# On Aggregating Labels from Multiple Crowd Workers to Infer Relevance of Documents

Mehdi Hosseini[1], Ingemar J. Cox[1], Nataša Milić-Frayling[2],
Gabriella Kazai[2], and Vishwa Vinay[2]

[1] Computer Science Department, University College London, UK
{m.hosseini,i.cox}@cs.ucl.ac.uk
[2] Microsoft Research, Cambridge. UK
{natasamf,v-gabkaz,vvinay}@microsoft.com

**Abstract.** We consider the problem of acquiring relevance judgements for information retrieval (IR) test collections through crowdsourcing when no true relevance labels are available. We collect multiple, possibly noisy relevance labels per document from workers of unknown labelling accuracy. We use these labels to infer the document relevance based on two methods. The first method is the commonly used majority voting (MV) which determines the document relevance based on the label that received the most votes, treating all the workers equally. The second is a probabilistic model that concurrently estimates the document relevance and the workers accuracy using expectation maximization (EM). We run simulations and conduct experiments with crowdsourced relevance labels from the INEX 2010 Book Search track to investigate the accuracy and robustness of the relevance assessments to the noisy labels. We observe the effect of the derived relevance judgments on the ranking of the search systems. Our experimental results show that the EM method outperforms the MV method in the accuracy of relevance assessments and IR systems ranking. The performance improvements are especially noticeable when the number of labels per document is small and the labels are of varied quality.

## 1 Introduction

In information retrieval (IR), test collections are typically used to evaluate the performance of IR systems. A test collection consists of a document corpus, a set of search queries, and a set of relevance judgments for each query. Relevance judgments indicate which documents in the corpus are relevant to a query and are usually created by employing human assessors.

Traditionally, the assessors are trained experts. However, as the corpus and the number of test queries grow, the cost of acquiring relevance labels from expert judges for a sufficiently large number of documents becomes prohibitive. In response to this problem, the IR community has recently been exploring the use of crowdsourcing services to obtain relevance judgments at scale, see e.g., the TREC Relevance Feedback track.

R. Baeza-Yates et al. (Eds.): ECIR 2012, LNCS 7224, pp. 182–194, 2012.

Web services, such as Amazon's Mechanical Turk (www.mturk.com), facilitate the collection of relevance judgments by temporarily hiring thousands of crowd workers. While the labels provided by the workers are relatively inexpensive to acquire, they vary in quality, introducing noise into the relevance judgments and, consequently, causing inaccuracies in the system evaluation [1]. In order to address the issue of noisy labels, it is common to collect multiple labels per document from different workers, in the hope that the consensus across multiple labels would lead to more accurate relevance assessments.

In this paper we assume that a set of labels is collected for each document from multiple crowd workers and that the accuracy of each worker is unknown, as is the true relevance of the documents. Our aim is to estimate both the true relevance of the documents and the accuracy of the workers. We evaluate two methods: the majority voting (MV) that has been frequently used for label aggregation in IR ([1],[2],[3]) and the probabilistic model for estimating both the relevance of the documents and the workers accuracy using the expectation maximization algorithm (EM) as in [19].

The main contributions of this paper are (1) the use of the EM based method to create relevance judgments for IR test collections and (2) empirical evidence that the EM method offers more reliable relevance estimations than the MV method, especially when labels collected for a document are few or varied in quality.

In the following section we discuss the related work. In Section 3 we present details of the EM method. In Section 4 we describe simulations with synthetic data and experiments with the crowdsourced labels from the INEX 2010 Book Search track to compare the MV and EM methods. We consider crowdsourced labels from two task designs that lead to different level of noise and observe their effect on estimating relevance judgments and system rankings. In Section 5 we summarize the results and conclude with a discussion of future research directions.

## 2 Related Work

Relevance assessment errors in IR test collections have been considered by the IR community since the early Cranfield experiments [4]. Voorhees [5] studied the effects of variability in relevance judgments on the stability of comparative IR systems evaluation. She considered three sets of relevance judgments for the TREC-4 test collection and observed a significant disagreement among them. She explored the effects of judgments on the ranking of the systems that participated in TREC-4 and observed no significant changes in the systems ranking. This was attributed to the stability of the average precision (AP) metric [6] that was used to evaluate the systems' performance. Indeed, AP is calculated based on *deep pools* of documents obtained from the participating systems. Thus, a few incorrect judgments in a ranked list do not significantly affect the values of AP and, therefore, do not perturb the ordering of the systems. In our experiments (in Section 4) we confirm that a deep pool of judged documents can reduce the effect of noisy crowdsourced labels in the systems evaluation.

Recent trends in IR evaluations involve the use of large numbers of topics to enhance the reliability of the evaluation [7] while reducing the pool depths and, with that, the cost of acquiring relevance judgments [8]. However, the use of the AP metric with shallow document pools becomes more sensitive to assessment errors and leads to significant changes in systems rankings [9]. This has motivated studies of the factors that cause assessment errors such as the level of assessors' expertise [10], the presentation of the documents for assessment, such as the sequence in which the documents are shown to the assessors ([1],[11]), and the assessors' behavior [9].

Awareness of the assessment errors has further increased with the use of crowdsourcing services to supplement or replace the traditional ways of collecting relevance judgments. In crowdsourcing, the relevance assessment task is expressed in terms of a *human intelligence task* (HIT) that is presented to crowd workers through a crowdsourcing platform to solicit their engagement, typically for a specified fee. The effectiveness of the crowdsourcing approach has been investigated in terms of various factors, including (*i*) the agreement with relevance judgments from trusted assessors [3], (*ii*) quality assurance techniques for detecting and removing unreliable workers [1], and (*iii*) the cost incurred due to redundant relevance assessments that are needed for quality assurance, e.g., [12] and [13].

The use of multiple labels per document to improve the quality of relevance judgments involves label aggregation across the assessors, e.g., by arriving at a consensus through majority voting ([2], [14]). The effectiveness of the consensus approach has been assessed by Kazai et al. [1] for IR tasks involving TREC and INEX test collections. Kumar and Lease [15] investigated the relationship between the document relevance and the workers' accuracy by using a set of documents with known relevance as training data for a naïve Bayes method. The trained model estimated the relevance of new documents by aggregating labels based on worker accuracy.

In this paper, we expand the existing body of research by applying the EM method from [19] to infer the relevance of documents from multiple, possibly noisy labels. In contrast to [15], we assume that no authoritative relevance assessments are available and estimate both the accuracy of the workers and the document relevance from the crowdsourced labels. Similar probabilistic models have been used in other research areas. For instance, Kasneci et al. [16] used a Bayesian probabilistic model in knowledge extraction systems to infer the relationships between entities from the input of multiple assessors. Welinder and Perona [17] proposed a probabilistic model for labeling images using crowd workers. Ipeirotis et al. [18] take a similar approach to identify systematic errors made by workers in the crowdsourcing experiments. Our aim is to infer true relevance judgments from crowdsourced labels and use them for evaluation of IR systems.

## 3 Aggregating Multiple Labels

In this section we describe the MV and EM methods that we use to aggregate multiple relevance labels from crowd workers and estimate both the true relevance judgments and the reliability of the workers.

Consider a set of $N$ documents and a set of $M$ workers that provide relevance labels for the documents. We assume that the relevance of a document is a discrete variable with values in $\{0,1,\dots,G\}$. If the relevance value of a document $i$ is $k$ ($k \in \{0,1,\dots,G\}$), then its $G$+1 dimensional vector $R_i$ is a binary vector with *k-th* component 1 and the rest 0, i.e., $R_{ik}$=1 and $R_{ij}$=0, $(\forall j \neq k)$. We now define a matrix $R$ of all the relevance vectors as $R \in \{0,1,\dots,G\}^{N\times(G+1)}$, comprising $N$ binary $R_i$ vectors.

Now consider a set of $M$ workers with the corresponding accuracies $A = \{a_1, a_2, \dots, a_M\}$, where $a_j$ represents the accuracy of the worker $j$. Both the document relevance $R$ and the workers accuracy $A$ are unknown to us. Instead, we have a set of relevance labels provided by the workers, i.e., $l_{ij}\in \{0,1,\dots,G\}$ is a relevance assessment of the document $i$ by the worker $j$. A worker may provide relevance labels for some or all the documents. Our goal is to estimate the true relevance value of the documents and the accuracy of the workers' assessments from a given set of labels $L$. We assume that each document receives at least one label and the accuracy of the labels is unknown. Thus, in contrast to [15], we assume no initial information regarding the workers' accuracy or the relevance of the documents.

### 3.1 Majority Voting

Consider a document $i$ with the corresponding labels provided by a set of workers. Let $n_{ig}$ be the number of times document $i$ is labeled as $g$, $g\in \{0,1,\dots,G\}$ by a set of workers. The majority voting assigns $g$ as the document's true relevance label if $n_{ig}$ is maximum. This technique is commonly used in IR experiments ( [1], [2], [3]).

### 3.2 Concurrent Estimation of Relevance and Accuracy

As an alternative to MV we consider the EM method for concurrent estimation of the document relevance and the workers accuracy. In this model the document relevance $R$ and the workers' accuracy $A$ are unknown variables and the labels $L$ provided by the workers are the observed data.

We take the same approach as [19] and consider the label aggregation model that assigns to each worker a $(G+1)\times(G+1)$ *latent confusion matrix* where $G$+1 is the number of different relevance grades. Each row refers to the true relevance value and each column refers to a relevance value assigned by a worker. Once the confusion matrix is calculated, we can determine the worker's expertise based on metrics such as accuracy, the true positive ratio and the true negative ratio [14].

Let $\pi_{kl}^{j}$ , $(\forall\ k\ \&\ l \in \{0,\dots,G\})$ be the probability that the worker $j$ provides a label $l$ given that $k$ is the true relevance value of an arbitrary document. The probability $\pi_{kl}^{j}$ is computed based on the confusion matrix for the worker $j$. One estimator of $\pi_{kl}^{j}$ is:

$$\pi_{kl}^{j} = \frac{\text{number of times worker } \boldsymbol{j} \text{ provides label } \boldsymbol{l} \text{ while the correct label is } \boldsymbol{k}}{\text{number of labels provided by worker } \boldsymbol{j} \text{ for documents of relevance } \boldsymbol{k}} \qquad (1)$$

where

$$\textstyle\sum_{l=0}^{G} \pi_{kl}^{j} = 1 \qquad (\forall\ \ k \in \{0,\dots,G\}, and\, j \in \{1,\dots,M\}).$$

Of course, the calculation of $\pi_{kl}^{j}$ assumes that $R$ is known. In the following we show how $\pi_{kl}^{j}$ and $R$ can be simultaneously estimated.

**Concurrent Estimations of R and $\boldsymbol{\pi}_{\mathbf{kl}}^{\mathbf{j}}$.** Let $p_k$ be the probability that a document drawn at random has a true relevance grade of $k$ ( $p_k = Pr[R_{ik} = 1]\,;\; i \in \{1, \dots, N\}$). Now let $n_{il}^{j}$ be the number of times worker $j$ provides label $l$ for document $i$; for our purpose $n_{il}^{j}$ is binary, so if a worker labels the document $n_{il}^{j} = 1$ otherwise $n_{il}^{j} = 0$. If $g$ is the true relevance grade of document $i$, $R_{ig}$=1, then the probability of the worker $j$ giving a grade $l$ is $\pi_{gl}^{j}$ and the probability of doing so $n_{il}^{j}$ times is $\left(\pi_{gl}^{j}\right)^{n_{il}^{j}}$. Thus, the number of labels of each grade $\{0,1, \dots, G\}$ provided by worker $j$ is distributed according to a multinomial distribution and its likelihood is proportional to

$$\Pr(n_{i0}^{j}, \dots, n_{iG}^{j};\, \pi_{g0}^{j}, \dots, \pi_{gG}^{j} | R_{ig} = 1) \propto \prod_{l=0}^{G} \left(\pi_{gl}^{j}\right)^{n_{il}^{j}}. \tag{2}$$

Under the assumption that $M$ workers independently label documents, the likelihood of labels provided for document $i$ when $R_{ig} = 1$ is also proportional to

$$\prod_{\mathrm{j=1}}^{\mathrm{M}} \Pr(n_{i0}^{j}, \dots, n_{iG}^{j};\, \pi_{g0}^{j}, \dots, \pi_{gG}^{j} | R_{ig} = 1) \propto \prod_{j=1}^{M} \prod_{l=0}^{G} \left(\pi_{gl}^{j}\right)^{n_{il}^{j}}$$

Since the value of $g$ is unknown, we compute the expectation of $\Pr(n_{i0}^{j}, \dots, n_{iG}^{j};\, \pi_{g0}^{j}, \dots, \pi_{gG}^{j})$ over all possible values of $g$, i.e., we compute the marginal probability over all possible values of $g$:

$$\sum_{k=0}^{G} p_k \prod_{j=1}^{M} \prod_{l=0}^{G} \left(\pi_{kl}^{j}\right)^{n_{il}^{j}}. \tag{3}$$

Also as the data from all documents are assumed to be independent, the joint probability distribution over all $N$ documents is

$$\prod_{i=1}^{N} \left( \sum_{k=0}^{G} p_k \prod_{j=1}^{M} \prod_{l=0}^{G} \left(\pi_{kl}^{j}\right)^{n_{il}^{j}} \right). \tag{4}$$

Equation (4) comprises mixtures of multinomial distributions. In order to estimate the quantities of interest, $p_k, \pi_{kl}^{j}$ and $R_{ig}$, we apply the expectation maximization (EM) [19]. In the EM algorithm we treat $\pi_{kl}^{j}$ and $p_k$as model parameters and $R_{ik}$ as missing data. The EM algorithm then involves the following steps:

1. Initialize $R_{ik}$ values, e.g., randomly choose $g$ and set $R_{ig} = 1$, and $R_{ik} = 0\ (\forall\, k \neq g)$.
2. Given the current estimate of $R_{ik}$, compute the *maximum likelihood* estimates of $\pi_{kl}^{j}$ and $p_k$, as

$$\hat{\pi}_{kl}^{j} = \frac{\sum_{i=1}^{N} R_{ik}\, n_{il}^{j}}{\sum_{l=0}^{G} \sum_{i=1}^{N} R_{ik}\, n_{il}^{j}} \;; \qquad \hat{p}_k = \frac{\sum_{i=1}^{N} R_{ik}}{N}.$$

3. Calculate the new estimate of $R_{ig}$ ($\forall g \in \{1, \dots, G\}$) based on $\hat{\pi}_{kl}^{j}$ and $\hat{p}_k$, as

$$\Pr\left(R_{ig} = 1 | n_{i1}^{\forall j}, \dots, n_{iG}^{\forall j};\, \pi_{g1}^{\forall j}, \dots, \pi_{gG}^{\forall j}\right) = \frac{p_g \prod_{j=1}^{M} \prod_{l=0}^{G} \left(\pi_{gl}^{j}\right)^{n_{il}^{j}}}{\sum_{k=0}^{G} p_k \prod_{j=1}^{M} \prod_{l=0}^{G} \left(\pi_{kl}^{j}\right)^{n_{il}^{j}}}. \qquad (5)$$

4. Repeat steps 2 and 3 until the results converge.
5. Finally, for each document $i$, set $R_{ig} = 1$ for the $g$ with the maximum probability as calculated in equation (5), and $R_{ik} = 0$ ($\forall\, k \neq g$).

Note that by combining $\hat{\pi}_{kl}^{j}$ values we can compute the accuracy of the worker $j$ or other statistics of interest, e.g., the true positive ratio. Accuracy is estimated as $\hat{a}_j \cong \frac{\sum_{l=0}^{G} \hat{\pi}_{ll}^{j}}{\sum_{l,k} \hat{\pi}_{lk}^{j}}$.

# 4 Experiments

In this section we describe a set of experiments that compare the aggregation of relevance labels based on the MV and EM methods and the implications for the IR systems evaluation. The experiments are based on both synthetic and crowdsourcing data collected for INEX 2010 Book Search evaluation track[1].

In the first experiment we use synthetic data and simulate the characteristics of the MV and EM methods. In the second experiment we assess the performance of the two methods based on crowdsourcing data. We then assess the accuracy of the MV and EM relevance assessments relative to the INEX official judgments. In the third experiment we investigate the impact of MV and EM relevance judgments on the system ranking using several performance metrics.

## 4.1 Experiment Design

In our experiments we use the test collection and crowdsourced relevance data from the INEX 2010 Book Search evaluation track [1]. The test collection comprises 50,239 books containing over 17 million scanned pages and 21 search topics with 169 judged pages per topic, on average. This amounts to 3,557 judged pages that serve as a gold standard set for IR systems evaluation. Each page is assigned a relevance judgment based on four grades {0,1,2,3}. In our experiments we assume that the relevance is binary and collapse labels {1,2,3} into label 1.

**Crowdsourcing Experiments**. Crowdsourced labels were collected for a subset of the test collection using the Mechanical Turk platform. The specific INEX task was

[1] http://www.inex.otago.ac.nz/tracks/books/books.asp

*prove it*: for a given search query, the user had to confirm whether the presented book page contains an answer to the search question. A search query and corresponding pages were presented to the crowd workers for relevance judgments in the form of HITs (Human Intelligence Tasks). Each HIT consisted of 10 pages including up to 3 pages judged as relevant by the INEX assessors. Two HIT designs, referred to as *simple* HIT and *full* HIT design, were used to control the workers' behavior and with that the label accuracy.

The *simple* HIT design included a minimal quality control using a single test question to capture random assignment of relevance labels by a worker. Furthermore, all the HITs were presented to the worker in a single batch, using the same generic HIT title, description, and keyword.

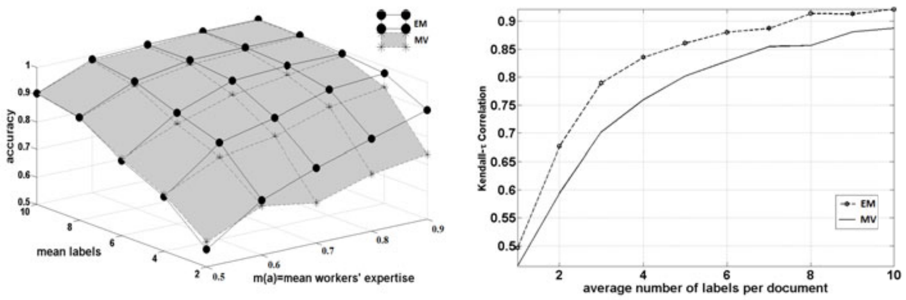
The *full* HIT design included several quality controls and qualified workers at different stages of the task. Since the HIT titles have an effect on the workers' recruitment, the full HITs were grouped into 21 topic-specific batches and included topic details in the title, description, and keywords. This was likely to attract workers interested in and knowledgeable about a particular topic. Each HIT included two test questions to detect sloppy behaviour: "Please tick here if you did NOT read the instructions" at the top of the HIT form and "I did not pay attention" as a relevance label option. Furthermore, to enforce the requirement that the workers needed to read a page before deciding about its relevance, a captcha was included asking them to enter the first word of the sentence that confirmed or refuted the relevance of the page.

On average, 6 labels from distinct workers were collected per document, 3 labels by simple HITs and 3 labels by full HITs. That amounts to 2179 labels for 727 topic-document pairs from simple HITs and 2060 labels for 683 topic-document pairs from full HITs. Also, 98% of topic-document pairs labelled in the full HIT were among those labelled in the simple HIT. The workers were paid $0.25 to complete a simple HIT and $0.50 for a full HIT.

For evaluation of the relevance labels obtained by the MV and EM methods we consider three commonly used measures [14]: (*i*) the *accuracy*—the proportion of judged documents that are assigned the correct relevance label, (*ii*) the *true positive ratio* (TPR)—the proportion of judged relevant documents that are correctly assigned the 'relevant' label, and (*iii*) the *true negative ratio* (TNR)—the proportion of judged non-relevant documents that are correctly assigned the 'non-relevant' label.

### 4.2 Simulation

We conduct simulations of multiple label aggregations to investigate the effects of (*i*) the number of labels collected for a document, and (*ii*) the level of the workers' expertise on the performance of the MV and EM methods. We consider a set of 1000 hypothetical documents with associated true relevance judgments. We also consider a set of 100 hypothetical workers, each with a particular level of expertise. We define workers' expertise as their accuracy of labeling a randomly chosen document. Similarly to [9], we randomly sample the workers' expertise from a Beta distribution and randomly assign documents to workers. We then apply the MV and EM methods to the collected labels in order to estimate the relevance of the documents. We use the measures defined in Section 4.1 to assess the performance of the two methods.



(a) (b)

**Fig. 1.** (a) Comparison of the judgment accuracy for MV and EM on synthetic data for varying number of labels per document and different levels of workers accuracy. (b) Kendall-τ correlation between the true and estimated workers expertise. Workers expertise is drawn from a beta distribution with the mean of 0.7.

We repeat the simulations by varying (*i*) the mean *m*(*a*) of the Beta distribution from which a worker's expertise is drawn, and (*ii*) the average number of labels collected per document. The results are shown in Figure 1(a) where the workers' mean expertise varies between 0.5 and 0.9 and the average number of labels varies between 1 and 10.

It can be seen that, when the workers average expertise is nearly random, *m*(*a*)=0.5, and the average number of labels per document is only 2, both methods exhibit poor accuracy. As the number of labels or the level of expertise increases, the performance of both methods improves. When the number of labels per document is small, e.g., 2 or 4 labels, but the workers average expertise is increased to 0.6 or higher, the EM method outperforms the MV. Finally, as the number of labels approaches 10 and the workers average expertise increases to 1, both methods obtain perfect accuracy. The simulations clearly show that the EM approach generally performs the same or better than MV.

We also assessed the performance of the MV and EM methods in estimating the workers accuracy. We use the gold standard set to determine the workers' true accuracy and compare with the estimated accuracies based on the relevance labels from the MV and EM methods. We compute Kendall-τ between the workers ranking induced by the EM or MV relevance judgments and the ranking based on gold standard set. Figure 1(b) shows that for the set of workers with *m*(*a*)=0.7, EM outperforms MV for a range of labels per document. We observed similar results for *m*(*a*)=0.6, 0.8, and 0.9.

### 4.3 Relevance Agreement with INEX Judgments

We apply the MV and EM methods to the labels collected from the two crowdsourcing experiments and compare the derived relevance judgments with the INEX gold standard set. We provide results for three sets of relevance judgments derived from: (*i*) the labels from the simple HITs, (*ii*) labels from the full HITs, and (*iii*) combined labels from both sets of HITs. For each of the sets we use samples of 1000, 1500, or 2000 labels to estimate the document relevance. The samples are randomly selected but guarantee that at least one label per document is included. We report average performance of the methods over 10 random trials.

**Table 1.** Comparison of MV and EM relevance judgments based on (*i*) accuracy, (*ii*) true positive ratio (TPR) and (*iii*) true negative ratio (TNR). INEX 2010 relevance judgements are used as the gold standard set. Statistically significant differences are marked by ‡.

| **HIT** | **~No. of Labels** | **accuracy** | | **TPR** | | **TNR** | |
|---|---|---|---|---|---|---|---|
| | | MV | EM | MV | EM | MV | EM |
| Simple | 1000 | 0.58 | 0.64‡ | 0.52 | 0.67‡ | 0.60 | 0.64‡ |
| | 1500 | 0.62 | 0.69‡ | 0.54 | 0.76‡ | 0.65 | 0.68‡ |
| | 2000 | 0.69 | 0.75‡ | 0.58 | 0.79‡ | 0.70 | 0.74‡ |
| Full | 1000 | 0.68 | 0.72‡ | 0.72 | 0.88‡ | 0.71 | 0.72 |
| | 1500 | 0.73 | 0.79‡ | 0.78 | 0.90‡ | 0.76 | 0.78‡ |
| | 2000 | 0.80 | 0.82 | 0.86 | 0.93‡ | 0.84 | 0.82 |
| Simple+Full | 1000 | 0.66 | 0.76‡ | 0.61 | 0.85‡ | 0.62 | 0.67‡ |
| | 1500 | 0.70 | 0.78‡ | 0.66 | 0.88‡ | 0.69 | 0.74‡ |
| | 2000 | 0.71 | 0.81‡ | 0.69 | 0.91‡ | 0.75 | 0.79‡ |

Note in case when there is an equal number of votes for relevant and non-relevant labels, the MV method will declare the associated document as non-relevant. We make the same assumption when the EM method estimates the relevance of a document as 0.5. This decision is based on the fact that the number of non-relevant documents for a query is typically higher than the number of relevant documents. The experimental results are shown in Table 1 for each of the three evaluation measures, the accuracy, TPR, and TNR. Statistically significant differences in the performance of the two methods are identified using a two sample z-test with a significance level of p=0.05.

As seen in Table 1, for the labels from the simple HIT task, the EM method significantly outperforms MV across all the samples and evaluation measures. The average improvement of EM over MV is 0.06 in accuracy, 0.19 in TPR, and 0.04 in TNR.

For the full HIT labels, the performance improvement of EM over MV is significant for most of the measures across the three samples. Only in three instances did we not get statistically significant differences. The average performance improvement of EM across the samples is smaller than for the simple HIT: 0.04 in accuracy, 0.12 in TPR, and 0.003 on TNR. This is expected since the labels from the full HITs are of higher quality due to more elaborate quality assurances tests. Indeed, there is 70% agreement between the full HIT labels and the INEX official judgments compared to 55% for the labels from the simple HITs.

This observation is consistent with the simulation results in Section 4.2: when the labels are provided by quality workers and the number of labels is large, both MV and

EM perform well. Indeed this is confirmed by the accuracy scores for the full HITs in Table 1. When the number of labels is 1000 or 1500, the accuracy of EM is significantly higher than that of MV. However, for a larger sample of 2000 labels there is no significant difference in the accuracy scores.

**Table 2.** Kendall-τ scores for MV and EM rankings of 10 runs from the INEX 2010 Book Search track by using the precision at 5 cut-off levels

| **HIT** | **P@10** | | **P@20** | | **P@30** | | **P@50** | | **P@100** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MV | EM | MV | EM | MV | EM | MV | EM | MV | EM |
| Simple | 0.45 | 0.62 | 0.55 | 0.71 | 0.77 | 0.89 | 0.91 | 0.98 | 0.90 | 0.99 |
| Full | 0.68 | 0.76 | 0.71 | 0.80 | 0.88 | 0.93 | 1.00 | 1.00 | 0.99 | 0.99 |

**Table 3.** Kendall-τ scores for MV and EM rankings of 10 runs from the INEX 2010 Book Search track. The average precision (AP) is calculated over all available judgments. stat-AP is calculated for the subsets of documents using corresponding relevance judgments.

| **HIT** | **statAP** | | | | | | **AP** | |
|---|---|---|---|---|---|---|---|---|
| | 10% | | 30% | | 50% | | | |
| | MV | EM | MV | EM | MV | EM | MV | EM |
| Simple | 0.58 | 0.67 | 0.64 | 0.77 | 0.91 | 0.91 | 0.84 | 0.91 |
| Full | 0.66 | 0.72 | 0.67 | 0.79 | 0.80 | 0.89 | 0.79 | 0.87 |
| INEX | 0.95 | | 0.96 | | 1.0 | | | |

Finally, we consider combined labels from the simple and full HITs. For each sample size 50% of labels are randomly selected from the simple HITs labels and 50% from the full HITs labels. For all three samples and performance measures, the EM method shows statistically significant improvements over MV. The average improvement across sample sizes is 0.09 in accuracy, 0.22 in TPR, and 0.05 in TNR. This is a larger improvement than for the simple HITs labels.

### 4.4 Impacts on Systems Ranking

We now observe the effect of MV and EM relevance judgments on the system ranking. For the crowdsourced labels collected from the simple and full HITs we apply MV and EM methods to create final sets of relevance judgments. These sets are used to rank the performance of 10 retrieval runs from the systems that participated in the INEX 2010 *prove it* task. We compare the runs based on the achieved precision at the top 10, 20, 30, 50 and 100 ranked documents. For consistency, when calculating the performance of runs based on INEX judgments, we consider only relevance judgments for the documents involved in the crowdsourcing experiments.

Table 2 summarizes the Kendall-τ correlations between the ranking of runs based on the INEX official judgments and the ranking obtained from MV or EM relevance judgments. We see for all cut-off levels, the rank correlation is higher for EM than for

MV. The average improvement for EM across the cut-off levels is 0.12 for simple HITs and 0.04 for the full HITs labels.

Generally, we see a considerable effect of the cut-off levels on Kendall-τ. This is expected since, when the cut-off level is small, e.g. p@10, even a few misjudged documents represents a high percentage of error and therefore significantly affects the ranking. As the cut-off level increases, for the same number of misjudged documents the percentage of error is relatively smaller and the ranking is not as affected.

We also explore the impacts of MV and EM on systems ranking when the average precision (AP) is used to evaluate the systems performance. The result is shown in Table 3. Once again EM outperforms MV for both the simple HITs and the full HITs labels. We investigate the effects of MV and EM on measuring AP with a shallow pool of documents, which is a common practice in IR experiments. We use statAP [8] to select subsets of 10%, 30% or 50% of documents labeled by the crowd workers. We apply MV and EM to the labels and then use the statAP metric to estimate the AP scores. For each sample size we calculate statAP based on the corresponding INEX judgments, MV judgments, and EM judgments and obtain systems rankings. In Table 3 we show the correlation between INEX systems ranking and the MV or EM systems rankings. The last row in Table 3 shows the Kendall-τ scores between the INEX systems ranking based on the statAP and the AP scores with the full set of the INEX judgments. Generally, we note that, as the sample size increases from 10% to 50%, the Kendall-τ scores increase correspondingly, similarly to the results in Table 2.

## 5 Summary and Concluding Remarks

In this paper, we consider the problem of creating relevance judgments using crowdsourcing experiments to collect multiple, possibly noisy relevance labels for documents. We assume that the workers' judgments are varied and of unknown accuracy. We also assume that the true relevance labels for documents are not available. We compare two methods for inferring document relevance from multiple labels. The MV method treats all the workers equally and assigns the relevance label that has received the most votes. The EM method simultaneously infers document relevance and workers' accuracy. We conduct a series of simulations with synthetic data and experiments with crowdsourced labels from the INEX 2010 Book Search track. Our experiments show that the relevance judgments inferred by the EM method are the better estimations of true document relevance and lead to more accurate systems ranking. EM performance improvements over MV are particularly noticeable when judgments are noisy and the number of relevance labels is small.

This research can be extended in several directions. In the evaluation of system performance we exploited the aggregation of noisy labels. However, the EM method provides estimation of the workers accuracy which can be used to grade workers and optimize the quality of additional labels by carefully selecting crowd worker. Furthermore, it can be used to compute workers' pay based on the quality of their work. Furthermore, our experiments were focused on the binary relevance judgments while the model supports multi-grade relevance. Thus, the future experiments will investigate the performance of the MV and EM methods for graded relevance and the sensitivity of the graded metrics, e.g., nDCG, to noise. Finally, the full potential of the EM method could be realized through an iterative model of selecting workers and collecting relevance labels. Thus, it is beneficial to extend the crowdsourcing experiments and evaluate the dynamic and real time collection of relevance judgments.

## References

[1] Kazai, G., Kamps, J., Koolen, M., Milic-Frayling, N.: 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, pp. 205–214 (2011)

[2] Alonso, O., Mizzaro, S.: Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In: SIGIR 2009: Workshop on the Future of IR Evaluation, Boston (2009)

[3] Smucker, M.D., Jethani, C.P.: Measuring assessor accuracy: a comparison of nist assessors and user study participants. In: Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, pp. 1231–1232 (2011)

[4] Cuadra, C.A., Katter, R.V.: Opening the Black Box of 'Relevance'. Journal of Documentation 23(4), 291–303 (1967)

[5] Voorhees, E.: Variations in relevance judgments and the measurement of retrieval effectiveness. Inf. Process. Manage. 36(5), 697–716 (2000)

[6] Buckley, C., Voorhees, E.: Evaluating evaluation measure stability. In: Proceedings of SIGIR, pp. 33–40 (2000)

[7] Carterette, B., Pavlu, V., Kanoulas, E., Aslam, J.A., Allen, J.: Evaluation Over Thousands of Queries. In: Proceedings of SIGIR, pp. 651–658 (2008)

[8] Carterette, B., Soboroff, I.: The effect of assessor error on IR system evaluation. In: Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 539–546 (2010)

[9] Aslam, J.A., Pavlu, V., Yilmaz, E.: A Statistical Method for System Evaluation Using Incomplete Judgments. In: Proceedings of SIGIR, pp. 541–548 (2006)

[10] Bailey, P., et al.: Relevance assessment: are judges exchangeable and does it matter. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 667–674 (2008)

[11] Scholer, F., Turpin, A., Sanderson, M.: Quantifying Test Collection Quality Based on the Consistency of Relevance Judgements. In: Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1063–1072 (2011)

[12] Winter, M., Duncan, W.: Financial incentives and the ”performance of crowd”. In: Proceedings of the ACM SIGKDD Workshop on Human Computation, pp. 77–85 (2009)

[13] Snow, R., O'Connor, B., Urafsky, D., Ng, A.Y.: 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, Stroudsburg, PA, USA, pp. 254–263 (2008)

[14] Smucker, M.D., Jethani, C.P.: The Crowd vs. the Lab: A Comparison of Crowd-Sourced and University Laboratory Participant Behavior. In: Proceedings of the SIGIR 2011 Workshop on Crowdsourcing for Information Retrieval, Beijing (2011)

[15] Kumar, A., Lease, M.: Modeling Annotator Accuracies for Supervised Learning. In: WSDM 2011 Workshop on Crowdsourcing for Search and Data Mining, Hong Kong (2011)

[16] Kasneci, G., Gael, J.V., Stern, D.H., Graepel, T.: CoBayes: bayesian knowledge corroboration with assessors of unknown areas of expertise. In: Proceedings of the Forth International Conference on Web Search and Web Data Mining, pp. 465–474 (2011)

[17] Welinder, P., Perona, P.: Online crowdsourcing: rating annotators and obtaining cost-effective labels. In: CVPR 2010: IEEE Conference on Computer Vision and Pattern, pp. 1526–1534 (2010)

[18] Ipeirotis, P.G., Provost, F., Wang, J.: Quality Management on Amazon Mechanical Turk. In: Proceedings of the ACM SIGKDD Workshop on Human Computation, pp. 64–67 (2010)

[19] Dawid, P., Skene, A.M.: Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. Applied Statistics 28(1), 20–28 (1979)

[20] Bernstein, Y., Zobel, J.: Redundant documents and search effectiveness. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management, pp. 736–743 (2005)