Examining the Limits of Crowdsourcing for Relevance Assessment

Paul Clough: Information School, University of Sheffield, UK; p.d.clough@sheffield.ac.uk; +441142222664
Mark Sanderson: School of Computer Science and Information Technology, RMIT University, Australia; mark.sanderson@rmit.edu.au; +61399259675
Jiayu Tang: Alibaba.com Limited, China; jiayu.tangjy@alibaba-inc.com; +8657185022088
Tim Gollins: The National Archives, Kew, Richmond, UK; Tim.Gollins@nationalarchives.gsi.gov.uk; +442088763444
Amy Warner: Royal Holloway, University of London, UK; amy.warner@rhul.ac.uk; +441784443892

Abstract. Evaluation is instrumental in the development and management of effective information retrieval systems and ensuring high levels of user satisfaction. Using crowdsourcing as part of this process has been shown to be viable. What is less well understood are the limits of crowdsourcing for evaluation, particularly for domain specific search. We present results comparing relevance assessments gathered using crowdsourcing with those gathered from a domain expert for evaluating different search engines in a large government archive. While crowdsourced judgments rank the tested search engines in the same order as expert judgments, crowdsourced workers appear unable to distinguish different levels of highly accurate search results in a way that expert assessors can. The nature of this limitation in crowd sourced workers for this experiment is examined and the viability of crowdsourcing for evaluating search in specialist settings is discussed.

Keywords: information search and retrieval; performance of systems

1 Introduction

Evaluation is critical for designing, developing and evolving effective Information Retrieval (IR) systems. Although a wide ranging topic, the main focus of evaluation in IR has been on measuring retrieval effectiveness: the ability of an IR system to discriminate between relevant and non-relevant documents. Here, the Cranfield methodology (Cleverdon, 1962) is commonly used, where human judges manually assess documents for relevance to a range of queries.

The work is commonly regarded as onerous, consequently large-scale evaluation campaigns, such as the Text REtrieval Conference (TREC), have generated distributed assessments over the past 20 years. Many have reported the successful use of crowdsourcing for generating relevance assessments (Carvalho et al., 2011). A topic that is less examined, however, are the limitations of a crowdsourcing approach.

In this paper we report the design of a crowdsourcing experiment using Amazon's Mechanical Turk (AMT) to gather relevance assessments for measuring two competing search engines for the UK Government's National Archives (TNA). This paper takes as its baseline the existing view that generating relevance assessments with crowdsourcing is viable. The question here is whether crowdsourced judgments for queries and documents covering a specialist domain will be accurate and correlate with the judgments of a domain expert. The effects of domain expertise on search result relevance judgments has shown up in general web search (Kinney et al., 2008) but, as far as we know, no work to date has studied this in a specialised domain (cultural heritage) or with crowdsourced workers.

Based on previous work to evaluate search at TNA (Sanderson & Warner, 2011), a set of queries was used to design a crowdsourcing experiment which could be carried out by an expert employee at TNA and a population of crowdsourced workers. The research questions addressed in this study were:

- RQ1: Do the judgments of a domain expert and crowdsourced workers agree and does the ranking of search systems based on the judgments remain stable?
- RQ2: What factors influence agreement between expert and crowdsourced judgments?

Section 2 describes related work; Section 3 describes our methodology and the AMT setup; Section 4 presents results comparing crowdsourced worker and expert judgments; Section 5 discusses the findings and provides recommendations; before Section 6 concludes.

2 Related Work

In this section we discuss related work focussing on the use of crowdsourcing in the evaluation of IR systems and broader crowdsourcing studies

2.1 Crowdsourcing and IR Evaluation

Ever since Cleverdon's design of IR evaluation based on manual relevance assessments, some speculated that the assessors would be biased in some way, which would significantly affect the accuracy of the measurement of retrieval effectiveness. A series of experiments were conducted to test this hypothesis (Cleverdon, 1970; Voorhees, 1998). Despite there being marked differences in the documents that different assessors judged as relevant or non-relevant, the differences did not affect the relative ordering of IR systems being measured using the different assessments.

Bailey et al. (2008) compared the relevance assessments of judges who were subject experts and those who were not. They found that different assessments resulted, which had an effect on the measurement of system effectiveness. Discriminating between systems that were "good" and "bad" was possible between both forms of judgements. However, distinctions between top performing systems were harder to discern based on assessments from the judges who were not subject experts. Kinney et al. (2008) also investigated domain expertise and found that non-specialist assessors made significant errors. Compared to experts, they disagreed on the underlying meaning of queries and subsequently there were effects on calculations of system effectiveness. They found that the rating accuracy of generalists was improved if domain experts provided descriptions of what the user, issuing a query, was seeking.

2.2 Crowdsourced Assessments

The success of using crowdsourcing for various natural language processing tasks was tested by Snow et al. (2008). For information retrieval Alonso & Mizzaro (2009) showed that crowdsourcing was a reliable way of providing relevance assessments, although creating a TREC-like experiment suitable for crowdsourcing required careful design and execution. In order to obtain reliable results, quality control is important (Kazai, 2011). Aggregating multiple assessments for each task is probably the most popular strategy to control quality and has been used extensively (e.g., Alonso & Mizzaro, 2009). Sorokin & Forsyth (2008) tried to encourage the participants to follow the task protocol by injecting gold standard data. If a crowdsource worker's responses deviated significantly from the gold standard, the standard would be shown to help the worker learn what was required.

3 Methodology

Our work is similar to Bailey et al. (2008): we compare assessments from crowdsourced workers with those from a domain expert; measuring the impact on the effectiveness scores and relative rankings of two search engines.

3.1 Experimental Design

The approach taken was to identify a set of queries, retrieve sets of results on the two tested search engines, present results to assessors to gather relevance judgments and measure the effectiveness of different ranked lists. The requests were generated by a member of the TNA search quality team who located, in a set of search logs, the most popular queries as determined by the number of times the query was entered. In total 48 queries were selected. This size of query set has been shown in the past to be sufficient for measuring the differences between retrieval systems (Voorhees, 2009).

Using Broder's query classifications (Broder, 2002), both navigational and informational queries were found in the logs. Here, the best answers for navigational queries were topic specific home pages found within the TNA site. Although commonly thought to exist only in mainstream web search, navigational queries have also found to be common in intranets (Zhu et al., 2007). So prevalent were the navigational queries in TNA's search logs, that half the selected queries were navigational and half informational; because it was known that being effective at these two query types was important.

The search requests had on average 1.6 words per query. To better understand the information need behind each request, the TNA team member examined retrieved documents. From this, a 1-2 sentence description of the information need was written to help assessors better understand the user's intent. The 48 queries were issued to two search engines (labelled system A and system B). The ten highest ranked documents were retrieved and judged. The retrieval effectiveness measures, precision at rank 10 (P@10) and Discounted Cumulative Gain (DCG) measured at rank 10 were used.

3.2 Design of the Mechanical Turk Experiment

Each job performed by the crowdsource workers (known as HITs, Human Intelligence Tasks) consisted of being shown a query, a description of the query intent, and 10 retrieved documents to be judged for relevance, which were output by either system A or B. Graded relevance assessments were used:0=not relevant, 1=partially relevant, and 2=highly relevant. Workers were encouraged to judge multiple queries. We aimed to gather 10 sets of judgments per query-system combination resulting in a total of 960 HITs (48 x 2 x 10). A HIT was completed by answering the following questions (using a 5-point Likert scale):

Q1) How difficult was this task?
Q2) How familiar were you with the subject of this query?
Q3) How confident were you in your judgments?
Q4) How satisfied would you be with these results?

The workers were offered 4¢ per completed HIT. The total cost of the experiment was $43.20 (including admin fees) with a total running time of 2 weeks. The experiment was run over two weeks and data was collected from 73 unique workers who produced 924 HITs (96.3% of the total available), the majority of which were obtained in the first week. The ten most active workers completed approximately 80% of the HITs. No gold standard data was used in the experiments; the data was manually checked for noise, which resulted in 91 HITs being eliminated from one worker. Indicators of noise included the short time to complete the tasks (compared to other workers) and the same values given to all documents and questionnaire results in response to multiple queries.

3.3 Worker Questionnaire Responses

For each of the four questions, ratings from the crowdsourced workers were averaged across responses for each query and compared with an expert assessor working for TNA (Table 1). The assessor held degrees in History (BA and MA) and was Head of Systems Development and Search at TNA for approximately 3 years. The role focussed on examining how people searched the TNA website and working with subject experts across the organisation to identify answers to search queries. In earlier roles while working for this organisation the assessor was sometimes asked to respond to enquiries relating to collections held at TNA. The assessor worked at TNA for around 6.5 years.

	Queries	Q1- difficulty 1=V. difficult; 5=V. easy	Q2 - familiarity 1=V. unfamiliar; 5=V. familiar	Q3 – confidence 1=Not at all confident; 5=V. confident	Q4 – satisfaction 1=V. unsatisfied; 5=V. satisfied
Expert	All	4.36	4.34	4.25	3.25
	Informational	4.10	4.13	4.00	3.04
	Navigational	4.63	4.56	4.50	3.46
	System A	4.54	4.44	4.44	3.90
	System B	4.19	4.25	4.06	2.60
Crowd-sourced workers	All	3.47	3.54	4.12	4.13
	Informational	3.48	3.43	4.04	4.05
	Navigational	3.47	3.65	4.20	4.21
	System A	3.42	3.57	4.18	4.18
	System B	3.52	3.51	4.05	4.08

Table 1: Questionnaire results for expert and crowdsourced worker responses

Table 1 shows average responses from the expert and crowdsourced workers. Statistically significant differences are found between the crowdsourced workers and the expert responses for Q1, Q2, and Q4, highlighting that the workers found the tasks more difficult, were less familiar with the subject of the query and more satisfied with results than the expert. Kinney et al. (2008) showed that lack of familiarity and confidence with more difficult tasks is likely to provide inaccurate judgments leading to disagreements between judges.

4 Results

We compared the effectiveness results obtained using the workers' judgments compared to those from the expert. The workers' relevance scores were averaged. First relative ranking of retrieval systems was examined; next correlations between the judgments were explored.

4.1 Comparing Effectiveness of System A and B

In the spirit of Bailey et al. (2008), we first determine to what extent the judgments of the workers and expert affected the ranking of the two systems. Table 2 shows retrieval effectiveness based on all queries for P@10 and DCG, together with DCG scores for informational and navigational queries. We observe that with both sets of

judgments, one can measure that system A is statistically significantly better than system B (using a paired t-test) and although the absolute scores differ, the relative ranking of systems remains stable.

		P@10	DCG		
			All (N=48)	Informational (N=24)	Navigational (N=24)
Expert	System A	0.81	7.43	6.59	8.26
	System B	0.51	4.39	4.66	4.11
	Average	0.66	5.91	5.63	6.19
Crowd-sourced workers	System A	0.77	6.40	5.60	7.21
	System B	0.63	5.32	4.93	5.71
	Average	0.70	5.86	5.27	6.46

Table 2: P@10 compared with DCG results across all queries, informational queries and navigational queries

From Table 2 we also observe that the difference between the two systems is greater based on the expert assessments than those from the crowdsourced workers. This is particularly apparent when considering query type (informational or navigational). The difference between system A and B is greater for navigational queries based on expert assessments. The crowdsourced workers rate system B as far better than it is compared to the expert causing the large differences in absolute scores. We also see from the questionnaire results (Table 1) that crowdsourced workers expressed similar levels of satisfaction with the results of system B (4.08) and system A (4.18), unlike the expert (respectively 2.60 and 3.90), confirming the workers are unable to detect poor performance.

As an aside, there has been a range of research conducted in the IR community on measuring system effectiveness based on a self-rated satisfaction measure polled from users (e.g., Al-Maskari et al., 2008). A large amount of that research has found that differences in 'satisfaction' for different systems are often insignificant. Many of those experiments have not considered the expertise of the users involved in the experiment.

4.2 Comparing Correlations

In this section we investigate the correlation between judgments to identify where differences occur. This helps in understanding the limitations of the crowdsourced judgments compared to the expert as there may be types of queries or search results where crowdsourced workers correlate poorer with experts than others. Such situations could be avoided in future evaluation exercises.

	Queries	DCG	P@10
System A	All (N=48)	0.492**	0.285*
	Informational	0.323	-0.029
	Navigational	0.485*	0.467*
System B	All (N=48)	0.601**	0.595**
	Informational	0.563**	0.523**
	Navigational	0.772**	0.786**

Table 3: Pearson correlation coefficient for retrieval effectiveness scores obtained between the expert and crowdsourced workers

Table 3 shows the Pearson correlation coefficient scores for P@10 and DCG scores measured across the queries. The correlations are between measures based on the crowdsourced workers' assessments and those from the expert. The DCG scores correlate better between groups of assessors than P@10 scores. We see that the correlation between the crowdsourced workers and the expert judgments are reasonable for system B: $r=0.601$ ($p<0.01$) for DCG and $r=0.595$ ($p<0.01$) for P@10. However, the correlation for system A, the better system, is lower: respectively 0.492 and 0.285. The judgments between the crowdsourced workers and the expert are more interchangeable for system B than A, despite the resulting differences in absolute scores.

When considering the subsets of queries, we observe that the correlation between worker and expert is better for navigational queries than for informational queries. It is particularly noticeable that the correlations for system A are markedly lower than those for system B. The lower correlation between expert and crowdsourced workers for system A, particularly for informational queries, suggests the results of a higher quality search engines are more difficult to assess using crowdsourcing. In addition, we find that correlations for system A are not statistically significant for informational queries and only significant at the 0.05 level for navigational queries. For the P@10 on informational queries, there is no correlation between crowdsourced workers and the domain expert. In general, especially for informational queries, the crowdsourced workers seem unable to distinguish between the results of system A.

Remembering that the crowdsourced worker assessments are composed of multiple judgments aggregated by computing an average score, to compute a level of inter-assessor agreement, we compute the standard deviation of P@10 scores across all crowdsourced workers for each query. A higher standard deviation score indicates more disagreement between the crowdsourced workers. The deviation is correlated with the difference between the worker and expert P@10 score. It was found that the correlation was strongly positive, $r=0.542$ ($p<0.01$), indicating that the greater the variation between the crowdsourced judgments, the greater the difference between the average crowdsourced worker score and the expert. We observed that many of the queries with high disagreement between the crowdsourced workers for system A were informational. The queries with the lowest standard deviation scores were navigational. This highlights queries which are more disagreeable and may provide an approach to improve correlation between the expert and crowdsourced workers, by filtering out cases where the crowdsourced worker scores do not agree.

5 Discussion and Recommendations

Many organisations need reliable and repeatable methodologies for evaluating their search services; using crowdsourcing potentially lowers the cost. In the experiment conducted here, the cost of the crowdsourced workers was $43.00 for 45 hours and 13 mins of assessor effort purchased on AMT. The expert assessor cost $106.02 for 3 hours and 5 mins of work. Our results suggest that crowdsourced assessments and those generated by an expert produce similar results when ranking system A and B, A was better. This agrees with previous work confirming variability of relevance assessment is not a major problem to test collection based evaluation (Cleverdon, 1970; Voorhees, 1998) and more recently that crowdsourcing is an valid alternative to gathering relevance judgments (Carvalho et al., 2011).

However, limits were found: correlations between the crowdsourced workers and the expert assessor were lower for certain kinds of queries. Informational queries, where the assessor must interpret content more, seemed to be more difficult to assess on a system that performed better (system A). Although the absolute scores for system B were higher based on crowdsourced worker judgments than the expert, it would seem likely that results from a poorer performing system were easier to judge.

In this particular comparison, both sets of assessments indicated strongly that TNA should choose system A as their search engine. If in future, TNA wishes to consider a system that was potentially more effective than system A, and a comparison between those two systems was required, it is not clear that using crowdsourced workers would provide assessments that enable accurate measurement of the differences between the two systems. Questionnaire feedback indicated that there were limitations in the domain expertise of the crowdsourced workers that lead to them making more inaccurate judgments. The implication is that a level of expertise is necessary to determine the subtle differences in results for better performing systems. In our case we see the crowdsourced workers rating system B far more highly than the expert resulting in smaller difference between system A and system B than exists in practice.

6 Conclusions and Future Work

This paper reported an investigation of two search engines at the UK National Archives to assess the viability of using crowdsourcing for gathering relevance assessments as a form of low-cost user-based evaluation. Our conclusions indicate that crowdsourced workers were good enough for the job (in hand) of correctly ranking the two search engines based on retrieval effectiveness. However, in these experiments, the utility of such workers in judging the more accurate search engine was less clear.

Addressing our first research question, we find strong positive correlations between search effectiveness measured based on judgments of a domain expert and crowdsourced workers. There are differences in the absolute scores given to each system, but the relative ranking of A and B remains stable between the assessor groups.

For the second research question, we investigated the correlations between the expert judgments and those averaged across the crowdsourced workers. We identified cases where disagreements occurred: for informational queries and for the better performing system.

We view this study as a starting point; there are many potential options for future work. Here, one of the most popular crowd sourcing systems was used, however, alternatives exist and could be tested for this task; for TNA's future evaluation experiments, crowds will be sourced using known groups of individuals with appropriate domain expertise and motivation. The queries used were popular ones, while representing a notable fraction of overall queries submitted to TNA's retrieval system, examining a larger number of the less frequent queries would also be of value. We also plan to experiment with different approaches for aggregating the

crowdsourced worker judgments and investigate the effects on correlations with the expert. Gathering expert judgments is time-consuming and impractical for repeated evaluation of retrieval systems at TNA. Therefore, we plan to identify ways of improving crowdsourced data based using the results from the current experiment as training data.

References

Al-Maskari, A., Sanderson, M., Clough, P., &Airio, E. (2008). The good and the bad system: does the test collection predict users' effectiveness? *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 59–66). New York, NY, USA: ACM. doi:http://doi.acm.org/10.1145/1390334.1390347

Alonso, O., & Mizzaro, S. (2009). Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation* (pp. 15–16).

Bailey, P., Craswell, N., Soboroff, I., Thomas, P., de Vries, A. P., & Yilmaz, E. (2008). Relevance assessment: are judges exchangeable and does it matter. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 667–674).ACM New York, NY, USA.

Broder, A. (2002). A taxonomy of web search. *SIGIR Forum*, *36*(2), 3–10. doi:10.1145/792550.792552

Carvalho, V. R., Lease, M., & Yilmaz, E. (2011).Crowdsourcing for search evaluation. *ACM SIGIR Forum* (Vol. 44, pp. 17–22).ACM.

Cleverdon, C. W. (1962). *Report on the Testing and Analysis of an Investigation Into the Comparative Efficiency of Indexing Systems*. ASLIB Cranfield Research Project. Cranfield, UK.

Cleverdon, C. W. (1970). *The effect of variations in relevance assessments in comparative experimental tests of index languages* (Cranfield Library Report No. 3). Cranfield Institute of Technology.

Kazai, G. (2011). In search of quality in crowdsourcing for search engine evaluation. *Advances in Information Retrieval*, 165–176.

Kinney, K. A., Huffman, S. B., & Zhai, J. (2008). How evaluator domain expertise affects search result relevance judgments. *Proceeding of the 17th ACM conference on Information and knowledge management* (pp. 591–598). ACM.

Sanderson, M., & Warner, A. (2011).Training students to evaluate search engines. *Teaching and Learning in Information Retrieval*, The Information Retrieval Series (Vol. 31). Springer.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 254–263).

Sorokin, A., & Forsyth, D. (2008). Utility data annotation with Amazon Mechanical Turk. *Computer Vision and Pattern Recognition Workshops, 2008.CVPRW'08. IEEE Computer Society Conference on* (pp. 1–8). IEEE.

Voorhees, E. M. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 315–323). ACM Press New York, NY, USA.

Voorhees, E. M. (2009). Topic set size redux. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (pp. 806–807).ACM.

Zhu, H., Raghavan, S., Vaithyanathan, S., &Löser, A. (2007). Navigating the intranet with high precision. *Proceedings of the 16th international conference on World Wide Web* (pp. 491–500).

7 Author Bios

Paul Clough is a Senior Lecturer at the University of Sheffield's Information School and member of the IR group. He received a B.Eng. (hons) degree from the University of York in 1998 and a Ph.D. from the University of Sheffield in 2002. His research interests mainly revolve around developing technologies to assist people with accessing and managing information. Paul has co-authored a book on multilingual information retrieval and written over 90 peer-reviewed publications in his research area. He is currently Scientific Director for the EU-funded PATHS (Personalised Access To cultural Heritage Spaces) project.

Mark Sanderson is a Professor at the School of Computer Science and Information Technology, RMIT University, Melbourne. He received B.Sc. (hons) and Ph.D. degrees in computer science from the University of Glasgow, Glasgow, U.K., in 1988 and 1997, respectively. Mark is an Associate Editor of Information Processing and Management and ACM Transactions on the Web; co-editor of Foundations and Trends in Information Retrieval; and in 2012 is Co Program Chair of ACM SIGIR.

Jiayu Tang is a Senior Research Engineer in Search Algorithms at Alibaba.com working on machine learning for ranking search results. He received a B.Sc. (hons) in Computer Science from Zhejiang University, China in 2004 and a Ph.D. in Computer Science from the University of Southampton in 2008. He has worked in various areas of information retrieval including geographic search and evaluation, as well as content-based image retrieval and automatic image annotation.

Tim Gollins is Head of Digital Preservation at The (UK) National Archives where he leads work on digital preservation, and cataloguing representing the archives nationally and internationally. He has a B.Sc. (hons) in Chemistry from the University of Exeter, an M.Sc. in Computer Science from University College London, and an M.Sc. in Information Management from the University of Sheffield. Tim has worked in the UK Civil Service in the fields of Information Assurance, User Requirements, Systems Design, Information Management, Information Security, Information Retrieval and Digital Preservation.

Amy Warner is Associate Director, eStrategy and Technical Services in the Library at Royal Holloway. She has a BA (hons) and an MA in History from Royal Holloway, University of London. She is interested in developing archival and library resource discovery systems and is currently systems lead on the Jisc funded Search25 Project.