

Correlation, Prediction and Ranking of Evaluation Metrics in Information Retrieval

Soumyajit Gupta¹, Mucahid Kutlu^{2*}, Vivek Khetan¹, and Matthew Lease¹

¹ University of Texas at Austin, USA

² TOBB University of Economics and Technology, Ankara, Turkey
smjtgupta@utexas.edu, m.kutlu@etu.edu.tr, vivek.khetank@utexas.edu,
ml@utexas.edu

Abstract. Given limited time and space, IR studies often report few evaluation metrics which must be carefully selected. To inform such selection, we first quantify correlation between 23 popular IR metrics on 8 TREC test collections. Next, we investigate prediction of unreported metrics: given 1 – 3 metrics, we assess the best predictors for 10 others. We show that accurate prediction of MAP, P@10, and RBP can be achieved using 2-3 other metrics. We further explore whether high-cost evaluation measures can be predicted using low-cost measures. We show RBP(p=0.95) at cutoff depth 1000 can be accurately predicted given measures computed at depth 30. Lastly, we present a novel model for ranking evaluation metrics based on covariance, enabling selection of a set of metrics that are most informative and distinctive. A *greedy-forward* approach is guaranteed to yield sub-modular results, while an *iterative-backward* method is empirically found to achieve the best results.

Keywords: Evaluation · Metric · Prediction · Ranking

1 Introduction

Given the importance of assessing IR system accuracy across a range of different search scenarios and user needs, a wide variety of evaluation metrics have been proposed, each providing a different view of system effectiveness [6]. For example, while *precision@10* (P@10) and *reciprocal rank* (RR) are often used to evaluate the quality of the top search results, *mean average precision* (MAP) and *rank-biased precision* (RBP) [32] are often used to measure the quality of search results at greater depth, when recall is more important. Evaluation tools such as `trec.eval` compute many more evaluation metrics than IR researchers typically have time or space to analyze and report. Even for knowledgeable researchers with ample time, it can be challenging to decide which small subset of IR metrics should be reported to best characterize a system’s performance. Since a few metrics cannot fully characterize a system’s performance, information is effectively lost in publication, complicating comparisons to prior art.

* Work began while at Qatar University.

To compute an unreported metric of interest, one strategy is to reproduce prior work. However, this is often difficult (and at times impossible), as the description of a method is often incomplete and even shared source code can be lost over time or difficult or impossible for others to run as libraries change. Sharing system outputs would also enable others to compute any metric of interest, but this is rarely done. While Armstrong *et al.* [2] proposed and deployed a central repository for hosting system runs, their proposal did not achieve broad participation from the IR community and was ultimately abandoned.

Our work is inspired in part by work on biomedical literature mining [23, 8], where acceptance of publications as the most reliable and enduring record of findings has led to a large research community investigating automated extraction of additional insights from the published literature. Similarly, we investigate the viability of predicting unreported evaluation metrics from reported ones. We show accurate prediction of several important metrics is achievable, and we present a novel ranking method to select metrics that are informative and distinctive.

Contributions of our work include:

- We analyze correlation between 23 IR metrics, using more recent collections to complement prior studies. This includes *expected reciprocal rank* (ERR) and RBP using graded relevance; key prior work used only binary relevance.
- We show that accurate prediction of a metric can be achieved using only 2 – 3 other metrics, using a simple linear regression model.
- We show accurate prediction of some high-cost metrics given only low-cost metrics (*e.g.* predicting RBP@1000 given only metrics at depth 30).
- We introduce a novel model for ranking top metrics based on their covariance. This enables us to select the best metrics from clusters with lower time and space complexity than required by prior work. We also provide a theoretical justification for metric ranking which was absent from prior work.
- We share³ our source code, data, and figures to support further studies.

2 Related Work

Correlation between Evaluation Metrics. Tague-Sutcliffe and Blustein [45] study 7 measures on TREC-3 and find R-Prec and AP to be highly correlated. Buckley and Voorhees [10] also find strong correlation using Kendall’s τ on TREC-7. Aslam *et al.* [5] investigate why R-Prec and AP are strongly correlated. Webber *et al.* [51] show that reporting simple metrics such as P@10 with complex metrics such as MAP and DCG is redundant. Baccini *et al.* [7] measure correlations between 130 measures using data from the TREC-(2-8) ad hoc task, grouping them into 7 clusters based on correlation. They use several machine learning tools including Principal Component Analysis (PCA) and Hierarchical Clustering Analysis (HCA) and report the metrics in particular clusters.

Sakai [41] compares 14 graded-level and 10 binary level metrics using three different data sets from NTCIR. Correlation between P(+)-measure, O-measure, and normalized weighted RR shows that they are highly correlated [40]. Correlation between precision, recall, fallout and miss has also been studied [19]. In

³ <https://github.com/smjtgupta/IR-corr-pred-rank>

addition, the relationship between F-measure, break-even point, and 11-point averaged precision has been explored [26]. Another study [46] considers correlation between 5 evaluation measures using TREC Terabyte Track 2006. Jones *et al.* [28] examine disagreement between 14 evaluation metrics including ERR and RBP using TREC-(4-8) ad hoc tasks, and TREC Robust 2005-2006 tracks. However, they use only binary relevance judgments, which makes ERR identical to RR, whereas we consider graded relevance judgments. While their study considered TREC 2006 Robust and Terabyte tracks, we complement this work by considering more recent TREC test collections (*i.e.* Web Tracks 2010-2014), with some additional evaluation measures as well.

Predicting Evaluation Metrics. While Aslam *et al.* [5] propose predicting evaluation measures, they require a corresponding retrieved ranked list as well as another evaluation metric. They conclude that they can accurately infer user-oriented measures (*e.g.* P@10) from system-oriented measures (*e.g.* AP, R-Prec). In contrast, we predict each evaluation measure given only other evaluation measures, without requiring the corresponding ranked lists.

Reducing Evaluation Cost. Lu *et al.* [29] consider risks arising with fixed-depth evaluation of recall/utility-based metrics in terms of providing a fair judgment of the system. They explore the impact of evaluation depth on truncated evaluation metrics and show that for recall-based metrics, depth plays a major role in system comparison. In general, researchers have proposed many methods to reduce the cost of creating test collections: new evaluation measures and statistical methods for incomplete judgments [3, 9, 39, 52, 53], finding the best sample of documents to be judged for each topic [11, 18, 27, 31, 37], topic selection [21, 24, 25, 30], inferring some relevance judgments [4], evaluation without any human judgments [34, 44], crowdsourcing [1, 20], and others. We refer readers to [33] and [42] for detailed review of prior work for low-cost IR evaluation.

Ranking Evaluation Metrics. Selection of IR evaluation metrics from clusters has been studied previously [7, 41, 51]. Our methods incur lower cost than these. We further provide a theoretical basis to rank the metrics using the proposed determinant of covariance criteria, which prior work omitted as an experimental procedure, or by inferring results using existing statistical tools. Our ranking work is most closely related to Sheffield [43], which introduced the idea of unsupervised ranking of features in high-dimensional data using the covariance information of the feature space. This enables selection and ranking of features that are highly informative yet less correlated with one another.

3 Experimental Data

To investigate correlation and prediction of evaluation measures, we use runs and relevance judgments from TREC 2000-2001 & 2010-2014 Web Tracks (WT) and the TREC-2004 Robust Track (RT) [48]. We consider only *ad hoc* retrieval. We calculate 9 evaluation metrics: AP, bpref [9], ERR [12], nDCG, P@K, RBP [32], *recall* (R), RR [50], and R-Prec. We use various cut-off thresholds for the metrics (*e.g.* P@10, R@100). Unless stated, we set the cut-off threshold to 1000. The cut-off threshold for ERR is set to 20 since this was an official measure in WT2014 [17]. RBP uses a parameter p representing the probability of a user

proceeding to the next retrieved page. We test $p = \{0.5, 0.8, 0.95\}$, the values explored by Moffat and Zobel [32]. Using these metrics, we generate two datasets.

Topic-Wise (TW) dataset: We calculate each metric above for each system for each separate topic. We use 10, 20, 100, 1000 cut-off thresholds for AP, nDCG, P@K and R@K. In total, we calculate 23 evaluation metrics.

System-Wise (SW) dataset: We calculate the metrics above (and GMAP as well as MAP) for each system, averaging over all topics in each collection.

4 Correlation of Measures

We begin by computing Pearson correlation between 23 popular IR metrics using 8 TREC test collections. We report correlation of measures for the more difficult TW dataset in order to model score distributions without the damping effect of averaging scores across topics. More specifically, we calculate Pearson correlation between measures across different topics. We make the following observations from the results shown in **Figure 1**.

- R-Prec has high correlation with bpref, MAP and nDCG@100 [45, 10, 5].
- RR is strongly correlated with RBP($p=0.5$), decreasing as its p parameter increases (while RR always stops with the first relevant document, RBP becomes more of a deep-rank metric as p increases). That said, later Figure 2 shows accurate prediction of RBP($p = 0.95$) even with low-cost metrics.
- nDCG@20, one of the official metrics of WT2014, is highly correlated with RBP($p=0.8$), connecting with Park and Zhang’s [36] noting $p=0.78$ is appropriate for modeling web user behavior.
- nDCG is highly correlated with MAP and R-Prec, and its correlation with R@K consistently increases as K increases.
- P@10 ($\rho = 0.97$) and P@20 ($\rho = 0.98$) are most correlated with RBP($p=0.8$) and RBP($p=0.95$), respectively.
- Sakai and Kando [38] report that RBP(0.5) essentially ignores relevant documents below rank 10. Our results are consistent: we see maximum correlation between RBP(0.5) and nDCG@K at $K=10$, decreasing as K increases.
- P@1000 is the least correlated with other metrics, suggesting that it captures a different effectiveness measure of IR systems than other metrics.

While a varying degree of correlation exists between many measures, this should not be interpreted to mean that measures are redundant and trivially exchangeable. Correlated metrics can still correspond to different search scenarios and user needs, and the desire to report effectiveness across a range of potential use cases is challenged by limited time and space for reporting results. In addition, showing two metrics are uncorrelated shows only that each captures a different aspect of system performance, and not whether each aspect is equally important or even relevant to a given evaluation scenario on interest.

5 Prediction of Metrics

In this section, we describe our prediction model and experimental setup, and we report results of our experiments to investigate prediction of evaluation measures. Given the correlation matrix, we can identify the correlated groups of

Test Set	Document Set	#Sys	Topics
WT2000 [22]	WT10g	105	451-500
WT2001 [49]	WT10g	97	501-550
RT2004 [48]	TREC 4&5*	110	301-450, 601-700
WT2010 [14]	ClueWeb'09	55	51-99
WT2011 [13]	ClueWeb'09	62	101-150
WT2012 [15]	ClueWeb'09	48	151-200
WT2013 [16]	ClueWeb'12	59	201-250
WT2014 [17]	ClueWeb'12	30	251-300

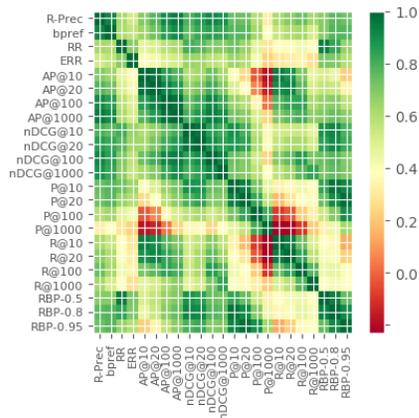


Fig.1: **Left:** TREC collections used. *RT2004 excludes the Congressional Record. **Right:** Pearson Correlation coefficients between 23 Metrics. Deep green entries indicate strong correlation, while red entries indicate low correlation.

metrics. The task of predicting an independent metric m_i using some other dependent metrics m_d under a linear regression model is $m_i = \sum_{k=1}^K \alpha^k m_d^k$.

Because a non-linear relationship could also exist between two correlated metrics, we also tried using a radial basis function (RBF) Support Vector Machine (SVM) for the same prediction. However, the results were very similar, hence not reported. We further discuss this at the end of the section.

Model & Experimental Setup. To predict a system’s missing evaluation measures using reported ones, we build our model using only the evaluation measures of systems as features. We use the SW dataset in our experiments for prediction because studies generally report their average performance over a set of topics, instead of reporting their performance for each topic. Training data combines WT2000-01, RT2004, WT2010-11. Testing is performed separately on WT2012, WT2013, and WT2014, as described below. To evaluate prediction accuracy, we report coefficient of determination R^2 and Kendall’s τ correlation.

Results (Table 1). We investigate the best predictors for 10 metrics: R-Prec, bpref, RR, ERR@20, MAP, GMAP, nDCG, P@10, R@100, RBP(0.5), RBP(0.8) and RBP(0.95). We investigate which K evaluation metric(s) are the best predictors for a particular metric, varying K from 1 – 3. Specifically, in prediction of a particular metric, we try all combinations of size K using the remaining 11 evaluation measures on WT2012 and pick the one that yields the best Kendall’s τ correlation. Then, this combination of metrics is used to predict the respective metric separately for WT2013 and WT2014. Kendall’s τ scores higher than 0.9 are bolded (a traditionally-accepted threshold for correlation [47]).

bpref: We achieve the highest τ correlation and interestingly the worst R^2 using only nDCG on WT2014. This shows that while predicted measures are not accurate, rankings of systems based on predicted scores can be highly correlated with the actual ranking. We observe the same pattern of results in prediction

Table 1: System-wise Prediction of a metric using varying number of metrics $K = [1 - 3]$. Kendall’s τ scores higher than 0.9 are bolded.

Predicted Metric	Independent Variables			WT2012		WT2013		WT2014	
				τ	R^2	τ	R^2	τ	R^2
bpref	nDCG	-	-	0.805	-0.693	0.885	0.079	0.915	-1.174
	nDCG	R-Prec	-	0.872	-0.202	0.850	0.094	0.824	-0.989
	nDCG	R-Prec	R@100	0.906	0.284	0.844	0.645	0.866	0.390
ERR	RR	-	-	0.764	-1.874	0.734	0.293	0.704	-1.004
	RR	RBP(0.8)	-	0.790	-1.809	0.777	0.392	0.714	-0.686
	RR	RBP(0.8)	R@100	0.796	-1.728	0.741	0.478	0.704	-0.473
GMAP	bpref	-	-	0.729	-1.216	0.704	-2.982	0.739	-1.034
	nDCG	RBP(0.5)	-	0.817	0.877	0.777	0.600	0.767	0.818
	nDCG	RBP(0.95)	RR	0.817	0.882	0.748	0.514	0.794	0.854
MAP	R-Prec	-	-	0.885	0.754	0.824	0.667	0.952	0.819
	R-Prec	nDCG	-	0.904	0.894	0.905	0.760	0.958	0.897
	R-Prec	nDCG	RR	0.924	0.916	0.901	0.779	0.947	0.922
nDCG	bpref	-	-	0.805	-2.101	0.885	-0.217	0.915	-2.008
	bpref	GMAP	-	0.803	-0.079	0.809	0.574	0.872	0.024
	bpref	GMAP	RBP(0.95)	0.794	-0.113	0.801	0.556	0.850	-0.032
P@10	RBP(0.8)	-	-	0.884	0.942	0.832	0.895	0.866	0.893
	RBP(0.8)	RBP(0.5)	-	0.941	0.994	0.882	0.966	0.914	0.988
	RBP(0.8)	RBP(0.5)	RR	0.946	0.994	0.885	0.968	0.914	0.987
RBP(0.95)	R-Prec	-	-	0.824	0.346	0.651	-0.786	0.607	-2.401
	bpref	P@10	-	0.911	0.952	0.718	0.873	0.728	0.591
	bpref	P@10	RBP(0.8)	0.911	0.967	0.720	0.868	0.744	0.639
R-Prec	R@100	-	-	0.899	0.708	0.871	0.624	0.935	0.019
	R@100	RBP(0.95)	-	0.909	0.952	0.820	0.882	0.820	0.759
	R@100	RBP(0.95)	GMAP	0.924	0.970	0.833	0.914	0.841	0.825
RR	RBP(0.5)	-	-	0.782	0.904	0.806	0.927	0.810	0.878
	RBP(0.5)	RBP(0.8)	-	0.869	0.918	0.809	0.919	0.820	0.942
	RBP(0.5)	RBP(0.8)	ERR	0.876	0.437	0.818	0.924	0.915	0.824
R@100	R-Prec	-	-	0.899	0.423	0.871	0.232	0.935	-1.075
	R-Prec	GMAP	-	0.899	0.433	0.871	0.238	0.940	-1.077
	R-Prec	RR	ERR	0.881	-0.104	0.823	0.355	0.935	-1.187

of RR on WT2012 and WT2014, R-prec on WT2013 and WT2014, R@100 on WT2013, and nDCG in all three test collections.

GMAP & ERR: Both seem to be the most challenging measures to predict because we could never reach $\tau = 0.9$ correlation in any of the prediction cases of these two measures. Initially, R^2 scores for ERR consistently increase in all three test collections as we use more evaluation measures for prediction, suggesting that we can achieve higher prediction accuracy using more independent variables.

MAP: We can predict MAP with very high prediction accuracy and achieve higher than $\tau = 0.9$ correlation in all three test collections using R-Prec and nDCG as predictors. When we use RR as the third predictor, R^2 increases in all cases and τ correlation slightly increases on average (0.924 vs. 0.922).

nDCG: Interestingly, we achieve the highest τ correlations using only bpref; τ decreases as more evaluation measures are used as independent variables. Even though we reach high τ correlations for some cases (e.g. 0.915 τ on WT2014 using only bpref), nDCG seems to be one of the hardest measures to predict.

P@10: Using RBP(0.5) and RBP(0.8), which are both highly correlated measures with P@10, we are able to achieve very high τ correlation and R^2 in all three test collections ($\tau = 0.912$ and $R^2 = 0.983$ on average). We reach nearly perfect prediction accuracy ($R^2 = 0.994$) on WT2012.

RBP(0.95): Compared to RBP(0.5) and RBP(0.8), we achieve noticeably lower prediction performance, especially on WT2013 and WT2014. On WT2012, which is used as the development set in our experimental setup, we reach high prediction accuracy when we use 2-3 independent variables.

R-Prec, RR and R@100: In predicting these three measures, while we reach high prediction accuracy in many cases, there is no independent variable group yielding high prediction performance on all three test collections.

Overall, we achieve high prediction accuracy for MAP, P@10, RBP(0.5) and RBP(0.8) on all test collections. RR and RBP(0.8) are the most frequently selected independent variables (10 and 9 times, respectively). Generally, using a single measure is not sufficient to reach $\tau = 0.9$ correlation. We achieve very high prediction accuracy using only 2 measures for many scenarios.

Note R^2 is sometimes negative, whereas theoretically the value of the coefficient of determination should lie in $[0, 1]$. R^2 compares the fit of the chosen model with a horizontal straight line (the null hypothesis); if the chosen model fits worse than a horizontal line, then R^2 will be negative⁴.

Although the empirical results might suggest that the relationship between metrics are linear because non-linear SVMs did not improve results much, the negative values of R^2 contradict this observation, as the linear model clearly did not fit well. Specifically, we tried out RBF SVM's using different kernel sizes of $\{0.5, 1, 2, 5\}$, without significant result changes as compared to linear regression. Additional non-linear models could be further explored in future work.

5.1 Predicting High-Cost Metrics using Low-Cost Metrics

In some cases, one may wish to predict a “high-cost” evaluation metric (i.e., requiring relevance judging to some significant evaluation depth D) when only “low-cost” evaluation metrics have been reported. Here, we consider prediction of Precision, MAP, nDCG, and RBP [32] for high-cost $D = 100$ or $D = 1000$ given a set of low-cost metric scores ($D \in \{10, 20, \dots, 50\}$): precision, bpref, ERR, infAP[52], MAP, nDCG and RBP. We include bpref and infAP given their support for evaluating systems with incomplete relevance judgments. For RBP we use $p = 0.95$. For each depth D , we calculate the powerset of the 7 measures

⁴ <https://stats.stackexchange.com/questions/12900/when-is-r-squared-negative>

mentioned above (excluding the empty set \emptyset). We then find which elements of the powerset are the best predictors of the high-cost measures on WT2012. The set of low-cost measures that yields the maximum τ score for a particular high-cost measure for WT2012 is then used for predicting the respective measure on WT2013 and WT2014. We repeat this process for each evaluation depth $D \in \{10, 20, \dots, 50\}$ to assess prediction accuracy as a function of D .

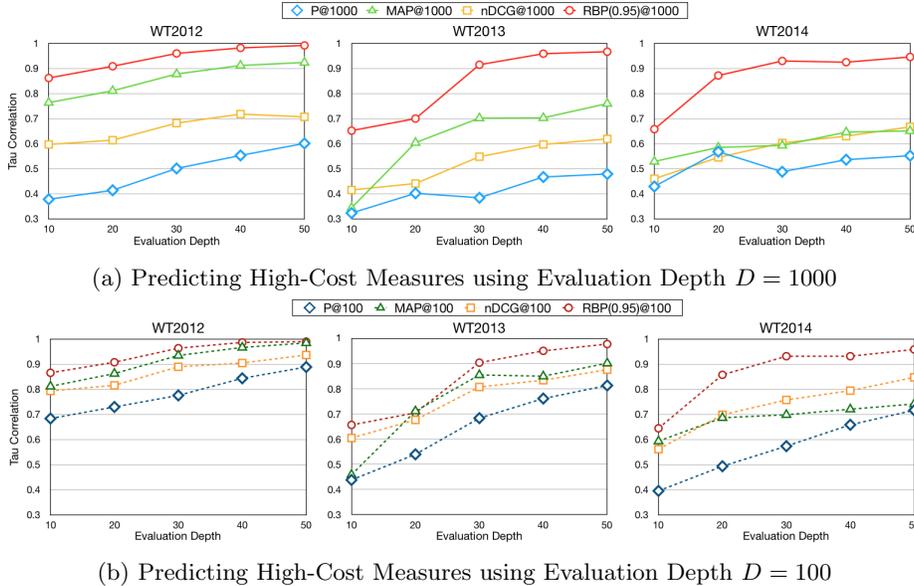


Fig. 2: Linear regression prediction of high-cost metrics using low-cost metrics

Figure 2 presents results. For depth 1000 (Figure 2a), we achieve $\tau > 0.9$ correlation and $R^2 > 0.98$ for RBP in all cases when $D \geq 30$. While we are able to reach $\tau = 0.9$ correlation for MAP on WT2012, prediction of P@1000 and nDCG@1000 measures performs poorly and never reaches a high τ correlation. As expected, the performance of prediction increases when evaluation depth of high-cost measures are decreased to 100 (Figure 2a vs. Figure 2b).

Overall, RBP seems the most predictable from low-cost metrics while precision is the least. Intuitively, MAP, nDCG and RBP give more weight to documents at higher ranks, which are also evaluated by the low-cost measures, while precision@ D does not consider document ranks within the evaluation depth D .

6 Ranking Evaluation Metrics

Given a particular search scenario or user need envisioned, one typically selects appropriate evaluation metrics for that scenario. However, this does not necessarily consider correlation between metrics, or which metrics may interest other researchers engaged in reproducibility studies, benchmarking, or extensions. In this section, we consider how one might select the most informative and distinctive set of metrics to report in general, without consideration of specific user needs or other constraints driving selection of certain metrics.

We thus motivate a proper metric ranking criteria to efficiently compute the top L metrics to report amongst the S metrics available, i.e., a set that best captures diverse aspects of system performance with minimal correlation across metrics. Our approach is motivated by Sheffield [43], who introduced the idea of unsupervised ranking of features in high-dimensional data using the covariance information in the feature space. This method enables selection and ranking of features that are highly informative and less correlated with each other.

$$\Omega^* = \arg \max_{\Omega: |\Omega| \leq L} \det(\Sigma(\Omega)) \quad (1)$$

Here we are trying to find the subset Ω^* of cardinality L such that the covariance matrix Σ sampled from the rows of and columns of the entries of Ω^* will have the maximum determinant value, among all possible sub-determinant of size $L \times L$. The general problem is NP-Complete [35]. Sheffield provided a backward rejection scheme that throws out elements of the active subset Ω until it is left with L elements. However, this approach suffers from large cost in both time and space (**Table 2**), due to computing multiple determinant values over iterations.

We propose two novel methods for ranking metrics: an *iterative-backward* method (Section 6.1), which we find to yield the best empirical results, and a *greedy-forward* approach (Section 6.2) guaranteed to yield sub-modular results. Both offer lower time and space complexity vs. prior clustering work [7, 51, 41].

Table 2: Complexity of Ranking Algorithms.

Algorithm	Time Complexity	Space Complexity
Sheffield [43]	$O(LS^4)$	$O(S^3)$
Iterative-Backward	$O(LS^3)$	$O(S^2)$
Greedy-Forward	$O(LS^2)$	$O(S^2)$

6.1 Iterative-Backward (IB) Method

IB (**Algorithm 1**) starts with a full set of metrics and iteratively prunes away the less informative ones. Instead of computing all the sub-determinants of one less size at each iteration, we use the adjugate of the matrix to compute them in a single pass. This reduces the run-time by a factor of S and completely eliminates the need for additional memory. Also, since we are not interested in the actual values of the sub-determinants, but just the maximum, we can approximate $\Sigma_{adj} = \Sigma^{-1} \det(\Sigma) \approx \Sigma^{-1}$ since $\det(\Sigma)$ is a scalar multiple.

Once the adjugate Σ_{adj} is computed, we look at its diagonal entries for values of the sub-determinants of size one less. The index of the maximum entry is found in Step 7 and it is subsequently removed from the active set. Step 9 ensures that adjustments made to rest of the matrix prevents the selection of correlated features by scaling down their values appropriately. We do not have any theoretical guarantees for optimality of this IB feature elimination strategy, but our empirical experiments found that it always returns the optimal set.

Algorithm 1 Iterative-Backward Method

- 1: **Input:** $\Sigma \in \mathbb{R}^{S \times S}$, L : number of channels to be retained
 - 2: Set counter $k = S$ and $\Omega = \{1 : S\}$ as the active set
 - 3: **while** $k > L$ **do**
 - 4: $\Sigma_{adj} \approx \Sigma^{-1}$ ▷ Approximate adjugate
 - 5: $i^* \leftarrow \arg \max_{i \in \Omega} \text{diag}(\Sigma_{adj}(i))$ ▷ Index to be removed
 - 6: $\Omega^{k+1} \leftarrow \Omega^k - \{i^*\}$ ▷ Augment the active set
 - 7: $\sigma_{ij} \leftarrow \sigma_{ij} - \sigma_{ii^*} \sigma_{i^*j} / \sigma_{i^*i^*}$, $\forall i, j \in \Omega$ ▷ Update covariance
 - 8: $k \leftarrow k - 1$ ▷ Decrement counter
 - 9: **Output:** Retained features Ω
-

6.2 Greedy-Forward (GF) Method

GF (**Algorithm 2**) iteratively selects the most informative features to add one-by-one. Instead of starting with the full set, we initialize the active set as empty, then grow the active set by greedily choosing the best feature at each iteration, with lower run-time cost than its backward counterpart. The index of the maximum entry is found in Step 6 and is subsequently added to the active set. Step 8 ensures that the adjustments made to the other entries of the matrix prevents the selection of correlated features by scaling down their values appropriately.

Algorithm 2 Greedy-Forward Method

- 1: **Input:** $\Sigma \in \mathbb{R}^{S \times S}$, L : number of channels to be selected
 - 2: Set counter $k = 0$ and $\Omega = \emptyset$ as the active set
 - 3: **while** $k < L$ **do**
 - 4: $i^* \leftarrow \arg \max_{i \notin \Omega} \sum_{j \notin \Omega} \sigma_{ij}^2 / \sigma_{ii}$ ▷ Index to be added
 - 5: $\Omega^{k+1} \leftarrow \Omega^k \cup \{i^*\}$ ▷ Augment the active set
 - 6: $\sigma_{ij} \leftarrow \sigma_{ij} - \sigma_{ii^*} \sigma_{i^*j} / \sigma_{i^*i^*}$, $\forall i, j \notin \Omega$ ▷ Update covariance
 - 7: $k \leftarrow k + 1$ ▷ Increment counter
 - 8: **Output:** Selected features Ω
-

A feature of this greedy strategy is that it is guaranteed to provide sub-modular results. The solution has a constant factor approximation bound of $(1 - 1/e)$, *i.e.* even under worst case scenario, the approximated solution is no worse than 63% of the optimal solution.

Proof. For any positive definite matrix Σ and for any $i \notin \Omega$:

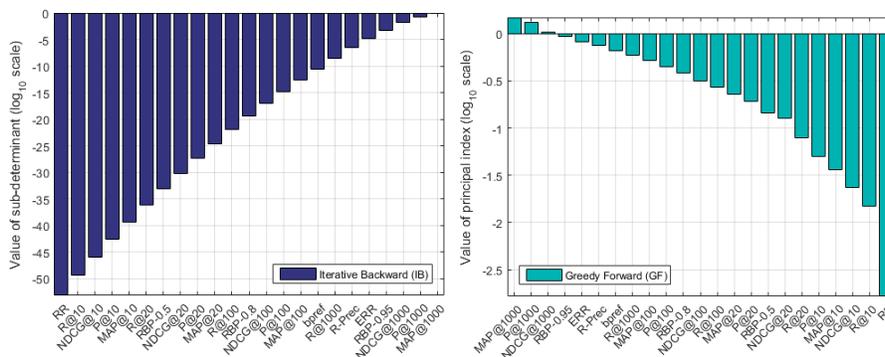
$$f_{\Sigma}(\Omega \cup \{i\}) = f_{\Sigma}(\Omega) + \frac{\sum_{j \notin \Omega} \sigma_{ij}^2}{\sigma_{ii}}$$

where σ_{ij} are the elements of $\Sigma(\notin \Omega)$ *i.e.* the elements of Σ not indexed by the entries of the active set Ω , and f_{Σ} is the determinant function $\det(\Sigma)$.

Table 3: Metrics are ranked by each algorithm as numbered below.

IB	1. MAP@1000	2. P@1000	3. NDCG@1000	4. RBP-0.95	5. ERR
	6. R-Prec	7. R@1000	8. bpref	9. MAP@100	10. P@100
	11. NDCG@100	12. RBP-0.8	13. R@100	14. MAP@20	15. P@20
	16. NDCG@20	17. RBP-0.5	18. R@20	19. MAP@10	20. P@10
	21. NDCG@10	22. R@10	23. RR	-	-
GF	1. MAP@1000	2. P@1000	3. NDCG@1000	4. RBP-0.95	5. ERR
	6. R-Prec	7. bpref	8. R@1000	9. MAP@100	10. P@100
	11. RBP-0.8	12. NDCG@100	13. R@100	14. MAP@20	15. P@20
	16. RBP-0.5	17. NDCG@20	18. R@20	19. P@10	20. MAP@10
	21. NDCG@10	22. R@10	23. RR	-	-

Hence, we have $f_{\Sigma}(\Omega) \geq f_{\Sigma}(\Omega')$ for any $\Omega' \subseteq \Omega$. This shows that $f_{\Sigma}(\Omega)$ is a monotonically non-increasing and sub-modular function, so that the simple greedy selection algorithm yields an $(1 - 1/e)$ -approximation. \square



(a) Iterative Backward. Left-to-Right: metrics discarded
 (b) Greedy Forward. Left-to-Right: metrics included

Fig. 3: Metrics ranked by the strategies. Positive values on the GF plot shows values computed by the greedy criteria were positive for the first three selections.

6.3 Results

Running the Iterative-Backward (IB) and Greedy-Forward (GF) methods on the 23 metrics shown in Figure 1 yields the results shown in **Table 3**. The top six metrics are the same (in order) for both IB and GF: MAP@1000, P@1000, NDCG@1000, RBP($p - 0.95$), ERR, and R-Prec. They then diverge on whether R@1000 (IB) or bpref (GF) should be at rank 7. GF makes some constrained choices that lead to swapping of ranks among some metrics (bpref and R@1000, RBP-0.8 and NDCG@100, RBP-0.5 and NDCG@20, P@10 and MAP@10). However, due to the sub-modular nature of the greedy method, the approximated solution is guaranteed to incur no more than 27% error compared to the true solution. Both methods assigned lowest rankings to NDCG@10, R@10, and RR.

Figure 3a shows the metric deleted from the active set at each iteration of the IB strategy. As irrelevant metrics are removed by the maximum determinant criteria, the value of the sub-determinant increases at each iteration and is empirically maximum among all sub-determinants of that size. **Figure 3b** shows the metric added to the active set at each iteration by the GF strategy. Here we add a metric that maximizes the greedy selection criteria. We can see that over iterations the criteria value steadily decreases due to proper updates made.

The ranking pattern shows that the relevant, highly informative and less correlated metrics (MAP@1000, P@1000, nDCG@1000, RBP-0.95) are clearly ranked at the top. While ERR, R-Prec, bpref, and R@1000 may not be as informative as the higher ranked metrics, they still rank highly because the average information provided by other measures (*e.g.* MAP@100, nDCG@100 *etc.*) decreases even more in presence of already selected features MAP@1000, nDCG@1000 *etc.* Intuitively, even if two metrics are informative, both should not be ranked highly if there exists strong correlation between them.

Relation to prior work. Our findings are consistent with prior work in showing that we can select best metrics from clusters, although we report lower algorithmic (time and space) cost procedures than prior work [7, 51, 41]. Webber *et al.* [51] consider only the diagonal entries of the covariance; we consider the entire matrix since off-diagonal entries indicate cross-correlation. Baccini *et al.* [7] use both Hierarchical Clustering (HCA) of metrics which lacks ranking, does not scale well, and is slow, having runtime $O(S^3)$ and memory $O(S^2)$ with large constants. Their results are also somewhat subjective and subject to outliers, while our ranking is computationally effective and theoretically justified.

7 Conclusion

In this work, we explored strategies for selecting IR metrics to report. We first quantified correlation between 23 popular IR metrics on 8 TREC test collections. Next, we described metric prediction and showed that accurate prediction of MAP, P@10, and RBP can be achieved using 2-3 other metrics. We further investigated accurate prediction of some high-cost evaluation measures using low-cost measures, showing RBP(p=0.95) at cutoff depth 1000 could be accurately predicted given other metrics computed at only depth 30. Finally, we presented a novel model for ranking evaluation metrics based on covariance, enabling selection of a set of metrics that are most informative and distinctive.

We proposed two methods for ranking metrics, both providing lower time and space complexity than prior work. Among the 23 metrics considered, we predicted MAP@1000, P@1000, nDCG@1000 and RBP(p=0.95) as the top four metrics, consistent with prior research. Although the timing difference is negligible for 23 metrics, there is a speed-accuracy trade-off, once the problem dimension increases. Our method provides a theoretically-justified, practical approach which can be generally applied to identify informative and distinctive evaluation metrics to measure and report, and applicable to a variety of IR ranking tasks.

Acknowledgements. This work was made possible by NPRP grant# NPRP 7-1313-1-245 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

1. Alonso, O., Mizzaro, S.: Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In: Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation. vol. 15, p. 16 (2009)
2. Armstrong, T.G., Moffat, A., Webber, W., Zobel, J.: Improvements that don't add up: ad-hoc retrieval results since 1998. In: Proceedings of the 18th ACM conference on Information and knowledge management. pp. 601–610. ACM (2009)
3. Aslam, J.A., Pavlu, V., Yilmaz, E.: 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. pp. 541–548. ACM (2006)
4. Aslam, J.A., Yilmaz, E.: Inferring document relevance from incomplete information. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. pp. 633–642. ACM (2007)
5. Aslam, J.A., Yilmaz, E., Pavlu, V.: A geometric interpretation of r-precision and its correlation with average precision. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 573–574. ACM (2005)
6. Aslam, J.A., Yilmaz, E., Pavlu, V.: The maximum entropy method for analyzing retrieval measures. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 27–34. ACM (2005)
7. Baccini, A., Déjean, S., Lafage, L., Mothe, J.: How many performance measures to evaluate Information Retrieval Systems? *Knowledge and Information Systems* **30**(3), 693 (2012)
8. de Bruijn, L., Martin, J.: Literature mining in molecular biology. In: Proceedings of the EFMI Workshop on Natural Language Processing in Biomedical Applications. pp. 1–5 (2002)
9. Buckley, C., Voorhees, E.M.: Retrieval evaluation with incomplete information. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 25–32. ACM (2004)
10. Buckley, C., Voorhees, E.M.: Retrieval system evaluation. TREC: Experiment and evaluation in information retrieval pp. 53–75 (2005)
11. Carterette, B., Allan, J., Sitaraman, R.: Minimal test collections for retrieval evaluation. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 268–275. ACM (2006)
12. Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: Proceedings of the 18th ACM conference on Information and knowledge management. pp. 621–630. ACM (2009)
13. Clarke, C., Craswell, N.: Overview of the TREC 2011 Web Track. In: TREC (2011)
14. Clarke, C., Craswell, N., Soboroff, I., Cormack, G.: Overview of the TREC 2010 Web Track. In: TREC (2010)
15. Clarke, C., Craswell, N., Voorhees, E.M.: Overview of the TREC 2012 Web Track. In: TREC (2012)
16. Collins-Thompson, K., Bennett, P., Clarke, C., Voorhees, E.M.: TREC 2013 Web Track Overview. In: TREC (2013)
17. Collins-Thompson, K., Macdonald, C., Bennett, P., Voorhees, E.M.: TREC 2014 Web Track Overview. In: TREC (2014)

18. Cormack, G.V., Palmer, C.R., Clarke, C.L.: Efficient construction of large test collections. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 282–289. ACM (1998)
19. Egghe, L.: The measures precision, recall, fallout and miss as a function of the number of retrieved documents and their mutual interrelations. *Information Processing & Management* **44**(2), 856 – 876 (2008), evaluating Exploratory Search Systems Digital Libraries in the Context of Users Broader Activities
20. Grady, C., Lease, M.: 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. pp. 172–179. Association for Computational Linguistics (2010)
21. Guiver, J., Mizzaro, S., Robertson, S.: A few good topics: Experiments in topic set reduction for retrieval evaluation. *ACM Transactions on Information Systems (TOIS)* **27**(4), 21 (2009)
22. Hawking, D.: Overview of the TREC-9 Web Track. In: TREC (2000)
23. Hirschman, L., Park, J.C., Tsujii, J., Wong, L., Wu, C.H.: Accomplishments and challenges in literature data mining for biology. *Bioinformatics* **18**(12), 1553–1561 (2002)
24. Hosseini, M., Cox, I.J., Milic-Frayling, N., Shokouhi, M., Yilmaz, E.: An uncertainty-aware query selection model for evaluation of IR systems. In: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. pp. 901–910. ACM (2012)
25. Hosseini, M., Cox, I.J., Milic-Frayling, N., Vinay, V., Sweeting, T.: Selecting a subset of queries for acquisition of further relevance judgements. In: Conference on the Theory of Information Retrieval. pp. 113–124. Springer (2011)
26. Ishioka, T.: Evaluation of criteria for information retrieval. In: Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on. pp. 425–431. IEEE (2003)
27. Jones, K.S., van Rijsbergen, C.J.: Report on the need for and provision of an “ideal” information retrieval test collection (british library research and development report no. 5266) p. 43 (1975)
28. Jones, T., Thomas, P., Scholer, F., Sanderson, M.: Features of disagreement between retrieval effectiveness measures. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 847–850. ACM (2015)
29. Lu, X., Moffat, A., Culpepper, J.S.: The effect of pooling and evaluation depth on IR metrics. *Information Retrieval Journal* **19**(4), 416–445 (2016)
30. Mizzaro, S., Robertson, S.: Hits hits TREC: exploring IR evaluation results with network analysis. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 479–486. ACM (2007)
31. Moffat, A., Webber, W., Zobel, J.: Strategic system comparisons via targeted relevance judgments. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 375–382. ACM (2007)
32. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)* **27**(1), 2 (2008)
33. Moghadasi, S.I., Ravana, S.D., Raman, S.N.: Low-cost evaluation techniques for information retrieval systems: A review. *Journal of Informetrics* **7**(2), 301–312 (2013)
34. Nuray, R., Can, F.: Automatic ranking of information retrieval systems using data fusion. *Information Processing & Management* **42**(3), 595–614 (2006)

35. Papadimitriou, C.H.: The largest subdeterminant of a matrix. *Bull. Math. Soc. Greece* **15**, 96–105 (1984)
36. Park, L., Zhang, Y.: On the distribution of user persistence for rank-biased precision. In: *Proceedings of the 12th Australasian document computing symposium*. pp. 17–24 (2007)
37. Pavlu, V., Aslam, J.: A practical sampling strategy for efficient retrieval evaluation. Tech. rep., College of Computer and Information Science, Northeastern University (2007)
38. Sakai, Tetsuya and Kando, N.: On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval* **11**(5), 447–470 (Oct 2008)
39. Sakai, T.: Alternatives to bpref. In: *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. pp. 71–78. ACM (2007)
40. Sakai, T.: On the properties of evaluation metrics for finding one highly relevant document. *Information and Media Technologies* **2**(4), 1163–1180 (2007)
41. Sakai, T.: On the reliability of information retrieval metrics based on graded relevance. *Information processing & management* **43**(2), 531–548 (2007)
42. Sanderson, M.: *Test collection based evaluation of information retrieval systems*. Now Publishers Inc (2010)
43. Sheffield, C.: Selecting band combinations from multispectral data. *Photogrammetric Engineering and Remote Sensing* **51**, 681–687 (1985)
44. Soboroff, I., Nicholas, C., Cahan, P.: Ranking retrieval systems without relevance judgments. In: *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. pp. 66–73. ACM (2001)
45. Tague-Sutcliffe, J., Blustein, J.: Overview of TREC 2001. In: *Proceedings of the third text retrieval conference (TREC-3)*. pp. 385–398 (1995)
46. Thom, J., Scholer, F.: A comparison of evaluation measures given how users perform on search tasks. In: *ADCS2007 Australasian Document Computing Symposium*. RMIT University, School of Computer Science and Information Technology (2007)
47. Voorhees, E.M.: Variations in relevance judgments and the measurement of retrieval effectiveness. *Information processing & management* **36**(5), 697–716 (2000)
48. Voorhees, E.M.: Overview of the TREC 2004 Robust Track. In: *TREC*. vol. 4 (2004)
49. Voorhees, E.M., Harman, D.: Overview of TREC 2001. In: *TREC* (2001)
50. Voorhees, E.M., Tice, D.M.: The TREC-8 Question Answering Track Evaluation. In: *TREC*. vol. 1999, p. 82 (1999)
51. Webber, W., Moffat, A., Zobel, J., Sakai, T.: Precision-at-ten considered redundant. In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. pp. 695–696. ACM (2008)
52. Yilmaz, E., Aslam, J.A.: Estimating average precision with incomplete and imperfect judgments. In: *Proceedings of the 15th ACM international conference on Information and knowledge management*. pp. 102–111. ACM (2006)
53. Yilmaz, E., Aslam, J.A.: Estimating average precision when judgments are incomplete. *Knowledge and Information Systems* **16**(2), 173–211 (2008)