On the Evaluation of the Quality of Relevance Assessments Collected through Crowdsourcing

Gabriella Kazai
Microsoft Research
7 JJ Thomson Ave
Cambridge, UK
gabkaz@microsoft.com

Natasa Milic-Frayling
Microsoft Research
7 JJ Thomson Ave
Cambridge, UK
natasamf@microsoft.com

ABSTRACT

Established methods for evaluating information retrieval systems rely upon test collections that comprise document corpora, search topics, and relevance assessments. Building large test collections is, however, an expensive and increasingly challenging process. In particular, building a collection with a sufficient quantity and quality of relevance assessments is a major challenge. With the growing size of document corpora, it is inevitable that relevance assessments are increasingly incomplete, diminishing the value of the test collections. Recent initiatives aim to address this issue through crowdsourcing. Such techniques harness the problem-solving power of large groups of people who are compensated for their efforts monetarily, through community recognition, or by the entertaining experience. However, the diverse backgrounds of the assessors and the incentives of the crowdsourcing models directly influence the trustworthiness and the quality of the resulting data. Currently there are no established methods to measure the quality of the collected relevance assessments. In this paper, we discuss the components that could be used to devise such measures. Our recommendations are based on experiments with collecting relevance assessments for digitized books, conducted as part of the INEX Book Track in 2008.

Keywords

Test collection construction, relevance judgments, incentives, social game, quality assessment.

1. INTRODUCTION

The established approach to constructing a test collection involves employing a single judge, usually the topic author, to assess the relevance of documents to a topic. Recent practices are, however, diversifying the ways in which relevance judgments are collected and used. In Web search the tendency is to use explicit judgments from a sample of the user population or to analyze user logs to infer relevance. An increasingly popular strategy is to use crowdsourcing. Amazon's Mechanical Turk service, for example, employs Internet users to complete 'human intelligence tasks', such as providing relevance labels, for micro-payments. Google's Image Labeler game [7] works by entertaining its participants who label images for free. Community Question Answering (cQA) services, such as Yahoo! Answers, reward the members who provide the best answers with 'points' which leads to increased status in the community. Participants of the Initiative for the Evaluation of XML retrieval (INEX) [3] contribute relevance assessments of highlighted passages in Wikipedia documents [6] or digitized books [5] in order to gain access to the full test set.

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.

Obtaining relevance judgments through a collective user effort, however, carries inherent risks regarding the quality of the collected data. For example, it has been shown that the different background knowledge of the assessors can lead to different conclusions in the evaluation [1]. A further critical factor is the incentive that motivates assessors to provide relevance judgments. For example, workers on Amazon's Mechanical Turk benefit from completing more jobs per time unit. Thus, the quality of their output may not be of their concern unless tied to the potential loss of their income. Studies have also shown that some members of the cQA community 'play the system' by colluding in order to increase their status. Similar problems of user tactics have been reported in reputation systems like eBay.

This raises the question of how to estimate the trustworthiness of relevance labels provided by members of the 'crowd' and how to evaluate the quality of the collected relevance data set. In this paper, we make recommendations based on the experiments conducted at the INEX 2008 Book Track.

2. COLLECTIVE ASSESSMENTS AT INEX

In 2008, the INEX Book Track [4] experimented with a method for the collective gathering of relevance assessments using a social game model [5]. The Book Explorers' game was designed to provide incentives for assessors to follow a predefined review procedure. It also made provisions for the quality control of the collected relevance judgments by facilitating the review and re-assessment of judgments and by enabling communication between judges. The game was based on two competing roles: explorers who discover and mark relevant content and reviewers who check the quality of the explorers' work. Both were rewarded points for their efforts. Disagreements between explorers and reviewers led to point deductions which could be recovered by re-assessing the pages under conflict (though agreement was not necessary).

In two pilot runs, several types of relevance data were collected: text regions highlighted on a page, relevance labels assigned to a page, comments recorded for a page, and the relevance degree assigned to the books. In total, 17 assessors judged 3,478 books and 23,098 pages across 29 topics, and marked a total of 877 highlighted regions. The assessment system recorded 32,112 navigational events, 45,126 relevance judgment events, and 2,970 'search inside the book' events.

In addition, as part of the assessment process assessors were asked to indicate their familiarity with their selected topics, as well as to record their familiarity with each book they judged before and after they browsed the book.

3. TRUST AND QUALITY CONTROL

Given the assessors' diverse backgrounds and intentions, the question arises as to what degree relevance assessments can be trusted. For example, assessors' desire to win may influence their work, leading to more labels but of lower quality. To incorporate the notion of reliability, we may associate a *trust weight* with each assessment. The final assessments can then be derived as weighted averages of the individual opinions. However, how can such trust weights be derived without an established ground-truth to compare with? In the following sections we discuss possible sources of evidence for computing the trust score.

3.1 Assessor agreement

We hypothesize that *judgments agreed upon by multiple assessors can be trusted more*. Agreement can suggest that the topic is less ambiguous, that the interpretation of the document and the relevance criterion is similar across the judges, but it may also signal collusion. Judges may collude in order to increase their scores. Disagreement can indicate an ambiguous topic, possible differences in the assessors' knowledge or interpretation of the relevance criterion, or may reflect their intention to reduce each others' scores. The trust weight will depend on being able to differentiate between these reasons.

In our data set, a total of 239 books were judged by multiple assessors (between 2-4) across 18 topics. The level of pairwise agreement between judges, based on binary relevance, was relatively high, around 80.7%. Out of 239 books, judges only disagreed on the relevance of 24 books. Their opinion differed only on the degree of relevance for 34 relevant books (71% by 1 degree, 20% by 2 degrees, 6% by 3 degrees and 3% by 4 degrees). At the page level, 4,622 pages were judged by multiple assessors with an agreement level of 57%.

Suggestive influence. The observed levels of agreement are relatively high compared to those reported elsewhere (i.e., around 33-49% for documents at TREC, and 27-57% for documents and 16-24% for elements at INEX). This high level of agreement could suggest collusion between judges or could simply reflect bias in their work. Since reviewers were shown the relevance labels that explorers assigned to a page, their own judgments could have been influenced by these opinions. Indeed, the majority of the multiple judgments were results of reviewers checking the explorers' work (74%). However, the reviewers were not aware of the relevance labels that explorers assigned to books.

Topic familiarity. The average difference between assessors' familiarity with the topics for books on which they agreed on (based on binary relevance) was 1.95 while for books on which they disagreed was 3.36. This shows that background knowledge does contribute to differences of opinions.

Collusion. Possible collusions may involve judges from the same institution who agree with each other. In the collected data, 6 books and 606 pages were judged by different members of the same group. Judges agreed on the relevance of all 6 books at the binary level and disagreed on the degree of relevance for 4 of the books. They also agreed on the (binary) relevance of all, expect 5, pages. This agreement may, however, be genuine rather than a result of collusion. The amount of time spent on assessing a page (dwell time) could provide a clue: it may be reasonable to expect that judges with similar levels of topic and book familiarity spend similar lengths of time assessing the same page. Collusion could thus be detected when judges consistently agree whilst having different averages for time spent on a page or book.

3.2 Annotations

Annotations, i.e., comments added to pages by assessors, could be used when considering the trustworthiness of the assessments. We hypothesize that *the judgments of annotated pages may be more trustworthy* since judges spent extra time and effort.

Comments were added by 9 of the 17 assessors to 227 pages in 98 books. The distribution of comments varied greatly, with an average of 25 comments per judge (σ=37, min=1, max=102). Two judges, in particular, made frequent use of this feature, adding 102 and 75 comments, respectively. This reflects commitment on their part and suggests that their judgments may be more trustworthy.

Most comments were explanations of relevance decisions or short summaries (76%), or qualitative statements about the relevance of the content (15%). We suspect that the comments may have acted as indirect messages, purposefully added by explorers to preempt possible challenges and thus penalty from reviewers. The presence of comments may thus also signal ambiguous content or questions about relevance. Furthermore, comments can also provide clues on the user background and the user task.

3.3 Learning effect

At the start of the assessment process, judges indicated their familiarity with their selected topics. However, although initially unfamiliar with a topic, a judge may learn about it during the review process. One way to assess this is to examine changes in the length of time that judges spend on assessing pages for a given topic. Assuming that judges learn, we expect them to become faster in assessing pages over time. This should be combined with their self-declared familiarity with the book and incorporated into the trust weight.

4. CONCLUSIONS

In this paper we draw attention to potential issues with collective relevance assessments through crowdsourcing, where judges with diverse backgrounds and intentions contribute data with varied levels of reliability. We discuss several sources of evidence that could be used to derive a trust weight for the judgments: topic familiarity and familiarity with the content being assessed, dwell time and changes in the patterns of dwell time, agreement between judges, and the presence and length of comments. However, other factors, such as the incentives that influence judges' behavior, also need to be considered. How to define the trust weight function based on these factors, taking into account their complex relationship, is the subject of our further research.

5. REFERENCES

[1] Alonso, O., Rose, D. and Stewart, B.. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15, 2008.

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

[3] Fuhr, N., Kamps, J., Lalmas, M., Malik, S., and Trotman, A. 2007. Overview of the INEX 2007 ad hoc track. In *Proc. of INEX'07*.

[4] Kazai, G., Doucet, A., Landoni, M. 2009. Overview of the INEX 2008 Book Track. In *Proc. of INEX'08*. LNCS Vol. 5613, Springer.

[5] Kazai, G., Milic-Frayling, N., Costello, J. 2009. Towards Methods for the Collective Gathering and Quality Control of Relevance Assessments. In *Proc. of SIGIR 2009*.

[6] Trotman, A. and Jenkinson, D. 2007. IR evaluation using multiple assessors per topic. In *Proc. of ADCS*.

[7] von Ahn, L. and Dabbish, L. 2004. Labeling images with a computer game. In *Proc. of SIGCHI 2004*, 319-326.