

Correlation, Prediction and Ranking of Evaluation Metrics in Information Retrieval

Soumyajit Gupta, Mucahid Kutlu, Vivek Khetan, and Matthew Lease



TOBB
UNIVERSITY OF
ECONOMICS AND TECHNOLOGY

ECIR 2019, Cologne, Germany,



- ▶ More than 100 metrics
- ▶ Limited time and space to report all

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

$$\text{AveP} = \frac{\sum_{k=1}^n (P(k) \times \text{rel}(k))}{\text{number of relevant documents}}$$

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}$$

$$F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{(\text{precision} + \text{recall})}$$

$$\text{nDCG}_p = \frac{DCG_p}{IDCG_p}$$



Which ones should we report?

	MAP	P@10	P@30	NDCG
QL	0.3043	0.5560	0.4980	0.5475
SRM	0.3110	0.5700	0.5060	0.5502
RQLM	0.3161 [‡]	0.5960 [‡]	0.5120	0.5601 [‡]
RW+RQLM	0.3132 [†]	0.5840 [‡]	0.5067	0.5579 [‡]
RM	0.3540 [‡]	0.5800 [‡]	0.5440 [‡]	0.5797 [‡]
RW+RQLM+RM	0.3617 ^{‡*}	0.6080 ^{‡*}	0.5580 [‡]	0.5866 ^{‡*}

Table 3: Retrieval performance on the TREC 2005 Terabyte Track queries (test).

Table 1: Top results for TREC-TB 2005

Run	p@20	CPUs	Time per query (ms)
MU05TBy3	0.5550	8	24
uwmtEwteD10	0.3900	2	27
MU05TBy1	0.5620	8	42
zetdist	0.5300	8	58
pisaEff4	0.3420	23	143

Taken from two different papers



If paper A reports metric X and paper B reports metric Y on the same collection, how can I know which one is better?

- ▶ Run them again on the collection
 - ▶ Do they share their code?
- ▶ Implement the methods
 - ▶ Is it well explained in the paper?
- ▶ Check if there is any common baseline used against and compare indirectly?

- ▶ Wouldn't be nice to predict a system performance based on metric X using its performance on other metrics as features?
- ▶ Here is the general idea
 - ▶ Build a classifier using only metric scores as features
 - ▶ Predict the unknown metric using the known ones
 - ▶ Compare systems based on predicted score with some confidence value

- ▶ Going back to our example:
 - ▶ Predict A's P@20 score using its MAP, P@10, P@30 and NDCG score
 - ▶ Compare A's predicted P@20 with B's actual P@20

	MAP	P@10	P@30	NDCG
QL	0.3043	0.5560	0.4980	0.5475
SRM	0.3110	0.5700	0.5060	0.5502
RQLM	0.3161 [‡]	0.5960 [‡]	0.5120	0.5601 [‡]
RW+RQLM	0.3132 [†]	0.5840 [‡]	0.5067	0.5579 [‡]
RM	0.3540 [‡]	0.5800 [‡]	0.5440 [‡]	0.5797 [‡]
RW+RQLM+RM	0.3617 ^{‡*}	0.6080 ^{‡*}	0.5580 [‡]	0.5866 ^{‡*}

Table 3: Retrieval performance on the TREC 2005 Terabyte Track queries (test).

Table 1: Top results for TREC-TB 2005

Run	p@20	CPUs	Time per query (ms)
MU05TBy3	0.5550	8	24
uwmtEwteD10	0.3900	2	27
MU05TBy1	0.5620	8	42
zetdist	0.5300	8	58
pisaEff4	0.3420	23	143

- ▶ **Goal:** investigate which K evaluation metric(s) are the best predictors for a particular metric
- ▶ **Training data:** System average scores over topics in WT2000-01, RT2004, WT2010-11 collections.
- ▶ **Test data:** WT2012, WT2013, and WT2014
- ▶ **Learning algorithms:** Linear Regression and SVM
- ▶ **Approach:**
 - ▶ For a particular metric, we try all combinations of size K using other evaluation metrics on WT2012
 - ▶ Pick the highest and apply it on WT2013 and WT2014

Prediction Results

Predicted Metric	Independent Variables			WT2012		WT2013		WT2014	
				τ	R^2	τ	R^2	τ	R^2
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



Which metrics should I report?

- ▶ Metrics do have correlation
 - ▶ Why do we need to report correlated ones?
- ▶ **Goal:** Report the most informative set of metrics
 - ▶ NP-Hard problem
- ▶ **Iterative Backward Strategy:**
 - ▶ Start with a full set of covariance of metrics
 - ▶ Iteratively prune less informative ones
 - ▶ Remove the one that yields maximum entropy without it
- ▶ **Greedy Forward Strategy**
 - ▶ Start with a empty set
 - ▶ Greedily add most informative ones
 - ▶ Pick the metric that is most correlated with all the remaining ones

Metrics ranked by each algorithm

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	-	-

- ▶ Quantified correlation between 23 popular IR metrics on 8 TREC test collections
- ▶ Showed that accurate prediction of MAP, P@10, and RBP can be achieved using 2-3 other metrics
- ▶ Presented a model for ranking evaluation metrics based on covariance, enabling selection of a set of metrics that are most informative and distinctive.

Thank you!

This work was funded by the Qatar National Research Fund, a member of Qatar Foundation.

