on 38 topics. Around 25% of the documents originally judged as relevant by the TREC assessors were re-assessed as non-relevant, and around 1% of the documents originally judged as non-relevant were re-assessed as relevant. Al-Maskari et al. [1] also ran an experiment on 56 TREC-8 topics using 56 participants in an interactive search task. The study found a 37% difference between TREC and non-TREC assessors. That is, out of the 2,262 documents judged relevant by the non-TREC assessors, 834 of the documents were judged non-relevant by the TREC assessors. In both studies, there is a clear difference between the TREC assessors who are topic originators, and the non-TREC assessors who often are not.

To address differences between TREC judges and secondary assessors, Bailey et al. [3] identified three classes of judges – gold, silver and bronze – as discussed in Section 1. Bailey et al. found that assessments generated by silver judges were often comparable to gold judges, but that extra care was needed when using bronze level judgments. However, the study did not prescribe exactly how this might be accomplished. In this study, we focus on different types of bronze level of assessors, as they now represent the most common class of judges outside of evaluation campaigns such as TREC which are being employed in large scale assessment gathering initiatives.

3 METHODS AND DATASETS

The TREC 7 and 8 datasets are used in this study. We focus on topics from these collections since they are widely believed to be among the most complete collections available [10], and provide a strong ground truth when attempting to quantify reliability in re-assessment exercises. Our work builds on two previous studies using the same topic configuration, and which provide further details about the user study configuration [5, 6]. We use 4 different topics: the 2 highest and 2 lowest performing topics from the dataset were selected using the average precision of each topic, averaged over the 110 runs submitted to TREC 2004 Robust track. This approach, called average-average-precision (AAP), was initially described by Carterette et al. [4], and used to quantify topic difficulty. Topic #365 (`el nino`) and #410 (`schengen agreement`) have the 2 highest AAP scores, and topic #378 (`euro opposition`) and #448 (`ship losses`) are the 2 lowest AAP scoring topics in the collection. For assessment, 30 documents were chosen for each topic, in proportion to an existing distribution of graded document relevance judgments made by Sormunen [14].

A total of 32, 40 and 43 assessors judged documents in the *Lab*, *CF-Topic* and *CF-Document* experimental groups, respectively. For all crowdsourcing experiments, a mandatory explanation of relevance assignment per document was required, and manually checked as a quality control, to ensure that crowdsourcing participants were performing assessments in good faith. A total of 10 assessors, 5 from *CF-Topic* and 5 from *CF-Document* failed the sanity check, and their data was removed from the final evaluation. All crowdsourcing experiments were conducted using the CrowdFlower platform in a manner similar to previously run studies [2].

The setup for the *CF-Document* group was designed to be as flexible as possible, with assessors free to judge any number of the 30 documents for any of the 4 topics which were assigned randomly by the system. This setup introduces challenges during final data analysis, however, since assessors judged an unequal number of documents, as shown in Figure 1, and a comparison of agreement between assessors with the same level of precision requires an incomplete balanced block design to be constructed as described by Fleiss [7]. This results in a sparse matrix of relevance scores for the maximum number of unique documents (30 per topic in our case) across the 121 unique assessors who contributed judgments.

Krippendorff's Alpha (α) is a chance-corrected measure of agreement, and not affected by differences in sample sizes or missing values, and therefore appropriate for analysis of our experimental data [8]. Cohen's Kappa (κ) which is more suited for categorical data [17] is also used to quantify assessment quality against a gold standard. The values produced by these metrics is between -1 and 1, where a level of 0 indicates agreement at the level predicted by chance, 1 signifies perfect agreement between raters, and a negative score occurs when agreement is less than what is expected by chance alone.

4 RESULTS AND DISCUSSION

Assessor reliability – measured by the mean pairwise agreement between each assessor and the Sormunen gold standard assessments – is used to assess the quality of the assessments from each experimental group. This analysis is then compared with a measure of assessment quality using only inter-rater agreement, in the absence of any ground truth.

Assessor Reliability. The pairwise overall average reliability score of the *Lab*, *CF-Topic* and *CF-Document* groups, measured using [Krippendorff's α, Cohen's κ] is [0.687, 0.581], [0.407, 0.236] and [0.561, 0.522] respectively. The κ scores are calculated on binary foldings of the 4-level graded relevance levels – non-relevant (0), marginally relevant (1), relevant (2) and highly relevant (3). The marginally relevant (1) and non-relevant (0) judgments are binarized as non-relevant and the others as relevant as recommended by Scholer and Turpin [12].

The results in Table 1 indicate *Lab* and *CF-Document* assessors are more reliable than *CF-Topic* assessors. The statistical significance of the differences is evaluated using an unpaired two-tailed t-test across the individual pairwise agreement scores, and reported in Table 2. For both α and κ, the overall pattern from highest to lowest reliability score measured using the Sormunen judgments

Table 1: Average pairwise agreement between judges and Sormunen gold standard judgments, measured across All and individual topics using Krippendorff's Alpha (α) on a 4-levels of ordinal scale and Cohen's Kappa(κ) on a binary scale.

	Krippendorff's Alpha (α)			Cohen's Kappa(κ)		
	Lab	CF-Topic	CF-Document	Lab	CF-Topic	CF-Document
All	0.687	0.407	0.561	0.581	0.236	0.522
`el nino`	0.843	0.531	0.725	0.761	0.277	0.599
`schengen agreement`	0.622	0.057	0.380	0.558	0.111	0.410
`euro opposition`	0.665	0.437	0.377	0.436	0.112	0.391
`ship losses`	0.617	0.561	0.704	0.565	0.416	0.666

Table 2: Statistical significance of Table 1 results, evaluated using an unpaired two-tailed t-test for all bronze assessors. Results for Krippendorff's Alpha (α) are shown below the diagonal line with ratings on a 4-level ordinal scale, while results for Cohen's Kappa (κ) are shown above the diagonal line with ratings on a binary scale, flattening 0 and 1 to 0; and 2 and 3 to 1.

	Lab	CF-Topic	CF-Document
Lab	$[\alpha = 0.687/\kappa = 0.581]$	$95\%_\kappa$ CI 0.211, 0.479 Lab_κ (M=0.581, SD=0.308) CF-Topic_κ (M=0.236, SD=0.244) $t(65)_\kappa = 5.082, p < 0.001$	$95\%_\kappa$ CI −0.099, 0.216 Lab_κ (M=0.581, SD=0.308) $\text{CF-Document}_\kappa$ (M=0.522, SD=0.347) $t(68)_\kappa = 0.739, p = 0.462$
CF-Topic	$95\%_\alpha$ CI 0.123, 0.435 Lab_α (M=0.687, SD=0.214) CF-Topic_α (M=0.407, SD=0.390) $t(65)_\alpha = 3.583, p < 0.001$	$[\alpha = 0.407/\kappa = 0.236]$	$95\%_\kappa$ CI −0.426, −0.144 $\text{CF-Document}_\kappa$ (M=0.522, SD=0.347) CF-Topic_κ (M=0.236, SD=0.244) $t(71)_\kappa = -4.026, p < 0.001$
CF-Document	$95\%_\alpha$ CI −0.142, 0.266 Lab_α (M=0.687, SD=0.214) $\text{CF-Document}_\alpha$ (M=0.561, SD=0.345) $t(68)_\alpha = 1.793, p = 0.077$	$95\%_\alpha$ CI −0.325, 0.018 CF-Topic_α (M=0.407, SD=0.390) $\text{CF-Document}_\alpha$ (M=0.561, SD=0.345) $t(71)_\alpha = -1.781, p = 0.079$	$[\alpha = 0.561/\kappa = 0.522]$

as a baseline is: *Lab*, *CF-Document* and *CF-Topic* respectively. One explanation for this trend might be that the *Lab* study is a more directed environment, and assessors know that they are being closely monitored the entire time. This could contribute to longer periods of focus, resulting in a higher overall agreement with the gold standard, and therefore a presumed higher overall quality of obtained judgments.

When comparing only the two crowdsourcing groups, the *CF-Document* assessors show higher reliability. This is a somewhat surprising result, since the judges assess fewer documents and therefore spend less time overall forming a notion of relevance for a particular topic. However, this lack of "domain knowledge" might be counteracted by task completion time: an assessor in *CF-Topic* had to judge all 30 documents to get paid, and when an assessor encounters long or difficult documents at the tail of an assessment list, the likely outcome is that the assessor becomes less motivated to get any single judgment exactly right. Fatigue and motivation are known to influence relevance judgment outcomes [9, 19], and perhaps contribute to the drop in quality. In contrast, *CF-Document* assessors may perceive that less effort is required on their behalf to judge a single topic-document pair before getting paid. These "micro" transactions could very well be a strong motivator for crowdsourced assessors, despite having an implicit startup cost in understanding the task at hand that is amortized when judging multiple documents for the same topic. We plan to study this phenomenon in more detail in future work.

Figure 2 and Figure 3 give further insight on the reliability levels (agreement with the gold standard) of individual *CF-Topic* and *CF-Document* assessors, respectively. Results for the *Lab* group were omitted due to space limitations; the reliability score for this group was consistently well above $\alpha > 0.2$, with no negative scores for any assessors. A number of assessors in *CF-Topic* showed lower levels of agreement with the gold standard than expected by chance alone for 2 of the topics as shown in Figure 2. Reliability for the other 2 topics in this group is similar to the trend observed for the *Lab* assessors. Only one assessor's relative performance in the *CF-Document* setup deviated significantly from the others, as shown in Figure 3. We plan to further investigate the reasons for why such low reliability scores were observed for some individual assessors in these groups in followup work. Note that all of these assessors passed manual sanity control measures, and appeared to be performing judgments in good faith.

Agreement. As can been seen in Table 3, overall agreement is higher in *Lab*, followed by *CF-Document* and *CF-Topic*, which are in the same relative order as the reliability scores when comparing against a gold standard, suggesting that inter-rater reliability is a reasonable proxy for the quality of judgments.

To further establish our belief of assessor reliability, we computed the median of the multiple assessments made for each document in each experimental group, and computed the Krippendorff's Alpha (α) agreement between individual assessors and this score, shown in Table 3 (right). The overall trend is again consistent with the findings of Table 1.

Table 3: Inter-rater agreement (left) and majority vote (right) measured between assessors in the Lab, CF-Topic and CF-Document groups using Krippendorff's alpha (α) across All and individual topics with ratings on a 4-level ordinal scale. The number of assessors for inter-rater agreement is shown in parenthesis next to each α value.

	Inter-rater agreement			Majority vote		
Topic	Lab	CF-Topic	CF-Document	Lab	CF-Topic	CF-Document
All	0.657 (32)	0.426 (35)	0.530 (121)	0.787	0.544	0.663
`el nino`	0.845 (8)	0.394 (8)	0.682 (31)	0.917	0.608	0.771
`schengen agreement`	0.634 (8)	0.170 (8)	0.500 (29)	0.691	0.436	0.542
`euro opposition`	0.565 (8)	0.464 (9)	0.431 (29)	0.867	0.537	0.599
`ship losses`	0.558 (8)	0.377 (10)	0.471 (32)	0.710	0.605	0.799

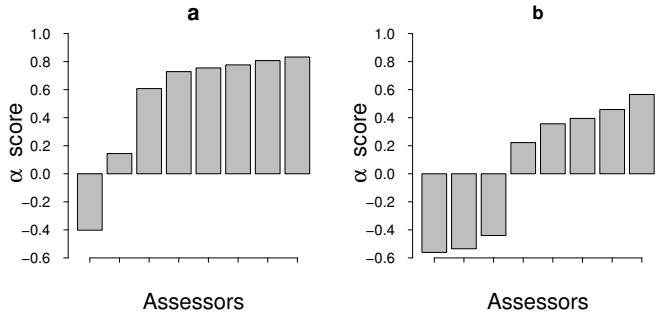

Figure 2: Reliability of *CF-Topic* assessors when compared with the Sormunen judgments using Krippendorff's Alpha (α) for the topics: (a) El nino; and (b) Schengen agreement.

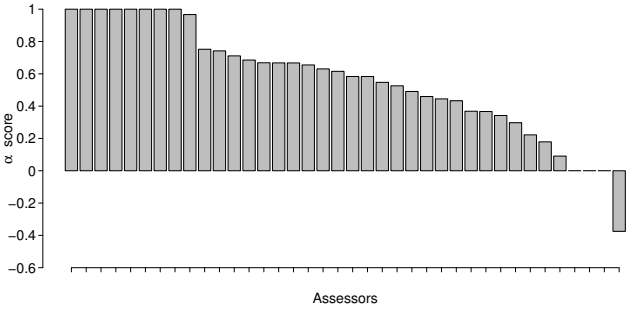

Figure 3: Reliability of *CF-Document* assessors when compared to the Sormunen judments using Krippendorff's Alpha (α).

Getting gold standard relevance labels is rarely possible in a live judging scenario, but it is possible to compute inter-rater agreement between assessors, and use this to establish the quality of assessments. Our experiments confirm that using agreement between judges to gauge the quality of relevance judgments collected is indeed one possible approach to controlling the quality of judgments gathered by bronze level assessors.

5 CONCLUSION

This study analyzed the quality of relevance judgments generated in three (of many possible) different sub-classes of bronze assessors, using Krippendorff's Alpha (α) and Cohen's Kappa (κ). The results of both metrics confirm the existence of assessment quality differences among the three sub-classes of bronze assessors, warranting further study. Nevertheless, inter-rater agreement can be a reliable tool to benchmark the quality of relevance judgments when gold standard judgments are not readily available.

Acknowledgment. This work was supported by the Australian Research Council's *Discovery Projects* Scheme (DP170102231 and DP140101587).

REFERENCES

[1] A. Al-Maskari, M. Sanderson, and P. Clough. 2008. Relevance judgments between TREC and Non-TREC assessors. In *Proc. SIGIR*. 683–684.
[2] O. Alonso and S. Mizzaro. 2012. Using crowdsourcing for TREC relevance assessment. *Inf. Proc. & Man.* 48, 6 (2012), 1053–1066.
[3] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A.P. de Vries, and E. Yilmaz. 2008. Relevance assessment: are judges exchangeable and does it matter. In *Proc. SIGIR*. 667–674.
[4] B. Carterette, V. Pavlu, H. Fang, and E. Kanoulas. 2009. Million Query Track 2009 Overview.. In *Proc. TREC*.
[5] T.T. Damessie, F. Scholer, and J.S. Culpepper. 2016. The Influence of Topic Difficulty, Relevance Level, and Document Ordering on Relevance Judging. In *Proc. ADCS*. 41–48.
[6] T.T. Damessie, F. Scholer, K. Järvelin, and J.S. Culpepper. 2016. The effect of document order and topic difficulty on assessor agreement. In *Proc. ICTIR*. 73–76.
[7] J.L. Fleiss. 1981. Balanced incomplete block designs for inter-rater reliability studies. *Applied Psychological Measurement* 5, 1 (1981), 105–112.
[8] A.F. Hayes and K. Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Comm. Methods and Measures* 1, 1 (2007), 77–89.
[9] G. Kazai, J. Kamps, and N. Milic-Frayling. 2013. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Inf. Retr.* 16, 2 (2013), 138–178.
[10] X. Lu, A. Moffat, and J. S. Culpepper. 2016. The effect of pooling and evaluation depth on IR metrics. *Inf. Retr.* 19, 4 (2016), 416–445.
[11] T. Saracevic. 2007. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *J. Amer. Soc. Inf. Sc. Tech.* 58, 13 (2007), 1915–1933.
[12] F. Scholer and A. Turpin. 2009. Metric and relevance mismatch in retrieval evaluation. In *Proc. AIRS*. 50–62.
[13] F. Scholer, A. Turpin, and M. Sanderson. 2011. Quantifying test collection quality based on the consistency of relevance judgements. In *Proc. SIGIR*. 1063–1072.
[14] E. Sormunen. 2002. Liberal relevance criteria of TREC-: Counting on negligible documents?. In *Proc. SIGIR*. 324–330.
[15] M. Stefano. 1997. Relevance: The whole history. *J. Amer. Soc. Inf. Sc.* 48, 9 (1997), 810–832.
[16] R. Tang and P. Solomon. 1998. Toward an understanding of the dynamics of relevance judgment: An analysis of one person's search behavior. *Inf. Proc. & Man.* 34, 2 (1998), 237–256.
[17] A.J. Viera and J.M. Garrett. 2005. Understanding interobserver agreement: the kappa statistic. *Fam. Med.* 37, 5 (2005), 360–363.
[18] E.M. Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Inf. Proc. & Man.* 36, 5 (2000), 697–716.
[19] J. Wang. 2011. Accuracy, agreement, speed, and perceived difficulty of users' relevance judgments for e-discovery. In *Proc. of SIGIR Inf. Ret. for E-Discovery Workshop.*, Vol. 1.