

Measuring the Reusability of Test Collections

Ben Carterette[†], Evgeniy Gabrilovich[‡], Vanja Josifovski[‡], Donald Metzler[†]

[†] Department of Computer & Information Sciences, University of Delaware, Newark, DE

[‡] Yahoo! Research, 2821 Mission College Blvd, Santa Clara, CA

carteret@cis.udel.edu | {gabr | vanjaj | metzler }@yahoo-inc.com

ABSTRACT

While test collection construction is a time-consuming and expensive process, the true cost is amortized by reusing the collection over hundreds or thousands of experiments. Some of these experiments may involve systems that retrieve documents not judged during the initial construction phase, and some of these systems may be “hard” to evaluate: depending on which judgments are missing and which judged documents were retrieved, the experimenter’s confidence in an evaluation could potentially be very low. We propose two methods for quantifying the reusability of a test collection for evaluating new systems. The proposed methods provide simple yet highly effective tests for determining whether an existing set of judgments is useful for evaluating a new system. Empirical evaluations using TREC datasets confirm the usefulness of our proposed reusability measures. In particular, we show that our methods can reliably estimate confidence intervals that are indicative of collection reusability.

Categories and Subject Descriptors

H.3.4 [Information Storage and Retrieval]: Performance Evaluation

General Terms

Algorithms, Experimentation

Keywords

evaluation, test collections, reusability

1. INTRODUCTION

Test collections lie at the heart of IR evaluation; they foster reproducible results and allow principled comparison of multiple retrieval systems. Test collections typically consist of a set of queries, a set of documents, and a set of relevance judgments. In some test collections, the judgments cover all possible query-document pairs. For instance, text categorization test collections normally provide an exhaustive list

of categories each document belongs to. However, this is not the case in most retrieval tasks, notably for those where the set of documents is so large that it is simply not feasible to judge every document for every query in the set.

The *pooling method* provides a way to focus judging effort on those documents least likely to be non-relevant [12]. Given a set of systems to be evaluated over the queries in the test collection, the top-scoring documents retrieved by the systems are pooled and judged for relevance to the queries that retrieved them. In test collections of realistic size, it is unlikely that pooling will find all the relevant documents in the corpus, but identifying and judging all such documents would be prohibitively expensive. When new systems are subsequently evaluated using the same test collection, practitioners are faced with one of the following two choices. One option is to collect judgments for the documents retrieved that were not previously judged. This can be costly and time consuming, especially when many new systems must be tested over a large test collection, as is the case for Web search. The other option is to only use existing judgments and effectively ignore newly retrieved documents that have not been previously judged. Evaluation can then be done either using compressed ranked lists [10], or by using evaluation metrics that can handle missing judgments [1, 3]. Depending on the number of queries and the number of unjudged documents retrieved, this approach may lead to a highly inaccurate measure of the system’s true performance.

We propose methods for quantifying the suitability of an existing set of judgments for evaluating new systems. Specifically, we show how estimates of *confidence intervals* for evaluation metrics such as precision or mean average precision (MAP) over the space of *possible* judgments for unjudged documents. The widths of these intervals provide clues as to the suitability of existing judgments to evaluate the new system. We also propose point estimates of reusability based on standard evaluation metrics and show that despite being less informative than the full confidence interval, they can provide quick and easy estimates of the interval width.

The main contributions of this paper are threefold. First, we introduce the concept of *reusability estimation*, which aims to quantify how useful a set of existing relevance judgments is for evaluating a new system. To the best of our knowledge, principled estimates of reusability of test collections have not been previously studied. Second, we propose two novel methods for quantifying reusability. The first method constructs confidence intervals for evaluation metrics using logistic regression, while the second converts standard information retrieval metrics into reuse metrics. Fi-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WSDM’10, February 4–6, 2010, New York City, New York, USA.

Copyright 2010 ACM 978-1-60558-889-6/10/02 ...\$10.00.

nally, we report the results of experimental evaluation using several TREC collections, which confirm that our methodology provides a reliable way for quantifying reusability and predicting system performance.

The remainder of the paper is organized as follows. Section 2 describes previous work related to evaluation and reusability. Section 3 presents our methodology for measuring the reusability of a test collection with respect to a new system. In Section 4 we evaluate our proposed methodology using TREC data. Section 5 concludes the paper and discusses possible directions for future work.

2. RELATED WORK

Reusability has long been a central concern about test collections. The original IR test collection, the Cranfield set of aerodynamic engineering abstracts, research questions, and relevance judgments, came to be reused to study retrieval tasks far beyond what it was designed for simply because it was easier to obtain it than to build a new collection. When TREC (the Text REtrieval Conference) began in 1992, part of its mission was to provide the wider research community with large test collections devoted to particular tasks that would be reusable (or “portable”) across research groups and over time [14]. There can be no doubt that it succeeded at that. However, in recent years, test collections have begun to become so large that their reusability is unclear. Pools tend to contain documents that exhibit certain properties such as containing a high proportion of title query terms, and many relevant documents that do not have those properties are left unjudged [2]. Reusing such collections will tend to favor systems that are like those that contributed the original documents.

Zobel was among the first to question the completeness of the pooling method and the effect of missing relevant documents on system evaluations [16]. He found that pooling could miss up to 50% of the relevant documents in the corpus, but relative orderings of systems would not be seriously affected. However, more recent collections seem to have serious bias problems [2, 4].

There are two general approaches to cope with the bias due to missing judgments. One is to modify existing or introduce new evaluation metrics. The *bpref* metric was introduced to deal with missing judgments by counting the number of non-relevant documents ranked above relevant documents [3]. Inferred average precision (infAP), as its name suggests, uses inferred precision values when judgments are missing [15]. Sakai introduced a suite of metrics based on compressing the ranked list to eliminate unjudged documents [10].

An alternative approach involves attempting to predict the relevance of unjudged documents for use in standard or new evaluation metrics. Jensen et al. addressed experimental repeatability, which is related to reusability) using judgments inferred from manually-built taxonomies [9]. Carterette addressed reusability in terms of whether two new systems could be reliably ranked relative to each other using relevance predictions based on very small sets of judgments. The predictions are used to calculate a probability that two systems are likely to swap after additional judgments [5]. Büttcher et al. used an SVM to predict the relevance of unjudged documents to find likely new relevant documents [4]. Although such estimates will have errors, it is unclear that they are any noisier than the judgments themselves.

This work is about assessing whether a particular system that did not contribute to a pool can be accurately evaluated given the judgments in that pool. We use both of the approaches above: we estimate the relevance of unjudged documents and use those estimates to calculate expectations for standard retrieval metrics. On top of that we add a *confidence interval* calculated using the predictions of relevance (note the difference from Carterette’s previous work, which only estimated swap probabilities [5]). When confidence intervals are wide and overlapping with confidence intervals for other systems, it is a red flag that more judgments are required before any reliable conclusion about the systems may be made. Because these confidence intervals are difficult to compute, we also introduce much simpler point estimates of their width, using a simple linear combination of features.

3. MEASURING REUSABILITY

The judgments from an existing test collection are often used to measure the *performance*, or effectiveness, of a new system. Classical information retrieval metrics include precision, recall, F1, mean average precision, R-precision, and DCG. These metrics assume that the relevance judgments are complete, that is, they require every document retrieved for every query to be judged, or else the metrics are undefined. That does not mean they are not useful, of course; depending on the experimental setting and assumptions made, metrics calculated without knowing all judgments can impart a great deal of information. Nevertheless, when using past judgments to evaluate new systems with classical metrics, problems can arise when unjudged documents are retrieved. In this case the judgments are said to be incomplete. In practice, nearly all collections are incomplete, so dealing with missing judgments is very important.

As discussed above, there are several ways to deal with unjudged documents. Such documents can be treated as non-relevant, based on the assumption that most documents are indeed non-relevant to any given query—a problematic assumption for recent large collections. Instead of assuming they are non-relevant, then, they can be ignored by forming condensed ranked lists [10]. However, recently it was shown that evaluations based on condensed ranked lists are biased when judgments are collected by pooling [11]. Several metrics have been proposed that overcome the problem of missing judgments by inferring the relevance of such items [1, 4, 7], but these approaches fail to quantify how accurate such evaluations actually are in the presence of missing data.

We propose a set of *reusability measures* that quantify the confidence that the existing test collection can be used to accurately evaluate the performance of a new system. Such measures are of theoretical and practical importance. The theory behind the measures can be used to develop more robust evaluation metrics. From a practical side, the measures can be used by IR practitioners to determine whether or not their existing test collection is sufficient to evaluate a new system, or if new judgments are needed.

We propose two types of reusability measures, each with their strengths and weaknesses, as we will describe shortly. Measures of the first type estimate a confidence interval for the metric of interest, such as mean average precision, by inferring the relevance of unjudged documents within a logistic regression framework. Measures of the second type compute a single scalar value that is distilled from classical and newly proposed evaluation metrics.

3.1 Interval Estimates of Reusability

Our first measure of reusability comes in the form of confidence intervals. More formally, suppose we have an existing set of judgments \mathcal{J} over the set of queries \mathcal{Q} and we wish to evaluate a new system on \mathcal{Q} (or some subset of \mathcal{Q}) according to metric m . Our goal is to estimate a confidence interval for m given \mathcal{J} and the ranked list of documents retrieved by the new system.

If the estimated confidence interval is wide, then we can say that \mathcal{J} is non-reusable. However, if the confidence interval is within some acceptable tolerance, as dictated by the underlying task, then we say that \mathcal{J} is reusable.

Confidence intervals are rather powerful in this situation, as they allow the practitioner to determine an acceptable level of uncertainty in their estimate. The uncertainty in the confidence intervals comes from unjudged documents being retrieved by the new system.

One can see, from a mathematical perspective, how such variance arises for common retrieval metrics. Carterette showed that the mean and variance for precision at k and average precision have analytical forms [6]. Given a query $Q \in \mathcal{Q}$, these analytical forms are:

$$E[\text{prec}@k] = \frac{1}{k} \sum_i p_i I(A_i \leq k)$$

$$\text{Var}[\text{prec}@k] = \frac{1}{k^2} \sum_i p_i q_i I(A_i \leq k)$$

$$E[AP] \approx \frac{1}{\sum_i p_i} \left(\sum_i a_{ii} p_i + \sum_{i,j} a_{ij} p_i p_j \right)$$

$$\begin{aligned} \text{Var}[AP] \approx & \frac{1}{(\sum_i p_i)^2} \left(\sum_i a_{ii} p_i q_i + \sum_{i,j} a_{ij} p_i p_j (1 - p_i p_j) \right. \\ & \left. + \sum_{i,j} 2a_{ii} a_{i,j} p_i p_j q_i + \sum_{i,j,k} 2a_{ij} a_{ik} p_i p_j p_k q_i \right) \end{aligned}$$

where the indexes i , j , and k go over the set of documents retrieved for Q , p_i is the probability that document i is relevant, $q_i = 1 - p_i$ is the probability that document i is non-relevant, A_i is the rank of document i , $I(A_i \leq q) = 1$ if the inequality is true and 0 otherwise, and $a_{ij} = 1/\max\{A_i, A_j\}$.

3.1.1 Modeling document relevance probability (p_i)

If document i is judged relevant, then $p_i = 1$; if it is judged non-relevant, then $p_i = 0$. There are several options for estimating p_i if document i is unjudged. The most naïve is to let $p_i = 0$ or $p_i = 0.5$ for unjudged documents. However, these estimates are unlikely to be accurate and may lead to poor estimates of the mean and variance. An alternative, which we adopt here, is to estimate p_i using a statistical model. We choose to model p_i using logistic regression, which is commonly used to model binary responses. Under this model, estimates for p_i have the form:

$$p_i = \frac{1}{1 + \exp[-\theta^T F(Q_i, D, \mathcal{J})]}$$

where θ is the model parameter vector and $F(Q_i, D, \mathcal{J})$ is a vector of features extracted for some query Q_i , document D , and set of relevance of judgments \mathcal{J} .

The model is trained as follows. Given an existing test collection, we first extract feature vectors for every (query,

judged document) pair. The target, or response, associated with each pair is the judgment, with non-relevant and relevant judgments corresponding to targets of 0 and 1, respectively. Finally, the model parameter vector θ is estimated using maximum likelihood. The trained model can then be used to estimate p_i for unjudged documents retrieved by a new system, thereby allowing us to effectively estimate the mean and variance of the metric under consideration.

3.1.2 Features

We explore two types of features in this work. The first type are *document similarity features*, which were originally proposed by Carterette and Allan [7]. For a given document i , we compute the cosine similarity between i and all of the judged documents. The general motivation behind these features is that if a given unjudged document is similar to one or more of the relevant documents, then the document itself is likely to be relevant. This is related to the well-known cluster hypothesis [13]. A similar argument can be made for non-relevant documents, as well.

Features of the second type are so-called *system features*, which quantify the effectiveness of a system and how complete the existing judgments are for the system. For every (query, document) pair we compute the following system features with respect to the existing judgments: the rank of the document, precision for known relevant documents at that rank, expected precision at that rank, and mean average reuse, which is a measure we will describe in more detail in Section 3.2. Each unique document may be associated with multiple feature vectors by virtue of having been ranked by more than one system. In these cases, the final probability of relevance p_i is obtained by averaging the values predicted by its feature vectors.

In addition to features that depend on both the query and the document, we also extract the following query-level features: fraction of relevant results retrieved, fraction of non-relevant results retrieved, fraction of unjudged results retrieved, and the mean average reuse of the query. Finally, for each query-level feature, we produce a system-level feature that is the mean of the query-level features computed over the entire set of queries.

Although we only consider these two types of features here, it is easy to include additional features, such as domain- or task-specific features, within the model. Additional features may improve the quality of the model estimates.

3.1.3 Confidence Intervals

Given a set of queries, it is common to report the mean of some metric over the entire set (e.g., mean average precision). We denote the mean of metric m by \bar{m} . Under the assumption that metrics are independent across queries, we compute the mean and variance of \bar{m} as follows:

$$E[\bar{m}] = \frac{1}{n} \sum_{i=1}^n E[m(Q_i)], \text{Var}[\bar{m}] = \frac{1}{n^2} \sum_{i=1}^n \text{Var}[m(Q_i)]$$

where $m(Q_i)$ is metric m evaluated on query Q_i .

It is straightforward to estimate confidence intervals for \bar{m} now that we have estimates for its mean and variance. The $100(1 - \alpha)\%$ confidence interval for \bar{m} is computed as follows:

$$\left[E[\bar{m}] - z_{\frac{\alpha}{2}} \sqrt{\frac{\text{Var}[\bar{m}]}{n}}, E[\bar{m}] + z_{\frac{\alpha}{2}} \sqrt{\frac{\text{Var}[\bar{m}]}{n}} \right]$$

where n is the number of queries and $z_{\frac{\alpha}{2}}$ is the value of z that satisfies $P(Z \leq z) = 1 - \frac{\alpha}{2}$, where Z is distributed according to a standard normal distribution. We note that these confidence intervals are only valid if \bar{m} is normally distributed, which is generally true by the Central Limit Theorem, assuming a large enough sample of queries. Studentized intervals can be used, if necessary (e.g., for $n < 30$).

Based on this formulation, it is easy to see that the two primary ways to tighten the bounds of the confidence interval are to lower the variance of the metric (i.e., obtain relevance judgments for unjudged documents) or increase the number of queries.

Although we primarily focus on precision at k and average precision in this paper, it should be noted that analytical forms for the means and variances of other retrieval metrics exist, including recall and NDCG [6]. Thus, our interval-based reusability measures can be easily applied to these metrics, as well. In general, our approach can be applied to any IR metric, including those that are not normally distributed and those that do not have analytical forms for their mean and variance. For such metrics, it may be possible to use bootstrap methods to estimate confidence intervals [8].

3.2 Point Estimates of Reusability

Confidence intervals are useful because they estimate the entire range of possible values of evaluation metrics for a new system based on the existing judgments. In general, interval estimates are more expressive and useful than point estimates. However, point estimates can be useful, not only because they provide a single number summary, but also because they are typically easier to compute. As we just showed, estimating confidence intervals can be somewhat involved, as it requires extracting features, estimating model parameters, computing means and variances, and so on. Therefore, we would like to develop point measures that can be used as proxies for confidence intervals. An ideal point measure for reusability would correlate strongly with the width of the estimated confidence intervals.

To compute point estimates of retrieval metrics, we define a set of novel features that directly quantify collection reusability. Specifically, we propose a methodology for converting standard precision-based evaluation metrics, such as precision at rank k and mean average precision, into reusability measures. Traditionally, the concept of precision has been used in information retrieval evaluation to determine the distribution of relevant documents in a set of retrieved documents. We can use a similar approach to define the reusability measure called *reuse* as the proportion of the *judged* documents that are retrieved by the new version of the system. Reuse at rank k for query Q is defined as:

$$reuse@k(Q) = \frac{judged@k(Q)}{k}$$

where $judged@k(Q)$ is the number of judged documents in the top k results for query Q using the new system.

While $reuse@k(Q)$ provides a measure that indicates the reusability of the judgments of the new version of the system, it suffers from many of the same problems as precision in standard performance evaluation, as it is not rank-aware. Various precision-based evaluation metrics are rank-aware, including average precision. We can easily convert average precision into a reusability measure, which we call *average*

reuse (AR), as follows:

$$AR(Q) = \frac{1}{judged(Q)} \sum_i reuse@i(Q)$$

where $judged(Q)$ is the number of documents judged for Q and i ranges over the judged document positions. We define the *mean average reuse* (MAR) as the mean of the AR values computed over a set of queries.

It should now be clear that any precision-based metric that uses binary relevance judgments can be easily converted into a reusability measure by assuming that judged documents are “relevant” (positive) and unjudged documents are “non-relevant” (negative).

To wit, our proposed reuse and average reuse measures ignore whether or not a retrieved document is relevant or not. The measures simply account for whether or not the document is judged. However, knowing whether the documents retrieved are relevant or not may be indicative of reusability. For example, if a system fails to retrieve many judged relevant documents, then we can assert with high confidence that the new system is bad. The opposite is not always true, however. A system that returns many judged relevant documents is not necessarily good, because its unjudged documents may actually be non-relevant. It depends on how many of its retrieved documents remain unjudged.

This is somewhat counterintuitive, but it follows from the fact that the proportion of relevant documents is very low. Most of the documents systems retrieve are non-relevant, and thus when a system fails to retrieve the relevant documents we know about, it is very unlikely that it retrieved many that we do not know about. There are exceptions, of course, but on average this is true. A system that retrieves many known relevant documents but still has many unjudged documents could go either way. It has established itself as being good at finding relevant documents, so there is reason to believe many of its unjudged documents are relevant. On the other hand, we know *a priori* that most unjudged documents are non-relevant. These conflicting states of knowledge produce low confidence in the system’s performance.

Therefore, we propose using traditional retrieval metrics calculated over judged relevant documents as point estimates of reusability as well. These include recall, precision at k , and MAP. We hypothesize that a combination of relevance-unaware measures like mean average reuse and relevance-aware measures like recall are good proxies for full-blown confidence intervals. We test this hypothesis in Section 4 by measuring the correlation between these two measures and the widths of confidence intervals estimated using the procedure described in Section 3.1.

4. EXPERIMENTAL EVALUATION

In this section we present experimental results demonstrating our ability to estimate and predict confidence intervals for different evaluation metrics and tasks. We show that a surprisingly small set of judgments is needed to rank new systems accurately *and* with high confidence, as long as those judgments came from a diverse set of systems.

Broadly speaking, our experimental procedure is to simulate three experiments. In the first experiment, a small set of runs contribute documents to a pool that is judged, and the pool is then used to evaluate those systems. In the second

