New Metrics to Encourage Innovation and Diversity in Information Retrieval Approaches

Mehmet Deniz Türkmen[1](✉), Matthew Lease[2], and Mucahid Kutlu[1]

[1] Department of Computer Engineering, TOBB University of Economics and Technology, Ankara, Turkey
{m.turkmen,m.kutlu}@etu.edu.tr

[2] School of Information, University of Texas at Austin, Austin, TX, USA
ml@utexas.edu

Abstract. In evaluation campaigns, participants often explore variations of popular, state-of-the-art baselines as a low-risk strategy to achieve competitive results. While effective, this can lead to local "hill climbing" rather than a more radical and innovative departure from standard methods. Moreover, if many participants build on similar baselines, the overall diversity of approaches considered may be limited. In this work, we propose a new class of IR evaluation metrics intended to promote greater diversity of approaches in evaluation campaigns. Whereas traditional IR metrics focus on user experience, our two "innovation" metrics instead reward exploration of more divergent, higher-risk strategies finding relevant documents missed by other systems. Experiments on four TREC collections show that our metrics do change system rankings by rewarding systems that find such rare, relevant documents. This result is further supported by a controlled, synthetic data experiment, and a qualitative analysis. In addition, we show that our metrics achieve higher evaluation stability and discriminative power than the standard metrics we modify. To support reproducibility, we share our source code.

Keywords: Evaluation · Metrics · Information retrieval

1 Introduction

Researchers must balance risk vs. reward in prioritizing methods to investigate. Higher-risk methods offer the potential for a larger impact, but with a greater chance of sub-baseline performance. In contrast, lower-risk methods are more likely to yield improvement but may be incremental. A popular strategy to straddle such risk is to investigate variants of popular state-of-the-art models (e.g., use of pre-trained language models, such as GPT-3 [1]). While this represents a low-risk strategy to achieve competitive results, it can lead to local "hill climbing" rather than exploring higher-risk, more radical departures from current state-of-the-art methods. Moreover, if many researchers build on similar baselines, this can limit the overall diversity of approaches being explored in the field.

J. Kamps et al. (Eds.): ECIR 2023, LNCS 13981, pp. 239–254, 2023.
https://doi.org/10.1007/978-3-031-28238-6_16

In this work, we investigate a novel class of "innovation" evaluation metrics that seek to promote greater diversity among participant methods in evaluation campaigns. Such community benchmarking and evaluation campaigns play an important role in assessing the current state-of-the-art and promoting continuing advancements. For participants, evaluation campaigns provide a valuable testing ground for novel methods, and evaluation metrics chosen by a campaign can galvanize community attention on particular aspects of system performance. Evaluation campaign metrics thus help to steer a field.

Whereas traditional IR metrics focus on ranking quality for the user, our innovation metrics instead reward exploration of more divergent, higher-risk ranking methods that find relevant documents missed by most other systems. The key intuition is that a system finding relevant documents missed by other systems must differ in approach. Specifically, we modify standard Precision@K and Average Precision metrics to reward retrieval of such "rare" relevant documents missed by other systems. A simple mixture-weight parameter controls the relative weight placed on such rarity, and setting this to zero reverts to the original metric. As such, evaluation campaigns adopting our metrics could easily control the extent to which they want to reward diversity of approaches vs. more standard user-oriented performance measures.

Experiments over four TREC collections show that our proposed metrics do yield different rankings of systems compared to the existing metrics. In particular, we observe a steady decrease in rank correlation with official system rankings as greater weight is placed on finding rare, relevant documents. This means that if our metrics were adopted in practice, participants would be incentivized to retrieve more diverse relevant documents, with the potential to spur further innovation in the field. Additional results show that our metrics provide higher discriminative power and evaluation stability than the standard Precision@K and Average Precision metrics that we modify.

Contributions. 1) We propose a novel class of "innovation" metrics to stimulate greater diversity of document ranking approaches for evaluation campaigns. Future work is expected to expand and improve upon our initial metrics. 2) We propose new generalizations of classic P@K and AP metrics via a simple user-specified mixture weight. This allows weighting document rarity or trivial reversion to the standard metric. 3) Results over four TREC collections show our metrics change system rankings, as well as providing higher discriminative power and evaluation stability than the standard metrics we modify. 4) We share our source code to support reproducibility and follow-on work[1].

Our article is organized as follows. Section 2 describes our proposed metrics. Section 3 then presents an initial, controlled study using synthetic data to show how retrieving rare vs. common documents affects system rankings. Next, Sect. 4 presents our main results with TREC collections, including a qualitative analysis in Sect. 4.6. We then present discussion and limitations in Sect. 5. Section 6 discusses related work, and we conclude in Sect. 7.

[1] https://github.com/mdenizturkmen/ecir2023.

2 Proposed IR Metrics

Retrieval of *rare* documents (that few or no other systems retrieve) indicates that a system's ranking algorithm diverges from that of other systems. In this section, we introduce our two "innovation" metrics that seek to promote exploration of different approaches by rewarding retrieval of such rare, relevant documents. Specifically, we adapt Precision@K (P@K) (Sect. 2.1) and Average Precision (AP) metrics (Sect. 2.2), introducing a linear interpolation parameter α that balances the original metric vs. innovation by varying the weight placed on document rarity. In both cases, setting $\alpha = 0$ reverts to the original metric.

2.1 Rareness-Based Precision@K ($\boldsymbol{P@K_{Rareness}}$)

We define our rareness-based precision-at-k as follows:

$$P@K_{Rareness} = \frac{1}{k}\sum_{i=1}^{k} Rel(d_i)\,(1 + \alpha\, R(d_i)) \tag{1}$$

where k is the rank cut-off value, $Rel(d_i)$ is a binary indicator function for whether d_i is relevant or not, $R(d_i)$ quantifies document rarity, and α is the aforementioned linear interpolation parameter. As noted earlier, setting $\alpha = 0$ reverts to the standard P@K formula. In the other direction, larger α values provide greater rewards for the retrieval of rare documents. Like the original P@K, only relevant documents contribute to the score (i.e., when $Rel(d_i) = 1$), so document rarity is immaterial when $Rel(d_i) = 0$. We define rarity $R(d)$ by:

$$R(d) = 1 - \frac{S_d}{S} \tag{2}$$

where S is the total number of systems and $S_d \geq 1$ is the number of those that retrieve document d. Rareness is bounded by $R(d) \in [0, \frac{(S-1)}{S}]$, minimized when a document is retrieved by all systems (i.e., $S_d = S$) and maximized when only one system retrieves d (i.e., $S_d = 1$). Therefore, as the number of systems S increases, retrieving rare documents becomes more valuable.

While α can be at any value, we recommend setting $\alpha \in [0, 1]$ yielding bounds of $P@K_{Rareness} \in [0, 2)$. The lower bound of $P@K_{Rareness} = 0$ occurs when all documents are non-relevant. The upper-bound is reached when $\alpha = 1$ and all retrieved documents are relevant and have maximal rarity $R(d) = \frac{(S-1)}{S}$, thus $P@K_{Rareness} = 2\frac{(S-1)}{S} < 2$.

2.2 Rareness Based Average Precision ($\boldsymbol{AP_{Rareness}}$)

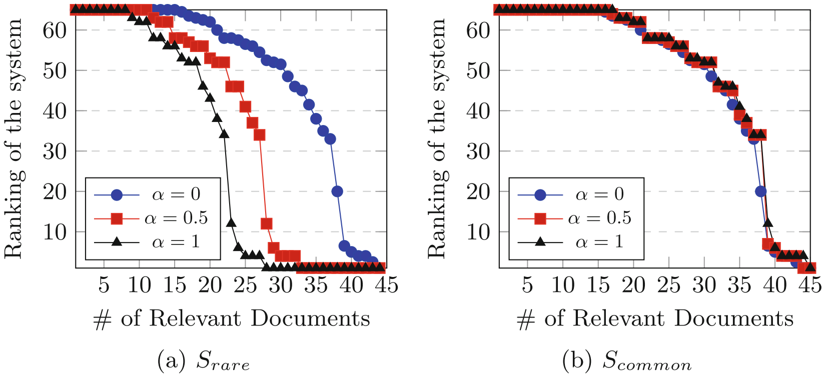
Assuming N_R relevant documents for a given topic, we define $AP_{Rareness}$ as:

$$AP_{Rareness} = \frac{1}{N_R}\sum_{i=1}^{k} Rel(d_i) P@K_{Rareness}(i) \tag{3}$$

When $\alpha = 0$, $P@K_{Rareness} = P@K$, and thus $AP_{Rareness} = AP$. $AP_{Rareness}$ directly inherits $P@K_{Rareness}$'s same lower-bound and upper-bound of $[0, 2)$.

3 Experiment with Synthetic Data

We first present a controlled, synthetic data experiment to explore the behavior of $P@K_{Rareness}$ for varying $\alpha \in [0,1]$ and numbers of D relevant documents retrieved. We contrast the evaluation of two hypothetical systems: S_{rare} vs. S_{common}, on a single topic (#1127540) from the Deep Learning Track 2020 (DLT20) [9], as if our hypothetical systems had participated with other real participants. While S_{rare} always retrieves simulated relevant documents found by no other system, S_{common} retrieves the most common, real relevant documents first. We include all official runs from DLT20's document ranking task.



(a) S_{rare}

(b) S_{common}

Fig. 1. Ranking of hypothetical systems, S_{rare} and S_{common}, for topic 1127540 of Deep Learning Track 2020 based on $P@100_{Rareness}$. Experiments vary rarity weight α as well as D, the number of relevant documents retrieved.

Figure 1 shows the $P@100_{Rareness}$ ranking of S_{rare} vs. S_{common}. Note that a lower rank indicates a better system, with the best system being ranked first (i.e., having rank 1). First, recall that when $\alpha = 0$, $P@K_{Rareness} = P@k$. In this case, both S_{rare} and S_{common} are seen to exhibit the same $P@K_{Rareness}$ curve, as expected, since no weight is placed on rarity. Second, we see that S_{common}'s ranking is largely unaffected by α since it always retrieves common (i.e., non-rare) relevant documents. In contrast, the ranking of S_{rare} noticeably changes across different α values. For example, it requires 28, 33, and 44 relevant documents to be ranked first when α is set to 1, 0.5, and 0, respectively.

Overall, the results above validate our expectations regarding the behavior of $P@k_{Rareness}$ under controlled conditions. It reverts toward standard $P@k$ at $\alpha = 0$, and results place greater emphasis on rarity as we move toward $\alpha = 1$.

4 Experiments with Real Data

In this section, we first describe our experimental setup (Sect. 4.1). Next, we compare our modified metrics vs. their original counterparts in terms of system rankings (Sect. 4.2), discriminative power (Sect. 4.3), and evaluation stability

(Sect. 4.4). We also assess how our metrics are affected by the number of systems (Sect. 4.5). Furthermore, we conduct qualitative analysis to better understand the nature of rarely-retrieved documents (Sect. 4.6).

4.1 Experimental Setup

We use trec_eval[2] for calculations of classical evaluation metrics. We set the cut-off threshold to 100 for all metrics we use including ours. We use four different TREC collections, including TREC-5 [11], TREC-8 [12], Web Track 2014 (WT14) [7], and Deep Learning Track 2020 (DLT20) [9]. We carry out our experiments using all official runs from ad-hoc search tasks of TREC-5, TREC-8, and WT14, and the document ranking task of DLT20.

4.2 System Rankings

We compare system rankings for $P@100_{Rareness}$ and $AP_{Rareness}$ against rankings based on P@100 and AP, respectively, in order to observe the impact of rewarding rarity. We report Kendall's τ rank correlation. Experiments with τ_{AP} [31] yielded similar results and so are omitted.

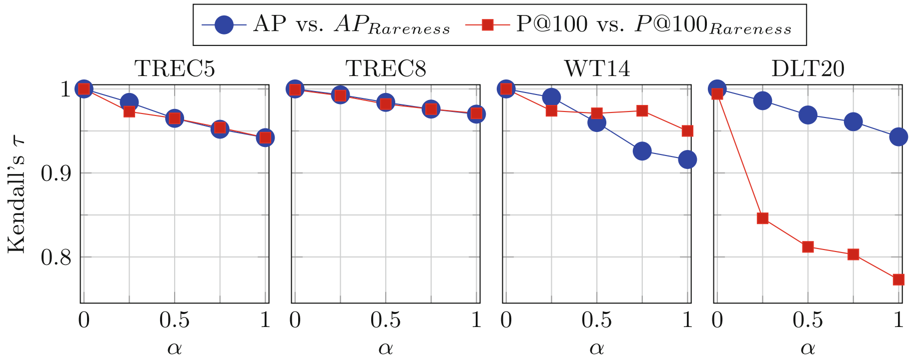


Fig. 2. Kendall's τ correlation between system rankings based on $P@100_{Rareness}$ vs. $P@100$ and system rankings based on AP vs. $AP_{Rareness}$.

Figure 2 shows Kendall's τ scores on four test collections for varying α. As expected, when α=0, our modified metrics revert to their unmodified forms, thus yielding perfect $\tau = 1$ rank correlation. We observe steady trends of decreasing rank correlation with increasing α. While Kendall's τ scores for comparisons against AP and P@K metrics are similar in TREC-5 and TREC-8, they diverge in WT14 and DLT20. For instance, when we compare P@100 vs. $P@100_{Rareness}$ in DLT20, Kendall's τ is lower than 0.9 (a traditionally-accepted threshold for acceptable correlation [29]) for $\alpha > 0$. However, we do not observe this when we

[2] https://trec.nist.gov/trec_eval/.

Table 1. Discriminative power of metrics for 95% and 99% significance thresholds. The highest score for each collection and significance threshold is written in **bold**. Note that the total number of system pairs are 1830, 8256, 406, 2016 for TREC-5, TREC-8, WT14, DTL20, respectively.

Metric		TREC-5		TREC-8		WT14		DLT20	
	α	95%	99%	95%	99%	95%	99%	95%	99%
P@100	0.0	598	406	3666	2973	218	174	61	15
$P@100_{Rareness}$	0.5	680	457	3778	3050	213	168	173	24
$P@100_{Rareness}$	1.0	728	476	3825	3079	213	168	237	82
AP	0.0	541	334	3731	2976	211	169	320	169
$AP_{Rareness}$	0.5	632	404	3915	3209	216	168	352	209
$AP_{Rareness}$	1.0	701	467	4048	3363	214	170	376	238

compare $AP_{Rareness}$ vs. AP. This suggests that DLT20 systems retrieve many rare, relevant documents at low ranks, causing large changes in system rankings when we use $P@100_{Rareness}$. Smaller changes occur with $AP_{Rareness}$ as the impact of documents is diminished due to their low ranks.

4.3 Discriminative Power

Discriminative power indicates how well a metric can tell systems apart. Zhou et al. [34] measure discriminative power by counting the number of significantly different system pairs. We apply this same method to measure the discriminative power of our proposed metrics, using Tukey's HSD test as the statistical hypothesis test. Table 1 shows the number of significantly different pairs for baseline and our proposed metrics when we use 95% and 99% significance thresholds.

We observe that our metrics have higher discriminative power than baselines. Increasing α tends to increase discriminative power across test collections.

4.4 Stability

If an evaluation methodology is reliable, the measured performance of systems should be stable, i.e., should not change dramatically under different conditions. In order to measure the stability of metrics, we adopt Buckley and Voorhees [3]'s approach. We first sample T topics and calculate system scores on the sampled topic set only. Next, we compare each pair of systems to see which performs better. After repeating this process R times, we assess the stability of the comparison over the R trials. For example, imagine one system outperforms another in 700/1000 trials, yielding a stability score of 0.7 for that pair. We take the average stability scores of all pairs as the overall metric stability. In our experiments, we arbitrarily set T to the half of the topic set in each collection (i.e., 22 ($= \lfloor 45/2 \rfloor$) for DLT20 and 25 ($= 50/2$) for the others). We set the number of trials $R = 1000$ but observed that the results largely converged

Table 2. Metric stability scores. Our metrics are most stable across collections.

Metric	α	TREC-5	TREC-8	WT14	DLT20
P@100	0.0	0.532	0.545	0.635	0.071
$P@100_{Rareness}$	0.5	0.642	0.628	0.716	0.128
$P@100_{Rareness}$	1.0	0.709	0.684	0.768	0.179
AP	0.0	0.513	0.580	0.433	0.425
$AP_{Rareness}$	0.5	0.578	0.623	0.517	0.444
$AP_{Rareness}$	1.0	0.633	0.656	0.585	0.466

after 100 trials. Results for baselines vs. proposed metrics are shown in Table 2. $P@100_{Rareness}$ and $AP_{Rareness}$ yield a higher stability score in all cases vs. their classic counterparts.

4.5 Impact of Number of Systems

As retrieval-rarity of documents depends on the participating systems, system scores and rankings might change when we use a different set of systems to calculate the rarity scores of documents. To test how scores of systems change as the systems to be evaluated vary, we conduct the experiment described in Algorithm 1. In particular, we first rank all systems [Line 1]. Then we randomly pick N number of systems [Line 5] and rank them [Line 6]. Subsequently, we get how these N systems are ranked initially (i.e., when all systems are used) [Line 7] and calculate the τ score between these two rankings [Line 8]. We repeat this process 1000 times [Lines 3–9] and calculate the average τ score [Line 10]. Table 3 shows the results for $N = 2^j, j \in [1-6]$ in TREC-8. We observe that correlation scores are generally very high, suggesting that rankings of systems are stable even though we use different sets of systems.

Algorithm 1. Experiment to Analyze Impact of Using N Participants

```
    Input: $P \leftarrow$ The whole participant list
           $N \leftarrow$ The number of selected systems
 1: $R_o \leftarrow$ rank systems in $P$
 2: $\tau_N \leftarrow 0$
 3: $trials \leftarrow 1000$
 4: for all $trials$ do
 5:     $p_N \leftarrow$ randomly sample N systems from P
 6:     $r_N \leftarrow$ rank systems in $p_N$
 7:     $R_N \leftarrow$ filter systems $\in p_N$ from $R_o$
 8:     $\tau_N \leftarrow \tau_N + \tau\_correlation(R_N, r_N)$
 9: end for
10: $\tau_N \leftarrow \tau_N \;/\; trials$
```

Table 3. Impact of number of systems based on the experimental setup explained in Algorithm 1. We use TREC-8 for this experiment.

Metrics	N = 2	N = 4	N = 8	N = 16	N = 32	N = 64
$P@100_{Rareness}(\alpha = 1)$	0.970	0.976	0.983	0.987	0.992	0.995
$AP_{Rareness}(\alpha = 1)$	0.970	0.984	0.985	0.989	0.993	0.996

4.6 Qualitative Analysis

To better understand the nature of rarely-retrieved documents, we conducted the following qualitative analysis. We randomly selected six TREC-8 topics, computed the rarity $R(d)$ of each relevant document d, and then selected five documents with varying rarity scores. We manually analyzed how document relevance changes depending on rarity. In general, while commonly retrieved documents appear focused on the search topic, rarely retrieved documents differ in focus but still contain relevant passages.

Table 4 presents manually analyzed documents for topic 431, whose narrative states the information need: "latest developments in robotic technology". The relevant document FBIS4-44815 with minimal rarity is entitled, "Germany: Automation, Robotics Seen as Keys in Industrial", which seems directly relevant to the information need. In contrast, relevant document FBIS3-38782 ($R(d) = 0.81$) only indirectly mentions that a robot can be used for underwater photography, with the title "BND Warns Against Nuclear Terrorists".

If rarely retrieved documents are less relevant, why reward their retrieval? First, while the observation above may hold when all systems are roughly comparable, this is not always true. For example, manual runs have long been advocated in evaluation campaigns because they tend to differ markedly from automated runs and find relevant documents that other systems miss. In general, we cannot tell whether outlier systems are brilliant or remedial without human labels [27]. Second, our goal in this work is to encourage systems that diverge from the pack, with the hope that such divergence will correspond to improvement. The nature of research is that some amount of failure often precedes success, and that making larger departures is important to create a potential for larger improvements. Third, even if we assume a user-centered view, finding additional, less relevant documents can still be important in various cases: when there are few relevant documents, in a "total recall" task setting [24] or pooling [28], or as input to a rank fusion ensemble model [19]. We discuss these further in the next section.

5 Discussion and Limitations

In this section, we discuss various aspects and limitations of our work: motivation and concept of "innovation" metrics (Sect. 5.1), proposed methods (Sect. 5.2), our experimental design and findings (Sect. 5.3), and potential impacts and directions for future work (Sect. 5.4).

Table 4. Analyzed documents for topic 431. The relevant content column corresponds to sentences that might fulfill the information need. If there are multiple useful sentences in a document, the most informative one is selected.

Document ID	Rareness	Relevant Content
FBIS3-38782	0.81	One of the trapeze-like wings had broken off, the nose was missing, and in the dull gray water of Lake Constance even a diving robot of the “Sear Rover” type could send only diffuse video pictures from 159 m below the surface of the lake
LA020889-0003	0.62	A Japanese robot named Wabot II tickles a keyboard to produce original music as part of an exhibit at the Chicago Museum of Science and Industry
LA102589-0109	0.41	The trucks, equipped with robotic arms that hoist and empty containers, will collect trash every week and recyclable items twice monthly
LA092189-0061	0.22	Industrially they use robots for welding, painting or picking and placing items, for example
FBIS4-44815	0.08	New applications for service robots are opening up also in medicine and rehabilitation, in care for the aged and handicapped, in bureaus and logistics, in municipal activities, in households, in hobbies and recreation

5.1 Concept and Motivation

We envision potential benefit from stimulating greater diversity in document ranking methods. In regard to evaluation campaigns, we suggest the field would benefit if participants built upon a wider range of existing methods and/or investigated more radical departures from those methods. While today’s evaluation campaigns are already healthy and vibrant, we believe it could be fruitful: 1) to reflect on, assess, and discuss as a community the ways in which we might further strengthen evaluation campaigns; 2) to focus on the diversity of approaches and innovation in particular, and how to promote higher-risk research with potential for greater gains; and 3) to operationalize metrics by which we might measure and optimize for such innovation in evaluation campaigns. Potential counter-arguments could be that: a) campaign steering committees are already doing (1) and don’t need larger community engagement in it; b) innovation is a complex construct that is best left to organic processes rather than trying to “force” it through explicit optimization; and c) one can argue that research construed as incremental is actually instrumental (i.e., small steps and minor variants can add up over time to large advances). Such discussion and debate seem healthy for a community, regardless of the outcome.

One controversial aspect of our work is the proposal of IR evaluation metrics that explicitly seek to optimize something other than retrieval quality for the user. In particular, the metrics we propose reward systems for retrieving relevant documents missed by other systems, but there is no obvious reason a user would prefer such rare relevant documents over common ones. In fact, less retrieved documents may tend to be less relevant on average and thus aptly lower-ranked (Sect. 4.6). In fact, prior work in meta-ranking (aka rank fusion) has exploited the number of systems that retrieve a given document as a useful feature in estimating document relevance [19]. However, our goal of promoting greater community diversity of ranking methods is not a user-oriented metric, but a field-oriented metric. Moreover, in seeking to promote higher-risk research, we may need to explore a variety of methods yielding sub-par results for the user before we discover a novel method that does provide a transformative advance. For example, years of research on (then) sub-par neural networks was necessary before yielding today's state-of-the-art deep learning methods [20].

Rewarding retrieval of rare relevant documents also has the potential to improve meta-ranking (aka rank fusion or ensemble ranking) and pooling [28]. For instance, ensemble models benefit from a diverse set of input systems that complement each other's shortcomings. Thus, including input systems that find unique relevant documents could boost ensemble performance. Pooling similarly benefits from the diversity of participating systems so that the pool finds as many relevant documents as possible. This helps to ensure that the pool is reusable for future systems using innovative approaches. Our metrics could thus encourage more diverse systems to improve meta-ranking and pooling. In the other direction, recall measures for those tasks might also be repurposed to measure and promote overall diversity and innovation of ranking approaches.

5.2 Proposed Metrics

The two specific innovation metrics we propose have a variety of limitations and represent only the tip of the iceberg of better innovation metrics. We expect future work will propose better metrics that surpass ours.

As noted above, the notion of innovation is a complex construct. Our metrics that reward retrieval of relevant documents missed by other systems are clearly crude metrics for quantifying such a complex construct. To the best of our knowledge, ours is the first metric for measuring and promoting such innovation, but the first effort seldom represents the only or best way. More sophisticated future work by others could model this construct with greater detail and fidelity.

While we have suggested combining $Rel(d_i)$ and $Rel(d) \cdot R(d)$ together into a single mixture for simplicity, an evaluation campaign could also use these as separate and complementary official metrics, akin to evaluating precision vs. recall separately rather than fusing them together into a single f-measure metric. On the other hand, our mixture approach can also be seen as an easy way to generalize existing metrics to consider additional aspects of performance. Because our modified metrics revert to their standard counterpart metrics when $\alpha = 0$, generalization allows use in that original, more restricted setting while also permitting greater flexibility in incorporating additional factors when $\alpha > 0$. While we focus on generalizing existing metrics to include consideration of document rarity, other researchers might incorporate other aspects of system performance into traditional metrics using similar linear mixtures.

At a more mundane level, because our metrics are bounded by $[0, 2)$, it may be useful to renormalize them to a more standard $[0, 1]$ range. While this might be done to values post hoc, hindsight instead suggests two minor revisions to formulas for future use. First, re-define rarity as $R'(d) = 1 - \frac{S_d - 1}{S - 1} \in [0, 1]$ for $S, S_d >= 1$, maximized when $S_d = S$. Second, re-define $P'@K_{rareness}$ as:

$$P'@K_{rareness} = \frac{1}{k} \sum_{i=1}^{k} \Big[(1 - \alpha) Rel(d_i) + \alpha \, Rel(d_i) \, R(d_i) \Big] \tag{4}$$

where we now constrain $\alpha \in [0, 1]$ as a probability. This mixture model formulation directly bounds $P'@K_{rareness} \in [0, 1]$.

Our metrics assume linearity in: 1) how we quantify rarity $R(d)$; and 2) the mixture model between the classic metric and rarity. If we consider IR's rich history exploring many variant functions for inverse-document frequency (IDF) to weight rare terms [26], one could imagine similarly exploring many other weighting functions for rarity. Regarding the mixture model, while we have assumed a fixed α across topics, future work might also investigate a hyperparameter approach (akin to Dirichlet smoothing [33]) to intelligently vary α per topic in relation to per topic factors, such as the number of relevant documents.

Yet another idea would be to incorporate document importance alongside rarity in the reward metric for innovation. Intuitively, finding a relevant document that other systems miss is more important when there are few relevant documents in total. As an example, assume for some topic that a given relevant document is only retrieved by a single system. If there are only two relevant documents in total, finding that second relevant document may be vital to satisfying a user's information need. On the other hand, if there were 100 relevant documents, finding the 100^{th} document may provide minimal further value. This would suggest extending the metric to consider the number of relevant documents for each topic.

Finally, our use of P@K and AP assumes binary relevance judgments. Future work could extend innovation metrics to graded relevance judgments.

5.3 Experimental Design and Findings

While we evaluated over four test collections to assess generality, we did not explore the properties of these test collections in detail, or how those varying properties could impact our findings. In addition, expanding our coverage to further test collections could further assess the robustness of findings. Finally, it could be useful to conduct a qualitative inspection of the meta-data descriptions of the best-performing systems (submitted by participants along with their TREC runs) in order to assess the correlation between system descriptions vs. which systems perform best when scored by our innovation metrics.

5.4 Expected Use and Impact

Imagine our metrics were adopted by an evaluation campaign and one or more participating systems sought to optimize them. Beyond the broad goal of promoting higher-risk research and accelerating field innovation, this would be expected to specifically lead to more diverse document rankings. Assuming a fixed evaluation budget (i.e., the number of documents that human judges will review), less overlap across document rankings would mean that we could only pool to a lower depth for the same cost. However, whether this would lead to a more or less complete document pool remains an open, empirical question, likely dependent on the setting of α used. For evaluation campaigns that permit participants to submit multiple runs and distinguish an "official" run (contributing to pooling) vs. additional runs (scored by the official run pool), whether official vs. additional runs would be used to set S_d would also impact subsequent findings.

A well-known issue in IR is the reusability of pools. A very different system might find relevant documents all other systems missed, but if it did not participate in the pool, it would be penalized in evaluation rather than rewarded. Similarly, when we quantify rarity $R(d)$ based on participating systems, there are questions of reusability for future systems evaluating on an existing pool. Moreover, we would expect that a system optimizing for such rarity would be even more likely to run into this problem in practice. Another common distinction made is between methods to create reusable test collections (e.g., pooling) vs. methods to efficiently rank a current set of systems (e.g., StatAP [21] and MTC [4]). Similarly, our rarity metrics will return different scores depending on the other participating systems in the pool. A limitation of our work is that we only rank systems participating in a shared-task, leaving study of reusability for future work.

6 Related Work

To the best of our knowledge, no existing IR evaluation metrics consider the innovativeness of systems. While we frame this *wrt.* rarely-retrieved documents, prior work has usefully designed metrics to evaluate systems reliably with missing judgments, such as Bpref [2] and infAP [30]. These metrics aim to predict the performance of systems with incomplete judgments. In contrast, our focus is to promote innovation in document ranking methods.

To handle missing judgments, a number of studies have explored how to select documents to be judged such as Move-To-Front [8] and MaxMean [16]. These studies aim to maximize the number of relevant documents because unjudged documents are assumed to be non-relevant. As a document is more likely to be relevant if retrieved by many systems, commonly-retrieved documents are more likely to be judged than rarely retrieved ones. However, in contrast to these document selection methods, we assign more weight to rarely-retrieved ones.

In modifying P@K and AP, we have followed standard practice in aggregating scores over topics using a simple arithmetic average. However, various other aggregate statistics have been proposed. Robertson [23] asserts that the impact of hard topic scores is diminished on the overall score with the arithmetic mean. He thus recommends geometric mean instead. Ravana and Moffat [22] show that geometric mean average precision (GMAP) is better at handling variability in topic difficulty than arithmetic mean average precision (MAP). Mizzaro [17] proposes normalized mean average precision (NMAP), which takes into account topic difficulty. He defines topic difficulty as 1-(average AP score). Unlike these studies, we focus on retrieval difficulty at the document level. In addition, prior studies on topic difficulty work on how to aggregate traditional IR metrics.

As noted earlier, while we assumed binary relevance judgments and modify only P@K and AP metrics, many other metrics exist, beyond binary relevance, that could be extended to innovation. Prominent examples include normalized discounted cumulative gain (nDCG) [14], and rank biased precision (RBP) [18], which assume that users will examine documents in the retrieval order and might stop examining whenever their information need is satisfied. Such rank-based metrics ascribe more weight to documents at higher ranks. Other important evaluation metrics include miss (i.e., the fraction of non-retrieved documents that are relevant) [13], fallout [15], expected reciprocal rank [5], weighted reciprocal rank [10], and O-measure [25].

Prior work has also proposed metrics rewarding the diversity within a single document ranking in relation to novelty and coverage of different topic facets. For instance, Zhai et al. [32] propose three metrics – subtopic recall metric (S-recall), subtopic precision (S-precision) and weighted subtopic precision (WS-precision) – that consider redundancy in ranked lists. Clarke et al. [6] extend nDCG by rewarding novelty and covering multiple topic aspects. In contrast, we quantify diversity across systems rather than within a single ranked list. In particular, we reward systems for retrieving relevant documents that other systems miss.

7 Conclusion

We propose a new class of IR evaluation metrics designed to promote exploration of higher-risk, more radical departures from current state-of-the-art methods. These “innovation metrics” reward retrieval of relevant documents missed by other systems. The key intuition is that finding relevant documents missed by other systems suggests a markedly different approach. More specifically, we generalize classic Precision@K and Average Precision metrics via a simple mixture-weight parameter controlling the relative reward for finding relevant documents other systems miss. Setting this to zero reverts to the original metric.

Experiments over four TREC collections show that our proposed metrics yield different system rankings compared to the existing metrics. In particular, we observe a steady decrease in rank correlation with official system rankings as reward increases for finding rare, relevant documents. These results are further supported by a controlled, synthetic data experiment, as well as qualitative analysis. Collectively, results suggest that if our metrics were adopted in practice, participants would be incentivized to retrieve more diverse relevant documents, with the potential to spur further innovation in the field. Finally, we also show that our metrics provide higher discriminative power and evaluation stability than the standard Precision@K and Average Precision metrics that we modify.

To the best of our knowledge, ours is the first proposal of IR evaluation metrics designed to explicitly measure and promote innovation in ranking methods. That said, the first attempt at any endeavor is seldom the only or best way to accomplish it. Our two proposed metrics have a variety of limitations and represent only the tip of the iceberg for imagining this new class of innovation metrics. Consequently, we expect future metrics will be proposed that surpass ours in better modeling the complex construct of innovation, and in doing so, will further advance the cause of promoting innovation in ranking methods.

Acknowledgments. We thank the reviewers for their valuable feedback. This research was supported in part by the Scientific and Technological Research Council of Turkey (TUBITAK) ARDEB 3501 (Grant No 120E514) and by Good Systems (https://goodsystems.utexas.edu), a UT Austin Grand Challenge to develop responsible AI technologies. Our opinions are our own.

References

1. Brown, T., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901 (2020)
2. 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 (2004)
3. Buckley, C., Voorhees, E.M.: Evaluating evaluation measure stability. In: ACM SIGIR Forum, vol. 51, pp. 235–242. ACM New York (2017)
4. 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 (2006)
5. 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 (2009)
6. Clarke, C.L., et al.: Novelty and diversity in information retrieval evaluation. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 659–666 (2008)
7. Collins-Thompson, K., Macdonald, C., Bennett, P., Diaz, F., Voorhees, E.M.: Trec 2014 web track overview. Technical report, MICHIGAN UNIV ANN ARBOR (2015)
8. 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 (1998)
9. Craswell, N., Mitra, B., Yilmaz, E., Campos, D.: Overview of the TREC 2020 deep learning track. arXiv preprint arXiv:2102.07662 (2021)
10. Eguchi, K., Oyama, K., Ishida, E., Kando, N., Kuriyama, K.: Overview of the web retrieval task at the third NTCIR workshop. In: NTCIR. Citeseer (2002)
11. Harman, D., Voorhees, E.: Overview of the fifth text retrieval conference (TREC-5). In: Harman, D., Voorhees, E. (eds.) Information Technology: The Fifth Text REtrieval Conference (TREC-5), National Institute of Standards and Technology Special Publication, pp. 500–238 (1996)

12. Hawking, D., Voorhees, E., Craswell, N., Bailey, P., et al.: Overview of the TREC-8 web track. In: TREC (1999)
13. Heine, M.: Information-retrieval from classical databases from a signal-detection standpoint-a review. Inf. Technol. Res. Dev. Appl. **3**(2), 95–112 (1984)
14. Järvelin, K., Kekäläinen, J.: IR evaluation methods for retrieving highly relevant documents. In: ACM SIGIR Forum, vol. 51, pp. 243–250. ACM, New York (2017)
15. Kraft, D.H., Bookstein, A.: Evaluation of information retrieval systems: a decision theory approach. J. Am. Soc. Inf. Sci. **29**(1), 31–40 (1978)
16. Losada, D.E., Parapar, J., Barreiro, A.: Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems. Inf. Process. Manag. **53**(5), 1005–1025 (2017)
17. Mizzaro, S.: The good, the bad, the difficult, and the easy: something wrong with information retrieval evaluation? In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 642–646. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78646-7_71
18. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. Inf. Syst. (TOIS) **27**(1), 1–27 (2008)
19. Nuray, R., Can, F.: Automatic ranking of information retrieval systems using data fusion. Inf. Process. Manag. **42**(3), 595–614 (2006)
20. Onal, K.D., et al.: Neural information retrieval: at the end of the early years. Inf. Retrieval J. **21**(2), 111–182 (2018)
21. Pavlu, V., Aslam, J.: A practical sampling strategy for efficient retrieval evaluation. Northeastern University, College of Computer and Information Science (2007)
22. Ravana, S.D., Moffat, A.: Exploring evaluation metrics: Gmap versus map. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 687–688 (2008)
23. Robertson, S.: On GMAP: and other transformations. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, pp. 78–83 (2006)
24. Roegiest, A., Cormack, G.V., Clarke, C.L., Grossman, M.R.: TREC 2015 total recall track overview. In: TREC (2015)
25. Sakai, T.: On the task of finding one highly relevant document with high precision. Inf. Media Technol. **1**(2), 1025–1039 (2006)
26. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. **24**(5), 513–523 (1988)
27. 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 (2001)
28. Spark-Jones, K.: Report on the need for and provision of an ‘ideal’ information retrieval test collection. Computer Laboratory (1975)
29. Voorhees, E.M.: Variations in relevance judgments and the measurement of retrieval effectiveness. Inf. Process. Manag. **36**(5), 697–716 (2000)
30. 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 (2006)
31. Yilmaz, E., Aslam, J.A., Robertson, S.: A new rank correlation coefficient for information retrieval. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 587–594 (2008)

32. Zhai, C., Cohen, W.W., Lafferty, J.: Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In: ACM SIGIR Forum, vol. 49, pp. 2–9. ACM New York (2015)
33. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to ad hoc information retrieval. In: ACM SIGIR Forum, vol. 51, pp. 268–276. ACM, New York (2017)
34. Zhou, K., Lalmas, M., Sakai, T., Cummins, R., Jose, J.M.: On the reliability and intuitiveness of aggregated search metrics. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, pp. 689–698 (2013)