Evaluating Retrieval Models through Histogram Analysis

Kriste Krstovski[†,§]
kriste@cs.umass.edu

David A. Smith[‡]
dasmith@ccs.neu.edu

Michael J. Kurtz[§]
kurtz@cfa.harvard.edu

[†]College of Information and Computer Sciences
University of Massachusetts Amherst
Amherst, MA, 01003

[‡]College of Computer and Information Science
Northeastern University
Boston, MA, 02115

[§]Harvard-Smithsonian Center for Astrophysics
Cambridge, MA, 02138

ABSTRACT

We present a novel approach for efficiently evaluating the performance of retrieval models and introduce two evaluation metrics: **Distributional Overlap** (DO), which compares the clustering of scores of relevant and non-relevant documents, and **Histogram Slope Analysis** (HSA), which examines the log of the empirical distributions of relevant and non-relevant documents. Unlike rank evaluation metrics such as mean average precision (MAP) and normalized discounted cumulative gain (NDCG), DO and HSA only require calculating model scores of queries and a fixed sample of relevant and non-relevant documents rather than scoring the entire collection, even implicitly by means of an inverted index. In experimental meta-evaluations, we find that HSA achieves high correlation with MAP and NDCG on a monolingual and a cross-language document similarity task; on four ad-hoc web retrieval tasks; and on an analysis of ten TREC tasks from the past ten years. In addition, when evaluating latent Dirichlet allocation (LDA) models on document similarity tasks, HSA achieves better correlation with MAP and NCDG than perplexity, an intrinsic metric widely used with topic models.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Retrieval models, Selection process

Keywords: efficient evaluation; retrieval models; topic models

1 Introduction

Evaluating retrieval models with ranking metrics, such as mean average precision (MAP) and normalized discounted cumulative gain (NDCG), requires computing, in the worst case, the relevance score of each query against each member of a collection of n documents and then sorting the results. In practice, of course, most retrieval systems achieve vastly better performance by: exploiting inverted indices containing sufficient statistics on features such as terms and n-grams; by performing lossless or lossy pruning of posting lists; and by keeping track of only the top k documents for each query. For large collections, however, building even one index can be costly, and evaluating multiple models may require the creation of multiple indices. Reindexing can become even more common when working with continuous representations, as in image retrieval or in using topic models for text.

Before building new indices and tuning other efficiency parameters of an IR system, researchers may want some validation that a new feature, such as skip n-grams or LDA [1] topics, will positively impact effectiveness on the target task. *Rescoring* ranked lists generated by a baseline system provides one such check on model validity; however, new models will be most useful when they identify relevant results outside the output of the baseline system.

To alleviate these drawbacks, this paper proposes a novel approach for efficiently evaluating retrieval models by analyzing the relationship between histograms computed over the empirical distribution of relevance scores of query relevant and non-relevant documents. We analyze the performance of two metrics based on these histograms—**Distributional Overlap** (DO) and **Histogram Slope Analysis** (HSA)—in three different settings: monolingual and cross-language document-similarity retrieval; ad-hoc web retrieval; and a meta-evaluation of the ranked lists of ten TREC tracks from the past ten years. HSA achieves better correlation with MAP and NDCG on full collections than does MAP or NCDG on the subcollections used by HSA. In addition, when evaluating LDA models on document similarity tasks, HSA achieves better correlation with MAP and NCDG than perplexity, an intrinsic metric widely used with topic models.

2 Histogram Analysis

Our evaluation approach is based, first of all, on the simple notion that a good retrieval model should cluster relevant documents together with higher scores than non-relevant

SIGIR'15, August 09 - 13, 2015, Santiago, Chile.

DOI: http://dx.doi.org/10.1145/2766462.2767821.

documents (We discuss below the relation to other such "cluster hypotheses".). Better retrieval models, under this hypothesis, will have a relatively small overlap between the scores of relevant and non-relevant documents. We use the volume of the distributional overlap between the histograms of query relevant and non-relevant documents to define our first evaluation metric—Distributional Overlap (DO).

Computing the log ratio between the histograms of the query relevant and non-relevant documents gives us the proportion of relevant documents that we would expect to find at certain relevance values. Across different retrieval models, as the performance of the model improves, the proportion of relevant documents found at higher relevance scores should increase thus making the slope of the histogram analysis steeper going down from higher to lower values on the relevance scale. We define this slope as an evaluation metric that we call Histogram Slope Analysis (HSA).

By computing DO and HSA from the full set of relevant documents and a relatively small sample of non-relevant documents, we are able to evaluate retrieval models much more efficiently than with metrics computed over ranked lists of documents. We simply need to summarize the empirical distribution of relevant and non-relevant documents using histograms with some fixed number of bins.

3 Computing DO and HSA

For a test collection of k queries $Q = q_1, q_2, q_3, ...q_k$ and sets of query relevant $R_{q_i} = r_1^i, r_2^i, r_3^i, ...r_m^i$ and non-relevant $NR_{q_i} = nr_1^i, r_2^i, r_3^i, ...r_n^i$ documents, we evaluate retrieval model's scoring function $Score_f$ to get $V_R = Score_f(q_i, R_{q_i})$ and $V_{NR} = Score_f(q_i, NR_{q_i})$. Relevance scores have range of values that vary depending on the scoring function of the retrieval model. Different retrieval models generate relevance values on different scales. In order to compare and rank their performance, relevance values generated by each model are normalized to a range of [0,1].

Using a set B of equally spaced bins, histograms are computed over the two sets of relevance scores to give $H_R = [h_1^r, h_2^r, h_3^r, ..., h_{|B|}^r]$ and $H_{NR} = [h_1^{nr}, h_2^{nr}, h_3^{nr}, ..., h_{|B|}^{nr}]$. The h_b^r and h_b^{nr} are the counts of the number of times relevant and non-relevant scores fall within the bin centered at b: $h_b^r = \#(V_R \in b)$, $h_b^{nr} = \#(V_{NR} \in b)$. Since in typical real document collections the portion of query relevant documents is significantly smaller than non-relevant documents, we use log scale for the frequency axis when creating the two histograms. We further define a set of supported bins $B' = \{b' : h_{b'}^r, h_{b'}^{nr} \neq 0\}$. Distributional Overlap measures the volume of the overlap between the two histograms:

$$DO = \sum_{b' \in B'} \log(\min(h_{b'}^r, h_{b'}^{nr})) \quad (1)$$

Figure 1 illustrates an example of computing DO. In this figure we show the joint histogram plots across two different document similarity models on the task of finding document translation pairs. Both similarity models are based on the Polylingual Topic Model (PLTM) [5] configured with number of topics T=50 and 500. We detail this experimental setup in §5.1.

For HSA, we take the log ratio of the two histograms for the bins $b' \in B'$ where they are both observed $O_{b'} = \log(\frac{H_{R_{b'}}}{H_{NR_{b'}}})$ and then fit a linear function $O_{b'} = \alpha + \beta b' + \epsilon_{b'}$

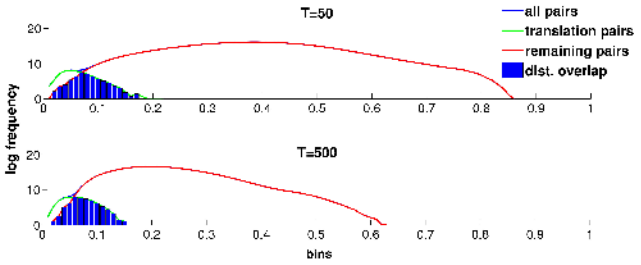

Figure 1: Computing DO to evaluate PLTM models on the task of retrieving document translation pairs. PLTM with 50 topics: acc.=94.3%; with 500 topics, acc.=99.3%.

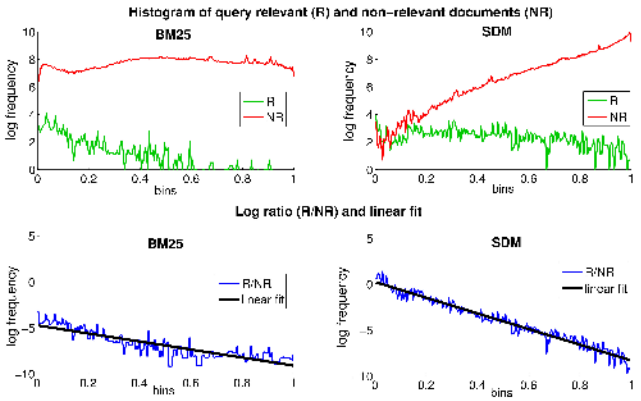

Figure 2: Computing HSA to evaluate ad-hoc retrieval models on the TREC Web Track 2009. NDCG(BM25)=0.106, NDCG(SDM)=0.293.

using linear least squares regression. We define HSA to be the estimated slope:

$$HSA = \hat{\beta} = \frac{\sum_{b' \in B'}(b' - \bar{B'})(O_{b'} - \bar{O})}{\sum_{b' \in B'}(b' - \bar{B'})^2} \quad (2)$$

Figure 2 shows an example of computing HSA across two retrieval models on the TREC Web Track 2009. In this example we compute HSA over the relevance scores of query relevant and non-relevant web pages obtained using BM25, which gave the worst NDCG, and the sequential dependence model (SDM), which achieved the highest NDCG. See §5.2 for experimental details. The first two subplots show the log histograms of the query relevant (R) and non-relevant documents (NR). The bottom two subplots show the log ratio of the empirical distributions of R and NR scores $O_{b'}$ along with the linear fit $\hat{O}_{b'}$.

4 Previous Work

In their original work, Jardine and van Rijsbergen [3] used histograms to define one of the prominent concepts in information retrieval: the cluster hypothesis, which states that "closely associated documents tend to be relevant to the same requests". The hypothesis introduced the cluster based retrieval [8] and many of its variants [2]. To measure the potential of the cluster based retrieval on a particular document collection, Jardine and van Rijsbergen [3] proposed the cluster hypothesis test. For a given collection and a set of queries the test measures the similarity between all query

relevant documents and between all query relevant and non-relevant documents. Aside from the cluster hypothesis test, researchers have proposed other measures of the cluster hypothesis [7, 9]. More recently, Raiber and Kurland [6] analyzed how these measures correlate with the performance of cluster based retrieval. While different measures exist for the cluster hypothesis they have not found their use in evaluating the performance of different retrieval models.

Although developed independently, the DO metric is consistent with the cluster hypothesis. Unlike cluster hypothesis tests, which asks whether two relevant documents are similar to each other, with DO we analyze the similarity of relevance scores between query relevant and non-relevant documents. Since DO is reflecting on the cluster hypothesis one may also consider DO as an intuitive implementation of the cluster hypothesis test in the space of query relevance values. As we shall see, however, HSA is better correlated with established ranking metrics than DO.

5 Experimental Results

The purpose of developing DO and HSA was to be able to evaluate and predict the performance of retrieval models. To demonstrate this ability we compare the values of DO and HSA with existing IR metrics and evaluate their predictive power by performing linear correlation using Pearson correlation coefficient (R). Using Spearman's rank correlation coefficient (ρ) we compute the correlation between the ranked list of models' performance sorted by existing IR metrics and the ranked list obtained using DO and HSA. We demonstrate the generality of our evaluation approach across different retrieval tasks, models, and scoring functions, which we group into three experimental setups: (1) document similarity models where the scoring function computes similarity between two documents, (2) ad-hoc retrieval where the scoring functions represents the relevance score of the document given a query, and (3) a meta-evaluation of ranked lists submitted to ten TREC tracks in the past ten years. With our last experimental setup, we also demonstrate that DO and HSA can be computed using the ranks of the retrieved documents as relevance values.

5.1 Document Similarity Tasks

We first showcase DO and HSA on two document similarity tasks: prior-art patent search [10] and the cross-language IR (CLIR) task of finding document translations [4]. Both tasks use topic models to retrieve similar documents. Experiments were performed with 7 different topic configurations. More specifically, the prior-art patent search task uses LDA with number of topics set to T=50, 100, 200, 500, 1k, 2k and 5k. While on the CLIR task PLTMs were configured with T=100, 200, 300, 400, 500, 700 and 1k. On the patent retrieval task, following the experimental setup of [10], model performance was evaluated using MAP computed over 372 queries and a test collection of 70k patents. In [4], performance of PLTMs was evaluated on a test collection of ∼14k English-Spanish Europarl speeches based on the percentage of true document translation pairs (out of the whole test set) that the model ranked as most similar. This metric is referred to as "percentage at rank one" (P@1).

Table 1 shows the correlation coefficients computed between our evaluation metrics and MAP and P@1. Topic models are typically evaluated intrinsically using perplexity on held-out data, which is considered as a good predictor of the model performance on an extrinsic task. Correlation coefficients were also computed for this metric in order to compare its predictive power with DO and HSA.

Correlation	MAP(LDA)			P@1(PLTM)		
	DO	HSA	Perp.	DO	HSA	Perp.
R	-0.75	0.92	-0.84	0.00	0.71	-0.63
ρ	0.64	0.93	0.89	0.11	0.64	0.25

Table 1: Predicting document similarity model performance using DO, HSA and perplexity: Pearson (R) and Spearman's (ρ) coef. computed over MAP and P@1.

Model	MAP[s]	DO[s]	HSA[s]	Perplexity[s]
LDA T=50	288.1	28.2	28.2	11.1
LDA T=100	224.8	33.0	33.0	27.2
LDA T=200	240.1	47.5	47.6	53.6
LDA T=500	345.9	99.2	99.2	143.1
LDA T=1k	405.6	176.3	176.3	3166.7
LDA T=2k	559.5	333.6	333.6	32930.0
LDA T=5k	1037.4	816.2	816.2	46340.0

Table 2: Absolute computation time for MAP, DO, HSA and perplexity for evaluating patent retrieval.

On both retrieval tasks, HSA exhibits better linear and rank correlation compared to perplexity, while the linear and rank correlation of all three metrics is higher with MAP compared to P@1. In practice this allows us to compute HSA only on the set of query relevant and non-relevant patents and predict the performance of the document similarity model without the need of processing all patents in the collection.

Table 2 shows the absolute computation times for each evaluation metric when computed on the patent search task across different LDA model configurations. Computation time between DO and HSA differs in the second step where computing the volume requires ∼10ms while computing the slope takes ∼60ms. It is evident from this table that, both DO and HSA, are the most efficient metrics to compute compared to MAP and perplexity. On the CLIR task, due to the nature of the evaluation metric, the computation time for MAP, DO and HSA, while being different for each metric, is equal across the different model configurations. While perplexity, as in the case with LDA, grows linearly with the number of topics. Computing DO and HSA on the PLTM model we achieve a relative speed improvement of 5.12 times over MAP.

5.2 Ad-Hoc Retrieval Tasks

We conducted experiments using query sets from 4 previous TREC Web Tracks (2009-2012). Experiments were performed on the ClueWeb09 Category-B with spam filtering (a threshold of 60 using the Waterloo spam scores) collection using the open source retrieval engine Galago[1] with 7 different retrieval models: BM25, BM25RF, RM, SDM and three QL models with various parameter settings. Table 3 and Table 4 show the correlation coefficient values for DO and HSA computed across three IR metrics: MAP, precision at ten (P@10), and NDCG. All evaluation metrics were

[1]http://www.lemurproject.org/galago.php

Web Track	HSA			DO		
	MAP	P@10	NDCG	MAP	P@10	NDCG
2009	0.89	0.91	0.95	0.84	0.88	0.94
2010	0.88	0.71	0.93	0.64	0.60	0.79
2011	0.91	0.93	0.99	0.77	0.81	0.92
2012	0.89	0.87	0.97	0.79	0.79	0.92

Table 3: Evaluating ad-hoc retrieval models using DO and HSA: Pearson (R) coef. computed across MAP, P@10 and NDCG.

Web Track	HSA			DO		
	MAP	P@10	NDCG	MAP	P@10	NDCG
2009	0.82	0.86	0.86	-0.86	-0.89	-0.89
2010	0.93	0.79	0.96	-0.46	-0.68	-0.50
2011	1.00	0.96	1.00	-0.82	-0.79	-0.82
2012	0.71	0.82	0.75	-0.46	-0.61	-0.39

Table 4: Evaluating ad-hoc retrieval models using DO and HSA: Spearman's (ρ) coef. computed across MAP, P@10 and NDCG.

computed using the top 10k retrieved documents and their relevance scores.

Across all evaluation sets, HSA has a high linear and rank correlation with all three IR metrics. While DO has a high linear correlation, its rank correlation is negative, since a large overlap between the distributions of relevant and non-relevant documents is undesirable.

5.3 TREC Ranked Lists

So far in our experiments, we have computed DO and HSA using a relatively large set of query based relevance scores (e.g. 70k patents and 10k ClueWeb documents). However, typical IR systems, across various tasks, are configured to return the top k documents, which is usually a relatively smaller percentage of all the documents in the collection. For example, across different TREC tracks, ranked lists submitted by participants typically consist of the top 1k retrieved documents. Relevance values obtained from such ranked lists are a very small subset on the values across the whole collection. To measure the correlation when DO and HSA are computed over such small sample sets of relevance values we used ranked lists submitted on ten TREC tracks from the past ten years (2004-2013). From each year's TREC we randomly picked a track and for the selected track we randomly chose 7 submitted ranked lists. Unlike previous experimental settings where we computed DO and HSA using the relevance values generated by the retrieval models, in this experimental setup we compute DO and HSA over the normalized values of the document ranks. This is due to the fact that in some instances the relevance scores in the ranked lists are not properly formatted or missing and moreover there is no information on the relevance function used. Table 5 and Table 6 show the linear and rank correlation coefficients computed across various TREC tracks. Results in these tables shown that over all ten tracks HSA has a high linear and rank correlation with MAP and NDCG.

6 Conclusion and Future Work

We presented two evaluation metrics, DO and HSA, that use a novel approach for evaluating retrieval models performance through histogram analysis. We showed that HSA has a high linear and rank correlation with MAP and NDCG while being more efficient to compute. These metrics can predict the performance of document retrieval models without the need for indexing and searching entire collections. In the future, we plan to explore the relationship between the number of query relevant and non-relevant score values used to compute HSA and its effect on the correlation coefficients.

TREC Track	HSA			DO		
	MAP	P@10	NDCG	MAP	P@10	NDCG
Microblog '13	0.98	0.95	0.93	-0.82	-0.80	-0.70
Medical '12	0.90	0.84	0.94	0.30	0.10	0.50
Web '11	0.75	0.86	0.79	0.80	0.51	0.90
Session '10	0.82	0.75	0.84	0.46	0.27	0.68
Chemical '09	0.85	0.95	0.82	0.66	0.84	0.65
Enterprise '08	0.93	-0.14	0.93	0.99	0.21	0.99
Million '07	0.85	0.93	0.88	0.87	0.89	0.94
Terabyte '06	0.97	0.98	1.00	0.94	0.97	0.97
Robust '05	0.75	0.60	0.74	0.82	0.83	0.91
Web '04	0.87	0.63	0.88	-0.41	-0.02	-0.36

Table 5: Evaluating TREC track submissions using DO and HSA: Pearson (R) coef. computed across MAP, P@10, and NDCG.

TREC Track	HSA			DO		
	MAP	P@10	NDCG	MAP	P@10	NDCG
Microblog '13	0.82	0.79	0.86	-0.53	-0.64	-0.51
Medical '12	0.96	0.82	0.96	-0.32	-0.14	-0.32
Web '11	0.71	0.86	0.75	-0.68	-0.43	-0.71
Session '10	0.71	0.86	0.64	-0.54	-0.32	-0.75
Chemical '09	0.82	0.86	0.79	-0.96	-0.89	-0.93
Enterprise '08	0.86	-0.14	0.75	-0.86	-0.04	-0.89
Million '07	0.86	0.68	0.86	-0.89	-0.89	-0.89
Terabyte '06	0.86	0.43	0.86	-0.18	-0.29	-0.12
Robust '05	0.78	0.57	0.79	-0.67	-0.54	-0.68
Web '04	0.68	0.39	0.71	0.00	-0.07	-0.07

Table 6: Evaluating TREC track submissions using DO and HSA: Spearman's (ρ) coef. computed across MAP, P@10, and NDCG.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[2] W. Croft. A model of cluster searching based on classification. *Information Systems*, 5(3):189–195, 1980.

[3] N. Jardine and C. J. van Rijsbergen. The use of hierarchical clustering in information retrieval. *Information Storage and Retrieval*, 7:217–240, 1971.

[4] K. Krstovski and D. A. Smith. Online polylingual topic models for fast document translation detection. In *WMT'11*, pages 252–261, 2013.

[5] D. Mimno, H. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum. Polylingual topic models. In *EMNLP'09*, pages 880–889, 2009.

[6] F. Raiber and O. Kurland. The correlation between cluster hypothesis tests and the effectiveness of cluster-based retrieval. In *SIGIR '14*, pages 1155–1158, 2014.

[7] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *CIKM '07*, pages 623–632, 2007.

[8] C. van Rijsbergen. *Automatic Information Structuring and Retrieval*. PhD thesis, University of Cambridge, 1972.

[9] E. M. Voorhees. The cluster hypothesis revisited. In *SIGIR '85*, pages 188–196, 1985.

[10] X. Xue and W. B. Croft. Transforming patents into prior-art queries. In *SIGIR '09*, pages 808–809, 2009.