ECIR 2019 Cologne

Best Student Paper award

Matthew Lease

Correlation, Prediction and Ranking of Evaluation Metrics in Information Retrieval

Soumyajit Gupta[1], Mucahid Kutlu[2⋆], Vivek Khetan[1], and Matthew Lease[1]

[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.

Soumyajit Gupta, Mucahid Kutlu, Vivek Khetan, and Matthew Lease. Correlation, Prediction and Ranking of Evaluation Metrics in Information Retrieval. In Proceedings of the 41st European Conference on Information Retrieval (ECIR), pages 636-651, 2019.
https://doi.org/10.1007/978-3-030-15712-8_41