Creation of Reliable Relevance Judgments in Information Retrieval Systems Evaluation

Experimentation through Crowdsourcing: A Review

Parnia Samimi and Sri Devi Ravana *

Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia

* Corresponding author. Tel.: +60 3 7967 6348; fax: +60 3 7957 9249.
E-mail addresses: sdevi@um.edu.my, sdravana@gmail.com.
Abstract

Test collection is used to evaluate the information retrieval systems in laboratory-based evaluation experimentation. In a classic setting, generating relevance judgments involves human assessors and is a costly and time consuming task. Researchers and practitioners are still being challenged in performing reliable and low-cost retrieval systems evaluation. Crowdsourcing as a novel method of data acquisition is broadly used in many research fields. It has been proven that crowdsourcing is an inexpensive and quick solution as well as reliable alternative for creating relevance judgments. One of the crowdsourcing applications in IR is to judge relevancy of query document pair. In order to have a successful crowdsourcing experiment, the relevance judgment tasks should be designed precisely to emphasize quality control. This paper is intended to explore different factors that have an influence on the accuracy of relevance judgments accomplished by workers and how to intensify the reliability of judgments in crowdsourcing experiment.

1. Introduction

In order to have an effective Information Retrieval (IR) system and user satisfaction, evaluation of system performance is crucial. IR evaluation is to measure whether the system addresses the information needs of the users. There are two approaches for evaluating the effectiveness of IR systems as shown in Figure 1: (i) user-based evaluation and (ii) system-based evaluation [1]. In the system-based evaluation method, a number of assessors prepare a set of data which can be reused in later experiments. A common system-based evaluation approach is test collections that is also referred to as the Cranfield experiments which is the beginning of today’s laboratory retrieval evaluation experiments [2]. The Text REtrieval Conference (TREC) was established in 1992 in order to support IR researches through providing the infrastructure for large scale evaluation of retrieval methodologies.
FIGURE 1: Classification of IR evaluation methods

Test collection consists of three components: (i) document corpus which is a set of large size documents, (ii) topics which are a collection of search queries, and (iii) relevance judgments which involve human assessors appointed by TREC. The relevance judgment is partial since not all documents from the corpus have been judged by assessors because the judging task is not only costly but also time consuming. For instance, a complete judgement collection for TREC-2010 Web needs expert assessors to assess 1 billion documents. By assuming that an expert assessor can assess two documents in a minute, 347,000 days are needed for judging 1 billion documents. Therefore, to assess more documents, a large number of human experts needs to be appointed which is costly [3].

In TREC, pooling method has been used to recognize a subset of documents for judging. Figure 2 shows the process of typical IR evaluation through test collection. The participating systems run their retrieval algorithms against the document corpus and topics in the test collection to generate a set of documents called runs. The systems 1, 2, ..., m are the contributing systems for the pool creation and a collection of top ranked documents retrieved by contributing systems for each topic has been selected for judgment. The documents in the pool are judged by human assessors to create relevance judgment set, and all of the other documents outside the pool are considered as non-relevant documents. Once the relevance
judgments are ready, whole systems runs retrieved by both contributing and non-contributing systems \((1, 2, \ldots, n)\) would be evaluated against relevance judgments or qrels to measure the accuracy and effectiveness of the retrieval systems through evaluation metrics. Each system gets a score for each topic and then to achieve the overall performance score for a system, scores are aggregated. Finally, as a result of each IR experiment, the system ranking is generated for all of the systems. The major drawback of test collections is the huge cost of relevance assessment which is conducted by human experts. It needs different resources including time, infrastructure and money while it does not scale up simply.

**FIGURE 2:** Typical IR evaluation process
The user-based evaluation method quantifies the satisfaction of users by monitoring the user’s interactions with the system. Table 1 illustrates the user-based evaluation methods which are divided into five groups [4].

**TABLE 1: User-based evaluation methods.**

<table>
<thead>
<tr>
<th>User-based methods</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human in the lab</td>
<td>This method involves human experimentation in the lab to evaluate the user-system interaction</td>
</tr>
<tr>
<td>Side-by-side panels</td>
<td>This method defined as collecting the top ranked answers generated by two IR systems for the same search query and representing them side by side to the users. To evaluate this method, in the eyes of human assessor, a simple judgment is needed to see which side retrieve better results.</td>
</tr>
<tr>
<td>A/B testing</td>
<td>A/B testing involves numbers of preselected users of a Web site to analyse their reactions to the specific modification to see whether the change is positive or negative.</td>
</tr>
<tr>
<td>Using clickthrough data</td>
<td>Using clickthrough data is to observe how frequently users click on retrieved documents for a given query.</td>
</tr>
<tr>
<td>Crowdsourcing</td>
<td>Crowdsourcing defined as outsourcing tasks, which formerly accomplished inside a company or institution by employees assigned externally to huge, heterogeneous mass of potential workers in the form of an open call through Internet.</td>
</tr>
</tbody>
</table>

The drawback of human in the lab method is the limitation of a small number of humans and a small set of information. Moreover, the cost of setting up and repeating the experiment is high. The side-by-side panels just allow the comparison of two systems but it is not applicable to multiple systems. Using clickthrough data seems attractive because of the low cost of collecting data, but it needs a precise setup since it can be noisy. The term crowdsourcing was coined by [Howe [5]] based on Web 2.0 technology in a Wired Magazine article. Recently, the use of the term “crowdsourcing” for relevance judgment is increased to conquer the high cost that current evaluation methods have through expert judges. Running experiments within low cost and fast turnaround make this approach very outstanding [4].

This paper discusses the issues related to using crowdsourcing for creating relevance judgment while giving advice on the implementation of successful crowdsourcing
experiments and presents a state-of-the-art review on available methods to enhance the quality of the judgments in crowdsourcing experiments. The organization of this paper is as follows: the introductory section explains IR evaluation process and its different methods; Section 2 elaborates crowdsourcing and its application in different fields especially IR; Section 3 explains how accuracy of crowdsourcing results measure; Section 4 highlights importance of task design in crowdsourcing as well as human features and monetary factors are explained in Section 5. The recently proposed methods on quality control of crowdsourcing will be reviewed in the Section 6. The section on conclusion presents a look into the future on the use of crowdsourcing for IR evaluation and some suggestions for further in related areas.

2. Introduction to Crowdsourcing

Crowdsourcing is an upcoming research field for Information Systems scholars to create noteworthy contributions. It can be applied widely in various fields of computer science and other disciplines to test and evaluate studies as shown and explained in Table 2 [6].

<table>
<thead>
<tr>
<th>Application</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural language processing</td>
<td>Crowdsourcing technology was used to investigate linguistic theory and language processing [7]</td>
</tr>
<tr>
<td>Machine learning</td>
<td>Automatic translation by using active learning and crowdsourcing was suggested to reduce the cost of language experts [8, 9].</td>
</tr>
<tr>
<td>Software engineering</td>
<td>The use of crowdsourcing was investigated to solve the problem of recruiting the right type and number of subjects to evaluate a software engineering technique [10].</td>
</tr>
<tr>
<td>Network event monitoring</td>
<td>Using crowdsourcing to detect, isolate and report service-level network events was explored which called CEM (Crowdsourcing Event Monitoring) [11].</td>
</tr>
<tr>
<td>Sentiment classification</td>
<td>The issues in training a sentiment analysis system using data collected through crowdsourcing was analysed [12].</td>
</tr>
<tr>
<td>Cataloguing</td>
<td>The application of crowdsourcing for libraries and archives was assessed [13].</td>
</tr>
</tbody>
</table>
Transportation plan

Using crowdsourcing was argued to enable the citizen participation process in public planning projects [14].

Information retrieval

To create relevance judgements, crowdsourcing was suggested as a feasible alternative [15].

There are various platforms for crowdsourcing, they are: Amazon Mechanical Turk (AMT), Microworkers, Crowdflower, MobileWorks, Crowd Computing Systems, ShortTask, CloudCrowd, ODesk, DoMyStuff, Livework, Clickworker, SmartSheet, Utest, Elance and vWorker. Currently, AMT is the most popular and largest marketplace and crowdsourcing platform. It is a profit-oriented marketplace which allows requesters to submit tasks and workers to complete them. Human Intelligence Tasks (HIT) or microtasks is the unit of work to be accomplished. Figure 3 presents the crowdsourcing scheme which includes multiple requesters who publish their tasks, workers who accomplish the tasks on a crowdsourcing platform.

FIGURE 3: Crowdsourcing scheme

Generally, the crowdsourcing process starts with publishing a task by requesters to the crowdsourcing platform. Workers select and work on a specific task of their interest and
complete it. The requesters assess the results of the tasks done by workers. If the results are acceptable to requesters, they would pay the workers. Otherwise the workers might be rejected because of performing the task carelessly. The flow of submitting and completing tasks via crowdsourcing is shown in Figure 4.

FIGURE 4: Flow of submitting and completing tasks via crowdsourcing

Simplicity is the main feature of crowdsourcing. Crowdsourcing platforms enable the requesters to have fast access to an on-demand, global, scalable workforce and the workers to choose thousands of tasks to accomplish online. In AMT, the requesters are able to create HITs via Application Programming Interface (API) which enable requesters to distribute tasks programmatically or using template via dashboard. Crowdsourcing platforms was also suggested for data collection as a viable choice [16]. Mason et al. stated three advantages of crowdsourcing platforms: (i) allowing large number of workers to take part in experiments with low payment, (ii) workers are from diverse language, culture, background, age and country and (iii) low cost at which the researches can be carried on [17]. Precise design and quality control are required to optimize work quality and conduct a successful crowdsourcing experiment. In general, there are three objectives that most researchers draw attention to in
this field of research especially in IR evaluation which are: (i) to determine the agreement percentage between experts and crowdsourcing workers, (ii) to find out factors that have effect on the accuracy of the workers’ results and (iii) to find the approaches to maximize the quality of workers’ results.

3. Accuracy and reliability of relevance judgments in Crowdsourcing

Researchers validate the use of crowdsourcing for creating relevance judgment in IR evaluation by measuring the agreement between crowdsourcing workers and human assessors to see whether the crowdsourcing is a reliable replacement for human assessors. The inter-rater or inter-annotator agreement is used to measure the performance in crowdsourcing experiments in order to analyse the agreement between crowdsourcing workers and human assessors. The score of homogeneity in the rating given by judges is the inter-rater agreement. There are various statistics used to calculate the inter-rater agreement. Table 3 shows four common methods which are suggested to calculate the inter-rater agreement between crowdsourcing workers and human assessors for relevance judgment in IR evaluation [18]:

<table>
<thead>
<tr>
<th>Methods</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint-probability of agreement (percentage agreement)</td>
<td>It is the simplest and easiest measure that is calculated by dividing number of times each rating (e.g. 1, 2, ... 5) is assigned by each assessor by the total number of ratings.</td>
</tr>
<tr>
<td>Cohen's kappa [19]</td>
<td>It is one of the statistical measures to calculate inter-rater agreement among raters. This measurement is more robust than percentage agreement since it takes into account the effects of random agreement between two assessors.</td>
</tr>
<tr>
<td>Fleiss' kappa [20]</td>
<td>This measurement extends the Cohen's kappa by considering the agreement among any number of raters instead of just two.</td>
</tr>
<tr>
<td>Krippendorff’s alpha [21]</td>
<td>This is calculated by considering the overall distribution of assessors regardless on which assessors produced the judgments.</td>
</tr>
</tbody>
</table>

The pioneer of using crowdsourcing for IR evaluation was Alonso et al. Alonso, Rose and Stewart [15]. Alonso et al. applied crowdsourcing through Amazon Mechanical Turk on
TREC data. The experimental results showed that crowdsourcing is a low cost, reliable and quick solution and an alternative for creating relevance judgment by experts but it is not a replacement for current methods as still there are several gaps and questions for future research. As an example, the scalability of this approach has not been investigated yet [18]. In another study, the differences between crowd-sourced workers and university students in judging the relevancy of a set of documents against a number of queries were assessed [22]. The results showed that crowdsourcing workers could perform almost twice as fast as students in judgments.

4. Experimental Design in Crowdsourcing

Beyond the workers’ attention, culture, preferences and skills, the presentation and properties of HITs have influence on the quality of the crowdsourcing experiment. The careful and precise task design, rather than filtering the poor quality work after accomplishing the task, is an important task since the complicated tasks distract cheaters [23]. It is possible that qualified workers accomplish erroneous tasks, if the user interface, instructions and design are of poor quality. Different phases in the implementation of crowdsourcing experiment are similar to software development as shown in Figure 5. The details of each phase are identified for creating relevance judgment task in this figure.
The impact of the two methods of HIT design, the full and simple design, was assessed by Kazai, Kamps, Koolen and Milic-Frayling [24] on the accuracy of labels. The full design method pre-qualified and restricts workers while the simple design method includes less quality control setting and restriction on the workers. However, the output of the full design method has a higher quality. In general, crowdsourcing is a useful solution for relevance assessment if the tasks are designed carefully and the methods for aggregating labels are selected appropriately. Experiment design is discussed based on three categories: (i) task definition, description and preparation, (ii) interface design and (iii) granularity.

4.1. Task definition, description and preparation. The information that is presented to the workers about crowdsourcing task is task definition. An important issue of implementing a crowdsourcing experiment is task description that is a part of task preparation. Clear instruction which is a part of task description is crucial to have a quick result usually in less than 24 hour. All workers should have a common understanding about the task and the words
used must be understandable by different workers [25]. Task description should consider the characteristics of workers such as language or expertise [26]. For example, to avoid jargon, plain English is suggested since the population is diverse [27]. As creating relevance judgment requires reading text, using plain English and simple words do help in a successful experiment. Having the phrase “I do not know” was suggested in response as it lets workers say that they do not have knowledge of to answer the question [25]. Getting user feedback at the end of the task by asking an open-ended question is recommended to improve the quality and avoid spammers. This kind of question can be optional. If the answer is useful, the requester can pay bonus [27]. In creating relevance judgments, one has to read text, hence instructions and questions should be readable, clear and presented with examples. An example of relevance judgment task is shown in Figure 6.

**Relevance judgment**

**Instruction**
Evaluate the relevance of a document to the given query.

**Task**
Please evaluate the relevance of the following document about Alzheimer.

*Dementia is a loss of brain function that occurs with certain diseases. Alzheimer's disease (AD), is one form of dementia that gradually gets worse over time. It affects memory, thinking, and behavior.*

Please rate the above document according to its relevance to Alzheimer as follows:

- o Highly relevant
- o Relevant
- o Not relevant

Please justify your answer:

**FIGURE 6:** An example of relevance judgment task
4.2. Interface design. The interface which users access and contribute to the tasks is user interface. Khanna, Ratan, Davis and Thies [28] suggested three design recommendations to have a doing well task completion along with paying less, they are: designing a user interface and instructions understandable by novice users, translating the instruction to the local language as well as English, and preparing a video tutorial. It also suggested the use of colours, highlights, bold, italics and typefaces to enhance comprehensibility [25]. Having verifiable questions within the task is also useful in validating if the workers are doing a good job which leads to a reduction in the number of worthless results since the users are aware that their answers are being analysed [29]. Another research work showed that more workers may be attracted to the user friendly and simple interface, and the quality of outcome may increase while reliable workers may be unenthusiastic about the unreasonably complex user interface which leads to delay [26].

4.3. Granularity. Tasks can be divided into three broad groups: routine, complex and creative tasks. The tasks which do not need specific expert are called routine such as creating relevance judgment. Complex tasks need some general skills like rewriting a given text. On the other hand, creative tasks needs specific skill and expertise such as research and development [30]. Submitting small tasks by splitting the long and complex tasks was recommended since small tasks tend to attract more workers [25]. However, some studies showed that novel tasks that need creativity tend to distract the cheaters and attract more qualified and reliable workers [23, 31]. In a separate study, Difallah, Demartini and Cudré-Mauroux [32] stated that designing complicated tasks is a solution to filter out the cheaters. On the other hand, money is the main reason why most of the workers accomplish the tasks. However, the workers who do the task because of the prospect of becoming a celebrity and fun are the least truthful while the workers who are provoked by fulfilment and fortune are the most truthful workers. Eickhoff and de Vries [23] found that the workers, who do the job
for entertainment, not for money, rarely cheat. So, designing and representing HITs in an exciting and entertaining way leads to better result along with cost efficiency.

5. Human Features and Monetary Factors in Crowdsourcing

Effect of human features like knowledge and monetary factors such as payment on accuracy and success of crowdsourcing experiments are discussed in this section. The discussion is based on three factors namely: (i) worker profile in crowdsourcing, (ii) payment in crowdsourcing, and (iii) compensation policy in crowdsourcing.

5.1. Worker profile in crowdsourcing. One of the factors that has an influence on the quality of results is worker profile. Worker profile consists of reputation and expertise. Requesters’ feedback about workers’ accomplishments creates reputation scores in the systems [33]. Requesters also should maintain a positive reputation since workers may refuse accepting HIT from poor requesters [16]. Expertise can be recognized by credentials and experience. Information such as language, location and academic degree are credentials while experience refers to knowledge that worker achieves through crowdsourcing system [26]. Location is another important factor that has strong effect on the accuracy of the results. A study conducted by Kazai, Kamps and Milic-Frayling [34] indicated that Asian workers performed poor quality work compared with American or European workers. The Asian mostly preferred simple design of HIT while American workers preferred Full design[34]. In crowdsourcing platforms such as AMT and Crowdflower, it is possible to limit the workers to the specific country while designing tasks. However, creating relevance judgment is a routine task and does not need specific expertise or experience.

5.2. Payment in crowdsourcing. Payment impacts on accuracy of the results in crowdsourcing since the workers who are satisfied with the payment accomplish the tasks more accurately than the workers who are not contented [31]. Monetary or non-monetary reasons can be the motivation for the workers of crowdsourcing platforms [35]. A study conducted by Ross,
Irani, Silberman, Zaldivar and Tomlinson [36] to find the motivation of the workers in crowdsourcing showed that money is the main incentive of 13% of Indian and 5% US workers. Another study carried out by Ipeirotis [37] reported that AMT was the main income of 27% of Indians and 12% of US workers. Kazai [38] stated that higher payment leads to better quality of work while Potthast, Stein, Barrón-Cedeño and Rosso [39] reported that higher payment only has effect on completion time rather than on the quality of results.

Kazai, Kamps and Milic-Frayling [31] analysed the effect of higher payment to the workers on accuracy of the results. Although higher payment encourages more qualified workers in contrast with lower payment which upsurges the sloppy workers, sometimes unethical workers are attracted by higher payment. In this case, strong quality control method can prevent poor result as well as filtering. In another study conducted by Mason and Watts [40], and Heer and Bostock [41], reported that by increasing payment, it leads to increase in quantity rather than quality while some studies showed that considering greater payment may only influence getting the task done faster but not better as increasing payment incentive speeds up work. However, a reasonable pay is a better cautious solution as high pay tasks attract spammers as well as legitimate workers [42]. Indeed, the payment amount should be fair and based on the complexity of the tasks.

Payment method is another factor which impacts on the quality of the outputs of the workers [40, 43]. For instance, a task which requires finding 10 words in a puzzle, more puzzles would be solved with method of payment per puzzle rather than payment per word [44]. It is better to pay workers after the work quality has been confirmed by measuring the accuracy and clarification that payment is based on the quality of the completed work. This can help in achieving better results from workers [25]. In an experiment conducted by Alonso et al., the workers were paid $0.02 for relevancy evaluation that took about one minute
consisting of judging the relevancy of one topic and document [18]. In another experiment, the payment was $0.04 for judging one topic against ten documents [45].

5.3. Compensation in crowdsourcing. A proper compensation policy and inducement has an impact on the quality of the results [46, 47]. There are two types of incentives, extrinsic incentives like monetary bonus such as extra payment, and intrinsic incentives such as personal enthusiasm [26]. Scsekic, Truong and Dustdar [46] categorized rewards into psychological, monetary and material. Mason and Watts [40] found that using non-financial compensation to motivate workers such as enjoyable tasks or social rewards, is a better choice than financial rewards and have more effect on quality. Malone, Laubacher and Dellarocas [48] also reported that using social rewards is like harnessing intrinsic motivation to improve the quality of work. The use of intrinsic along with extrinsic incentives was recommended to motivate more reliable workers to perform a task. Furthermore, considering the task requirements, properties, and social factors such as income and interests of workers according to the profiles of requesters and workers are vital to select proper compensation policy [26]. Offering bonus for writing comments and justification or paying bonus to the qualified and good workers who perform the tasks accurately and precisely was recommended for relevance evaluation task [25].

In addition to task design, human features and monetary factors which are important to successful crowdsourcing experiment, quality control is another vital part of crowdsourcing experiment which will be elaborated in the following section.

6. Quality control in Crowdsourcing

Managing the workers’ act is a crucial part of crowdsourcing because of the sloppy workers who accomplish the tasks carelessly or spammers who trick and complete the tasks inaccurately. Kazai, Kamps and Milic-Frayling [49] determined the different worker types by
using different factors including label accuracy, HIT completion time and fraction of useful labels through behavioral observation as explained below (see Table 4). This can help to develop methods for HIT design to attract the proper workers for the job. Table 5 shows different types of workers classified based on their average precision [50].

**TABLE 4: Worker types based on their behavioral observation**

<table>
<thead>
<tr>
<th>Workers</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>diligent</td>
<td>Completed tasks precisely with a high accuracy and longer time spent on tasks</td>
</tr>
<tr>
<td>competent</td>
<td>Skilled workers with high accuracy and fast work.</td>
</tr>
<tr>
<td>sloppy</td>
<td>Completed tasks quickly without considering quality.</td>
</tr>
<tr>
<td>incompetent</td>
<td>Completed tasks in a time with low accuracy.</td>
</tr>
<tr>
<td>spammer</td>
<td>Did not deliver useful works.</td>
</tr>
</tbody>
</table>

**TABLE 5: Worker types based on their average precision**

<table>
<thead>
<tr>
<th>Workers</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>proper</td>
<td>Completed tasks precisely.</td>
</tr>
<tr>
<td>random spammer</td>
<td>Gave worthless answer.</td>
</tr>
<tr>
<td>semi-random spammer</td>
<td>Answered incorrectly on the most questions while answering correctly on few questions, hoping to avoid detecting as a spammer.</td>
</tr>
<tr>
<td>uniform spammer</td>
<td>Repeated answers.</td>
</tr>
<tr>
<td>sloppy</td>
<td>Not precise enough in their judgments.</td>
</tr>
</tbody>
</table>

In 2012, to detect random spammers, for each worker \( w \), Vuurens et al. [51] calculated the \( \text{RandomSpam} \) score as the average squared ordinal distance \( \text{ord}_{ij}^2 \) between their judgments \( j \in J_w \) and judgments by other workers on the same query-document pair \( i \in J_{j,w} \) as shown in Equation 1 below:

\[
\text{RandomSpam}_w = \frac{\sum_{j \in J_w} \sum_{i \in J_{j,w}} \text{ord}_{ij}^2}{\sum_{j \in J_w} |J_{j,w}|}
\] (1)
In addition, to detect the uniform spammers, the *UniformSpam* score is calculated as shown in Equation 2:

\[
UniformSpam_w = \frac{\sum_{s \in S} |s| (f_{s,J_w} - 1) (\sum_{j \in J_{s,w}} \sum_{i \in J_{j,w}} disagree_{ij})^2}{\sum_{s \in S} \sum_{j \in J_{s,w}} |J_{j,w}|}
\]  

(2)

Where \(disagree_{ij}\) is the number of disagreements between judgments \(f_{s,J_w}\) which happen within label sequence \(s\) and judgments \(J_{j,w}\) and \(S\) is a set of all possible label sequences \(s\) with length \(|s| = 2\) or \(3\) while \(f_{s,J_w}\) is the frequency at which label sequence \(s\) occurs within worker \(w\)’s time-ordered judgments \(J_w\).

Quality control in crowdsourcing area is defined as the extent to which the provided outcome fulfills the requirements of the requester. Quality control approaches can be categorized into: (i) design-time approaches which employed while the requester preparing the task before submission, and (ii) runtime approaches which used during or after doing the tasks. Another classification for quality control methods are: (i) filtering workers, and (ii) aggregating labels that discuss in this section [52]. In other words, filtering workers are design-time approaches and aggregating labels methods can be considered as run-time approaches [26].

6.1. Design-time methods. There are three approaches to select workers: (i) credential-based, (ii) reputation-based, and (iii) open-to-all. It is also possible to use the combination of the three approaches. Credential-based approach is used in systems where users are well profiled. However, this approach is not practical in crowdsourcing systems since users are usually unknown. Reputation-based approach selects workers based on their reputation like AMT by using approval rate parameter. Open-to-all approach such as in Wikipedia allows any worker to contribute and it is easy to use and implement but increase the number of unreliable workers [26]. Requesters are also able to implement their own qualification methods [17] or combine different methods. In total, there are various methods of filtering workers which
identify the sloppy workers at first and then exclude them. We have categorized the methods into five categories. Table 6 presents a list of these methods:

### TABLE 6: Design-time methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>Platform</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qualification test</td>
<td>Qualification test is a set of questions which the workers must answer to qualify for doing the tasks</td>
<td>AMT</td>
<td>IQ test [53]</td>
</tr>
<tr>
<td>Honey pots or gold standard data</td>
<td>Creating pre-defined questions with known answers is honey pots [54]. If the workers answer these questions correctly, they are marked as appropriate for this task.</td>
<td>Crowdflower</td>
<td>-</td>
</tr>
<tr>
<td>Qualification settings</td>
<td>Some qualification settings are set when creating HITs.</td>
<td>AMT, Crowdflower</td>
<td>using approval rate</td>
</tr>
<tr>
<td>Trap questions</td>
<td>This method is about designing HITs along with a set of questions with known answers to find unreliable workers [56].</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CAPTCHAs and reCAPTCHA</td>
<td>CAPTCHAs is an anti-spamming technique to separate computers and humans to filter automatic answers. A text of the scanned page is used to identify whether a human inputs data or spam is trying to trick the web application as only a human can go through the test [57]. reCAPTCHA is the development of CAPTCHAs [58].</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

When using qualification test, it is still possible that workers perform task carelessly. Moreover, there are two issues that researchers encounter with qualification test: (i) tasks with qualification test may take longer to complete as most workers prefer tasks without qualification test, and (ii) the cost of developing and maintaining the test continuously. Honey pots or gold standard is used when the results are clear and pre-defined [54]. The honey pots are faster than qualification test as it is easy and quick to identify workers who answer the questions at random. Combining qualification test and honey pots is also possible for quality control [25].
Qualification settings such as filtering workers especially based on origin have an important impact on the cheater rates. For instance, some assume that workers from developed countries have lesser cheater rates [23]. Setting approval rate has been used by AMT. AMT provides a metric called the approval rate to pre-filter the workers. It represents percentage of assignments the worker has performed and confirmed by the requester, over all tasks the worker has done. Generally, it is the total rating of each worker in the system. It is also possible to limit the tasks to the master workers. In AMT, the group of people who accomplish HITs with a high degree of accuracy across a variety of requesters are called master workers. Master workers expect to receive higher payment. Limiting the HITs to the master workers gets better quality results [31]. If the high approval rate is considered for quality control, it would take a longer time to complete experiment since the worker population may be decreasing [27]. In general, the setting should be done according to the task type. If the task is complicated, the approval rate and other setting should be set precisely or the task is assigned to master workers only. If the task is routine, ordinary workers with simple setting can be effective. Trap questions are also used to detect careless workers who do not read instruction of the task carefully. Kazai, Kamps, Koolen and Milic-Frayling [24] designed two trap questions to avoid this situation: “Please tick here if you did NOT read the instructions” and at each page “I did not pay attention”. All of the unreliable workers may not be detected by trap questions but it showed strange behavior and it can be effective both in discouraging and identifying spammers. In crowdsourcing, there are some programs or bots designed to accomplish HITs automatically [59] and these kinds of data definitely are of poor quality [17]. CAPTCHAs and reCAPTCHA are used in crowdsourcing experiment to detect superficial workers and random clicking on answers [28, 31]. These methods are easy to use and inexpensive.

It is also possible to use a combination of the above methods to filter workers. For example, a real time strategy applying different methods to filter workers was proposed in
recruiting workers and monitoring the quality of their works. At first, a qualification test was used to filter the sloppy workers, and the completion time of HITs was calculated to reflect on the truthfulness of the workers. Then, a set of gold questions were used to evaluate the skills of workers [60]. However, by using these methods to filter workers, there is the possibility of lesser quality being recruited to do a work. The quality control methods used after filtering workers in run time are explained in subsequent section.

6.2. Run-time methods. Although design time techniques can intensify quality, there is the possibility of low quality because of misunderstanding while doing the tasks. Therefore, runtime techniques are vital to high quality results. Indeed, quality of the results would be increased by applying both approaches [26]. In crowdsourcing, if we assume that one judgment per each example or task called single labeling method, the time and cost may be saved. However, the quality of work is dependent on an individual’s knowledge, and in order to solve this issue of single labeling methods, integrating the labels from multiple workers was introduced [61, 62]. If labels are noisy, multiple labels can be desirable to single labeling even in the former setting when labels are not particularly low-cost [63]. It leads to have more accuracy for relevance judgments [64]. An important issue related to multiple labels per example is how to aggregate labels accurately and efficiently from various workers into a single consensus label. Run-time quality control methods are explained in Table 7.

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Voting (MV)</td>
<td>MV is a straightforward and common method which eliminates the wrong results by using the majority decision [30, 61, 65].</td>
</tr>
<tr>
<td>Expectation Maximization (EM) algorithm</td>
<td>Expectation Maximization (EM) algorithm measures the worker quality by estimating the accurate answer for each task through labels completed by different workers using maximum likelihood. The algorithm has two phases: (i) the correct answer is estimated for each task through multiple labels submitted by different workers, accounting for the quality of each worker (ii) comparing the assigned responses to the concluded accurate answer in order to estimate quality of each worker [66].</td>
</tr>
<tr>
<td>Method</td>
<td>Description</td>
</tr>
<tr>
<td>----------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Naive Bayes (NB)</td>
<td>Following EM, NB is a method to model the biases and reliability of single workers and correct them in order to intensify the quality of the workers’ results. According to gold standard data, a small amount of training data that labeled by expert was used to correct the individual biases of workers. The idea is to recolibrate answers of workers to be more matched with experts. An average of four inexpert labels for each example is needed to emulate expert level label quality. This idea helps to improve annotation quality [65].</td>
</tr>
<tr>
<td>Observation of the pattern of responses</td>
<td>Looking at the pattern of answers is another effective way of filtering unreliable responses as some untrustworthy workers have a regular pattern for example selecting the first choice of every question.</td>
</tr>
<tr>
<td>Probabilistic Matrix Factorization (PMF)</td>
<td>Using Probabilistic Matrix Factorization (PMF) that indues a latent feature vector for each worker and example to infers unobserved worker assessments for all examples [67]. PMF is an standard method in collaborative filtering through converting a crowdsourcing data to collaborative filtering data to predict unlabeled labels from workers [68, 69].</td>
</tr>
<tr>
<td>Expert review</td>
<td>Expert review uses experts to evaluate workers [70].</td>
</tr>
<tr>
<td>Contributor evaluation</td>
<td>The workers are evaluated according to quality factors such as their reputation, experience, or credentials. If the workers have enough quality factors, the requester would accept their tasks. For instance, Wikipedia would accept the article wrote by administrators without evaluation [26]. As an instance, the tasks that submitted by the workers who have a higher approval rate or master workers would be accepted without doubt.</td>
</tr>
<tr>
<td>Real-time support</td>
<td>The requesters give workers feedback about the quality of their work in real time while workers are accomplishing the task. This helps workers to amend their works and the results showed that self-assessment and external feedback improve the quality of the task [47]. Another real time approach was proposed by Kulkarni et al. where requesters can follow the workers workflows while solving the tasks. A tool called Turkomatic was presented which employees workers to do tasks for requesters. While workers are doing task, requesters are able to monitor the process and view the status of the task in real time [71].</td>
</tr>
</tbody>
</table>

The MV is a better choice for routine tasks which are paid a lower payment since it is easy to implement and achieve reasonable results depending on truthfulness of the workers [52]. The drawback of this method is that the consensus label is measured for a specific task without considering the accuracy of the workers in other tasks. Another drawback of this method is that MV considers all workers are equally good. For example, if there is only one
expert and the others are novices and the novices give the same inaccurate responses, the MV considers the novices’ answer as the correct answer because they are the majority.

In EM algorithm, a set of estimated accurate answers for each task and a set of matrixes that include the list of workers errors (in that the replies might not be the corrected answers) are the outputs. The error rate of each worker can be accessed by this confusion matrix. However, to measure the quality of a worker, the error rate is not adequate since workers may have completed the task carefully but with bias. For example, in labeling websites, parents with young children are more conservative in classifying the websites. To prevent this situation, a single scalar score can be assigned to each worker corresponding to the completed labels. The scores lead to separation of error rate from worker bias and satisfactory treatment of the workers. Approximately five labels are needed for each example for this algorithm for it to become accurate [72]. Recently Carpenter [73] and Raykar et al. [74], proposed a bayesian version of the EM algorithm by using confusion matrix. A probabilistic framework was proposed by Raykar, Yu, Zhao, Valadez, Florin, Bogoni and Moy [74] in the case of no gold standard with multiple labels. A specific gold standard is created repeatedly by the proposed algorithm and the performances of workers are measured and then the gold standard is refined according to the performance measurements. Hosseini, Cox, Milić-Frayling, Kazai and Vinay [64] evaluated the performance of EM and MV for aggregating labels in relevance assessment. His experiment results showed that the EM method is a better solution to problems related to the reliability of IR system ranking and relevance judgment especially in the case of small number of labels and noisy judgments. Another study conducted by Vuurens, de Vries and Eickhoff [50] compared MV and EM methods which stated that EM outperformed MV when 60% of workers were spammers. However, a combination of these two methods could produce more accurate relevance judgments.

Considering the time which takes to complete the task is an example of observation of the pattern of responses which used to determine random answers of unreliable workers [29].
As tasks that were completed fast deemed to have poor quality, completion time is a robust method of detecting the sloppy workers. Soleymani and Larson [75] used crowdsourcing for affective annotation of video in order to improve the performance of multimedia retrieval systems. The completion time of each HIT was compared to video length in order to evaluate quality. Zhu and Carterette [56] assessed the time that each worker spent on judgment in order to control quality. Three types of pattern were found: (i) normal pattern whereby the workers begin slowly and get faster when they learn about the task, (ii) periodic pattern, a peculiar behavior since some of the judgments are done fast and some slow, and (iii) interrupted pattern refers to interruption in the middle of doing tasks. This method in combination with other methods of quality control can be effective in crowdsourcing experiments since this method alone may not be able to detect all sloppy workers. Another method of quality control, expert review, is commonly applied in practice, but the drawbacks of this method are the high cost of employing experts and time consuming. This approach is not applicable in IR evaluation since crowdsourcing is used to lower the cost of hiring experts for creating relevance judgment. Therefore, if the expert review method is used to aggregate judgments, the cost would be increased.

The drawbacks of using crowdsourcing to create relevance judgment are: (i) each worker judges a small number of examples; (ii) to decrease cost, few judgments are collected per example. Therefore, the judgments are imbalanced and sparse as each worker assesses a small number of examples. The MV is vulnerable to this problem while EM indirectly addresses this problem and PMF tackles this issue directly [67].

Different quality control methods applied in crowdsourcing experiment in different areas of study were listed and discussed both in design-time approaches and run-time approaches. Although successful experiments are crucial to both approaches but the important point is that quality control method should be well-suited to crowdsourcing platform.
7. Conclusion and future work

Crowdsourcing is a new phenomenon which is being applied to various fields of study. IR evaluation experiments are one of the crowdsourcing applications. Although recent studies found that crowdsourcing is a viable alternative to the creation of relevance judgment, it needs precise design and appropriate quality control methods to assure the quality of outcomes since the workers of crowdsourcing platforms come from diverse cultural and linguistic background. One of the most important issues in crowdsourcing experiments is the untrustworthiness of workers who try to earn money without paying due attention to the task. Another issue is confusion or misunderstanding that happened because of the poor quality design. Quality control is a vital element in crowdsourcing experiment not only in designing phase but also in run-time phase. It should be applied in all stages of the experiment from planning to monitoring. User interface and instruction should be comprehensible as well as motivating and exciting as innovative tasks are less attractive to sloppy workers. The payment should be reasonable but monetary or non-monetary rewards can be useful for successful experiment. Even though there are efforts at developing crowdsourcing platforms in recent years, there are many limitations and more studies are needed to refine the various approaches, they are:

- currently a random topic-document is shown to the workers for them to judge, workers should be provided the freedom to choose their topic of interest for relevance judgment.
- to scale up creating relevance judgments for a larger set of topic-document to find out whether crowdsourcing is a replacement for hired assessors in terms of creating relevance judgments.
- to design a more precise grading system for workers rather than approval rate which requesters can find out each workers’ grade in each task type. For example, the requesters are able to know which workers have a good grade in creating relevance judgments.
• to access reliable personal information of workers such as expertise, language and background by requesters to decide which type of workers are allowed to accomplish the tasks.
• trying different aggregating methods in creating relevance judgments and evaluate the influence on correlation of judgments with TREC experts.

Crowdsourcing remains an interesting area that requires further research and development. This paper gives an insight into issues related to crowdsourcing experiments in all stages including design, implementation and monitoring. Hence, further research and development could be conducted to enhance the reliability and accuracy of relevance judgments in IR experimentation.

Acknowledgment

This research was supported by High Impact Research Grant UM.C/625/1/HIR/MOHE/FCSIT/14 from University of Malaya and Ministry of Education, Malaysia.

References

55. Crowdflower.
SIGIR conference on Research and development in information retrieval: ACM, pp. 1095-1096.


