On the Information Difference between Standard Retrieval Models

Peter B. Golbus and Javed A. Aslam
College of Computer and Information Science
Northeastern University
Boston, MA, USA
pgolbus,jaa@ccs.neu.edu

ABSTRACT

Recent work introduced a probabilistic framework that measures search engine performance information-theoretically. This allows for novel meta-evaluation measures such as **Information Difference**, which measures the magnitude of the difference between search engines in their ranking of documents. for which we have relevance information. Using Information Difference we can compare the *behavior* of search engines—which documents the search engine prefers, as well as search engine *performance*—how likely the search engine is to satisfy a hypothetical user. In this work, we **a)** extend this probabilistic framework to precision-oriented contexts, **b)** show that Information Difference can be used to detect similar search engines at shallow ranks, and **c)** demonstrate the utility of the Information Difference methodology by showing that well-tuned search engines employing different retrieval models are more similar than a well-tuned and a poorly tuned implementation of the same retrieval model.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*

Keywords

Information Retrieval; Search Evaluation

1. INTRODUCTION

One of the ways search engines are compared is by computing the magnitude of the difference between their performance using some IR evaluation measure. These measures attempt to quantify the satisfaction of a hypothetical user of a search engine. This is crucial information about a search engine, but it does not necessarily provide a great deal of insight into search engine *behavior*—which documents does the search engine prefer and why? For example it is possible for two search engines to retrieve wildly different documents, or to rank similar documents in a very different order, and yet receive the same score from an evaluation metric, even a diversity metric such as ERR-IA [3]. It is equally possible for two ranked lists to be highly similar and yet for one system to have a much greater performance than the other.

Recently, Golbus and Aslam [2] introduced a probabilistic framework that measures performance information-theoretically. The advantage of this approach is that it provides additional novel interpretations beyond simply estimating user satisfaction. For example, the authors defined **Information Difference**, which measures the magnitude of the difference between systems in their ranking of documents for which we have relevance information. The authors demonstrated that Information Difference can be used to detect similar search engines whereas performance deltas cannot.

However, the probabilistic framework underlying Information Difference required a recall-oriented approach and was evaluated at rank 1000. This leads to the concerns that Information Difference detected the similarities between systems based on the uninteresting long "tail" of the ranked lists. In this work, we **a)** adapt the probabilistic framework to precision-oriented tasks at shallow ranks. We demonstrate that **b)** even when evaluated at rank 20, Information Difference is still able to detect similar systems with high accuracy. Therefore, Information Difference relies on whether systems choose the same highly relevant documents at the top. Finally, we **c)** demonstrate the utility of the Information Difference methodology by showing that well-tuned search engines employing different retrieval models are more similar than a well-tuned and a poorly-tuned implementation of a retrieval model, *i.e.* that a well-tuned instantiation of BM25 is more similar to a well-tuned instantiation of a language model than it is to a poorly tuned instantiation of BM25.

2. INFORMATION DIFFERENCE

In this section, we provide an overview of the Information Difference methodology described in [2].

Mathematically, one can view a search system as providing a *total ordering* of the documents ranked and a *partial ordering* of the entire collection, where all ranked documents are preferred to unranked documents but the relative preference among the unranked documents is unknown. Similarly, one can view relevance assessments—commonly referred to as QRELs—as providing a partial ordering of the entire collection: in the case of binary relevance assessments, for exam-

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

http://dx.doi.org/10.1145/2600428.2609528.

ple, all judged relevant documents are preferred to all judged non-relevant and unjudged documents, but the relative preferences among the relevant documents and among the non-relevant and unjudged documents is unknown. Thus, mathematically, one can view retrieval evaluation as comparing the partial ordering of the collection *induced by the search system* with the partial ordering of the collection *induced by the relevance assessments.*

Golbus and Aslam described a *probabilistic framework* within which to compare two such orderings, defined in terms of three things: (1) a sample space of objects, (2) a distribution over this sample space, and (3) random variables over this sample space. For example, Golbus and Aslam defined a new evaluation measure, **Relevance Information Correlation** in the following way. Let the sample space, $\Omega = \{(d_i, d_j) \mid rel(d_i) \neq rel(d_j)\}$, be the set of all ordered pairs of judged documents with different relevance grades. Let $P = U$ be the uniform probability distribution over all such pairs of documents. We define a QREL variable Q over ordered pairs of documents as

$$Q\left[(d_i, d_j)\right] = \begin{cases} 1 & \text{if } rel(g_i) > rel(g_j) \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The ranked list variable R is computed by truncating the list at the last retrieved relevant document. Let r_i represent the rank of document d_i.

$$R\left[(d_i, d_j)\right] = \begin{cases} 1 & \text{if } r_i < r_j \\ 0 & \text{if neither } d_i \text{ nor } d_j \text{were retrieved} \\ -1 & \text{otherwise.} \end{cases} \tag{2}$$

Relevance Information Correlation is the mutual information between the QREL variable Q and the truncated ranked list variable R.

$$RIC(System) = I(R_{System}; Q). \tag{3}$$

This quantity is estimated via Maximum Likelihood for a given QREL and system.

This definition is inherently recall-oriented. In this work, we propose a precision-oriented version, RIC@k. RIC@k differs from RIC in two ways. First, we normalize with respect to the maximum possible $RIC@k$ of an ideal ranked list, as with nDCG. Second, we alter the probability distribution so as to give more weight to documents with higher relevance grades. To do so, we begin by observing that evaluation metrics can be viewed as inducing probability distributions over ranks. For example, Carterette [1] defines the probability of stopping at a rank k according to $nDCG$ as

$$P_{DCG}(k) = \frac{1}{\log_2(k+1)} - \frac{1}{\log_2(k+2)}. \tag{4}$$

Imagine a QREL with R_{gmax} documents relevant at the highest grade. According to the QREL these documents are equally likely to appear at ranks one through R_{gmax}, but have zero probability of appearing anywhere else. Therefore, in any ideal ranked list, the probability associated with one of these documents will be $P_{DCG}(k)$ for some k with $1 \leq k \leq R_{gmax}$. Therefore, we define the probability of a *document* as the average probability of the ranks at which the document can appear in an ideal list. If R_g is the number of documents that are relevant at grade g, then for a document d with such that $rel(d) = g$, the minimum rank

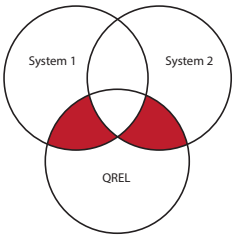

Figure 1: Information Difference corresponds to the symmetric difference between the intersections of the systems with the QREL in information space.

for this document in an ideal list

$$k_{min} = \sum_{i=g+1}^{gmax} R_i, \tag{5}$$

i.e. after all of the documents with higher relevance grades, and the maximum rank is

$$k_{max} = k_{min} + R_g. \tag{6}$$

Then the probability associated with the document is

$$\begin{aligned} P(d) &= \alpha \frac{\sum_{i=k_{min}}^{k_{max}} \frac{1}{\log_2(i+1)} - \frac{1}{\log_2(i+2)}}{R_g} \\ &= \alpha \frac{\frac{1}{\log_2(k_{min}+1)} - \frac{1}{\log_2(k_{max}+2)}}{R_g}, \end{aligned} \tag{7}$$

where α is a normalizing constant. Note that the probability of non-relevant documents is *non-zero*, and that this definition can also be used for binary relevance.

RIC requires us to define a probability distribution over document pairs, whereas Equation 7 defines a probability for documents. To create the appropriate distribution, we assume that each document in the pair is chosen independently,

$$P(d_i, d_j) = \beta P(d_i) P(d_j) \tag{8}$$

where β is a normalizing constant that ensures that $P(d_i, d_j)$ forms a distribution.

We define $RIC@k$ by normalizing by the ideal ranked list, as in nDCG, and computing mutual information with respect to the probability distribution defined in 8.

$$RIC@k(S) = \frac{I(R_S; Q)}{I(R_{ideal}; Q)} \tag{9}$$

Information Difference is inspired by the Boolean Algebra symmetric difference operator as applied to information space, corresponding to the symmetric difference between the intersections of the systems with the QREL (see Figure 1). For two systems S_1 and S_2,

$$id(S_1, S_2) = I(R_{S_1}; Q \mid R_{S_2}) + I(R_{S_2}; Q \mid R_{S_1}), \tag{10}$$

with Q and R defined as in Equations 1 and 2 respectively. We also define $id@k$ by using the probability distribution described in Equation 8, and by normalizing with respect to an ideal system.

$$id@k(S_1, S_2) = \frac{I(R_{S_1}; Q \mid R_{S_2}) + I(R_{S_2}; Q \mid R_{S_1})}{I(R_{ideal}; Q)}. \tag{11}$$

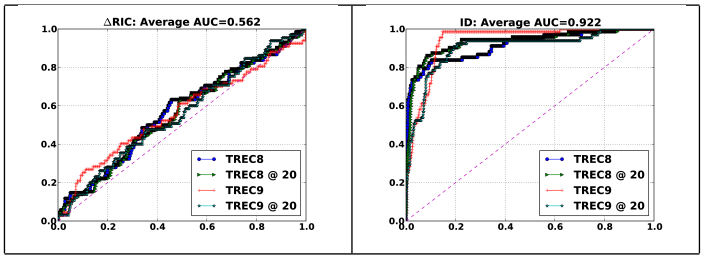

Figure 2: ROC curve of ΔRIC (left) and Information Difference (right) when used to predict whether systems with similar performance are in fact "similar" (as described in Section 3.1).

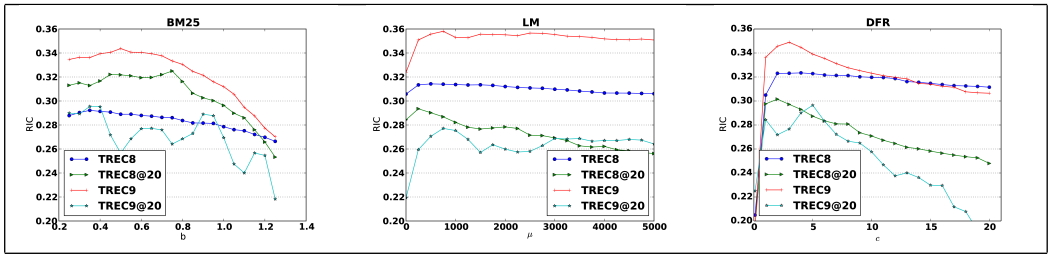

Figure 3: Performance as a function of retrieval model parameters

3. ANALYSIS

In this section, we employ the framework described in Section 2 to demonstrate the utility of the Information Difference framework. In Section 3.1, we show that Information Difference can be be used to determine whether systems are similar, even at shallow ranks, whereas performance deltas cannot. In Section 3.2, we demonstrate the use of Information Difference as a tool for meta-evaluation by performing a simplistic experiment concerning the similarity between search engines.

3.1 Detecting Similarity

We wish to detect whether two systems are similar. As a proxy for similarity, we will consider two systems submitted to TREC[1] to be "similar" iff they were submitted by the same group. It is reasonable to assume that the majority of these systems were different instantiations of the same underlying technology, although there will be many instances where this is not the case at all. We repeat the experiment first performed by Golbus and Aslam [2] to show that the results also hold when performed in a precision-oriented fashion (at rank 20), and are therefore not dependent on the long and uninteresting "tail" of the ranked lists.

Using the same construction, we sorted all the systems submitted to TREC 8 and TREC 9 by RIC and RIC@20, and separated them into twenty equal-sized bins. By construction, each bin contained systems with small differences in performance. All systems within each bin are compared to one another using Information Difference and performance delta. Figure 2 shows the resulting ROC curves when Information Difference and performance delta are used to predict whether two systems meet our proxy for similarity described above. We also report the area under the ROC curves averaged over all four conditions (TREC 8 vs TREC 9; RIC vs RIC@20). It is quite evident that Information Difference, with an average AUC greater than 0.9, is quite capable of detecting similarities as reported by our proxy, even at shallow ranks. It is equally evident that with an average AUC less than 0.6, performance delta is not capable of detecting similarities. This result is obvious—Information Difference was constructed for just this purpose, whereas performance delta is not. However, we believe that this result is also important. Since performance delta has traditionally been used to measure the similarities between search engines, our notions of which systems are similar may be false.

3.2 Measuring Differences

In this section, we will attempt to determine whether the choice of retrieval model has a bigger impact on the *behavior* (rather than the *performance*) of a search engine than does parameter tuning. To perform this experiment, we use a standard, state-of-the-art search engine, in this case the Terrier search engine [4], to create highly simple search engines, *i.e.* without query expansion, pseudo-relevance feedback, etc., and analyze the results when our systems are run over TRECs 8 and 9. We compare **1)** a query-prediction language model (LM) with Dirichlet smoothing, **2)** BM25, and **3)** Divergence from Randomness (DFR) models,[2] across a range of 21 different, evenly spaced, "reasonable" parameter values (see Figure 3), and the "best" of these observed parameter values which achieves the maximum performance (see Table 1).[3] As we can see from Table 1, the three models perform relatively consistently with one another. Figure 3 shows that, with the exception of BM25 and DFR at rank 20, each model has reasonably consistent performance with itself as parameters are tuned.

[1]The annual, NIST-sponsored Text REtrieval Conference (TREC) creates test collections commonly used in academic research.

[2]We use In$_e$B2 for RIC and PL2 for RIC@20.

[3]Our goal is to compare search engines during the *evaluation phase*, when relevance assessments have already been used. Therefore we employ the *best* parameters for *these* queries, rather than *optimal* parameters applicable to *future* queries.

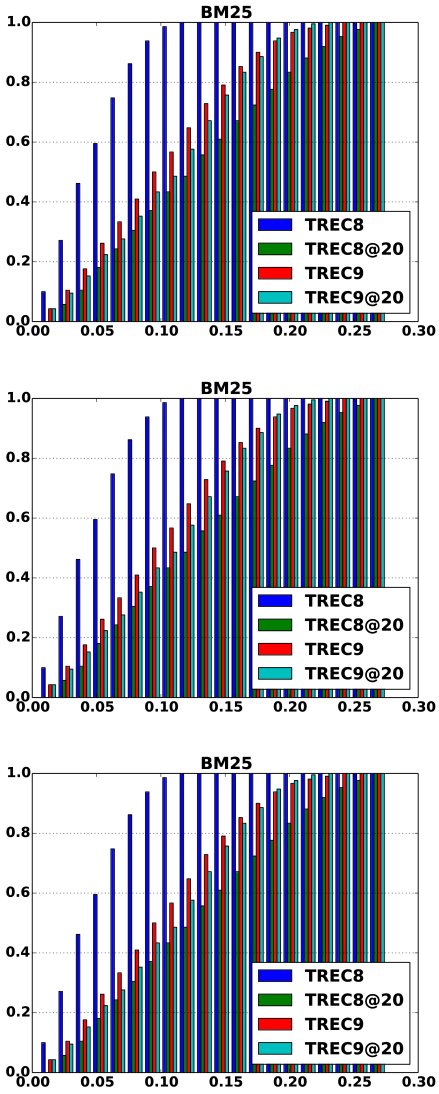

Figure 4: Cumulative histograms of Information Difference between parameterizations of a standard retrieval model.

Using Information Difference we can now measure the similarity of these models are in terms of their *behavior*, rather than their *performance*. Recalling that a small Information Difference implies a high degree of similarity, consider Table 2, which shows the Information Difference between the best performing retrieval models. As a point of reference, using an Information Difference threshold of 0.1 would have achieved a roughly 92% average accuracy on the classification task described in Section 3.1. Therefore, it is quite likely that Information Difference would have failed in this case and considered these retrieval models to be the same system.

Now we consider the difference between instantiations of a single model. We instantiate each model with all 21 parameter values and compute the Information Difference between each pair of instantiations. Figure 4 shows cumulative histograms of the Information Difference between all 21 choose 2 pairs of these model instantiations. These histograms show that there is far more difference in behavior within a model across parameterizations than their is across models with the best parameterization. For example, consider the largest Information Difference between models, which is between our language model and BM25 on TREC 8 at rank 20. The Information Difference of 0.149 is smaller than roughly 40% of the pairs of BM25 models. The smallest Information Difference is between BM25 and DFR on TREC 9 at rank 20. The Information Difference of 0.056 is smaller than all but roughly 45% of the pairs of LM models.

RIC	TREC8	TREC8@20	TREC9	TREC9@20
BM25	0.292	0.325	0.344	0.295
LM	0.314	0.294	0.358	0.277
DFR	0.323	0.302	0.349	0.296

Table 1: Best observed performance of standard retrieval models.

ID		TREC8	TREC8@20	TREC9	TREC9@20
LM	DFR	0.076	0.110	0.088	0.122
LM	BM25	0.068	0.149	0.092	0.127
BM25	DFR	0.077	0.091	0.108	0.056

Table 2: Information Difference between standard retrieval models with "best" parameters (Section 3.2).

4. CONCLUSION

The probabilist framework developed by Golbus and Aslam leads to interesting, novel, and highly interpretable meta-evaluations within the same context as traditional evaluation. As originally introduced, this framework was only applicable in deeply judged, recall-oriented contexts, greatly reducing its practical value. In this work, we extended this probabilistic framework to precision-oriented contexts, which are far more prevalent. We also showed two novel applications of Information Difference, a tool developed within this framework. We showed that Information Difference can be used to detect similar search engines at shallow ranks. We also showed that Information Difference can be used as a tool for meta-evaluation by showing that well-tuned search engines employing different retrieval models are more similar than a well-tuned and a poorly-tuned implementation of the same retrieval model.

5. REFERENCES

[1] Ben Carterette. System effectiveness, user models, and user utility: A conceptual framework for investigation. In *SIGIR '11*.

[2] Peter B. Golbus and Javed A. Aslam. A mutual information-based framework for the analysis of information retrieval systems. In *SIGIR '13*.

[3] Peter B. Golbus, Javed A. Aslam, and Charles L.A. Clarke. Increasing evaluation sensitivity to diversity. *Information Retrieval*, 16(4), 2013.

[4] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *OSIR '06*.