Computers in Human Behavior 64 (2016) 793–804

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

Effect of verbal comprehension skill and self-reported features on reliability of crowdsourced relevance judgments

Parnia Samimi [a], Sri Devi Ravana [a, *], Yun Sing Koh [b]

[a] Department of Information Systems, University of Malaya, Kuala Lumpur, Malaysia
[b] Department of Computer Science, University of Auckland, Auckland, New Zealand

ARTICLE INFO

Article history:
Received 22 July 2015
Received in revised form
6 July 2016
Accepted 27 July 2016
Available online 8 August 2016

Keywords:
Information retrieval evaluation
Crowdsourcing
Verbal comprehension
Relevance judgment
Reliability

ABSTRACT

Test collection is extensively used to evaluate information retrieval systems in laboratory-based evaluation experimentation. In a classic setting of a test collection, human assessors involve relevance judgments, which are both a costly and time-consuming task, and scales poorly. Researchers are still being challenged in performing reliable and low-cost evaluation of retrieval systems. Crowdsourcing as a novel method provides a cost-effective and quick solution for creating relevance judgments. However, crowdsourcing comes with the risk of a heterogeneous mass of potential workers who create the relevance judgments with varied levels of accuracy. It is, therefore, essential to understand the factors that affect the reliability of crowdsourced judgments. In this article, we measured various cognitive characteristics of workers, and explored the effects of these characteristics on judgment reliability, in comparison with a human gold standard. We discovered a significant correlation between judgment reliability and the level of verbal comprehension skill. This association conveys an idea for improving the reliability of judgments by discriminating workers into various groups according to their cognitive abilities and to filter out (or to include) certain group(s) of workers. Aside from that, we also discovered a significant association between reliability of judgments and self-reported difficulty of judgment as well as confidence in the task.

1. Introduction

Effective Information Retrieval (IR) system is supposed to return highly relevant information according to a user's query (user satisfaction). In this regard, and to assure about the level of effectiveness of an IR system, evaluation of the performance of this system is critically important. IR evaluation is used to assess how effectively an IR system addresses the information needs of the users. Test collection model such as Text Retrieval Conference (TREC)[1] is used by the system-based experiments to measure system performance, which is reproducible method with lower cost. Test collections consist of document corpora and search topics with respective relevance judgments. In traditional method, test collections are created under controlled conditions: expert searchers create topics and the documents retrieved by various IR systems are pooled to be judged by trusted human assessors. The compiled set of documents, topics and relevance labels are then used to compute performance metrics across IR systems, e.g., *precision* and *recall*. Formerly, relevance judgments set has been created by hiring human experts who are trained to interpret topics precisely and judge their relevancy to documents. As the size and diversity of test collections have massively increased, hiring expert assessors appeared expensive and burdensome for performing judgments. Indeed, major challenges of TREC-like test collection approach are time and cost for relevance judgments which makes that an unsuitable approach for scaling up (Alonso & Mizzaro, 2012).

The recent growth of the test collections has led to adapt crowdsourcing methods for creating relevance judgments. Based on Web 2.0 technology, the crowdsourcing term was invented by Howe in a Wired Magazine article (Howe, 2006). Crowdsourcing is outsourcing tasks, carried out by a massive heterogeneous mass of potential workers in the form of an open call via Internet, which were formerly performed inside an organization by employees. Crowdsourced workers are appointed through online web services such as Crowdflower[2] and work online to accomplish repetitive cognitive piece-work known as Human Intelligence Task (HIT) at

* Corresponding author.
E-mail addresses: sdevi@um.edu.my, sdravana@gmail.com (S.D. Ravana).

[1] http://trec.nist.gov/.

[2] https://www.crowdflower.com/.

http://dx.doi.org/10.1016/j.chb.2016.07.058
0747-5632/

low cost, with many workers potentially working in parallel to rapidly accomplish a task. Crowdsourcing is an efficient method particularly for tasks in which human participations are necessary, such as creating relevance judgments in IR evaluation (Alonso, Rose, & Stewart, 2008; Grady & Lease, 2010; Kazai, Kamps, Koolen, & Milic-Frayling, 2011). The main feature that makes this approach attractive is its flexibility, low cost and fast outcome (Alonso & Mizzaro, 2012). Despite the popularity of crowdsourcing in creating relevance judgments, its reliability has been questioned for various reason. For instance, do the workers have adequate expertise for a given task (Quinn & Bederson, 2011)? Are demographics and personality traits of workers affect the quality of crowdsourced relevance judgments (Kazai, Kamps, & Milic-Frayling, 2012)? Moreover, the quality of the final relevance judgments is highly subjective to how a worker is interested and incentive in performing a given task (Kazai, Kamps, & Milic-Frayling, 2011). According to Li, Zhao, and Fuxman (2014), reliability of workers is a long-lasting issue in crowdsourcing and therefore it is important to find a way to screen workers based on their levels of quality. Besides, the importance of rigorous mechanisms in order to prevent exploitation of performed tasks was emphasized (Gadiraju, Kawase, Dietze, & Demartini, 2015).

While a number of quality control methods are proposed to reduce noise that created during or after completion of a given task, a lot still needs to be learnt about the workers themselves and the role of individual differences in reliability of crowdsourced relevance judgment. Cognitive performance or cognitive abilities is of individual differences. Cognitive abilities are principally brain-based skills, concerning learning, remembering, problem-solving, and attention and mindfulness (Ekstrom, French, Harman, & Dermen, 1976). This study focuses on selected cognitive ability, “verbal comprehension skill” to determine its relationship with reliability of crowdsourced relevance judgment. Verbal comprehension skill is “the ability of reading, processing and understanding text in English”. We examine the cognitive capacities of the workers, through tests provided in a questionnaire format. In this study, the reliability of the workers is compared to that of the expert assessors, both directly as the overlap of relevance assessments, and indirectly by comparing the system effectiveness based on the evaluation obtained from the experts and from worker assessors. Furthermore, it may be similarly important to find out the workers’ experience with the task. In fact, we assess whether the self-reported competence can be introduced as a useful information to estimate the quality of workers. Specifically, this study addresses the following research questions:

- R1 Does the verbal comprehension skill of a worker have an effect on the reliability of crowdsourced relevance judgments?
- R2 Does the verbal comprehension skill have an effect on IR systems performance rankings in IR evaluation experimentation.
- R3 How does a worker's (i) topic knowledge, (ii) perceived difficulty of the task, and (iii) confidence in correctness, relate to the reliability of the worker's relevance judgments?

2. Related works

2.1. Crowdsourcing in IR evaluation

Alonso et al. (2008) are the pioneers of using crowdsourcing for obtaining relevance judgments in IR evaluation through Amazon Mechanical Turk (AMT) on TREC data. Crowdsourcing method in IR evaluation has been adapted massively in recent years (Alonso & Mizzaro, 2009; Grady & Lease, 2010; Kazai, 2014; Kazai, Kamps, & Milic-Frayling, 2013; Lease & Kazai, 2011; Lease & Yilmaz, 2013; Zuccon et al., 2012). A comprehensive experiment validated the use of crowdsourcing for creating relevance judgments (Alonso & Mizzaro, 2012). The experimental results show that crowdsourcing is low-cost, reliable and quick solution for creating relevance judgments. However, it is not a replacement for current methods (human experts) because of several gaps and shadows, which are left for future research. For instance, scalability of crowdsourcing has not been fully investigated yet, although the reproducibility of crowdsourced evaluation was investigated in a study (Blanco et al., 2011). In this study after a period of six months and with the use of different evaluation measures and system rankings, the crowdsourcing experiment was repeated and produced a similar output, showing that crowdsourcing experiments can be repeated over time in a reliable manner. Despite some differences in judgments between human expert and workers, the system ranking was the same.

In 2011, Kazai et al. investigated the relationship between workers' behavioural patterns, their personality profiles, and the accuracy of their judgments. The difference was based on behavioural observation including (i) label accuracy, (ii) HIT completion time and (iii) fraction of useful labels. The study investigated, whether the behaviour and personality of workers are able to influence the label accuracy through designing two different HITs namely Full Design (FD), a strict quality control, and Simple Design (SD), reduced the quality control compared with FD. The study correlated the worker types and personality trait information, with the accuracy of labels, considering the ‘Big Five’ personality dimensions (John, Naumann, & Soto, 2008) (namely openness, conscientiousness, extraversion, agreeableness and neuroticism). Using behavioural patterns method, various types of workers (spammer, sloppy, incompetent, competent, and diligent) were identified and as a result a strong correlation between the accuracy of judgments and the openness trait were reported (Kazai, Kamps, & Milic-Frayling, 2011). The impact of task design on the quality of labels has been assessed in several researches. In a study for book search evaluation with two different HIT design, FD and SD. The FD leads to higher label quality compared with SD. Moreover, it was reported that crowdsourcing is a useful method for creating relevance judgments for IR evaluation, but tasks design needs to be done carefully, as different HIT designs lead to a significant difference in agreement between crowdsourcing and the gold set (Kazai, Kamps, Koolen, et al., 2011).

The effects of the level of pay, effort to complete tasks, and qualification needed to do the tasks, on the quality of the labels were investigated while correlating them with various human factors (Kazai et al., 2013). Variety of information including perceived task difficulty, satisfaction with the offered pay, motivation, interest, and familiarity with the topic, were obtained from the workers to see how they influence label quality, along with aspects of the task design. A higher level of payment led to high quality of an output. However, this may also attract unethical workers to participate. On the other hand, higher efforts for HITs increased the probability of inaccurate labels, but enticed workers with higher performances. In addition, when the number of judgments that need to be made in a HIT increased, it led to increase productivity. Since achieving fewer judgments per HIT, decreased the possibility of detecting low quality judgments due to workers' limited exposure. Lower effort HITs had a faster overall task completion. Limiting HITs to workers that were more reliable increased the quality of the results. Therefore, a simple pre-filtering, such as filling captcha fields, helps to find unreliable workers. Obviously, the pre-filtering application was not sufficient but aided to enhance the quality of the labels. Earning money was the main reason and