Active Content-Based Crowdsourcing Task Selection

Piyush Bansal
International Institute of
Information Technology
Hyderabad, India
piyush.bansal@research.iiit.ac.in

Carsten Eickhoff
ETH Zurich, Switzerland
Dept. of Computer Science
ecarsten@inf.ethz.ch

Thomas Hofmann
ETH Zurich, Switzerland
Dept. of Computer Science
thomas.hofmann@inf.ethz.ch

ABSTRACT

Crowdsourcing has long established itself as a viable alternative to corpus annotation by domain experts for tasks such as document relevance assessment. The crowdsourcing process traditionally relies on high degrees of label redundancy in order to mitigate the detrimental effects of individually noisy worker submissions. Such redundancy comes at the cost of increased label volume, and, subsequently, monetary requirements. In practice, especially as the size of datasets increases, this is undesirable. In this paper, we focus on an alternate method that exploits document information instead, to infer relevance labels for unjudged documents. We present an active learning scheme for document selection that aims at maximising the overall relevance label prediction accuracy, for a given budget of available relevance judgements by exploiting system-wide estimates of label variance and mutual information.

Our experiments are based on TREC 2011 Crowdsourcing Track data and show that our method is able to achieve state-of-the-art performance while requiring 17% - 25% less budget.

Keywords

Crowdsourcing; Active learning; Relevance Assessment

1. INTRODUCTION

Owing to the wide-spread adoption of the Internet, crowdsourcing has been able to leverage the potential of a globally distributed and diverse workforce to efficiently create and enrich academic datasets. This has led to a noticeable decrease in the overall time, and monetary cost involved in corpus creation. While, traditionally, a group of domain experts were employed for a task, crowdsourcing involves multiple workers who may not hold significant domain expertise. The recurring challenge of quality control (QC), which arises due to an untrained workforce, has usually been addressed by assigning the same task to multiple crowd workers, and then aggregating these multiple individual relevance assessments to arrive at a more reliable final annotation label. While this leads to a better accuracy, it can also inflate the overall cost and time requirements. In more realistic applied settings, such as search engine evaluation, demanding redundant budget and time resources in terms of person hours is impractical. Even for large corporations, hiring human assessors to judge relevance of billions of documents is infeasible.

In such settings, employing techniques that can be used to infer the relevance of a significant proportion of all the documents, from a small subset of human relevance judgements can prove to be very useful. In the past, document similarity has been shown to be an effective signal in inferring relevance of unjudged documents from a small subset of manually judged documents [7]. However, to the best of our knowledge, no attempt has been made to optimally select this seed set of documents. In this paper, we rely on inter-document similarity for cost-optimal document selection such that we reduce the overall budget requirement to be able to infer relevance labels of unjudged documents as accurately as the state-of-the-art methods at significantly reduced cost.

The novel contributions of our work are threefold – 1) We present a systematic overview of the spread of crowdsourcing relevance labels across a distributed textual similarity space, demonstrating its usefulness to actively selecting documents for search result evaluation. 2) We discuss and present two information theoretic criteria to optimally select documents for relevance judgement, thereby minimising the overall budget necessary to achieve the same accuracy as a set of competitive baseline methods. 3) In a series of experiments based on historic submissions to the TREC 2011 Crowdsourcing Track, we demonstrate the merit of our methods in terms of maximising accuracy for a fixed crowdsourcing budget.

The remainder of the paper is structured as follows: Section 2 presents a detailed overview of the related work in crowdsourcing relevance assessment, quality control and active learning techniques. Section 3 starts with a formal description of the problem statement and then moves on to presenting the necessary preliminaries as well as related approaches that will later serve as performance baselines. Our main contribution resides in Section 4, where we describe the intuition and theoretical foundations of our techniques. Our experiments and results are described in Section 5. Finally, we conclude and discuss future directions to our work in Section 6.

CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA
© 2016 ACM. ISBN 978-1-4503-4073-1/16/10. . . $15.00
DOI: http://dx.doi.org/10.1145/2983323.2983716

2. RELATED WORK

Evaluation has been an essential part of the development and maintenance of search engines and other Information Retrieval (IR) systems. Commonly, the evaluation strategies fall into two main categories – 1) *Test collections* based evaluation that relies on creation of document corpora and 2) *User studies* based evaluation which relies on interactive IR system usage. A significant amount of test collections based evaluation strategies and methods have been inspired by the Cranfield experiments [6] and involve creation of datasets that contain document corpora and search topics along with corresponding relevance judgements. In the Information Retrieval community, one of the most widely known efforts in creation of such test collections can be attributed to the Text REtrieval Conference (TREC) [33]. For the purpose of relevance assessment required for creation of such test collections, TREC has been know to employ trained editorial judges who have prior experience as intelligence analysts. However, this approach has scalability issues and can even be prohibitively expensive for large-scale corpus creation and annotation. Acknowledging these issues, previous work has focused on proposing more robust measures and statistical techniques for retrieval evaluation in cases of incomplete test collections or judgements [2, 3] and selecting the right subset of documents for evaluation [5].

Jaochims *et al.* [15] have focused on user studies based evaluation by utilising the implicit feedback from clickthrough data, thus attempting to mitigate the efforts required for test collection creation.

In recent years, the rise of crowdsourcing [14] has opened new avenues for efficiently utilising and leveraging the potential of a digitally-connected and globally distributed human workforce. Crowdsourcing has also proved useful for creation of large-scale test collections, with recent experiments [1, 12, 19, 20] suggesting that aggregated labels of multiple untrained crowd workers can reach a quality comparable to that of a single highly-trained NIST assessor.

While this establishes crowdsourcing as a widely accepted mechanism for corpus creation and annotation tasks, the challenge of Quality Control (QC) has received much attention in the recent past. [23] identifies the most important factors which are detrimental for quality of the data collection, which include – (1) *Human Factors* : where the focus is on improving the interface, and overall design of the Human Intelligence Task (HIT) borrowing ideas from the field of Human-Computer Interaction (HCI) as in [16, 26], (2) *Annotation Guidelines* : where the focus is on developing concrete and unambiguous annotation guidelines to make sure that the collected data is consistent as in [30], and (3) *Worker Reliability* : which has received maximum attention. A fair amount of past work has focused on cheat-detection by deploying honey-pot questions [10, 13] or employing gamification to provide entertainment based incentives to improve data collection quality [11].

Worker behavior and demographics have also been exploited for worker reliability estimation. Kazai *et al.* [17, 18] study worker reliability by grouping workers into five different classes. There have also been studies to determine worker-topic affinity by analysing the social media profiles of crowd workers [9]. Employing these ideas in selecting the most reliable workers for a particular task has been shown to work well by reducing the number of relevance judgements required to obtain a highly confident single annotation label. However, in more practical crowdsourcing scenarios that involve online platforms such as Amazon Mechanical Turk and CrowdFlower, it can be hard to have control over the rapidly changing crowd workforce [8].

Active learning techniques have shown the promise to ensure data quality, while minimising the overall costs involved. Yan *et al.* [36] try to jointly select the data point and the worker based on a worker reliability model. Their approach of optimally selecting data points is iterative in nature. At each iteration, given a set of data points that have been selected and corresponding relevance judgements from crowd workers have been obtained, their decision to select subsequent data point relies on the relevance judgements obtained so far. This poses an additional operational constraint on the overall crowdsourcing process, where we have to wait for the relevance judgements for the previously selected data points to arrive in order to allocate new tasks to workers. However, in real-world scenarios, this can be impractical for a wide variety of crowdsourcing tasks. In this paper, our data-point selection decisions do not rely on the relevance judgements that we collect from crowd workers, thus getting rid of the above discussed operational constraint. Also, due to the iterative nature of the task selection process in Yan *et al.* [36], the batch sizes of HITs tend to be smaller. Wang *et al.* [34] have shown that smaller batch sizes of HITs tend to have longer per-HIT completion times than large ones. In this work, both our information-theoretic task-selection criteria allow for optimal task selection before the tasks are submitted as HITs, thereby enabling us to submit all our selected tasks to crowd workers in a single large batch, achieving greater speed and efficiency.

Our work is closely related to the recent work of Davtyan *et al.* [7], in which the authors exploit document content for efficient label aggregation of crowdsourcing votes. This is done by propagating label information using document similarities. Their argument is that ''Similar documents tend to be similarly relevant towards a given query Q". In our work, we exploit the document content not only for label aggregation, but also for active document selection. In doing so, we show that we need 17% - 25% fewer worker judgements in order to achieve the same accuracy as the best-performing methods in [7] for budget-constrained scenarios.

3. METHODOLOGY

In this section, we first describe our problem statement in detail, and then move on to discussing various components of our experimental setup.

3.1 Problem description

In this subsection, we formally describe the problem statement, and briefly present the preliminaries required in the course of the paper. Formulating the traditional crowdsourcing process as a modular three-step algorithm, we highlight those modules that our work focuses on.

For a given topic t, consider a set of documents $\mathcal{D}$, which need to be assessed for their relevance to this topic by some workers belonging to the set of all workers $\mathcal{W}$. For document $d_i \in \mathcal{D}$, we seek a binary label $r_{ij} \in \{0, 1\}$ from a crowd worker that denotes the relevance of this document w.r.t. topic t. 0 denotes "non-relevance" and 1 denotes "relevance" of a document d_i w.r.t. topic t. All the votes (relevance judgements) denoted by r_{ij} are collected in a set $\mathcal{R}$ and are then passed on to a "label-aggregation" step, which in most crowdsourcing applications is given by some form of (weighted) majority voting. In such a setting, crowdsourcing can be viewed as shown in Algorithm 1.

Algorithm 1 Generic Crowdsourcing Process

```
procedure RELEVANCEASSESSMENT($\mathcal{D}, t$)
    for $b \leftarrow 1...B$ do
        $d_i \leftarrow PickDocument(\mathcal{D}, \mathcal{R})$;
        $r_{ij} \leftarrow RequestVote(d_i, \mathcal{W})$;
        $\mathcal{R} \leftarrow \mathcal{R} \cup r_{ij}$;
    $\mathcal{L} \leftarrow AggregateVotes(\mathcal{D}, \mathcal{R})$;
    return $\mathcal{L}$;
```

Where $\mathcal{L}$ is the set containing final labels for all the documents $d_i \in \mathcal{D}$, and B represents a budget parameter indicating the number of available votes. Typically, the method **AggregateVotes** does not take document information into account, *e.g.*, majority voting, but Davtyan *et al.* [7], exploit textual similarity to demonstrate that in budget-constrained situations, using document information can enhance the overall accuracy of the process.

Our experiments rely on the set of relevance judgements collected for the TREC 2011 CrowdSourcing Track [24]. The ground truth labels are annotated by NIST assessors and are made available as a part of this dataset. For all our experiments, relevance judgements are sampled randomly from the pool of available judgements to prevent any selection bias.

As previously mentioned, our analysis focuses on budget-constrained scenarios in which the number of documents to be judged can be greater than the total number of relevance judgements one can assign to crowd workers (For a given topic, the average number of votes per document ≤ 1). For a given budget B (*e.g.*, a total of 50 votes to judge relevance of 100 documents for a given topic t, or say, an average of 0.5 votes per document), our goal is to select a set of documents such that we are able to achieve maximum accuracy in terms of the proportion of final labels matching the ground truth labels.

3.2 Baseline

Davtyan *et al.* present three realisations of **AggregateVotes** - *MergeEnoughVotes (MEV)* which borrows relevance judgements from the single nearest neighbour of a document (the most similar document to the given document), until they have accumulated a fixed number of votes V (V was empirically set to 1 in their experiments) to assign a final relevance label to the document. *MajorityVoteWithNearestNeighbor (MVNN)* which accumulates votes from all the neighbors if the similarity between the neighbors and the given document is above a certain threshold, and *GaussianProcessAggregation (GP)*, which propagates relevance information across the document space in a distance-based discounting scheme. In a first series of experiments we empirically confirmed the authors' findings of GP consistently and significantly outperforming the remaining methods. Due to its superior performance, and simplicity of analysis, in this paper, we rely on GP as our **AggregateVotes** strategy.

To specify a GP, mean and covariance functions are necessary. For our experiments, we assume a constant mean function $M(\cdot)$, and a linear covariance function $K(\cdot, \cdot)$ –

$$M(d) = c \tag{1}$$

and

$$K(d, d') = d.d' \tag{2}$$

where the linear covariance function is the scalar dot product between two document representations. For each document d, representing a random variable $d \in \mathcal{D}$, the mean μ_d is given by Equation 1, where $\mathcal{D}$ refers to the set of all documents. For simplicity of notation, we also denote the mean vector of some set of variables $\mathcal{A}$ by $\mu_{\mathcal{A}}$, where the entry for element d of $\mu_{\mathcal{A}}$ is $M(d)$, given by Equation 1. We explain these document vectors in detail in Section 3.3.

Using GP, we formulate the problem of label aggregation as a two-class classification task. We train the classifier using the available relevance judgements, and then use the resulting model to predict the labels for documents which have not, so far, been judged by any worker.

Our focus here lies on cost-optimal data acquisition. Hence, we focus on different realisations of the **PickDocument** method from Algorithm 1. All other components remain unaltered throughout the course of our experiments. In typical crowdsourcing scenarios, **PickDocument** merely involves selecting documents that have the lowest number of relevance judgements so far. This can be formally written as:

$$d_i = PickDocument(\mathcal{D}, \mathcal{R}) = \arg\min_{\mathcal{D}}(r_i) \tag{3}$$

where r_i denotes the number of relevance judgements for document d_i. If there is no unique d_i satisfying the above condition, ties are broken by random selection of a document from the set of optimal candidates.

3.3 Document representations

In the work of Davtyan *et al.* [7], the authors represent documents as tf-idf vectors. It has been long established that such term-based one-hot representations suffer from sparsity, and fail to capture latent semantics of the underlying text. In [22], Le *et al.* describe *Paragraph Vectors*, an unsupervised learning algorithm that establishes vector representations for variable-length pieces of text such as sentences and documents. We follow their work, and represent documents as fixed-length dense document vectors having 100 dimensions. Our corpus consists of Web documents from the ClueWeb09-T11Crowd collection, which is a subset of the full ClueWeb09 dataset [4]. Pennington *et al.* [28] showed distributed text representations to capture more semantic information when the models are trained on Wikipedia text, as opposed to other large corpora such as the Common Crawl. This is attributed to the structured, and comprehensive nature of articles on Wikipedia. Hence, we trained the model on the latest Wikipedia dump, and used the trained model to infer document vectors for the documents in our corpus. The chosen settings of a single training epoch and a hidden layer size of 100 have been empirically found to yield satisfying results. While it is conceivable that further fine-tuning of hyperparamaters should result in additional performance benefits, broad parameter sweeps can be costly and go beyond the scope of this work.

We demonstrate the usefulness of our embeddings in preserving the notion of relevancy by plotting inner, and outer similarities across the document pairs as originally proposed by Van Rijsbergen *et al.* [32] to forecast retrieval effectiveness by evaluating the separability of relevant and non-relevant documents. We use cosine similarity as formulated in Equation 4 as the similarity metric, which is also used for the rest

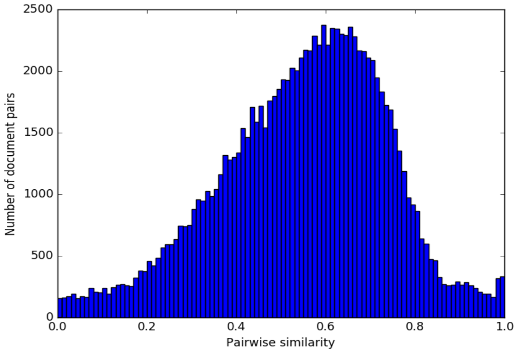


Figure 1: (Inner) Similarity between relevant documents averaged across topics.

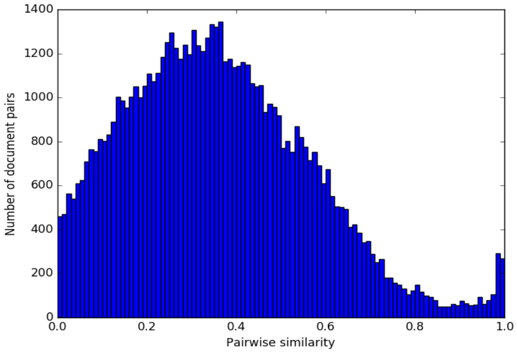


Figure 2: (Outer) Similarity between relevant and non-relevant documents averaged across topics.

of the experiments throughout our work:

$$Similarity(A,B) = \frac{d_a.d_b}{||d_a||||d_b||} \quad (4)$$

where d_a and d_b refer to the document vectors corresponding to documents A and B.

Figures 1 and 2 demonstrate the results of this experiment. The majority of the distribution mass lying to the left in case of "Outer" similarity and to the right in case of "Inner" similarity validates the choice of our document embeddings.

We also conduct a more direct empirical evaluation of our choice of document representation by comparing the performance of various label aggregation methods - *MergeEnoughVotes (MEV)*, *MajorityVoteWithNearestNeighbor (MVNN)* and *GaussianProcessAggregation (GP)* as proposed by Davtyan *et al.* [7] across the two types of document representations. All other parameters were kept static while comparing the methods. Figure 3 shows a detailed comparison for each of these methods. We observe that using the document embeddings achieves a consistent positive gain in performance across all label aggregation techniques as compared with tf-idf document vectors. An interesting observation to make here is that *MVNN* is a highly "localised" label aggregation method, as it exclusively draws information from a very proximal neighbourhood. *MEV* is relatively less "localised" as it can borrow relevance judgements even from far-away neighbours (as long as there are no closer neighbouring documents which have been sampled) and *GP* is the least "localised" technique as it considers all the available relevance judgements no matter how far they are away. We see that the usefulness of our document embeddings also correlates with this trend, resulting in greater performance gains for less "localized" methods. This further supports our choice of *GP* as a promising realisation of **AggregateVotes**.

4. DOCUMENT SELECTION METHODS

In this section, we present a range of realisations of the **PickDocument** method as described in Algorithm 1. Since we deal with budget-constrained scenarios, we may not always be able to afford relevance judgements for all the documents in our corpus (recall that for a given topic, the number of votes per document can be ≤ 1). In such a setting, our goal is to select a subset of documents that maximally speeds up learning. As such, we can model our task as a standard active learning problem.

4.1 Active learning for document sampling

In general active learning scenarios, unlabeled data is available, and at each iteration the algorithm must select an example (a document), and request a relevance assessment for it. The objective is to maximise overall classification accuracy, at a fixed cost or budget. Active learning strategies can be classified into two main types depending on the way in which these examples are made available. When examples can be chosen from an unlabeled dataset, this is referred to as *pool-based* active learning. In contrast, when a decision to label an example has to be made sequentially as each example becomes available, this is referred to as *online* active learning. In this paper, we focus on pool-based active learning.

Active learning strategies can further be classified into various categories depending on the underlying criteria that they optimise for. Our method falls into the class of *Uncertainty Sampling*, where the goal is to select an unlabelled data point that has maximum uncertainty given the current classification model. In the following subsections, we present two different realisations of *Uncertainty Sampling* – (1) Variance-based and (2) Mutual-Information-based.

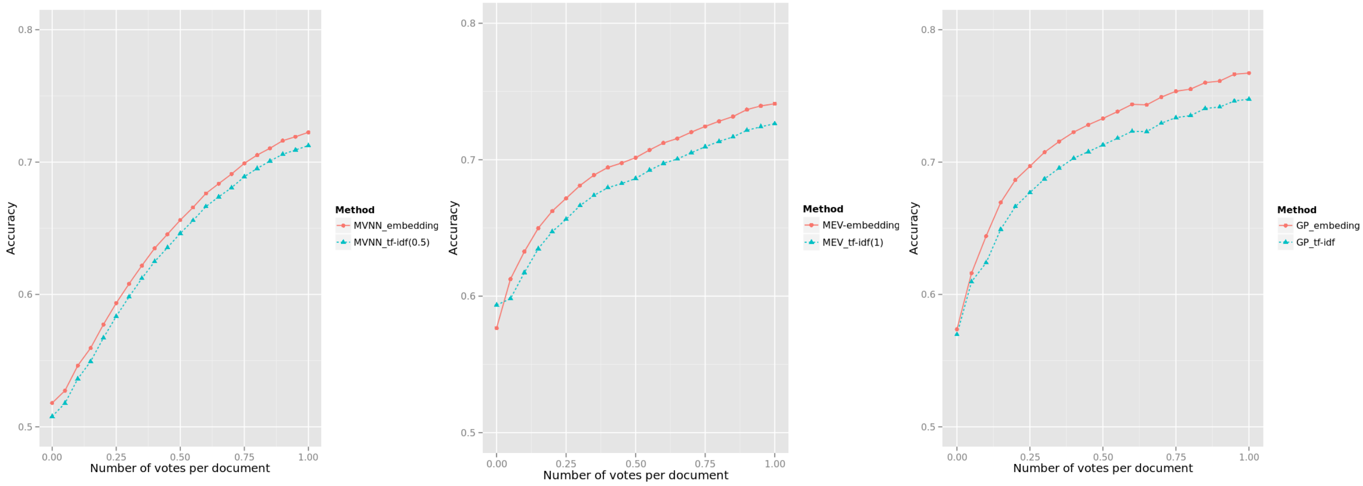
4.2 Variance-based sampling

Consider a set of documents $\mathcal{D}$, from which a subset $\mathcal{A}$ of documents has been chosen. Given a set of observed relevance judgements $x_{\mathcal{A}}$ corresponding to the finite subset $\mathcal{A} \subset \mathcal{D}$, we can predict the relevance of every other document $y \in \mathcal{D}$ conditioned on the set of observed relevance judgements, $P(X_y|x_{\mathcal{A}})$. More formally, the distribution of X_y given the observed relevance judgements is a Gaussian whose conditional mean $\mu_{y|\mathcal{A}}$, and conditional variance ${\sigma^2}_{y|\mathcal{A}}$ are given by the equations:

$$\mu_{y|\mathcal{A}} = \mu_y + \Sigma_{y\mathcal{A}}\Sigma^{-1}_{\mathcal{A}\mathcal{A}}(x_{\mathcal{A}} - \mu_{\mathcal{A}}) \quad (5)$$

$${\sigma^2}_{y|\mathcal{A}} = K(y,y) - \Sigma_{y\mathcal{A}}\Sigma^{-1}_{\mathcal{A}\mathcal{A}}\Sigma_{\mathcal{A}y} \quad (6)$$

where $\Sigma_{y\mathcal{A}}$ is a covariance vector with one entry for each document $d \in \mathcal{A}$ having value $K(y,d)$ according to Equation 2, and μ_y along with $\mu_{\mathcal{A}}$ are as described in Subsection 3.2. $\Sigma_{\mathcal{A}\mathcal{A}}$ refers to the covariance matrix, which follows



(a) MVNN with similarity threshold = 0.5.

(b) MEV with minimum 1 vote.

(c) Gaussian processes (GP).

Figure 3: Comparison of various label aggregation methods in tf-idf vectors and dense distributed document embeddings.

the notation where $\Sigma_{\mathcal{GH}}$ refers to the covariance matrix with each element having indices (i, j) given by $K(g, h)$ and g, h are i^{th}, j^{th} elements of sets $\mathcal{G}$ and $\mathcal{H}$ respectively.

Observing that the differential (continuous) entropy of a Gaussian random variable Y conditioned on a set of variables $\mathcal{A}$ is given by –

$$H(Y|\mathcal{A}) = \frac{1}{2} log(2\pi e \sigma^2_{Y|\mathcal{A}}) \tag{7}$$

and applying the chain rule to Equation 7, it can be written as:

$$H(\mathcal{A}) = H(Y_k|Y_1, ..., Y_{k-1}) + ... + H(Y_2|Y_1) + H(Y_1) \tag{8}$$

where $\mathcal{A}$ refers to the set $\{Y_1, Y_2, ..., Y_k\}$.

Hence, in light of Equations 7 and 8, the variance-based sampling technique is equivalent to an approximate solution to the following problem –

$$\arg \max_{\mathcal{A}:|\mathcal{A}|=B} H(\mathcal{A}), \tag{9}$$

that is, selecting the documents that have maximum joint entropy. Krause *et al.* [21] show that directly optimising the above entropy based criterion is NP-complete. Algorithm 2 represents a greedy approach which results in an approximate solution to this criterion. We start with an empty set of documents $\mathcal{A} = \emptyset$, and at each step we greedily add documents to it until $|\mathcal{A}| = B$, where B is our budget. The greedy rule selects a document from $\mathcal{D} \setminus \mathcal{A}$, which has maximum variance according to Equation 6. This heuristic intuitively makes sense, as we try to pick documents that we are most *uncertain* about, given those documents that we have already selected. At this point, it is also worth noting that posterior variance in case of Gaussian Processes, as given by Equation 6, does not depend on the previously collected relevance judgements $x_{\mathcal{A}}$. This allows us to optimally select subsequent documents for relevance assessment without having to wait for the previous relevance judgements. This property also holds for our next document selection method, described in Subsection 4.3.

We describe the experiments and results for this technique later in Section 5.

Algorithm 2 Variance-based sampling

```
procedure PICKDOCUMENT($\mathcal{D}, \mathcal{R}$)
    $\mathcal{A} \leftarrow \emptyset$
    for $b \leftarrow 1...B$ do
        if $|\mathcal{A}| == 0$ then
            $\mathcal{A} \leftarrow pickRandomDocument(\mathcal{D})$;
        else
            $variance \leftarrow getConditionalVariance(\mathcal{D} \setminus \mathcal{A})$;
            ▷ variance for candidate documents is given by Equation 6.
            $\mathcal{A} \leftarrow \mathcal{A} \cup \arg\max_{\mathcal{D} \setminus \mathcal{A}}(variance)$;
    return $\mathcal{A}$
```

4.3 Mutual-Information-based sampling

In case of the variance-based sampling technique, we were concerned about the entropy at the documents that have been sampled, as opposed to directly maximising the prediction quality over the full space of interest (the set of all documents). Hence, in this second sampling technique, we try to select a subset of documents that minimises the uncertainty over the rest of the space. More formally, our selection criterion is

$$\arg \max_{\mathcal{A}:|\mathcal{A}|=B} H(\mathcal{D} \setminus \mathcal{A}) - H(\mathcal{D} \setminus \mathcal{A}|\mathcal{A}), \tag{10}$$

as opposed to Equation 9. Intuitively, this is equivalent to selecting a set of documents $\mathcal{A}$ that maximally minimise the entropy over the rest of the space $\mathcal{D} \setminus \mathcal{A}$. Equation 10 is equivalent to maximising *mutual information* between $\mathcal{A}$ and $\mathcal{D} \setminus \mathcal{A}$, which we denote as $I(\mathcal{A}, \mathcal{D} \setminus \mathcal{A})$ or $F(\mathcal{A})$.

Again, Krause *et al.* [21] note that optimising this mutual-information-based criterion is NP-complete. However, they show that the function $\mathcal{A} \mapsto I(\mathcal{A}, \mathcal{D} \setminus \mathcal{A})$ is *submodular*. Intuitively, this refers to the notion of diminishing returns observed when adding a document Y to a small set of documents $\mathcal{A}$ gives us more new information than adding a document Y to a larger existing set of documents $\mathcal{A}'$. The authors also propose a greedy approximate solution to this general problem, which due to Nemhauser *et al.* [27] has a performance guarantee of $(1-1/e)OPT$, where OPT is the value of optimal subset of size B.

The greedy algorithm selects the document Y that provides maximum increase in mutual information at every step. Formally, we pick a document Y such that $F(\mathcal{A} \cup Y) - F(\mathcal{A})$ is maximal, where $F(\mathcal{A}) = I(\mathcal{A}, \mathcal{D} \setminus \mathcal{A})$.

$$\begin{aligned}
&F(\mathcal{A} \cup Y) - F(\mathcal{A}) = \\
&H(\mathcal{A} \cup Y) - H(\mathcal{A} \cup Y|\bar{\mathcal{A}}) - [H(\mathcal{A}) - H(\mathcal{A}|\bar{\mathcal{A}} \cup Y)] \\
&\qquad = H(Y|\mathcal{A}) - H(Y|\bar{\mathcal{A}})
\end{aligned} \tag{11}$$

where $\bar{\mathcal{A}}$ refers to $\mathcal{D} \setminus (\mathcal{A} \cup Y)$. Using Equation 6 for our definition of entropy, this can also be written in the following algorithmic form (Algorithm 3), using the same notational conventions as in Section 4.2:

Algorithm 3 Mutual-Information-based sampling

```
procedure pickDocument($\mathcal{D}, \mathcal{R}$)
    $\mathcal{A} \leftarrow \emptyset$
    for $b \leftarrow 1...B$ do
        $Y^* \leftarrow \arg\max_{Y \in \mathcal{D} \setminus \mathcal{A}} \frac{\sigma^2{}_Y - \Sigma_{Y\mathcal{A}}\Sigma^{-1}_{\mathcal{A}\mathcal{A}}\Sigma_{\mathcal{A}Y}}{\sigma^2{}_Y - \Sigma_{Y\bar{\mathcal{A}}}\Sigma^{-1}_{\bar{\mathcal{A}}\bar{\mathcal{A}}}\Sigma_{\bar{\mathcal{A}}Y}}$;
        $\mathcal{A} \leftarrow \mathcal{A} \cup Y^*$;
    return $\mathcal{A}$
```

5. EXPERIMENTS

In this section, we describe our experimental setup in detail. We compare the performance of the two active selection strategies presented in the previous section. We also contrast these results against those of our non-active baseline, and establish that the mutual-information-based strategy outperforms both the baseline, and the variance-based sampling strategy. Later in this section, we also draw more intuitive explanations of the results.

5.1 Data

The source of our data is the TREC 2011 Crowdsourcing Track [24], which aimed at testing the effectiveness of crowdsourcing as a means of search engine evaluation. The collection of documents that was used for this purpose is a subset of the full ClueWeb09 dataset [4], known as ClueWeb09-T11Crowd. Each document in the collection is a uniquely identified Web page represented by the page URL, the content of the page – which is comparable to what a person would see if they visited that URL on a browser, along with the page's HTML source. For simplicity, we, however, only utilise the full text content of the page, ignoring its structure and layout.

We use the original TREC 2011 relevance judgments. The collection consists of labels for a total of 30 unique topics, such as "free email directory" or "growing tomatoes". For every topic, there is a set of approximately 100 documents that require relevance labels. For every document, there are on average 15 relevance judgments from individual workers, although some topic-document pairs have fewer affiliated judgments. For two topics (20644 and 20922) there were documents with as few as one single vote. Since such singleton "pools" of votes are very brittle and make for a poor representation of human knowledge, we exclude these two outlier topics from our investigation, leaving us with 28 functional ones. This treatment is similar to the one advocated in [7].

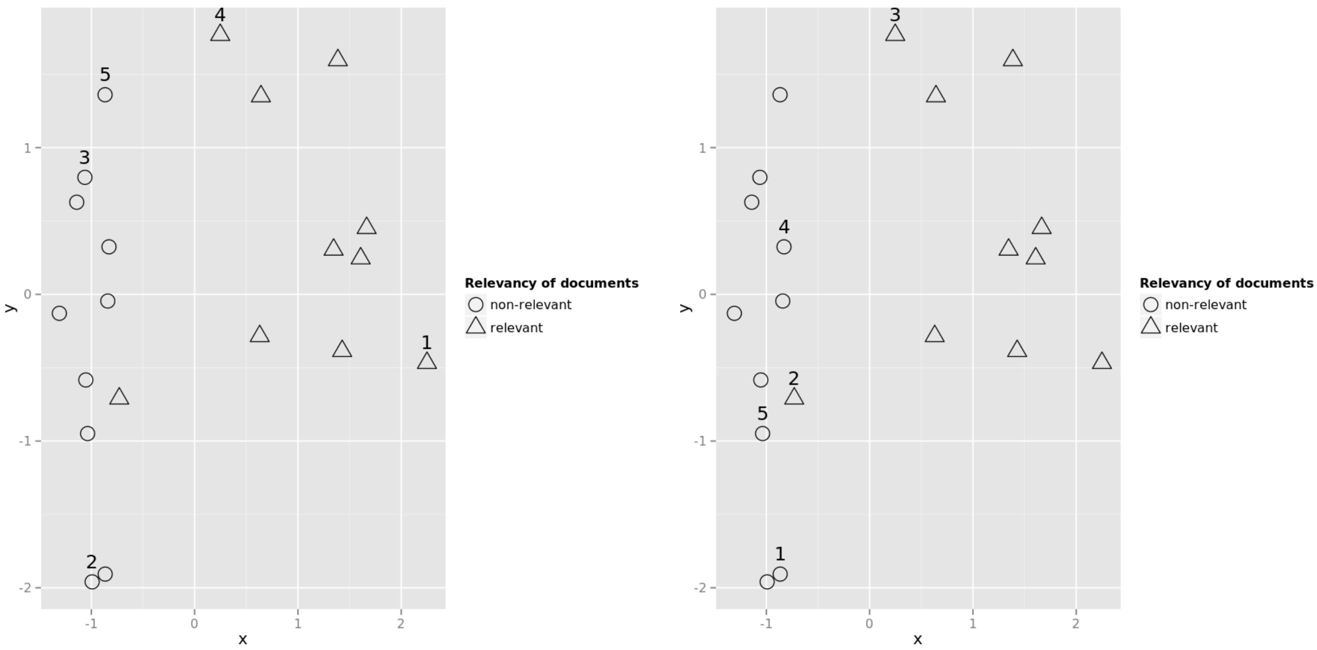
The 2011 Crowdsourcing track involved two different evaluation strategies based on the benchmark annotation that was used. One was based on the expert judgements from NIST assessors, and the other on the consensus labels gathered across all participating teams. In this work, we compare against labels from the more reliable NIST assessors as ground truth. It consists of 395 relevance judgments and most topics contain ten to twenty documents with such ground truth labels.

After the track participants submitted the aggregated judgments they were evaluated using the benchmark sets. For every submission, precision, recall, accuracy and specificity were measured. For the available ground truth labels, both relevance classes have similar orders of magnitude - 68% of 395 labels are "relevant". While the "relevant" class is dominant across all topics, there is only one case in which it represents as much as 80% of labels. We follow [7] in concentrating on accuracy as a performance measure, since it is expressive of the classifier's performance in case of balanced class sizes. We use the same performance measure across various document selection strategies in order to compare all candidate and baseline methods.

5.2 Results and Discussion

Our experimental setup is based on the generic crowdsourcing algorithm presented in Section 3.1. We compare our active document selection methods against the baseline **PickDocument** realisation that is described by Equation 3 and according to which a document is selected randomly among the currently least-frequently judged ones. We obtain votes for a document by randomly sampling relevance judgements from the pool of available judgements. Figure 5 displays accuracy as a function of budget (normalised by the number of documents in a topic) averaged across all the topics. To account for chance variation, each plotted performance represents the mean accuracy across 60 individual runs of the crowdsourcing process. We keep the **AggregateVotes** method of Algorithm 1 static across all crowdsourcing runs while evaluating the active document selection methods, in order to ensure comparability of our findings.

There are several interesting trends to note. Firstly, the active document selection relying on variance, referred to as *ActiveGPVariance* in Figure 5 and described as Algorithm 2 suffers in the experiment onset, when the average number of votes per document ≤ 0.25. Ramakrishna *et al.* [29] and Mackay [25] note that variance-based approaches tend to repeatedly gather data that lies at the "edge" of the input space. This is attributed to the fact that estimated variances are typically high towards the boundaries of an interpolation region. This is also evident in Figure 4, which demonstrates the sampling behavior of the variance based strategy, by contrasting it against that of the mutual-information based strategy. For this experiment, we randomly selected a small subset of 20 documents (out of which 10 are relevant to a particular topic, and the remaining 10 are not), and represented those documents in 2 dimensions by applying multidimensional scaling to the document vectors. Having constructed this *toy-dataset*, we run both our active sampling techniques in order to visualise the differences in their sampling behaviors. In Figures 4a and 4b, the first five documents sampled by both the active sampling tech-



(a) Sampling order of a variance-based strategy.

(b) Sampling order of a mutual-information-based strategy.

Figure 4: Comparison of sampling behaviours of the active sampling techniques.

niques have been annotated with numbers that represent the sequence in which these documents were sampled. Since the variance-based sampling technique starts out by picking points that have high variance, it ends up selecting points that lie towards the "edges" of input space, as is visible in Figure 4a. The information gained by sampling these "outlier" documents cannot be well extrapolated to the majority of other documents which are closer towards the center of the input space, thereby explaining the poor performance of variance-based sampling in the early stage. This, however, is mitigated after we have sampled a few documents (average number of votes per document > 0.25), and the new documents that we select do not lie exclusively on the outer envelope of the input space anymore. On the other hand, the random sampling technique has no such bias towards selection of "outlier" points.

It can also be seen that all the techniques have converging performance when approaching 1-vote-per-document. This is due to the fact that in each iteration, we only select a document that has not been selected before - so when we approach 1-vote-per-document, we have sampled a vote for nearly every document, and selecting new documents does not contribute as much to the accuracy owing to diminishing returns. At 1-vote-per-document, we have exactly the same document selection situation for active as well as random document selection approaches. This phenomenon can also be observed when we are close to 0-votes-per-document.

Table 1 presents savings as percentages of the available budget when reaching a level of accuracy identical to that of the baseline method. It can be seen that our best-performing method using mutual information for active document selection achieves as much as 25.8% savings at 0.5 votes per document. Additionally, we ascertain the statistical significance of our results using a Wilcoxon signed rank test [35] at $\alpha < 0.05$. Overall best-performing methods at significance level are indicated by a hash symbol, and performances that are statistically significantly better than the baseline are highlighted by using boldface. Table 2 presents these results in detail, demonstrating the effectiveness of our method for crowdsourcing procedures that only have a constrained access to budget.

Table 1: Budget vs. savings for active document selection techniques.

Method	**Budget (in votes per document)**	**Savings (as the percentage of budget)**
Variance based sampling	0.25	2.6%
	0.50	15.4%
	0.75	14.6%
Mutual Information based sampling	0.25	17.4%
	0.50	25.8%
	0.75	23.3%

Table 2: Accuracy and statistical significance for active document selection techniques.

Method / **Budget**	*GP Baseline*	*Variance Based*	*Mutual Information Based*
0.25	0.696	0.699	**0.715**#
0.50	0.732	**0.740**	**0.749**#
0.75	0.753	**0.760**	**0.762**#

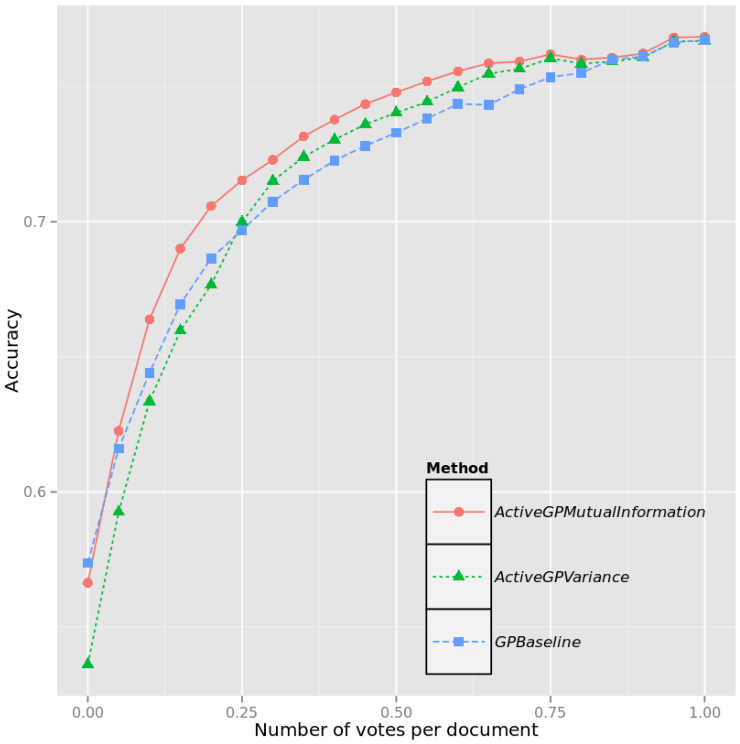


Figure 5: Learning curves of various document selection techniques.

6. CONCLUSIONS AND FUTURE WORK

In this work, we presented two information theoretic criteria which are employed towards the goal of active document selection. Focusing on budget-constrained scenarios, we demonstrate that our methods require 17% - 25% reduced budget in order to achieve the same accuracy as competitive baseline methods. To the best of our knowledge, this represents the first demonstration of using textual similarity for active document selection in crowdsourced document relevance assessment.

We additionally show the usefulness of semantics-preserving document embedding spaces for capturing, and subsequently exploiting, document similarities. To this end, we noted a consistent performance improvement across all methods when using dense doc-2-vec representations instead of sparse tf-idf vectors.

There are several promising directions for future work. In this paper, we have assumed that relevance judgements collected from crowd workers are noise-free. Although there have been attempts to model noisy sensor placement in the past [31], the current problem requires a different formulation of noise modeling, as the noise in our case follows a Bernoulli distribution. Analysing our current crowdsourcing procedure in light of Bernoulli noise appears to be a promising direction for the future.

While, in this paper, we exclusively experiment on textual documents, our method is applicable to any type of document that can be projected into a semantics-preserving embedding space. Experiments involving different document types, such as images or videos can leverage the ideas we described in this work for a broader range of applications.

7. ACKNOWLEDGEMENTS

We would like to thank Martin Davtyan and Hamed Hassani for their insightful feedback, questions and suggestions.

8. REFERENCES

[1] O. Alonso, D. E. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15, Nov. 2008.

[2] J. A. Aslam, V. Pavlu, and E. Yilmaz. A statistical method for system evaluation using incomplete judgments. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 541–548, New York, NY, USA, 2006. ACM.

[3] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 25–32, New York, NY, USA, 2004. ACM.

[4] J. Callan, M. Hoy, C. Yoo, and L. Zhao. Clueweb09 data set, 2009.

[5] B. Carterette, J. Allan, and R. Sitaraman. Minimal test collections for retrieval evaluation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 268–275, New York, NY, USA, 2006. ACM.

[6] C. Cleverdon. Readings in information retrieval. In K. Sparck Jones and P. Willett, editors, *Readings in Information Retrieval*, chapter The Cranfield Tests on Index Language Devices, pages 47–59. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[7] M. Davtyan, C. Eickhoff, and T. Hofmann. Exploiting document content for efficient aggregation of crowdsourcing votes. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 783–790. ACM, 2015.

[8] D. E. Difallah, M. Catasta, G. Demartini, and P. Cudré-Mauroux. Scaling-up the crowd: Micro-task pricing schemes for worker retention and latency improvement. In *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

[9] D. E. Difallah, G. Demartini, and P. Cudré-Mauroux. Pick-a-crowd: tell me what you like, and i'll tell you what to do. In *Proceedings of the 22nd international conference on World Wide Web*, pages 367–374. International World Wide Web Conferences Steering Committee, 2013.

[10] C. Eickhoff and A. P. de Vries. Increasing cheat robustness of crowdsourcing tasks. *Information retrieval*, 16(2):121–137, 2013.

[11] C. Eickhoff, C. G. Harris, A. P. de Vries, and P. Srinivasan. Quality through flow and immersion: gamifying crowdsourced relevance assessments. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 871–880. ACM, 2012.

[12] C. Grady and M. Lease. Crowdsourcing document relevance assessment with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 172–179, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[13] M. Hirth, T. Hoßfeld, and P. Tran-Gia. Cheat-detection mechanisms for crowdsourcing. *University of Würzburg, Tech. Rep*, 474, 2010.

[14] J. Howe. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Crown Publishing Group, New York, NY, USA, 1 edition, 2008.

[15] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 154–161, New York, NY, USA, 2005. ACM.

[16] 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*, pages 205–214. ACM, 2011.

[17] G. Kazai, J. Kamps, and N. Milic-Frayling. Worker types and personality traits in crowdsourcing relevance labels. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1941–1944. ACM, 2011.

[18] G. Kazai, J. Kamps, and N. Milic-Frayling. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information retrieval*, 16(2):138–178, 2013.

[19] G. Kazai and N. Milic-Frayling. On the evaluation of the quality of relevance assessments collected through crowdsourcing. In *SIGIR Workshop on Future of IR Evaluation*. Association for Computing Machinery, Inc., July 2009.

[20] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 453–456, New York, NY, USA, 2008. ACM.

[21] A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *The Journal of Machine Learning Research*, 9:235–284, 2008.

[22] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*, 2014.

[23] M. Lease. On quality control and machine learning in crowdsourcing. *Human Computation*, 11:11, 2011.

[24] M. Lease and G. Kazai. Overview of the trec 2011 crowdsourcing track. In *Proceedings of the text retrieval conference (TREC)*, 2011.

[25] D. J. MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

[26] C. C. Marshall and F. M. Shipman. The ownership and reuse of visual media. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, pages 157–166. ACM, 2011.

[27] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functionsâĂŤi. *Mathematical Programming*, 14(1):265–294, 1978.

[28] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

[29] N. Ramakrishnan, C. Bailey-Kellogg, S. Tadepalli, V. Pandey, et al. Gaussian processes for active data mining of spatial aggregates. In *SDM*, pages 427–438. SIAM, 2005.

[30] M. Sabou, K. Bontcheva, L. Derczynski, and A. Scharl. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *LREC*, pages 859–866, 2014.

[31] N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

[32] C. J. van Rijsbergen and K. SPARCK JONES. A test for the separation of relevant and non-relevant documents in experimental retrieval collections. *Journal of Documentation*, 29(3):251–257, 1973.

[33] E. M. Voorhees and D. K. Harman. *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*. The MIT Press, 2005.

[34] J. Wang, S. Faridani, and P. G. Ipeirotis. Estimating the completion time of crowdsourced tasks using survival analysis models. *Crowdsourcing for Search and Data Mining (CSDM 2011)*, page 31, 2011.

[35] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

[36] Y. Yan, G. M. Fung, R. Rosales, and J. G. Dy. Active learning from crowds. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1161–1168, 2011.