Semiautomatic Evaluation of Retrieval Systems Using Document Similarities

Ben Carterette & James Allan
{carteret, allan}@cs.umass.edu

Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003

ABSTRACT

Taking advantage of the well-known *cluster hypothesis* that "closely associated documents tend to be relevant to the same request", we can use inter-document similarity to provide more accurate and robust evaluation of retrieval systems. Using our method, we are able to accurately rank retrieval systems with up to 99% fewer relevance judgments than collected for the TREC conferences, and significantly more accurately than other algorithms given the same number of judgments.

Categories and Subject Descriptors: H.3 Information Storage and Retrieval; H.3.4 Systems and Software: Performance Evaluation

General Terms: Experimentation, Measurement

Keywords: information retrieval, evaluation, test collections, clustering

1. INTRODUCTION

Test collection construction is an important part of information retrieval experimentation. But it is expensive—although documents and queries are relatively easy to come by, relevance judgments are much harder. Assessors must be hired to read and judge documents, a process that is very resource-intensive. As a result, there has been a great deal of work on test collection construction and evaluation in the presence of incomplete or imperfect relevance judgments. Some of the work on the latter includes Buckley & Voorhees' *bpref* evaluation measure [2], Yilmaz & Aslam's inferred average precision [12], and Carterette et al.'s expected average precision [4]. For the former, there are two sub-fields: intelligent selection of documents for human assessors to judge (e.g. Cormack et al.'s Move-to-Front pooling [6], Carterette et al.'s algorithm [4], and Aslam et al.'s unbiased sampling method [1]) and automatic evaluation without human assessors (e.g. Joachims' method based on clicks [9] and Jensen's method based on assigning relevance from web taxonomies [8]). In this work we unite manual and automatic assessments of relevance with estimation methods for incomplete and imperfect judgments.

To do so we take advantage of van Rijsbergen's *cluster hypothesis*. The cluster hypothesis says that "closely associated documents tend to be relevant to the same request" [11]. In other words, a document that is similar to other relevant documents is likely to be relevant as well. The cluster hypothesis says nothing about documents that are nonrelevant or dissimilar, but these provide evidence as well: although there are many more ways a document can be nonrelevant than relevant, it still may be the case that a document similar to other nonrelevant documents is itself likely to be nonrelevant. Using this hypothesis to create probabilities that unjudged documents are relevant, we can estimate the differences between retrieval systems.

Our main contribution in this work is a model for evaluation based on document similarities. The most closely related previous work is that of Jensen [8], who used manual web taxonomies to assign relevance to documents for web evaluation purposes. Our work is complementary to his, showing how uncertainty due to these types of automatic relevance judgments can be incorporated formally.

2. ESTIMATING AVERAGE PRECISION

In previous work, we showed that average precision could be estimated by treating it as a random variable with a distribution over possible judgments of relevance [4]. Although the original work assumed that the probability of relevance for each document was a uniform $\frac{1}{2}$, one of the advantages of our approach is that we can model many different sources of evidence for the relevance of documents. There has been other work on estimating evaluation metrics, particularly by Aslam et al. [1], but to the best of our knowledge this cannot incorporate multiple sources of evidence for relevance.

Specifically, we showed that average precision can be written as a quadratic equation over X_i, Bernoulli random variables for the relevance of documents i:

$$AP = \frac{1}{\sum X_i} \sum_{i=1}^{n} \sum_{j \geq i} a_{ij} X_i X_j$$

where a_{ij} is a constant derived from the ranks of documents i and j[1]. The distribution of AP converges to normal, so it

[1] See Carterette et al. [4] for details

CIKM'07, November 6–8, 2007, Lisboa, Portugal.

can be described by its expectation and variance alone.

Calculating expectation and variance involves summing over all possible assignments of relevance. Since there are 2^n possible relevance assignments, this is intractable in practice. However, it can be approximated as follows:

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

$$Var[AP] \approx \frac{1}{\left(\sum p_i\right)^2} \left(\sum_i^n a_{ii}^2 p_i (1-p_i) + \sum_{j>i} a_{ij}^2 p_i p_j (1 - p_i p_j) \right.$$
$$\left. + \sum_{i \neq j} 2 a_{ii} a_{ij} p_i p_j (1-p_i) + \sum_{k>j \neq i} 2 a_{ij} a_{ik} p_i p_j p_k (1-p_i) \right)$$

The error in these approximations is a negligible $\mathcal{O}(n2^{-n})$.

2.1 Application to MAP

Assuming topics are independent, we can easily extend this to mean average precision (MAP), the mean of average precisions over a set of topics T. MAP is also normally distributed with expectation and variance:

$$\mathcal{E}MAP = \frac{1}{T} \sum_{t \in T} E[AP_t] \qquad (1)$$

$$\mathcal{V}MAP = \frac{1}{T^2} \sum_{t \in T} Var[AP_t]$$

Instead of assuming that all unjudged documents are nonrelevant (the conventional assumption), systems can then be ranked by $\mathcal{E}$MAP, taking into account any information we have about the relevance of documents.

In addition to ranking documents, we would also like to be able to estimate our confidence in our ranking. A measure of confidence allows us to quantify how good we believe a ranking of systems to be. Define ΔMAP to be the difference in mean average precisions for two different systems over the same set of topics. We will then define the confidence in the sign of the difference of mean average precision to be

$$\text{confidence} = P(\Delta MAP < 0) = \Phi\left(\frac{-E[\Delta MAP]}{\sqrt{Var[\Delta MAP]}} \right)$$

where $\Phi(X)$ is the normal cumulative density function with zero mean and unit variance.

Note that the expressions for expectation and variance of average precision depend upon knowing the probability of relevance p_i of each document. In our earlier work, we used a uniform $\frac{1}{2}$ probability for each document [4]. More recently, we have shown that better estimates of probability can provide much more accurate evaluation [3]. In the next section, we turn to the task of estimating the probability of relevance of documents.

3. ESTIMATING RELEVANCE

How well we can estimate average precision depends on how well we can estimate relevance. The advantage of our model is that it can incorporate any type of evidence for relevance that can be modeled.

The *cluster hypothesis* gives us an idea for a type of evidence: the similarity of documents to other relevant documents. The cluster hypothesis says "closely associated documents tend to be relevant to the same request" [11]. As we acquire judgments and learn about which documents are relevant, we may be able to take advantage of the cluster hypothesis to estimate the relevance of unjudged documents, which we can use in our evaluation.

We will model the probability of relevance of a document conditional on its similarity to other documents. This is similar to the approach Diaz [7] takes in regularizing retrieval scores to be re-ranked, substituting "probabilities of relevance" for Diaz's "retrieval scores". In addition, since we require our outputs to be probabilities, we will use logistic regression to fit the model rather than the least-squares approach Diaz used.

3.1 Regularization with Logistic Regression

In our logistic regression model, the log-odds of relevance is modeled as a weighted sum of similarities:

$$\log \frac{p_i}{1-p_i} = \beta_0 + \sum_{j=1}^n \beta_j \text{sim}(i,j)$$

where p_i is $P(X_i = 1)$, the probability of relevance of document i.

The weights are found by maximizing the likelihood of the data. The log-likelihood is

$$\log \mathcal{L}(\beta) = \sum \left(y_i \log p_i + (1-y_i) \log(1-p_i) \right) + \lambda \beta^T \beta$$

To help avoid overfitting, λ is a penalization parameter that can also be seen as a prior; the greater λ is, the stronger the prior, and the closer to 0 the trained coefficients will be.

The response y_i is the prior probability of relevance that we wish to regularize. Given some set of judgments $\mathcal{J} = R \cup N$ (relevant and nonrelevant judgments), let

$$y_i = 1 \quad \forall i \in R$$
$$y_i = 0 \quad \forall i \in N$$
$$y_i = \frac{|R|+1}{|R|+|N|+2} \quad \forall i \notin \mathcal{J}$$

If no judgments have been made, then $y_i = \frac{1}{2}$, the uniform probability of relevance. The vector $y = (y_1, y_2, ...)$ of judgments and plus-one-smoothed probabilities is the set of scores to be regularized.

3.2 Similarity

A well-known and commonly-used measure of similarity is the cosine of the angle between document vectors:

$$\text{sim}(i,j) = \cos(i,j) = \frac{\sum_{t \in V} w_{i,t} w_{j,t}}{\sqrt{\sum_{t \in V} w_{i,t}^2} \sqrt{\sum_{t \in V} w_{j,t}^2}}$$

where V is the vocabulary and $w_{i,t}$ is the term weight of term t in document i. Term weights are generally a combination of term frequency in the document and document frequency in the collection.

4. EXPERIMENTS

4.1 Data

We obtained the retrieval runs submitted to TREC ad hoc tracks in 1994 and 1996–1999 (TRECs 3 and 5–8). Each run ranks at most 1000 documents to 50 topics. The number of runs in each set varies from 40 in TREC-3 to 129 in TREC-8. We also obtained the NIST *qrels* files to use as the “true” relevance judgments. These contain relevance judgments for the top 50 or 100 documents retrieved by nearly every system.

Because of the computational cost of calculating the variance of MAP, we truncated all ranked lists after 100 documents. We are therefore only computing both true MAP and $\mathcal{E}$MAP over the top 100.

4.2 Selecting Documents to Judge

We compare two algorithms for selecting documents to judge: *minimal test collections* (MTC), presented in our earlier work [4], and a simple pooling method we call *incremental pooling* (IP). MTC takes as input a minimum confidence level α; it select documents for judging by adaptively reweighting based on previous judgments until that level of confidence is reached. We will target a minimum confidence of $\alpha = 95\%$ in our experiments. IP simply pools documents, orders them by the highest rank at which they were retrieved, then asks for judgments down the list.

4.3 Calculating Similarity

We indexed TREC disks 1–5 with Indri, using the Krovetz stemmer and the default list of stopwords included in the Indri package. Cosine similarities between all documents in the pool of depth 100 were pre-computed and stored on disk.

4.4 Estimating Relevance

To solve the logistic regression problem, we use our own R implementation of TR-IRLS [10], a conjugate gradient descent algorithm for iteratively reweighted least squares. We used the TREC-4 collection to train the regularization parameter λ; $\lambda = 1$ was selected for providing the best combination of few judgments and accurate evaluation.

The baseline we compare to is using the plus-one-smoothed estimates alone, without any similarity information. MTC+sim refers to the MTC algorithm plus similarity-based relevance probabilities; MTC+one refers to MTC with plus-one-smoothed relevance probabilities.

4.5 Evaluation

We wish to compare MTC to IP, and for MTC compare the probability estimation methods MTC+sim and MTC+one. To evaluate, we will look at the following statistics:

1. the rank correlation between a ranking with an incomplete set of judgments and the true ranking.
2. the number of judgments needed by MTC+sim and MTC+one to reach 95% confidence.
3. the accuracy at predicting the sign of ΔMAP.
4. the accuracy at predicting the sign of ΔMAP for the pairs that have a statistically significant difference.

The last statistic has been proposed by Cormack & Lyman [5] as an alternative to Kendall’s τ rank correlation, as the cost of missing the significant differences is much greater than the cost of missing non-significant differences.

For each experiment we run multiple trials with randomized orderings of systems and topics. Using the same randomization for two algorithms allows us to evaluate the significance of our results using paired hypothesis tests.

5. RESULTS AND DISCUSSION

The τ correlation, number of judgments needed by MTC+sim to reach 95% confidence, the accuracy at predicting the sign of ΔMAP, and the accuracy on pairs with a significant difference by a paired one-sided t-test are shown in Table 1. With 99% fewer judgments than in the *qrels*, we are able to achieve about 90% accuracy at identifying the sign of the difference in MAP between two systems. Although the τ correlations appear low, this method is doing an excellent job at identifying the significant differences. For example, on the TREC-3 set, we made only 951 judgments total (19 per topic, 23 per system, or less than one judgment per topic per system) and correctly identified 95% of the significant differences between systems.

Table 1 also shows the pairwise accuracy and τ correlation when using IP to judge the same number of documents judged by MTC+sim. The correlations and accuracies are close to MTC+sim, reinforcing that the pooling method works well, but the differences are statistically significant ($p < 0.001$). MTC+one (not shown) requires more judgments to reach 95% confidence and has slightly lower τ correlations and accuracy than MTC+sim. Its results are significantly better than IP, but significantly worse than MTC+sim.

Figure 1 plots the true ranking and the estimated ranking by $\mathcal{E}$MAP for each TREC. Manual runs are highlighted with boxes around their points.

5.1 Number of Judgments

Figure 2 shows how τ correlation changes for all three algorithms as documents are judged for one of the TREC-5 trials. τ correlation increases steadily, jumping fairly fast during the first hundred judgments. MTC+sim shows the greatest rate of increase, followed by MTC+one (which is much more variable), followed by incremental pooling.

5.2 “Tyranny of the Majority”

A concern about using document similarity to estimate relevance is the problem of the “tyranny of the majority”: most of the submitted runs are automatic, using variants of bag-of-words approaches to rank documents. Our similarity measure also relies on a bag-of-words approach, and thus it is reasonable to wonder whether we are actually doing a good job, or if we have managed to do well simply by doing well on those documents that were retrieved by the bulk of the submissions.

To answer this, we point to Figure 1. The manual runs (the runs that have retrieved the most different documents) are generally the best systems submitted, and therefore are on the right-hand side of the plot. For example, the rightmost ten points in Figure 1(a) are manual runs. Manual runs are ranked well by our approach, suggesting that it is not dominated by such an effect.

6. CONCLUSION

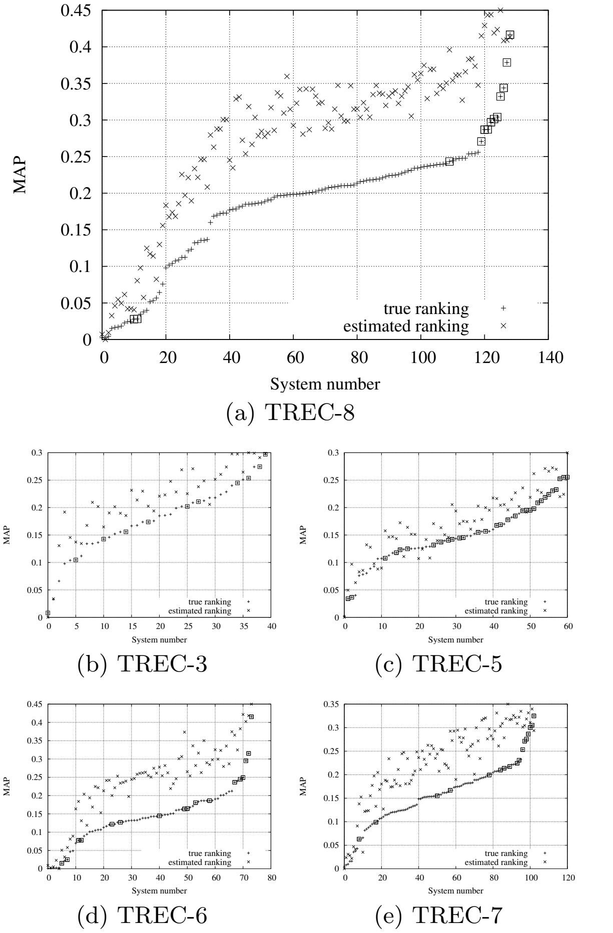
We have shown how document similarities can be used to evaluate retrieval systems with greatly reduced effort. The resulting similarity-based test collections provide more accurate evaluation results than the same number of pooled judgments and better confidence estimates than a simpler method of probability estimation.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval and in part by the Defense Ad-

		MTC+sim		IP	
	judgments	all systems	significant	all systems	significant
TREC	(% dec)	τ (accuracy)	accuracy	τ (accuracy)	accuracy
3	951 (99%)	0.774 (0.887)	0.947	0.697 (0.839)	0.922
5	829 (99%)	0.778 (0.894)	0.973	0.743 (0.860)	0.954
6	1592 (98%)	0.811 (0.893)	0.994	0.771 (0.880)	0.981
7	743 (99%)	0.768 (0.895)	0.952	0.730 (0.867)	0.931
8	1794 (98%)	0.836 (0.918)	0.978	0.738 (0.886)	0.946

Table 1: The first column shows the number of judgments needed to reach 95% confidence and the percent decrease from the full set of *qrels*. The next two columns show rank correlations and pairwise accuracy between the true ranking and the estimated ranking by $\mathcal{E}$MAP for the full set of systems and the subset of pairs with significant differences. All MTC+sim results are statistically significant improvements over IP.



(a) TREC-8

(b) TREC-3

(c) TREC-5

(d) TREC-6

(e) TREC-7

Figure 1: True and estimated rankings for the five TREC collections. Manual runs are highlighted with boxed points.

vanced Research Projects Agency (DARPA) under contract number HR001-06-C-0023. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

7. REFERENCES

[1] J. A. Aslam, V. Pavlu, and E. Yilmaz. A statistical method for system evaluation using incomplete judgments. In *Proc. of SIGIR*, pages 541–548, 2006.

[2] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of SIGIR*, pages 25–32, 2004.

[3] B. Carterette. Robust test collections for retrieval evaluation. In *Proc. of SIGIR*, pages 55–62, 2007.

[4] B. Carterette, J. Allan, and R. K. Sitaraman. Minimal test collections for retrieval evaluation. In *Proceedings of SIGIR*, pages 268–275, 2006.

[5] G. V. Cormack and T. R. Lyman. Power and bias of subset pooling strategies. In *Proceedings of SIGIR*, pages 837–838, 2007.

[6] G. V. Cormack, C. R. Palmer, and C. L. Clarke. Efficient Construction of Large Test Collections. In *Proceedings of SIGIR*, pages 282–289, 1998.

[7] F. Diaz. Regularizing ad hoc retrieval scores. In *Proceedings of CIKM*, pages 672–679, 2005.

[8] E. C. Jensen. *Repeatable Evaluation of Information Retrieval Effectiveness in Dynamic Environments.* PhD thesis, Illinois Institute of Technology, 2006.

[9] T. Joachims. Evaluating retrieval performance using clickthrough data. In *Text Mining*, pages 79–96. 2003.

[10] P. Komarek and A. Moore. Making logistic regression a core data mining tool: a practical investigation of accuracy, speed, and simplicity. Technical Report CMU-RI-TR-05-27, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, May 2005.

[11] C. J. van Rijsbergen. *Information Retrieval.* Butterworths, London, UK, 1979.

[12] E. Yilmaz and J. Aslam. Estimating average precision with incomplete and imperfect relevance judgments. In *Proceedings of CIKM*, pages 102–111, 2006.

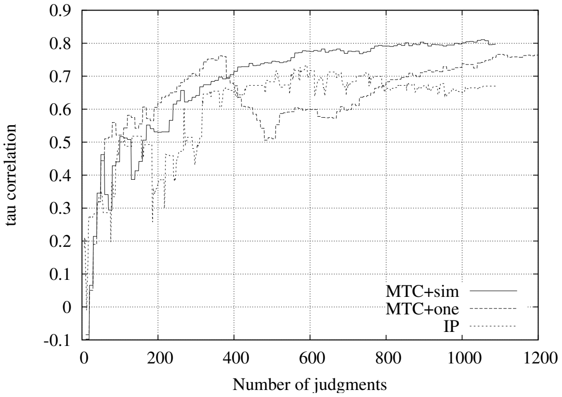


Figure 2: Kendall's τ correlation increases with the number of judgments.