Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 135641, 13 pages
http://dx.doi.org/10.1155/2014/135641

*Review Article*

# 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, 50603 Kuala Lumpur, Malaysia*

Correspondence should be addressed to Sri Devi Ravana; sdevi@um.edu.my

Received 30 December 2013; Accepted 8 April 2014; Published 19 May 2014

Academic Editors: L. Li, L. Sanchez, and F. Yu

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 evaluation of retrieval systems. 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 a 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, which 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.

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 judgment 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 need 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 $1, 2, \ldots, m$ are the contributing systems for the pool creation and a collection of

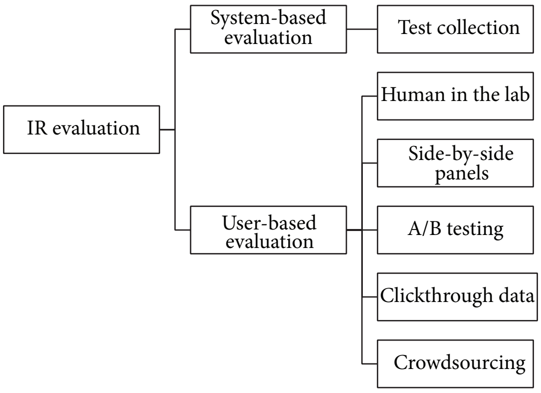


Figure 1: Classification of IR evaluation methods.

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 nonrelevant documents. Once the relevance judgments are ready, the whole set of runs retrieved by both contributing and noncontributing 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 scores are aggregated to achieve the overall performance score for a system. 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.

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].

The drawback of human involvement in the user-based evaluation experiments in labs is the high cost of experimentation set up and the difficulty in repeating the experiment. In addition, this method limited to a small set of information needs accomplished by a small number of human subjects. 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 conceived by Howe based on Web 2.0 technology in a Wired Magazine article [5]. Recently, the use of the term "crowdsourcing" for relevance judgment is increasing to conquer the high cost that current evaluation methods have through expert judges. Running experiments within low cost and fast turnaround makes 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 is measured; 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 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 study 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].

There are various platforms for crowdsourcing such as Amazon Mechanical Turk (AMT) (https://www.mturk.com/mturk/welcome), Crowdflower [16], ODesk (https://www.odesk.com/), and Elance (https://www.elance.com/). 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 and workers who accomplish the tasks on a crowdsourcing platform.

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.

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 pragrammatically or use template via dashboard. Crowdsourcing platforms were also suggested for data collection as a viable choice [17]. Mason and Suri [18] stated three advantages of crowdsourcing platforms: (i) allowing a 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. Precise design and quality control are required to optimize work quality and conduct a successful crowdsourcing experiment. In general, there are

Table 1: User-based evaluation methods.

| User-based methods | Description |
|---|---|
| Human in the lab | This method involves human experimentation in the lab to evaluate the user-system interaction |
| Side-by-side panels | This method is 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 retrieves better results |
| A/B testing | A/B testing involves numbers of preselected users of a website to analyse their reactions to the specific modification to see whether the change is positive or negative |
| Using clickthrough data | Clickthrough data is used to observe how frequently users click on retrieved documents for a given query |
| Crowdsourcing | Crowdsourcing is defined as outsourcing tasks, which was 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 |

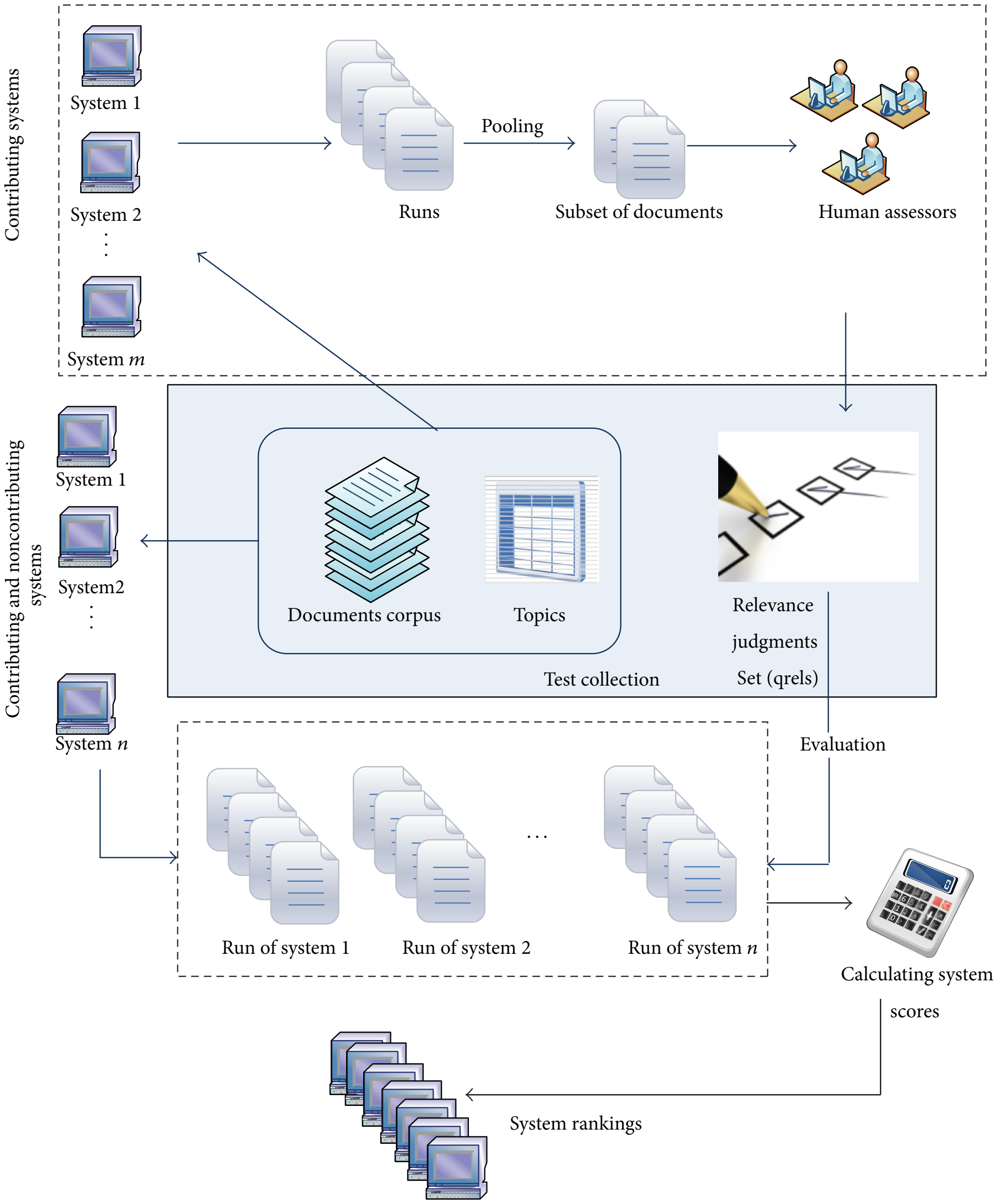


Figure 2: Typical IR evaluation process.

Table 2: Different applications of crowdsourcing.

| Application | Description |
|---|---|
| Natural language processing | Crowdsourcing technology was used to investigate linguistic theory and language processing [7] |
| Machine learning | Automatic translation by using active learning and crowdsourcing was suggested to reduce the cost of language experts [8, 9] |
| Software engineering | 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] |
| Network event monitoring | Using crowdsourcing to detect, isolate, and report service-level network events was explored which was called CEM (crowdsourcing event monitoring) [11] |
| Sentiment classification | The issues in training a sentiment analysis system using data collected through crowdsourcing were analysed [12] |
| Cataloguing | The application of crowdsourcing for libraries and archives was assessed [13] |
| Transportation plan | Use of crowdsourcing was argued to enable the citizen participation process in public planning projects [14] |
| Information retrieval | To create relevance judgments, crowdsourcing was suggested as a feasible alternative [15] |

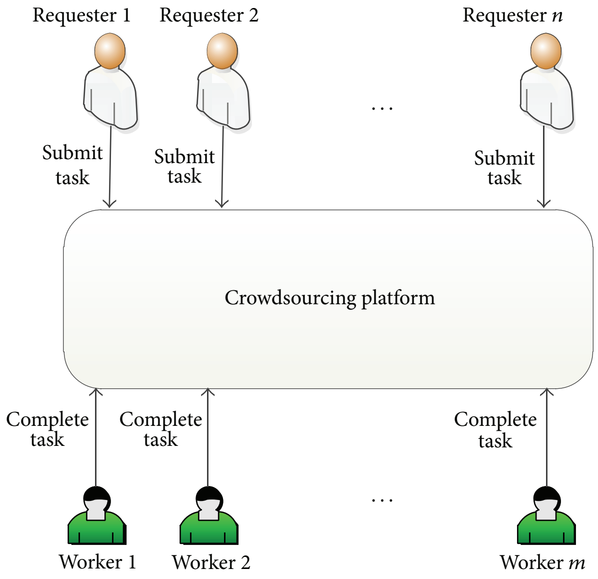


Figure 3: Crowdsourcing scheme.

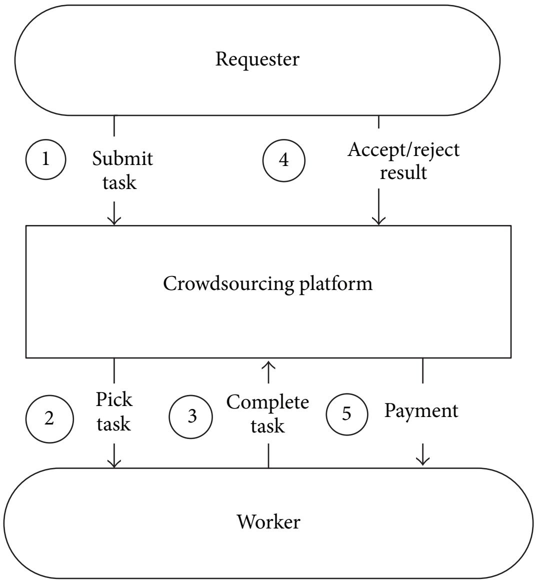


Figure 4: Flow of submitting and completing tasks via crowdsourcing.

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 an effect on the accuracy of the workers' results, and (iii) to find the approaches to maximize the quality of workers' results. A survey of crowdsourcing was done in terms of social, academic, and enterprise [19] while assessing different factors which cause motivation and participation of volunteers including the following:

(i) role-oriented crowdsourcing (leadership and ownership),

(ii) behaviour-oriented crowdsourcing (attribution, coordination, and conflict),

(iii) media-oriented crowdsourcing (mobile computing and ubiquitous computing).

However, this study looks from the academic aspect of crowdsourcing specifically focusing on factors that contribute to a successful and efficient crowdsourcing experiment in the area of information retrieval evaluation.

## 3. Accuracy and Reliability of Relevance Judgments in Crowdsourcing

The use of crowdsourcing in creating relevance judgment for an IR evaluation is generally validated through measuring the agreement between crowdsourcing workers and human assessors. This is to see whether crowdsourcing is a reliable replacement for human assessors. In crowdsourcing experiments, the interrater or interannotator agreement is used to measure the performance and to analyse the agreement between crowdsourcing workers and human assessors.

The score of homogeneity in the rating list given by judges is the interrater agreement. Various statistical methods are used to calculate the interrater agreement. Table 3 summarizes four common methods suggested to calculate the interrater agreement between crowdsourcing workers and human assessors for relevance judgment in IR evaluation [20].

Alonso et al. [15], pioneers of crowdsourcing in IR evaluation, applied crowdsourcing through Amazon Mechanical Turk on TREC data. Their results showed that crowdsourcing is a low cost, reliable, and quick solution and could be considered an alternative for creating relevance judgment by experts but it is not a replacement for current methods. This is due to the fact that there are still several gaps and questions that should be addressed by future researches. The scalability of crowdsourcing approach in IR evaluation has not been investigated yet [20].

## 4. Experimental Design in Crowdsourcing

Beyond the workers' attention, culture, preferences, and skills, the presentation and properties of HITs have an 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 [24]. It is possible that qualified workers accomplish erroneous tasks if the user interface, instructions, and design have a poor quality. Different phases in the implementation of crowdsourcing experiment are similar to software development. The details of each phase are identified for creating relevance judgment task as follows.

*Planning*

(i) Choosing a crowdsourcing platform.

(ii) Defining a task.

(iii) Preparing a dataset.

*Design and Development*

(i) Selecting topics.

(ii) Choosing a few documents for each topic (relevant and nonrelevant).

(iii) Task definition, description, and preparation.

*Verifying*

(i) Debugging by using internal expert team.

(ii) Take note of shortcoming of the experiment.

(iii) Using external inexpert to test.

(iv) Reviewing and clearing the instruction.

(v) Comparing the results of experts with inexpert teams.

*Publishing*

(i) Monitoring/work quality.

The impact of the two methods of HIT design, the full and simple design, was assessed on the accuracy of labels [25]. The full design method prequalifies 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 hours. All workers should have a common understanding about the task and the words used must be understandable by different workers [26]. Task description should consider the characteristics of workers such as language or expertise [27]. For example, to avoid jargon, plain English is suggested since the population is diverse [28]. As creating relevance judgments 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 allows workers to convey that they do not have sufficient knowledge to answer the question [26]. 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 [28]. 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 below.

*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:

(i) Highly relevant

(ii) Relevant

(iii) Not relevant

Please justify your answer: — — — — — —

Table 3: Statistics for calculating the interrater agreement.

| Methods | Description |
|---|---|
| Joint-probability of agreement (percentage agreement) [20] | The simplest and easiest measure based on dividing number of times for each rating (e.g.,1, 2, . . . , 5), assigned by each assessor, by the total number of the ratings |
| Cohen's kappa [21] | A statistical measure to calculate interrater agreement among raters. This measurement is more robust than percentage agreement since this method considers the effects of random agreement between two assessors |
| Fleiss' kappa [22] | An extended version of Cohen's kappa. This measurement considers the agreement among any number of raters (not only two) |
| Krippendorff's alpha [23] | The measurement is based on the overall distribution of assessors regardless of which assessors produced the judgments |

*4.2. Interface Design.* Users or workers access and contribute to the tasks via user interface. Three design recommendations were suggested 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 [29]. It also suggested the use of colours, highlights, bold, italics, and typefaces to enhance comprehensibility [26]. Verifiable questions such as common-knowledge questions within the task are also useful to validate the workers. It leads to a reduction in the number of worthless results since the users are aware that their answers are being analysed [30]. 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 [27].

*4.3. Granularity.* Tasks can be divided into three broad groups: routine, complex, and creative tasks. The tasks which do not need specific expertise 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 need specific skill and expertise such as research and development [31]. Submitting small tasks by splitting the long and complex tasks was recommended since small tasks tend to attract more workers [26]. However, some studies showed that novel tasks that need creativity tend to distract the cheaters and attract more qualified and reliable workers [24, 32]. In a separate study, Difallah et al. [33] 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 fulfillment and fortune are the most truthful workers. Eickhoff and de Vries [24] 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 lead to better results 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 is 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 have 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 [34]. Requesters also should maintain a positive reputation since workers may refuse accepting HIT from poor requesters [17]. Expertise can be recognized by credentials and experience. Information such as language, location, and academic degree is credentials while experience refers to knowledge that worker achieves through crowdsourcing system [27]. Location is another important factor that has a strong effect on the accuracy of the results. A study conducted by Kazai et al. [35] indicated that Asian workers performed poor quality work compared with American or European workers. The Asians mostly preferred simple design of HIT while American workers preferred full design. 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 [32]. Monetary or nonmonetary reasons can be the motivation for the workers of crowdsourcing platforms [36]. A study conducted by Ross et al. [37] to find the motivation of the workers in crowdsourcing showed that money is the main incentive of 13% of Indian and 5% of US workers. Another study carried out by Ipeirotis [38] reported that AMT was the main income of 27% of Indians and 12% of US workers. Kazai [39] stated that higher payment leads to better quality of work while Potthast et al. [40] reported that higher payment only has an effect on completion time rather than on the quality of results.

Kazai et al. [32] 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. Other studies 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 [41, 42]. However, a reasonable pay is a better cautious solution as high pay tasks attract spammers as well as legitimate workers [43]. 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 [41, 44], 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 [45]. 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 [26]. In an experiment conducted by Alonso and Mizzaro [20], the workers were paid $0.02 for relevancy evaluation that took about one minute consisting of judging the relevancy of one topic and document. In another experiment, the payment was $0.04 for judging one topic against ten documents [46].

*5.3. Compensation in Crowdsourcing.* A proper compensation policy and inducement has an impact on the quality of the results [47, 48]. There are two types of incentives, extrinsic incentives like monetary bonus such as extra payment and intrinsic incentives such as personal enthusiasm [27]. Rewards were categorized into psychological, monetary, and material [47]. Mason and Watts [41] found that using nonfinancial compensation to motivate workers such as enjoyable tasks or social rewards is a better choice than financial rewards and have more effect on quality. It was also reported that using social rewards is like harnessing intrinsic motivation to improve the quality of work [49]. 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 is vital in selecting proper compensation policy [27]. 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 [26].

In addition to task design, human features and monetary factors are important to have a successful crowdsourcing experiment; quality control is another vital part of crowdsourcing experiment which will be elaborated on 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. Different worker types by using different factors including label accuracy, HIT completion time, and fraction of useful labels were determined through behavioral observation [50] as explained below (see Table 4). This can help in developing 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 [51].

In 2012, to detect random spammers, for each worker $w$, Vuurens and de Vries [52] calculated the RandomSpam score as the average squared ordinal distance (it shows how the judgment of one worker is different from the other worker), $ord_{ij}^{2}$ 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 below:

$$\text{RandomSpam}_w = \frac{\sum_{j \in J_w} \sum_{i \in J_{j,\overline{w}}} ord_{ij}^{2}}{\sum_{j \in J_w} \left| J_{j,\overline{w}} \right|}. \tag{1}$$

In addition, to detect the uniform spammers, the *UniformSpam* score is calculated as shown below:

$$\text{UniformSpam}_w = \frac{\sum_{s \in S} |S| \cdot \left( f_{s,J_w} - 1 \right) \cdot \left( \sum_{j \in J_{s,w}} \sum_{i \in J_{j,\overline{w}}} \text{disagree}_{ij} \right)^2}{\sum_{s \in S} \sum_{j \in J_{s,w}} \left| J_{j,\overline{w}} \right|}, \tag{2}$$

where $\text{disagree}_{ij}$ is the number of disagreements between judgments $J_{s,w}$ 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 were employed while the requester prepared the task before submission and (ii) run-time approaches which were used during or after doing the tasks. Another classification for quality control methods is (i) filtering workers and (ii) aggregating labels that are discussed in this section [53]. In other words, filtering workers are design-time approaches and aggregating labels methods can be considered as run-time approaches [27].

*6.1. Design-Time Methods.* There are three approaches to select workers: (i) credential-based, (ii) reputation-based,

Table 4: Worker types based on their behavioral observation.

| Workers | Description |
|---|---|
| Diligent | Completed tasks precisely with a high accuracy and longer time spent on tasks |
| Competent | Skilled workers with high accuracy and fast work |
| Sloppy | Completed tasks quickly without considering quality |
| Incompetent | Completed tasks in a time with low accuracy |
| Spammer | Did not deliver useful works |

Table 5: Worker types based on their average precision.

| Workers | Description |
|---|---|
| Proper | Completed tasks precisely |
| Random spammer | Gave a worthless answer |
| Semirandom spammer | Answered incorrectly on most questions while answering correctly on few questions, hoping to avoid detection as a spammer |
| Uniform spammer | Repeated answers |
| Sloppy | Not precise enough in their judgments |

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 increases the number of unreliable workers [27]. Requesters are also able to implement their own qualification methods [18] 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.

When using qualification test, it is still possible that workers perform tasks 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 predefined [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 [26].

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 [24]. Setting approval rate has been used by AMT. AMT provides a metric called the approval rate to prefilter 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 [32]. If the high approval rate is considered for quality control, it would take a longer time to complete the experiment since the worker population may be decreasing [28]. In general, the setting should be done according to the task type. If the task is complicated, the approval rate and other settings 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 the instructions of the task carefully. Kazai et al. [25] 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 [55] and these kinds of data definitely are of poor quality [18]. CAPTCHAs and reCAPTCHA are used in crowdsourcing experiment to detect superficial workers and random clicking on answers [29, 32]. 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 [56]. 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 the subsequent section.

*6.2. Run-Time Methods.* Although design-time techniques can intensify quality, there is the possibility of low quality

Table 6: Design-time methods.

| Method | Description | Platform | Example |
|---|---|---|---|
| Qualification test | Qualification test is a set of questions which the workers must answer to qualify for doing the tasks | AMT | IQ test [57] |
| Honey pots or gold standard data | Creating predefined questions with known answers is honey pots [54]. If the workers answer these questions correctly, they are marked as appropriate for this task | Crowdflower [16] | — |
| Qualification settings | Some qualification settings are set when creating HITs | AMT, Crowdflower | Using approval rate |
| Trap questions | This method is about designing HITs along with a set of questions with known answers to find unreliable workers [58] | — | — |
| CAPTCHAs and reCAPTCHA | CAPTCHAs are an antispamming technique to separate computers and humans to filter automatic answers. A text of the scanned page is used to identify whether human inputs data or spam is trying to trick the web application as only a human can go through the test [59]. reCAPTCHA is the development of CAPTCHAs [60] | — | — |

because of misunderstanding while doing the tasks. Therefore, run-time techniques are vital to high quality results. Indeed, quality of the results would be increased by applying both approaches [27]. In crowdsourcing, if we assume that one judgment per each example or task is 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 having more accuracy for relevance judgments [64]. An important issue related to multiple labels per example is how to aggregate labels accurately and effieciently from various workers into a single consensus label. Run-time quality control methods are explained in Table 7.

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 [53]. 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 [65]. Recently Carpenter [66] and Raykar et al. [67] proposed a Bayesian version of the EM algorithm by using confusion matrix. A probabilistic framework was proposed by Raykar et al. [67] 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 et al. [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 compared MV and EM methods which stated that EM outperformed MV when 60% of workers were spammers [51]. However, a combination of these two methods could produce more accurate relevance judgments.

Considering the time taken to complete the task is an example of observation of the pattern of responses which was used to determine random answers of unreliable workers [30]. 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. In another study, the time that each worker spent on judgment was assessed in order to control quality [58]. Three types of patterns 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, which 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

Table 7: Run-time methods.

| Method | Description |
|---|---|
| Majority voting (MV) | MV is a straightforward and common method which eliminates the wrong results by using the majority decision [31, 61, 68] |
| Expectation maximization (EM) algorithm | 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 and (ii) comparing the assigned responses to the concluded accurate answer in order to estimate quality of each worker [69] |
| Naive Bayes (NB) | 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 labeled by expert was used to correct the individual biases of workers. The idea is to recalibrate 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 [68] |
| Observation of the pattern of responses | 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 |
| Probabilistic matrix factorization (PMF) | Using probabilistic matrix factorization (PMF) that induces a latent feature vector for each worker and example to infer unobserved worker assessments for all examples [70]. PMF is a standard method in collaborative filtering through converting crowdsourcing data to collaborative filtering data to predict unlabeled labels from workers [71, 72] |
| Expert review | Expert review uses experts to evaluate workers [73] |
| Contributor evaluation | 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 written by administrators without evaluation [27]. For instance, the tasks that are submitted by the workers who have a higher approval rate or master workers would be accepted without doubt |
| Real-time support | 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 [48]. Another real-time approach was proposed by Kulkarni et al. [74] 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 the task, requesters are able to monitor the process and view the status of the task in real time |

method are the high cost of employing experts and being 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; and (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 [70].

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

## 7. Conclusion and Future Work

Crowdsourcing is a novel phenomenon being applied to various fields of studies. One of the applications of the crowdsourcing is in IR evaluation experiments. Several recent studies show that crowdsourcing is a viable alternative to the creation of relevance judgment; however, it needs precise design and appropriate quality control methods to be certain about the outcomes. This is particularly imperative since the workers of crowdsourcing platforms come from diverse cultural and linguistic backgrounds.

One of the important concerns in crowdsourcing experiments is the possibility of untrustworthiness of workers trying to earn money without paying due attention to their tasks. Confusion or misunderstanding in crowdsourcing experiments is another concern that may happen due to low quality in the experimental designs. Conclusively, the quality control in crowdsourcing experiments is not only a vital element in the designing phase but also in run-time phase. The quality control and monitoring strategies should be considered and examined during all stages of the experiments. Moreover, the user interface and instruction should be comprehensible, motivating, and exciting to the users. The payment should be reasonable but monetary or nonmonetary rewards can be also considerable. Although there are efforts in developing crowdsourcing experiments in IR area in recent years, there are several limitations (some are listed below) pending to refine the various approaches:

(i) currently, a random topic-document is shown to the workers to judge; workers should be provided with the freedom to choose their topic of interest for relevance judgment,

(ii) 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,

(iii) to design a more precise grading system for workers rather than approval rate where 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,

(iv) 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,

(v) trying different aggregating methods in creating relevance judgments and evaluating the influence on correlation of judgments with TREC experts.

Crowdsourcing is an exciting research area with several puzzles and challenges that require further researches and investigations. This review shed a light on some aspects of using crowdsourcing in IR evaluation with insight into issues related to crowdsourcing experiments in some of the main stages such as design, implementation, and monitoring. Hence, further research and development should be conducted to enhance the reliability and accuracy of relevance judgments in IR experimentation.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## 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

[1] E. M. Voorhees, "The philosophy of information retrieval evaluation," in *Evaluation of Cross-Language Information Retrieval Systems*, pp. 355–370, Springer, 2002.

[2] C. Cleverdon, "The Cranfield tests on index language devices," *Aslib Proceedings*, vol. 19, no. 6, pp. 173–194, 1967.

[3] S. I. Moghadasi, S. D. Ravana, and S. N. Raman, "Low-cost evaluation techniques for information retrieval systems: a review," *Journal of Informetrics*, vol. 7, no. 2, pp. 301–312, 2013.

[4] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: the Concepts and Technology Behind Search*, Addison-Wesley, 2011.

[5] J. Howe, "The rise of crowdsourcing," *Wired Magazine*, vol. 14, no. 6, pp. 1–4, 2006.

[6] Y. Zhao and Q. Zhu, "Evaluation on crowdsourcing research: current status and future direction," *Information Systems Frontiers*, 2012.

[7] R. Munro, S. Bethard, V. Kuperman et al., "Crowdsourcing and language studies: the new generation of linguistic data," in *Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 122–130, Association for Computational Linguistics, 2010.

[8] V. Ambati, S. Vogel, and J. Carbonell, "Active learning and crowd-sourcing for machine translation," in *Language Resources and Evaluation (LREC)*, vol. 7, pp. 2169–2174, 2010.

[9] C. Callison-Burch, "Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, vol. 1, pp. 286–295, Association for Computational Linguistics, August 2009.

[10] K. T. Stolee and S. Elbaum, "Exploring the use of crowdsourcing to support empirical studies in software engineering," in *Proceedings of the 4th International Symposium on Empirical Software Engineering and Measurement (ESEM '10)*, no. 35, ACM, September 2010.

[11] D. R. Choffnes, F. E. Bustamante, and Z. Ge, "Crowdsourcing service-level network event monitoring," in *Proceedings of the 7th International Conference on Autonomic Computing (SIGCOMM '10)*, pp. 387–398, ACM, September 2010.

[12] A. Brew, D. Greene, and P. Cunningham, "Using crowdsourcing and active learning to track sentiment in online media," in *Proceedings of the Conference on European Conference on Artificial Intelligence (ECAI '10)*, pp. 145–150, 2010.

[13] R. Holley, "Crowdsourcing and social engagement: potential, power and freedom for libraries and users," 2009.

[14] D. C. Brabham, "Crowdsourcing the public participation process for planning projects," *Planning Theory*, vol. 8, no. 3, pp. 242–262, 2009.

[15] O. Alonso, D. E. Rose, and B. Stewart, "Crowdsourcing for relevance evaluation," *ACM SIGIR Forum*, vol. 42, no. 2, pp. 9–15, 2008.

[16] Crowdflower, https://http://www.crowdflower.com/.

[17] G. Paolacci, J. Chandler, and P. G. Ipeirotis, "Running experiments on amazon mechanical turk," *Judgment and Decision Making*, vol. 5, no. 5, pp. 411–419, 2010.

[18] W. Mason and S. Suri, "Conducting behavioral research on amazon's mechanical turk," *Behavior Research Methods*, vol. 44, no. 1, pp. 1–23, 2012.

[19] Y. Pan and E. Blevis, "A survey of crowdsourcing as a means of collaboration and the implications of crowdsourcing for interaction design," in *Proceedings of the 12th International Conference on Collaboration Technologies and Systems (CTS '11)*, pp. 397–403, May 2011.

[20] O. Alonso and S. Mizzaro, "Using crowdsourcing for TREC relevance assessment," *Information Processing and Management*, vol. 48, no. 6, pp. 1053–1066, 2012.

[21] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[22] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[23] K. Krippendorff, "Estimating the reliability, systematic error and random error of interval data," *Educational and Psychological Measurement*, vol. 30, no. 1, pp. 61–70, 1970.

[24] C. Eickhoff and A. P. de Vries, "Increasing cheat robustness of crowdsourcing tasks," *Information Retrieval*, vol. 16, no. 2, pp. 121–137, 2013.

[25] G. Kazai, J. Kamps, M. Koolen, and N. Milic-Frayling, "Crowdsourcing for book search evaluation: Impact of HIT design on comparative system ranking," in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*, pp. 205–214, ACM, July 2011.

[26] O. Alonso, "Implementing crowdsourcing-based relevance experimentation: an industrial perspective," *Information Retrieval*, vol. 16, no. 2, pp. 101–120, 2013.

[27] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H. R. Motahari-Nezhad, E. Bertino, and S. Dustdar, "Quality control in crowdsourcing systems: issues and directions," *IEEE Internet Computing*, vol. 17, no. 2, pp. 76–81, 2013.

[28] O. Alonso and R. Baeza-Yates, "Design and implementation of relevance assessments using crowdsourcing," in *Advances in Information Retrieval*, pp. 153–164, Springer, 2011.

[29] S. Khanna, A. Ratan, J. Davis, and W. Thies, "Evaluating and improving the usability of mechanical turk for low-income workers in India," in *Proceedings of the 1st ACM Symposium on Computing for Development (DEV '10)*, no. 12, ACM, December 2010.

[30] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with mechanical turk," in *Proceedings of the 26th Annual CHI Conference on Human Factors in Computing Systems (CHI '08)*, pp. 453–456, ACM, April 2008.

[31] M. Hirth, T. Hoßfeld, and P. Tran-Gia, "Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms," *Mathematical and Computer Modelling*, vol. 57, no. 11-12, pp. 2918–2932, 2013.

[32] G. Kazai, J. Kamps, and N. Milic-Frayling, "An analysis of human factors and label accuracy in crowdsourcing relevance judgments," *Information Retrieval*, vol. 16, no. 2, pp. 138–178, 2013.

[33] D. E. Difallah, G. Demartini, and P. Cudré-Mauroux, "Mechanical cheat: spamming schemes and adversarial techniques on crowdsourcing platforms," in *Proceedings of the CrowdSearch Workshop*, Lyon, France, 2012.

[34] L. De Alfaro, A. Kulshreshtha, I. Pye, and B. T. Adler, "Reputation systems for open collaboration," *Communications of the ACM*, vol. 54, no. 8, pp. 81–87, 2011.

[35] G. Kazai, J. Kamps, and N. Milic-Frayling, "The face of quality in crowdsourcing relevance labels: demographics, personality and labeling accuracy," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pp. 2583–2586, ACM, 2012.

[36] L. Hammon and H. Hippner, "Crowdsourcing," *Wirtschaftsinf*, vol. 54, no. 3, pp. 165–168, 2012.

[37] J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson, "Who are the crowdworkers? Shifting demographics in mechanical turk," in *Proceedings of the 28th Annual CHI Conference on Human Factors in Computing Systems (CHI '10)*, pp. 2863–2872, ACM, April 2010.

[38] P. Ipeirotis, "Demographics of mechanical turk," 2010.

[39] G. Kazai, "In search of quality in crowdsourcing for search engine evaluation," in *Advances in Information Retrieval*, pp. 165–176, Springer, 2011.

[40] M. Potthast, B. Stein, A. Barrón-Cedeño, and P. Rosso, "An evaluation framework for plagiarism detection," in *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 997–1005, Association for Computational Linguistics, August 2010.

[41] W. Mason and D. J. Watts, "Financial incentives and the performance of crowds," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 2, pp. 100–108, 2010.

[42] J. Heer and M. Bostock, "Crowdsourcing graphical perception: using mechanical turk to assess visualization design," in *Proceedings of the 28th Annual CHI Conference on Human Factors in Computing Systems (CHI '10)*, pp. 203–212, ACM, April 2010.

[43] C. Grady and M. Lease, "Crowdsourcing document relevance assessment with mechanical turk," in *Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 172–179, Association for Computational Linguistics, 2010.

[44] J. Chen, N. Menezes, J. Bradley, and T. A. North, "Opportunities for crowdsourcing research on amazon mechanical turk," *Human Factors*, vol. 5, no. 3, 2011.

[45] A. Kittur, B. Smus, S. Khamkar, and R. E. Kraut, "CrowdForge: crowdsourcing complex work," in *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST '11)*, pp. 43–52, ACM, October 2011.

[46] P. Clough, M. Sanderson, J. Tang, T. Gollins, and A. Warner, "Examining the limits of crowdsourcing for relevance assessment," *IEEE Internet Computing*, vol. 17, no. 4, pp. 32–338, 2012.

[47] O. Scekic, H. L. Truong, and S. Dustdar, "Modeling rewards and incentive mechanisms for social BPM," in *Business Process Management*, pp. 150–155, Springer, 2012.

[48] S. P. Dow, B. Bunge, T. Nguyen, S. R. Klemmer, A. Kulkarni, and B. Hartmann, "Shepherding the crowd: managing and providing feedback to crowd workers," in *Proceedings of the 29th Annual CHI Conference on Human Factors in Computing Systems (CHI '11)*, pp. 1669–1674, ACM, May 2011.

[49] T. W. Malone, R. Laubacher, and C. Dellarocas, "Harnessing crowds: mapping the genome of collective intelligence," MIT Sloan School Working Paper 4732-09, 2009.

[50] G. Kazai, J. Kamps, and N. Milic-Frayling, "Worker types and personality traits in crowdsourcing relevance labels," in *Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM '11)*, pp. 1941–1944, ACM, October 2011.

[51] J. Vuurens, A. P. de Vries, and C. Eickhoff, "How much spam can you take? an analysis of crowdsourcing results to increase accuracy," in *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR '11)*, pp. 21–26, 2011.

[52] J. B. P. Vuurens and A. P. de Vries, "Obtaining high-quality relevance judgments using crowdsourcing," *IEEE Internet Computing*, vol. 16, no. 5, pp. 20–27, 2012.

[53] W. Tang and M. Lease, "Semi-supervised consensus labeling for crowdsourcing," in *Proceedings of the SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR '11)*, 2011.

[54] J. Le, A. Edmonds, V. Hester, and L. Biewald, "Ensuring quality in crowdsourced search relevance evaluation: the effects of training question distribution," in *Proceedings of the SIGIR Workshop on Crowdsourcing for Search Evaluation*, pp. 21–26, 2010.

[55] R. M. McCreadie, C. Macdonald, and I. Ounis, "Crowdsourcing a news query classification dataset," in *Proceedings of the SIGIR Workshop on Crowdsourcing for Search Evaluation (CSE '10)*, pp. 31–38, 2010.

[56] T. Xia, C. Zhang, J. Xie, and T. Li, "Real-time quality control for crowdsourcing relevance evaluation," in *Proceedings of the 3rd IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC '12)*, pp. 535–539, 2012.

[57] G. Zuccon, T. Leelanupab, S. Whiting, E. Yilmaz, J. M. Jose, and L. Azzopardi, "Crowdsourcing interactions: using crowdsourcing for evaluating interactive information retrieval systems," *Information Retrieval*, vol. 16, no. 2, pp. 267–305, 2013.

[58] D. Zhu and B. Carterette, "An analysis of assessor behavior in crowdsourced preference judgments," in *Proceedings of the SIGIR Workshop on Crowdsourcing for Search Evaluation*, pp. 17–20, 2010.

[59] L. Von Ahn, M. Blum, N. J. Hopper et al., "CAPTCHA: using hard AI problems for security," in *Advances in Cryptology—EUROCRYPT 2003*, pp. 294–311, Springer, 2003.

[60] L. Von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum, "reCAPTCHA: human-based character recognition via web security measures," *Science*, vol. 321, no. 5895, pp. 1465–1468, 2008.

[61] V. S. Sheng, F. Provost, and P. G. Ipeirotis, "Get another label? Improving data quality and data mining using multiple, noisy labelers," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08)*, pp. 614–622, ACM, August 2008.

[62] P. Welinder and P. Perona, "Online crowdsourcing: rating annotators and obtaining cost-effective labels," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition—Workshops (CVPRW '10)*, pp. 25–32, June 2010.

[63] P. G. Ipeirotis, F. Provost, V. S. Sheng, and J. Wang, "Repeated labeling using multiple noisy labelers," *Data Mining and Knowledge Discovery*, vol. 28, no. 2, pp. 402–441, 2014.

[64] M. Hosseini, I. J. Cox, N. Milić-Frayling, G. Kazai, and V. Vinay, "On aggregating labels from multiple crowd workers to infer relevance of documents," in *Advances in Information Retrieval*, pp. 182–194, Springer, 2012.

[65] P. G. Ipeirotis, F. Provost, and J. Wang, "Quality management on amazon mechanical turk," in *Proceedings of the Human Computation Workshop (HCOMP '10)*, pp. 64–67, July 2010.

[66] B. Carpenter, "Multilevel bayesian models of categorical data annotation," 2008.

[67] V. C. Raykar, S. Yu, L. H. Zhao et al., "Learning from crowds," *The Journal of Machine Learning Research*, vol. 11, pp. 1297–1322, 2010.

[68] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 254–263, Association for Computational Linguistics, October 2008.

[69] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Applied Statistics*, vol. 28, pp. 20–28, 1979.

[70] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," *Advances in Neural Information Processing Systems*, vol. 20, pp. 1257–1264, 2008.

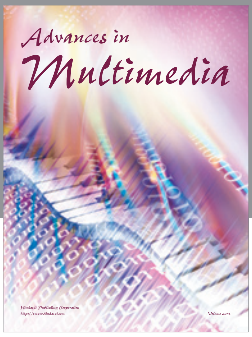
[71] H. J. Jung and M. Lease, "Inferring missing relevance judgments from crowd workers via probabilistic matrix factorization," in *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1095–1096, ACM, 2012.

[72] H. J. Jung and M. Lease, "Improving quality of crowdsourced labels via probabilistic matrix factorization," in *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, 2012.

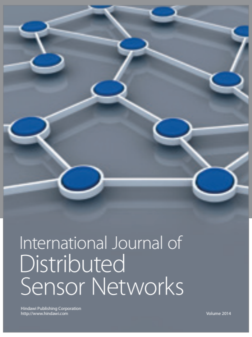
[73] A. J. Quinn and B. B. Bederson, "Human computation: a survey and taxonomy of a growing field," in *Proceedings of the 29th Annual CHI Conference on Human Factors in Computing Systems (CHI '11)*, pp. 1403–1412, ACM, May 2011.

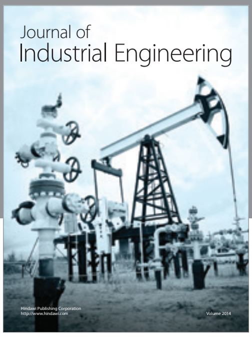
[74] A. Kulkarni, M. Can, and B. Hartmann, "Collaboratively crowdsourcing workflows with turkomatic," in *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW '12)*, pp. 1003–1012, February 2012.

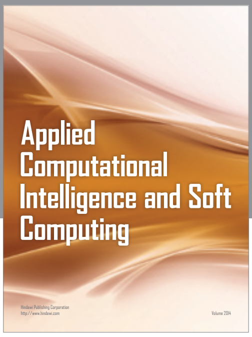
[75] M. Soleymani and M. Larson, "Crowdsourcing for affective annotation of video: development of a viewer-reported boredom corpus," in *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Search Evaluation (CSE '10)*, pp. 4–8, 2010.

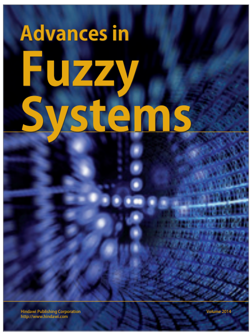

Advances in
Multimedia

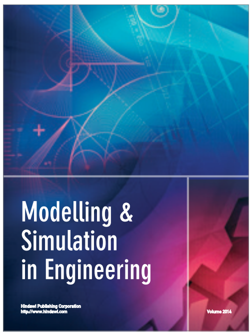
The Scientific
World Journal

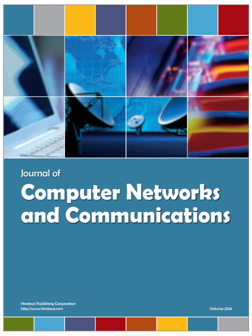

International Journal of
Distributed
Sensor Networks

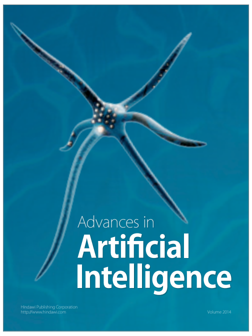

Journal of
Industrial Engineering

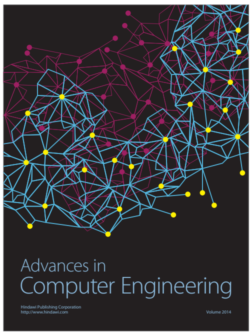

Applied
Computational
Intelligence and Soft
Computing


Advances in
Fuzzy
Systems

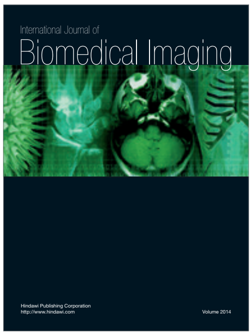

Modelling &
Simulation
in Engineering

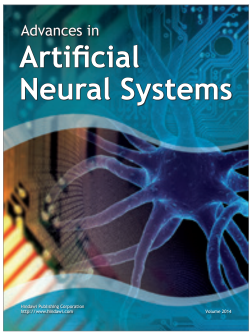

Journal of
Computer Networks
and Communications

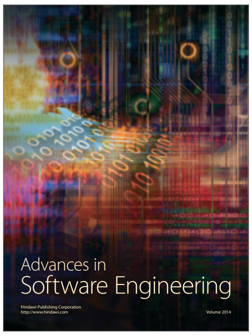

Advances in
Artificial
Intelligence


Advances in
Computer Engineering

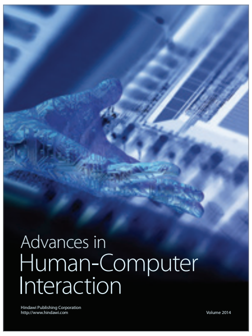
International Journal of
Computer Games
Technology

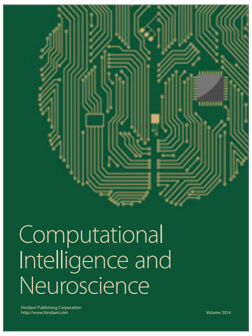

International Journal of
Biomedical Imaging


Advances in
Artificial
Neural Systems


Advances in
Software Engineering

Journal of
Robotics


Advances in
Human-Computer
Interaction


Computational
Intelligence and
Neuroscience

International Journal of
Reconfigurable
Computing

Journal of
Electrical and Computer
Engineering