The Effect of Document Order and Topic Difficulty on Assessor Agreement

Tadele T. Damessie
RMIT University
Melbourne, Australia
tadeletedla.damessie@rmit.edu.au

Falk Scholer
RMIT University
Melbourne, Australia
falk.scholer@rmit.edu.au

Kalervo Järvelin
University of Tampere
Tampere, Finland
kalvero.jarvelin@uta.fi

J. Shane Culpepper
RMIT University
Melbourne, Australia
shane.culpepper@rmit.edu.au

ABSTRACT

Human relevance judgments are a key component for measuring the effectiveness of information retrieval systems using test collections. Since relevance is not an absolute concept, human assessors can disagree on particular topic-document pairs for a variety of reasons. In this work we investigate the effect that document presentation order has on inter-rater agreement, comparing two presentation ordering approaches similar to those used in IR evaluation campaigns: decreasing relevance order and document identifier order. We make a further distinction between "easy" topics and "hard" topics in order to explore system effects on inter-rater agreement. The results of our pilot user study indicate that assessor agreement is higher when documents are judged in document identifier order. In addition, there is higher overall agreement on easy topics than on hard topics.

Keywords

Experimentation; measurement; relevance; assessor agreement; ordering effects

1. INTRODUCTION

Test collections are widely used for the evaluation of information retrieval system effectiveness. A key component of the approach is a set of relevance judgments, indicating which documents are considered to be appropriate answers in response to a topic. Differences in human judgments can potentially lead to alternative conclusions about system effectiveness, and so the question of judgment consistency is an important consideration for the generality of the conclusions that can be drawn from a test collection. Judgment consistency can be measured by calculating the *inter-rater agreement* between different human assessors who are asked to judge the relevance of common topic-document pairs in response to an information need. Given an information need, disagreement can exist between assessors as to which of the documents presented are relevant [4], and previous work has shown that topic familiarity [1], topic knowledge [11], document-specific factors [7], and the degree of relevance of documents that are presented early in the judging process [16], are all factors that can affect agreement.

In this work, we investigate the impact of presentation order on assessors when making relevance judgments, as measured by inter-rater agreement. We also analyze the stability of agreement in the context of topic difficulty, from both the system and user perspective. This issue has direct implications in information retrieval evaluation, and the design and construction of test collections [12]. Two document orderings are considered: decreasing relevance order, where documents are sorted from most to least relevant; and document identifier order, where documents are sorted by document identifier. We refer to these two orderings as *Rel order* and *DocID order* in the remainder of this paper. *Rel order* is similar to the approach used by the NTCIR evaluation campaign, where a pool of documents to be judged is sorted in decreasing *expected* relevance order, based on the number of participating systems that retrieved a document [13]. Here we use existing relevance judgments to ensure decreasing relevance order. *DocID order* is similar to that used by the TREC evaluation campaign, where the distribution of relevant documents can vary widely from topic to topic. Given these orderings, our main research questions are:

Research Quesiton (1): *Does the presentation order of documents for relevance judging affect inter-rater agreement?*

Research Quesiton (2): *Does topic difficulty influence inter-rater agreement with respect to the different presentation orderings?*

2. RELATED WORK

Several previous studies have investigated factors that influence assessor agreement. Bailey et al. [1] showed that agreement may be influenced by topic familiarity and topic origination. Their analysis demonstrated that assessors who are neither topic originators nor topic experts show more variation in their relevance judgments than those who are either topic originators or topic experts or both. Other studies have also shown a correlation between assessor experience and high agreement [3, 14].

Sormunen [17] compared the assessments of 5,271 documents from 38 topics chosen from TREC-7 and TREC-8 by a team of 9 master's students using a 4-point ordinal relevance scale with previous ratings from NIST assessors on a binary relevance scale. Of the TREC-relevant documents (2,772), 13% were re-assessed as

ICTIR'16 , September 12–16, 2016, Newark, DE, USA

ACM 978-1-4503-4497-5/16/09 ...$15.00.
http://dx.doi.org/10.1145/2970398.2970431.

highly relevant, 26% as relevant, 36% as marginally relevant and 25% as non-relevant. When the distribution of relevance agreement between TREC and the rejudged documents were compared, 25% of the documents rated relevant by TREC assessors were re-judged as non-relevant, 36% were judged to be marginally relevant, and 1% originally found to be non-relevant in the TREC assessment were judged relevant by the Sormunen assessors. These results clearly show the existence of disagreement between the two sets of assessors.

Huang and Wang [9] and Eisenberg and Barry [6] investigated the relationship between document order and relevance judgments. They found that document relevance is underestimated when documents are ordered from high relevance to low relevance, and document significance is overestimated if the order of documents is in reverse relevance order. Scholer et al. [16] studied the impact of the relevance of documents that are seen early in the judging process on the levels of relevance assigned to later documents, and concluded that presenting documents of varying relevance levels to assessors early on the judging process allowed assessors to calibrate their relevance thresholds.

Voorhees [19] studied the impact of assessor disagreement on retrieval system evaluation in terms of changes in system effectiveness rankings. Despite the existence of assessor disagreement, the study concluded that the relative effectiveness of systems is broadly stable. Though the study acknowledged the presence of disagreement between assessors, it did not investigate what causes the disagreement, or the implications on document ranking. In this work, we do not investigate how disagreement influences document ranking, but we do analyze the impact of document order on assessor agreement. Scholer et al. [15] studied assessment errors in relevance judgment files (or qrels) and demonstrated that inconsistencies in document assessments increase with time between judgments. However, they did not investigate how the ordering of documents affects assessor judgments.

Xu and Wang [20] examined cognitive aspects of relevance such as learning, subneed scheduling, and fatigue in a simulated retrieval task, and concluded that order effect on these cognitive aspects is weak at a document list length of 40. Xu and Wang suggested the need for further research on order effects and cognition.

In contrast to previous work, our current work focuses on a different aspect of inter-rater agreement – does document presentation ordering, or varying system topic difficulty, lead to a measurable impact on inter-rater agreement?

3. METHODOLOGY

To study the effect of document ordering on assessor agreement, we carried out a small scale user study using 120 documents, and 4 topics from the TREC-7 and TREC-8 collections that were judged both by NIST assessors on a binary scale, and later by Sormunen [17] using a 4-point scale. The graded relevance judgments from Sormunen are used as the ground truth in this experiment. The grades of the scale are: *Highly relevant (3)*, *Relevant (2)*, *Marginally relevant (1)*, and *Not relevant (0)*.

Query and Document Selection. For our study, we selected a mixture of hard and easy search topics, since we hypothesized that topic difficulty may have an effect on agreement. Following the approach of Carterette et al. [2] in the TREC Million Query track, we classified topic difficulty based on the per-topic *Average-Average-Precision* (AAP) scores (that is, average precision for each individual topic across a set of retrieval systems). We refer to this as *system topic difficulty* in the remainder of this paper. In our experiment, AAP was calculated for the topics of the 2004 Robust track (which included the TREC-7 and TREC-8 topics with dual binary and ordinal relevance judgments) for the 110 runs that participated in the track. From this ordering, we selected the two highest and the two lowest AAP scoring topics: #365 *el nino* (0.723), #378 *euro opposition* (0.046), #410 *schengen agreement* (0.643), #448 *ship losses* (0.025).

For each of the chosen topics, 30 documents were selected for judging in our user study. The selection process was designed so that the distribution of documents at all four relevance levels in the sample was the same as the relevance distribution of the full set of documents available for each topic. For example, consider topic 365. There are a total of 198 documents, of which 33 are relevant, and the remaining 165 non-relevant. Out of the 33 relevant documents, 24 were judged as marginally relevant, 8 were judged relevant, and 1 was found to be highly relevant in the relevance judgement file (qrel) file. Given this distribution, the proportional selection was 25, 4, 1 and 0 for non-relevant, marginally relevant, relevant and highly relevant respectively. However, it is important to include at least one document of each relevance level whenever possible, in order to ensure that there is a clear distinction between relevant and non-relevant documents; therefore a minimum of 1 document at each relevance level was included in the final selection.

Assessment Interface. An online assessment system was developed to gather relevance assessments from participants. At the top of the screen, the system displayed a search topic, including the title, description and narrative of the official TREC topic statement. Below this, a single document was displayed. An assessor could enter a response by clicking on a radio button, to indicate their relevance assessment on the 4-point Sormunen scale. After selecting a relevance level, the user could click a button to record their judgment and move on to the next document. The system did not allow users to go back and change the ratings given to previous documents as presentation order is the key control variable in the study. Thus, a strict judging ordering was enforced by the assessment tool.

In addition to judging the 30 documents, we also required each assessor to rate their familiarity with the topics, the clarity of the topic description, and their confidence in identifying relevant documents, on a 5-point scale, both before and after the assessment exercise.

User study. A total of 16 participants were recruited from RMIT University to take part in the experiment, and were between the ages of 25 and 35. The study was approved by the RMIT University Ethics Board. All participants were computer science students, and all indicated familiarity with online searching in a pre-experiment questionnaire.

After arriving at the lab where the study was conducted, each participant was given an introduction to the experiment and a brief explanation of the online judging system. A hard copy of the definitions for each of the four relevance level was given to each judge at the start of the experiment, and was available for reference throughout the study. The main task required each participant to judge the relevance of documents for two topics. Participants were given up to one hour to judge each topic (one at the system easy and hard level), with a short break between the two topics. Topics and conditions were rotated when assigned to participants to control for possible ordering and learning effects [10]. The main factors being investigated in this study are document ordering – *Rel order* and *DocID order* – and topic difficulty – *easy* and *hard*. As participants judged documents for two topics, this led to eight combinations. The same process is repeated for the remaining two topics, giving

Table 1: Inter-rater agreement measured using Krippendorff's α for *Rel order* and *DocID order* presentation of documents, with ratings on a 4-level ordinal scale.

	Relevance	*DocID*	Assessors
All	0.570	0.700	16
Easy	0.668	0.746	8
Hard	0.473	0.655	8
Easy (`el nino`)	0.842	0.858	4
Easy (`schengen agreement`)	0.548	0.656	4
Hard (`euro opposition`)	0.428	0.612	4
Hard (`ship losses`)	0.417	0.672	4

a total number of 16 combinations (and assessors) for our experiment.

4. RESULTS AND DISCUSSION

Krippendorff's alpha(α) is a chance-corrected measure of rater agreement that takes into account the type of data (ordinal, nominal, interval or ratio) being measured, and adjusts to different sample and group sizes [5, 8]. The value of the α coefficient is bounded between -1 and 1, where zero indicates the absence of agreement (that is, observed agreement is equal to the level of agreement expected by chance), while 1 indicates perfect agreement between assessors. A negative value indicates that disagreement surpasses what is expected by random chance.

The agreement results from our user study as measured using Krippendorff's α are shown in Table 1. The overall level of agreement across all four topics is 0.57 for *Rel order* and 0.7 for *DocID order*: assessors in our study agreed more on relevance when they were shown documents in a *DocID order* based on a TREC document ID than when shown documents in decreasing relevance order. Splitting the topics into easy and hard groups (rows 2 and 3 of the table) shows that this effect is consistent: *DocID order* presentation leads to higher agreement in both cases. However, the difference in α is larger for the hard topics, suggesting that the choice of ordering plays a larger role when topics are difficult, and it would seem that documents for hard topics are harder to agree on than documents for the easy topics.

The results may appear surprising at first glance. Intuitively, when documents are shown in decreasing relevance order, one might expect that it is easier for assessors to recognize similar sources of evidence that are presented close together, and that they therefore give more similar ratings. However, an alternative interpretation is that variation plays an important role identifying relevant documents. For example, after seeing a number of non-relevant documents, a subsequent document that includes some relevant material may become easier to spot. This would lead to higher overall agreement between assessors.

To investigate this further, Figure 1 shows the presentation order and judging results for each of the 8 topic and ordering combinations. Each plot shows the 30 documents that were presented to assessors along the x-axis, and the corresponding relevance levels on the y-axis. The purple line shows the ground truth (Sormunen) relevance label, while the four colored bars show the judgments made by each of the four assessors for a particular topic-ordering combination. (Note since each experimental participant only judged two topics, the colors do not represent the same assessor in each graph.)

A particular feature that becomes apparent from the plots is that the *DocID order* for two topics, `el nino`; Figure 1(a) and Figure 1(b); and `ship losses`; Figure 1(g) and Figure 1(h), cluster relevant documents towards the end of the list, approaching a *reverse* relevance ordering. This is an artifact of following the TREC convention of ordering documents by the document ID string. A similar clustering effect is also present in the TREC judgments [15].

Vakkari and Sormunen [18] reported that test subjects are able to recognize highly relevant documents quite consistently, but tend to err on marginal and non-relevant ones. Sormunen [17] also found inconsistency of assessment between neighboring relevance levels. This concern motivates the need to assess user agreement on a binary scale in addition to a graded relevance scale. When using a binary relevance, the overall trends are similar to those shown for graded relevance, with *DocID order* leading to higher agreement than *Rel order* ($\alpha = 0.557$ for *Rel order*, and $\alpha = 0.673$ for *DocID order*).

Agreement between our study participants and the Sormunen judgments can be measured by computing Krippendorff's α between these two groups. The trend for this comparison is also consistent with the findings reported in this work (mean pairwise $\alpha = 0.668$ for *Rel order*, and $\alpha = 0.705$ for *DocID order*). We note that for some documents, the majority of our participants disagree with the Sormunen ratings, as can be seen in the plots in Figure 1; we plan to investigate the sources of this disagreement further in future work.

Finally, we briefly return to the problem of system versus user difficulty. Our study assumes the two are correlated; while this might not always be true, or post-hoc questionnaire provides some evidence that the two are aligned for the queries used in this study. Assessors were asked to answer the question "How easy was it to identify relevant documents for the search topic?" after completing assessments for each topic, with responses made on a five-point Likert scale, ranging from "Extremely easy (4)" to "Not easy at all (0)". The boxplot in Figure 2 shows the distribution of responses for all 16 assessors, aggregated by system difficulty. As can be seen in the plot, system and user topic difficulty align for the selected topics. We plan to investigate the distinction between system and user difficulty further in future work.

5. CONCLUSION

Relevance judgments are a key component of test collections, and the order in which documents are presented to assessors may influence the judging outcomes. In this work we investigated the influence of two common document orderings – *Rel order* and *DocID order* – on judgement consistency for easy and hard topics, using Krippendorff's α as a measure inter-rater agreement. We also consider the subtle distinction between system and user difficulty, as both can play an important role in the assessment process. The results of our pilot user study show that agreement tends to be higher when documents are presented in *DocID order*. A possible explanation for this effect is that a more mixed presentation of relevance ordering helps to create a "surprise" effect when items of more starkly different relevance levels follow each other, and this surprise effect, being relatively easier to spot, leads to greater overall rating consistency. Interestingly, topic difficulty can amplify this effect. We plan on using the lessons learned in this study to design a more comprehensive comparison of alternative document presentation orderings, and the effects of query difficulty on inter- and intra-rater agreement.

Acknowledgment. This work was supported by the Australian Research Council's *Discovery Projects* Scheme (DP140102655 and DP140103256). Shane Culpepper is the recipient of an Australian Research Council DECRA Research Fellowship (DE140100275).

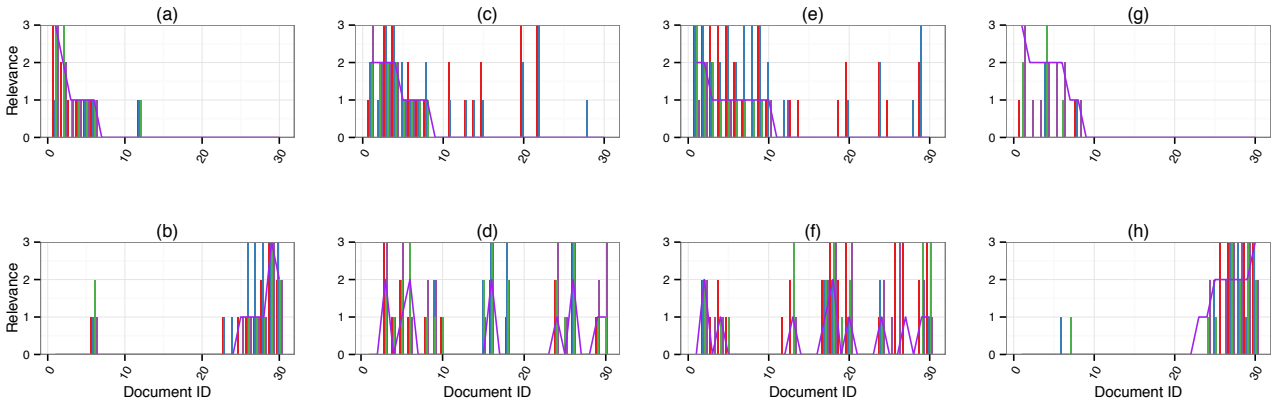

Figure 1: Document ordering and judgment results for the topic Easy (`el nino`) (a) & (b) and (`schengen agreement`) (c) & (d); Hard (`euro opposition`) (e) & (f) and (`ship losses`) (g) & (h) depicting *Rel order* (first row) and *DocID order* (second row).

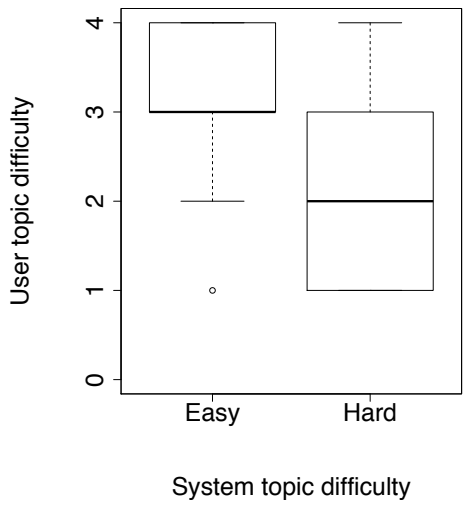

Figure 2: Relationship between system and user topic difficulty.

References

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

[2] B. Carterette, V. Pavlu, H. Fang, and E. Kanoulas. Million query track 2009 overview. In *TREC*, 2009.

[3] C. A. Cuadra, R. V. Katter, E. H. Holmes, and E. M. Wallace. *Experimental Studies of Relevance Judgments. Final Report. 3 Volumes.* System Development Corporation, 1967.

[4] D. Davidson. The effect of individual differences of cognitive style on judgments of document relevance. *J. Amer. Soc. Inf. Sc.*, 28(5): 273–284, 1977.

[5] K. De Swert. Calculating inter-coder reliability in media content analysis using Krippendorff's alpha. *Center for Politics and Communication*, 2012.

[6] M. Eisenberg and C. Barry. Order effects: A study of the possible influence of presentation order on user judgments of document relevance. *J. Amer. Soc. Inf. Sc.*, 39(5):293–300, 1988.

[7] C. D. Gull. Seven years of work on the organization of materials in the special library. *American Documentation*, 7(4):320–329, 1956.

[8] A. F. Hayes and K. Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89, 2007.

[9] M.-H. Huang and H.-Y. Wang. The influence of document presentation order and number of documents judged on users' judgments of relevance. *J. Amer. Soc. Inf. Sc. Tech.*, 55(11):970–979, 2004.

[10] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1–2):1–224, 2009.

[11] A. M. Rees and D. G. Schultz. A field experimental approach to the study of relevance assessments in relation to document searching. *Final Report to the National Science Foundation*, I, 1967.

[12] T. Sakai and N. Kando. Are popular documents more likely to be relevant? a dive into the ACLIA IR4QA pools. In *Proc. EVIA*, pages 8–9, 2008.

[13] T. Sakai and C. Lin. Ranking retrieval systems without relevance assessments: Revisited. In *Proc. EVIA*, pages 25–33, 2010.

[14] T. Saracevic. Effects of inconsistent relevance judgments on information retrieval test results: A historical perspective. *Library Trends*, 56(4):763–783, 2008.

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

[16] F. Scholer, D. Kelly, W.-C. Wu, H. S. Lee, and W. Webber. The effect of threshold priming and need for cognition on relevance calibration and assessment. In *Proc. SIGIR*, pages 623–632. ACM, 2013.

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

[18] P. Vakkari and E. Sormunen. The influence of relevance levels on the effectiveness of interactive information retrieval. *J. Amer. Soc. Inf. Sc. Tech.*, 55(11):963–969, 2004.

[19] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716, 2000.

[20] Y. Xu and D. Wang. Order effect in relevance judgment. *J. Amer. Soc. Inf. Sc. Tech.*, 59(8):1264–1275, 2008.