Effect of Written Instructions on Assessor Agreement

William Webber
College of Information Studies
University of Maryland
United States of America
wew@umd.edu

Bryan Toth and Marjorie Desamito *
Eleanor Roosevelt High School
Greenbelt, Maryland
United States of America
bryan.n.toth@gmail.com
magicaura2000@yahoo.com

ABSTRACT

Assessors frequently disagree on the topical relevance of documents. How much of this disagreement is due to ambiguity in assessment instructions? We have two assessors assess TREC Legal Track documents for relevance, some to a general topic description, others to detailed assessment guidelines. We find that detailed guidelines lead to no significant increase in agreement amongst assessors or between assessors and the official qrels.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and software—*performance evaluation*.

Keywords

Retrieval experiment, evaluation, e-discovery

General Terms

Measurement, performance, experimentation

1. INTRODUCTION

Assessors frequently disagree on the relevance of a document to a topic. Voorhees [2000] finds that TREC adhoc assessors have mutual F1 scores of around 0.6, while Roitblat et al. [2010] report mutual F1 as low as 0.35 for professional e-discovery reviewers. Such low agreement is of serious practical concern in e-discovery, where large-scale, delegated manual review is still widely used. Possible causes of disagreement include assessor error and ambiguity in instructions. We examine whether detailed relevance guidelines increase agreement amongst assessors and with the guideline author, and find no significant increase in either form of agreement

2. METHODS AND MATERIALS

We measure inter-assessor agreement by Cohen's κ, for which 1 is perfect and 0 is random agreement [Cohen, 1960]. Unassessable documents (too long or misrendered) are ignored. A trial experiment of 75 documents per treatment, on Topic 301 from the TREC 2010 Legal Track, indicated that a sample size of 215 documents per treatment, with even proportions relevant and irrelevant, was required to achieve 80% power for a true κ delta of 0.23, being the difference in agreement with official assessments between first and second tercile assessors in the TREC 2009 Legal Track.

*Work performed while interns at the University of Maryland.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

Message type	Messages	Documents		
		rel	irrel	unass
>1 relevant document	5	13	12	0
Relevant appealed	170	170	82	4
" unappealed	38	38	7	0
Irrelevant appealed	58	0	78	1
" unappealed returned	32	0	44	1
" " unreturned	16	0	17	1
Total	319	221	240	7

Table 1: Number and types of messages and documents sampled from Topic 204 for re-assessment. A message is classed as "relevant" if it contains a single relevant document (body or attachment). Counts of relevant, irrelevant, and unassessable documents are using the official, post-appeal assessments.

Topic 204 from the interactive task of the TREC 2009 Legal Track [Hedin et al., 2009] was used for the full experiment. The corpus is the EDRM Enron emails. Whole messages were sampled, but each email body and attachment was separately assessed. A stratified sample was taken, as described in Table 1. The strata were divided evenly and randomly into two batches. Each batch was assessed in document id order, with the parts of a message being assessed sequentially, as in TREC.

At TREC, a senior lawyer called the topic authority develops the topic, writes the detailed guidelines, and adjudicates appeals against first-round assessments. The appeals process for this topic was thorough [Webber, 2011], and the majority of sampled documents were appealed; we regard the assessments as an accurate representation of the topic authority's conception of relevance. We measure agreement for each batch between the two experimental assessors and the official, post-appeal assessments.

The latter two authors of this paper acted as experimental assessors. Each assessor assessed all documents in each batch. For the first batch, assessors were given the 42-word topic statement to guide their assessments; for the second, they received the 5-page detailed relevance guidelines. A third pass was then made, in which the two assessors jointly reviewed both batches, in light of the detailed guidelines, and tried to agree on a conception of relevance.

3. EXPERIMENTAL RESULTS

Table 2 shows the results of our experiments. The provision of detailed assessment guidelines (Batch 2) did not improve agreement, significantly or otherwise, over topic-only instructions (Batch 1), either amongst assessors or with the official assessments, in either the full or the trial experiment. Message-level analysis (in

	Full experiment			Trial experiment		
Batch	A v B	A v O	B v O	A v B	A v O	B v O
1	0.519	$0.454^{\mathbf{ab}}$	0.710	0.229	0.557	0.417
2	0.528	0.555	0.637	0.275	0.439	0.294
Jnt-1	*0.992*	$0.677^{\mathbf{a}}$	0.686	–	–	–
Jnt-2	*0.950*	$0.665^{\mathbf{b}}$	0.674	–	–	–

Table 2: Cohen's κ values between official and two experimental assessors, for full and trial experiments, on single-assessed Batch 1 (with topic statement only), single-assessed Batch 2 (with detailed guidelines), and (for full experiment only) joint-assessed Batches 1 and 2 (with topic guidelines and consultation between assessors). Columnar value pairs significantly different at $\alpha = 0.05$ (excepting inter-experimenter joint review) are marked by superscripts.

Assessors	Confidence interval
A v. B	$[-0.155, 0.173]$
A v. Official	$[-0.061, 0.263]$
B v. Official	$[-0.211, 0.065]$

Table 3: Two-tailed 95% normal-approximation confidence intervals on the true change in κ between Batch 1 and Batch 2 amongst different assessor pairs, for the full experiment.

which a message is relevant if any part of it is relevant) gives similar results. Inter-assessor κ values are high for the full experiment's joint assessment, since assessors reached agreement on all save a handful of documents (1 for Batch 1, and 5 for Batch 2). Assessor A's agreement with the official assessments increases significantly under joint review, but this may be due to Assessor A's assessments moving closer to Assessor B's; Assessor A's self-agreement on Batch 1 is 0.399 post-consultation, whereas Assessor B's is 0.739.

Table 3 gives 95% confidence intervals on the true change in κ values with the addition of assessor guidelines. A substantial improvement is still plausible in agreement between Assessor A and the official assessments, but not for Assessor B and official, nor for inter-assessor agreement.

Agreement between the original TREC assessors and the authoritative assessment on the documents examined in our experiment is 0.102 for Batch 1 and 0.024 for Batch 2, much lower than for our experimental assessors; however, this is a biased comparison, since sampling was heavily weighted towards appealed documents. Over the 7,289 documents sampled for assessment at TREC, though, the original assessors achieved a κ of 0.320, still well below that of the experimental assessors. The relatively high reliabilty of the assessor is reflected in their high mutual F1 scores (Table 4).

Qualitatively, the experimental assessors described the full experiment topic description by itself as being clear, and the detailed guidelines as being very clear and easy to relate to the documents. As can be seen in Table 2, agreement for this topic was generally higher than for the trial experiment.

4. DISCUSSION

Our initial, seemingly common-sense, hypothesis was that more detailed instructions would raise agreement between assessors and the authoritative conception of relevance, and therefore amongst assessors themselves. The results of this experiment have failed to confirm this hypothesis, or even to show a general trend in this direction. The only significant improvement occurred when Assessor A consulted with Assessor B, but that may be attributable to the former's assessments moving closer to the latter's. Indeed, confidence intervals indicate that a substantial increase in agreement is not plausible, except possibly between one assessor and the official view. We can conclude that, for this topic and these assessors, the provision of more detailed assessment guidelines did not lead to any marked increase in assessor reliability.

Batch	A v B	A v O	B v O
1	0.679	0.648	0.828
2	0.769	0.791	0.823

Table 4: Assessor mutual F1 scores for the full experiment.

It is also notable that our experimental assessors, who were high school students with no legal training, appear to have produced assessments much more in line with the authoritative conception of relevance than the original TREC assessors, who were legally trained, professional document reviewers.

Our findings are not reassuring for the widespread practice of using delegated manual review in e-discovery. If assessors do no better with detailed guidelines than with a general outline of the topic, then there is an irreducible loss of signal in transmitting the relevance conception of an authoritative reviewer into the minds of other assessors. E-discovery practice is moving towards the use of statistical classification tools [Grossman and Cormack, 2011]; it may well be that the lawyer overseeing a case is better able to convey their conception of relevance by personally training a machine classifier, than by instructing delegated human reviewers.

Acknowledgments.

Venkat Rangan of Symantec eDiscovery provided the TIFF images of Enron documents used in the TREC 2009 Legal Track assessments. Maura Grossman and Gord Cormack advised on the choice of TREC topics. This material is based upon work supported by the National Science Foundation under Grant No. 1065250. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Maura R. Grossman and Gordon V. Cormack. Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Richmond Journal of Law and Technology*, 17(3):11:1–48, 2011.

Bruce Hedin, Stephen Tomlinson, Jason R. Baron, and Douglas W. Oard. Overview of the TREC 2009 legal track. In Ellen Voorhees and Lori P. Buckland, editors, *Proc. 18th Text REtrieval Conference*, pages 1:4:1–40, Gaithersburg, Maryland, USA, November 2009. NIST Special Publication 500-278.

Herbert L. Roitblat, Anne Kershaw, and Patrick Oot. Document categorization in legal electronic discovery: computer classification vs. manual review. *Journal of the American Society for Information Science and Technology*, 61(1):70–80, 2010.

Ellen Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5): 697–716, September 2000.

William Webber. Re-examining the effectiveness of manual review. In *Proc. SIGIR Information Retrieval for E-Discovery Workshop*, pages 2:1–8, Beijing, China, July 2011.