Ranking Interruptus: When Truncated Rankings Are Better and How to Measure That

Enrique Amigó
UNED NLP & IR Group
Madrid, Spain
enrique@lsi.uned.es

Stefano Mizzaro
University of Udine
Udine, Italy
mizzaro@uniud.it

Damiano Spina
RMIT University
Melbourne, Australia
damiano.spina@rmit.edu.au

ABSTRACT

Most of information retrieval effectiveness evaluation metrics assume that systems appending irrelevant documents at the bottom of the ranking are as effective as (or not worse than) systems that have a stopping criteria to *truncate* the ranking at the right position to avoid retrieving those irrelevant documents at the end. It can be argued, however, that such truncated rankings are more useful to the end user. It is thus important to understand how to measure retrieval effectiveness in this scenario. In this paper we provide both theoretical and experimental contributions. We first define formal properties to analyze how effectiveness metrics behave when evaluating truncated rankings. Our theoretical analysis shows that de-facto standard metrics do not satisfy desirable properties to evaluate truncated rankings: only Observational Information Effectiveness (OIE) – a metric based on Shannon's information theory – satisfies them all. We then perform experiments to compare several metrics on nine TREC datasets. According to our experimental results, the most appropriate metrics for truncated rankings are OIE and a novel extension of Rank-Biased Precision that adds a user effort factor penalizing the retrieval of irrelevant documents.

CCS CONCEPTS

• **Information systems** → **Retrieval effectiveness**.

KEYWORDS

Information retrieval, Evaluation, Evaluation measures, Ranking cutoff

ACM Reference Format:
Enrique Amigó, Stefano Mizzaro, and Damiano Spina. 2022. Ranking Interruptus: When Truncated Rankings Are Better and How to Measure That. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3477495.3532051

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07.
https://doi.org/10.1145/3477495.3532051

1 INTRODUCTION

Let us imagine the following Information Retrieval (IR) system outputs:

`A : 11111 11111.` `B : 11111 11111 00000 00000.`

System `A` retrieves ten documents as a response to a query, all of them relevant (`1`). System `B` answers the same query by retrieving twenty documents, of which the same first ten documents, in the same order, as `A`, and then also ten irrelevant documents (`0`). Offline evaluation of IR systems consists of comparing system outputs against a ground truth containing relevance judgments [17, 33, 35, 39, 40] using effectiveness metrics – or, alternatively, assessing preference judgments [13, 15, 16, 21]. When computing traditional IR effectiveness metrics – e.g., Average Precision (AP), Normalized Document Cumulative Gain (NDCG), or Rank-Biased Precision (RBP) – on the two rankings, no difference is found: the two systems are considered equally effective. The same happens also when cutoff metrics (AP@N, NDCG@N, RBP@N) are used, for any cutoff N. However, in a *truncated*[1] *ranking* scenario it can be argued that system `A` is more useful to the user, since it avoids him/her the inspection of irrelevant documents. Such an inspection could be considered not only useless, but even harmful [12, 24]. Continuing with the example, another system `C` might return the same 19 documents of `B`, and then a relevant document in position 20 in place of an irrelevant one (`C : 11111 11111 00000 00001`). Most standard effectiveness metrics, when the cutoff is appropriate, correctly capture that `C` is better than `B`. But what is of interest here is that `C` is considered also more effective than `A` since "a higher amount of relevance is returned". However, it can be argued that the effort required to the user of `C` to inspect the documents in positions 11–19 might be higher than the benefit provided by the relevant document in position 20. The effort would be even higher, and the benefit lower, if the rankings were 100 documents long and the irrelevant documents to be inspected were 89 instead of 9. None of the standard metrics would reward system `A` for avoiding such an effort; the best that can be obtained is by using a cutoff metric, at cutoff @10 (or lower), that would consider `A` and `C` (and `B`) as equally effective.

Leaving the example aside and speaking in more general terms, traditional effectiveness metrics evaluate unlimited or fixed length lists of documents sorted by relevance. It is assumed that it is the user her/himself who stops her/his exploration at some point. However, truncating the ranking automatically, or even not giving an answer (i.e., limiting to zero the length of the ranking), can save

[1] To avoid confusion, we reserve "cutoff @N" to its traditional use (for metrics considering only the first part of the ranking) and we speak of "truncated rankings of length n" when the IR system returns a ranking with fewer documents.

time to the user. In such cases it may be more effective to invite the user to review their search criteria. This is especially pertinent in scenarios where the communication channel is limited – as in search on small screens [18] or spoken conversational search [10, 12, 37]. A parallel can also be drawn with question answering systems: in this case there is a clear benefit from knowing that the system does not have a correct answer [29]. Furthermore, we can find automatically truncated rankings in multiple scenarios, such as recommendation of variable-sized lists of products in an online sales application, contacts in a social network, ranking of trending topics, etc.

There is previous work in the literature proposing metrics that capture truncated rankings such as terminal document based metrics [25], Rank-Biased Utility (RBU) [9], or Observational Information Effectiveness (OIE) [5]. However, none of the above metrics have had a significant uptake in the community. We understand that these metrics require further comparative analyses, behavior studies, and parameter fixing before they can be applied to truncated ranking problems in practice.

To the best of our knowledge, this is the first paper that makes an in-depth study of metrics oriented to the evaluation of truncated rankings from both theoretical and empirical perspectives. We provide several contributions. We start from an axiomatic perspective [3] by defining a set of seven formal properties that should be satisfied by effectiveness metrics aimed at evaluating truncated rankings (Contribution 1). We then use these properties to analyze both existing truncated ranking metrics and a set of novel metrics that we propose. These are utility-based metrics that we obtain by adding an user effort factor to the traditional metrics such as RBP, Expected Reciprocal Rank (ERR), and Discounted Cumulative Gain (DCG) (Contribution 2). Such an analysis shows that neither the existing metrics, nor the novel utility-based ones that we propose, satisfy all the desirable properties, with the single exception of one metric, OIE (Contribution 3). We then perform some experiments over nine TREC datasets. The results confirm the theoretical analysis and suggest that OIE and the extended utility RBP are the best candidates for truncated ranking evaluation (Contribution 4). Finally, we propose guidelines for metric selection in different truncated ranking scenarios (Contribution 5).

The rest of the paper is structured as follows. In Section 2 we discuss the properties that the effectiveness evaluation metrics should satisfy. In Section 3 we analyze existing metrics in the literature, as well as some novel ones that we propose, categorizing them into different groups and analyzing their formal properties. Let us remark that we do not include a section explicitly describing related work as it is often done: we rather discuss it in the above two sections. In Section 4 we describe the experiments and results. We describe a guideline for metric selection in Section 5 and conclude in Section 6.

2 FORMAL PROPERTIES OF METRICS

We conduct our analysis on the basis of a number of formal properties – derived from the literature – that effectiveness metrics should satisfy. Depending on the particular scenario in which the metrics are applied, different properties may or may not be desirable; we list the properties for classical non-truncated rankings in Section 2.2 and the properties that are specific for truncated rankings in Section 2.3.

2.1 Notation

We formalize the output of a document retrieval system as an ordered list of documents $\vec{d} = (d_1, \ldots, d_n)$ of length n, extracted from the collection $\mathcal{D}$. We formalize the graded relevance of a document $d \in \mathcal{D}$ as $\texttt{Rel}(d) \in [0, 1]$; we also use superscripts to represent relevance and non relevance, so that $\texttt{Rel}(d_i^r) = 1$ and $\texttt{Rel}(d_i^{\neg r}) = 0$. We use $\mathcal{M}(\vec{d})$ to denote the score given by applying an evaluation metric $\mathcal{M}$ to a given ranking $\vec{d}$. In order to express formal constraints, we use $\vec{d}_{i \leftrightarrow j}$ to denote the ranking resulting from swapping the two documents in positions i and j.

2.2 Properties for Non-Truncated Rankings

The first property reflects the fact that relevant documents should appear at the top of the system output ranking. This idea has been captured by Ferrante et al. [20] as the *Replacement* and *Swapping* properties, by Moffat [27] as *Convergence*, and by Amigó et al. [8, 9] as the *Priority* property.

Property 1 (Priority). *Swapping documents in concordance with their relevance increases the ranking quality score. Being $k > 0$:*

$$\begin{aligned} &\textit{If } \texttt{Rel}\left(d_{i+k}\right) > \texttt{Rel}\left(d_i\right) \\ &\textit{then } \mathcal{M}\left(\vec{d}_{i \leftrightarrow i+k}\right) > \mathcal{M}\left(\vec{d}\right). \end{aligned}$$

The second property states that the effect of a document in the system quality score depends on the depth of the documents in the ranking. This notion is captured as the *Top-weightedness* property by Moffat [27] and as *Deepness* by Amigó et al. [8].

Property 2 (Top-Weightedness). *Correctly swapping contiguous items has more effect in early ranking positions:*

$$\begin{aligned} &\textit{If } \texttt{Rel}(d_i) = \texttt{Rel}(d_j) < \texttt{Rel}(d_{i+1}) = \texttt{Rel}(d_{j+1}) \\ &\textit{then } \mathcal{M}\left(\vec{d}_{i \leftrightarrow i+1}\right) > \mathcal{M}\left(\vec{d}_{j \leftrightarrow j+1}\right) \end{aligned}$$

where $i < j$.

The next property reflects that the effort spent by the user to inspect a long (deep) list of search results is limited [8, 9]. In other words, there is an area of the ranking that may never get explored by the user.

Property 3 (Deepness Threshold). *There exists a value n large enough such that, retrieving one single relevant document at the top of the ranking is better than retrieving n relevant documents after n irrelevant documents. Being $\texttt{Rel}(d_i^r) = 1$ and $\texttt{Rel}(d_i^{\neg r}) = 0$:*

$$\exists n \in \mathbb{N}^+ . \mathcal{M}\left(d_1^r, \ldots\right) > \mathcal{M}\left(d_1^{\neg r}, \ldots, d_n^{\neg r}, d_1^r, \ldots, d_n^r\right).$$

On the other hand, we can assume that there exists a (short) ranking area which is always explored by the user. In other words, at least a few documents are inspected by the user with a minimum effort. This means that, at the top of the ranking, the amount of

captured relevant documents is more important than their relative rank positions [8].[2]

Property 4 (Shallowness Threshold). *There exists a value m small enough such that retrieving one single relevant document in the first position is worse than m relevant documents after m irrelevant documents:*

$$\exists m \in \mathbb{N}^{+}.\, \mathcal{M}\left(d_1^r, \ldots\right) < \mathcal{M}\left(d_1^{\neg r}, \ldots, d_m^{\neg r}, d_1^r, \ldots, d_m^r\right). \quad (1)$$

2.3 Properties for Truncated Rankings

The following properties are directly related with the truncation point decided by the system. First, adding noise at the bottom of the ranking should decrease effectiveness [8].

Property 5 (Confidence). *Appending an irrelevant document to the bottom of the ranking decreases the score:*

$$\mathcal{M}\left(d_1, .., d_n\right) > \mathcal{M}\left(d_1, .., d_n, d^{\neg r}\right). \quad (2)$$

Another aspect that affects the adequacy of the ranking truncation point is the existence or not of relevant documents in the collection. It can be assumed that the more relevant documents there are in the collection (for a given query), the more of them the ranking system should cover. For example, if a system retrieves a single document, a metric should penalize the fact that there are many more relevant documents that have not been retrieved. This aspect is reflected by the Recall property [9].

Property 6 (Recall). *The existence of non-retrieved relevant documents in the collection decreases the score. Being $\vec{d}$ a ranking containing relevant documents and $\mathcal{M}^{\mathcal{D}}(r)$ a metric applied to $\vec{d}$ and a document collection $\mathcal{D}$, and being d^r a relevant document which is not in $\mathcal{D}$:*

$$\mathcal{M}^{\mathcal{D}}\left(\vec{d}\right) > \mathcal{M}^{\mathcal{D}\cup\{d^r\}}\left(\vec{d}\right).$$

The third truncation-related aspect is the document redundancy. A relevant document found by the user should have more weight if it is the first relevant document the user finds. In a way, two relevant documents should cover the same information need, so the effectiveness of the ranking should be sensitive to this redundancy [9]. The last property reflects this aspect.

Property 7 (Redundancy). *A relevant document found by the user increases effectiveness to a greater extent if it is the first one.*

$$\mathcal{M}\left(d_1^{\neg r}, .., d_n^{\neg r}, d_{n+1}^r\right) - \mathcal{M}\left(d_1^{\neg r}, .., d_n^{\neg r}, d_{n+1}^{\neg r}\right)$$
$$> \mathcal{M}\left(d_1^{\neg r}, .., d_{n-1}^{\neg r}, d_n^r, d_{n+1}^r\right) - \mathcal{M}\left(d_1^{\neg r}, .., d_{n-1}^{\neg r}, d_n^r, d_{n+1}^{\neg r}\right).$$

There are other properties in the literature that focus on the scale on which the metric scores is defined [19, 27]. These properties affect the aggregation of scores and statistical significance tests. We leave them for future work and we focus on the effectiveness of single system outputs.

3 METRICS ANALYSIS

Table 1 shows the properties satisfied by some of the most popular metrics relevant to the truncated ranking scenario and is described in the following.

[2]The Shallowness Threshold property has been called "Closeness Threshold" by Amigó et al. [8]; we rename it to highlight its symmetry with Deepness Threshold.

Table 1: Formal properties satisfied by effectiveness metrics.

		Non-Truncated				*Truncated*		
	Metric	Priority	Top-W.	Deep. Th.	Shall. Th.	Confidence	Recall	Redundancy
full ranking metrics	Spearman	✓			✓	✓		
	Kendall	✓			✓	✓		
	AUC-ROC	✓			✓	✓		
binary relevance	P@N			✓	✓			
	R@N			✓	✓		✓	
	R-p						✓	
	RR							✓
	F-measure					✓	✓	✓
	AP	✓	✓				✓	
graded relevance	NDCG	✓	✓		✓		✓	
	Q-measure	✓	✓		✓		✓	
	BPref	✓	✓		✓		✓	
probabilistic user model based	ERR	✓	✓	✓				✓
	RBP	✓	✓	✓	✓			
	iRBU	✓	✓	✓	✓			✓
terminal document based	NDCGT	✓	✓		✓	✓	✓	
	ERRT	✓	✓	✓		✓		✓
	RBPT	✓	✓	✓	✓	✓		
utility based	Flat Utility	✓			✓	✓		
	RBU	✓	✓	✓	✓	✓		✓
	DCGU	✓	✓		✓	✓		
	ERRU	✓	✓	✓		✓		✓
	RBPU	✓	✓	✓	✓	✓		
information based	OIE	✓	✓	✓	✓	✓	✓	✓

3.1 Traditional IR Metrics

In general, as new evaluation metrics have been proposed, they have covered the formal properties described previously. Ordinal correlation coefficients such as Spearman or Kendall capture Top-Weightedness but not Deepness Threshold – since the relevance of documents have the same effect regardless of their position in the ranking. The same applies to *Area Under the ROC Curve* (AUC-ROC) which averages the effectiveness of a classification systems across threshold settings and it is equivalent to the Wilcoxon test of ranks [26].[3] We will call this family full ranking metrics. They are not appropriate metrics for the evaluation of IR systems.

The first generation of ranking metrics in IR assumed a binary document relevance. Metrics such as Precision at N (P@N), Recall at N (R@N), Reciprocal Rank (RR), and R-Precision (R-p) focus exclusively on the top of the ranking. These metrics do not satisfy

[3]In passing, we remark that Top-Weightedness is instead captured by the (area under the) Precision-Recall curve, and by the related AP metric that we discuss later.

Priority, as they are not sensitive to the relevance of documents in the intermediate ranking area. However, P@N and R@N satisfy both Deepness Threshold and Shallowness Threshold: they assume that there exists a certain deepness limit when the user explores the ranking and an area which is always explored at the top of the ranking. In the case of P@N and R@N this region is the first N documents. The metric R-Precision (R-p) combines the P@N and R@N ideas, but these two properties are no longer satisfied. Finally, RR considers only the first relevant document in the ranking, i.e., it satisfies Redundancy in the extreme case. F-measure satisfies Confidence, Recall, and Redundancy. However, the relative order of documents is not considered and none of the first four properties are captured. The AP metric solves the this drawback by averaging Precision across ranking positions (recall levels). Unlike top ranking oriented metrics, AP satisfies Priority.

The metrics in the next generation, namely, NDCG [23], Q-measure [30], and Binary Preference (BPref) [11], proposed between 2002 and 2004, capture graded relevance.[4] However, these metrics, just like AP, do not satisfy Deepness Threshold. That is, retrieving a million of relevant documents after position one million is still better than one relevant document in the first position, even though it is clear that the user will never inspect this area of the ranking.

The Deepness Threshold property is captured by the next generation of metrics (2008–2009) which incorporate a probabilistic model of the user behavior: Expected Reciprocal Rank (ERR) [14] and Rank-Biased Precision (RBP) [28]. In addition, ERR is able to capture Redundancy and RBP is able to capture Shallowness Threshold. More recently, the intentwise Rank-Biased Utility (iRBU) metric proposed by Sakai and Zeng [32] – which is a component of the diversity metric Rank-Biased Utility (RBU) [9] – combines elements from ERR and RBP, capturing the first four properties. In principle, ERR, RBP, and iRBU do not comply with Recall. However, in order to satisfy Recall, they can be normalized to the ideal ranking score in the same way as in NDCG.

3.2 Terminal Document Based Metrics

Liu et al. [25] proposed a variant on the metrics for evaluating truncated rankings. Their approach consists of adding an extra terminal document at the end of the ranking with a relevance proportional to the number of relevant documents captured throughout the ranking:

$$\texttt{Rel}(d_{n+1}) = \begin{cases} 1 & \text{if } R = 0 \\ \sum_{i=1}^{n} \frac{\texttt{Rel}(d_i)}{R} & \text{otherwise,} \end{cases}$$

where R represents the document relevance sum in the collection. The idea is that, the more the terminal document appears later in the ranking, the more the discount function penalizes the truncated ranking. The authors experimented with several metrics such as NDCG, RR, and RBP, thus defining their truncated variants NDCGT, RRT, and RBPT. They investigated the impact of truncation on performance of individual systems.

The terminal document metric variants inherit the formal properties of the original metric, but also capture Confidence (see Table 1). However, Liu et al. observed that, despite the aggressive truncation, there are systems that obtain little improvement. We corroborate this effect in our experiments.

3.3 Utility-based Metrics

Utility-based metrics are based on the idea of progressively both: (i) adding *gain* (depending on whether we find relevant or non-relevant documents in the ranking) and (ii) subtracting *cost* (represented as a constant value e that should be consistent with the relevance gain $\texttt{Rel}(d_i)$). The simplest implementation of this idea is the Flat Utility metric: $\mathrm{U}(\vec{d}) = \sum_{k=1}^{n} (\texttt{Rel}(d_i) - e)$. Utility metrics are applied in classification and filtering problems. However, this metric scheme does not take top-heaviness into account (and therefore does not satisfy Top-Weightedness).

The Rank-Biased Utility metric (RBU) is similar to iRBU but incorporates the e factor into ranking evaluation metrics [9]. The e factor is also affected by ranking position discount. This approach is applicable to any metric that can be expressed as the sum across ranking positions i of a gain function $\mathrm{Gain}(d_i)$ and a ranking position discount $f(i)$:

$$\mathcal{M}\left(\vec{d}\right) = \sum_{i=1..n} (\mathrm{Gain}(d_i) - e) \cdot f(i).$$

In this paper, we propose to apply the same scheme to different metrics, obtaining the following family:

$$\mathrm{RBU}(\vec{d}) = (1-p) \sum_{i=1}^{n} p^{i-1} \left(\texttt{Rel}(d_i) \prod_{j=1}^{i-1} (1 - \texttt{Rel}(d_j)) - e \right)$$

$$\mathrm{DCGU}(\vec{d}) = \sum_{i=1}^{n} \frac{\texttt{Rel}(d_i) - e}{\log_2(i+1)}$$

$$\mathrm{ERRU}(\vec{d}) = \sum_{i=1}^{n} \frac{1}{i} \left(\texttt{Rel}(d_i) \prod_{j=1}^{i-1} (1 - \texttt{Rel}(d_j)) - e \right)$$

$$\mathrm{RBPU}(\vec{d}) = (1-p) \sum_{k=1}^{n} (\texttt{Rel}(d_k) - e) p^{k-1}$$

(where n is the truncated ranking length).

The new proposed version of the metrics satisfies Confidence, and the other formal properties satisfied by each metric are directly inherited from the corresponding non effort-oriented version (see Table 1). Unfortunately, since DCGU can assume negative values, it cannot be normalized as in NDCG. Therefore, Recall is not captured anymore. In our experiments we will fix $e = \frac{1}{20}$. This e value setting means that the benefit of finding a relevant document in the first position is equal to the effort of inspecting the first 20 documents.

In order to better understand the behavior of utility based metrics, let us consider a ranking that returns irrelevant documents in positions 1–10 and relevant documents from position 11 onwards. Figure 1 shows the metric score (vertical axis) with $e = \frac{1}{20}$ at each ranking truncation (horizontal axis) of this. For the sake of visibility, the obtained metric scores are scaled between -1 and 1, keeping the zero value invariant.[5] Note that the focus of this analysis is the

[4] Note that, although originally BPref was not proposed for graded relevance, it does accept it.

[5] To scale the scores between -1 and 1, we divide by the maximum absolute value of the truncated rankings scores: $f(x) = \frac{x}{\max(|\text{score}|)}$.

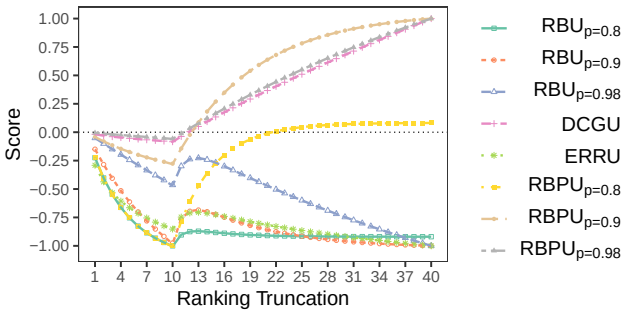

Figure 1: Utility based metric scores (vertical axis) with $e = \frac{1}{20}$ at each ranking truncation (horizontal axis) for a ranking containing relevant documents after rank position 10. Scores are normalized w.r.t. the highest score magnitude across ranking positions. Note that some curves overlap, e.g., RBPU$_{p=0.8}$ with RBU$_{p=0.8}$ in $[1..10]$ or RBPU$_{p=0.98}$ with DCGU.

horizontal line Score = 0 – which represents the boundary between effort penalty and gain – as well as the ratio between scores at different positions: neither are affected by normalization.

Throughout the first 10 positions, all metrics penalize the effort required to explore these 10 irrelevant documents. This penalty is higher in metrics that assume less patience on the part of the user (ERRU, RBPU$_{p=0.8}$, and the RBU variants). For all of them apart from RBU$_{p=0.98}$, scanning 10 irrelevant documents is never compensated by what can be found beyond position 10. The penalty of irrelevant documents at the first positions is lower for metrics that assume more patience on the part of the user, i.e., RBU$_{p=0.98}$, DCGU, RBPU$_{p=0.9}$ and RBPU$_{p=0.98}$. In addition, the metric RBU$_{p=0.98}$ reflects a progressive decrease in effectiveness after position 15 due to the effect of Redundancy, since it uses the same gain function as ERR. The behavior of DCGU is very similar to RBPU$_{p=0.9}$ in this scenario.

3.4 Information-based Metrics

The metric Observational Information Effectiveness (OIE) [5] can be shown to be adequate to evaluate truncated rankings. OIE aims to quantify via the Observational Information Quantity (OIQ) [22] how much information the system provides about the documents in the ranking (i.e., system output), how much information the ground truth provides (amount of relevant documents in the collection), and the information they jointly generate. The more the system output and the ground truth are dissimilar, the more their joint information is large.

3.4.1 Observational Information Quantity (OIQ). Quantifying the informativeness of IR system outputs and human assessments from an information theory perspective requires a framework to measure the specificity of documents in terms of ranking position and human relevance judgments in the ground truth. OIE is based on the notion of OIQ which particularizes the Shannon's Information Content (IC) to non-discrete random variables. Given a continuous random variable X, the OIQ of a random outcome x is the minus logarithm of the survival function:

$$\mathrm{OIQ}(x) = -\log(P(X \geq x)).$$

That is, the larger the outcome x, the higher is its OIQ. In our scenario, let x be the relevance of a certain document d, and let X be the relevance distribution of documents in the collection. Then, the higher the relevance x of the document d, the higher is its OIQ. The OIQ of the system output is analogous but considering the relevance estimated by the system. OIQ can be extended to bivariate with random variables X and Y:

$$\mathrm{OIQ}(x, y) = -\log(P(X \geq x, Y \geq y)).$$

Therefore, we can estimate the OIQ of a document regarding both the system output and the ground truth. Then, just like in traditional information theory, the Observational Entropy is the expected OIQ of values in the distribution: $H(X) = E[\mathrm{OIQ}(X)]$ and $H(X, Y) = E[\mathrm{OIQ}(X, Y)]$.

For evaluation purposes, documents can be interpreted as observation outcomes of two random variables $\Gamma = \{S, G\}$, i.e., their relevance according to the system and to the ground truth, respectively. The Observational Entropy $H(S)$ of the document collection $\mathcal{D}$ under S represents how much information the system provides. $H(G)$ represents how much information the ground truth provides, and $H(S, G)$ represents how much information they provide together. These probabilistic notions can be approximated as follows:

$$H(S) = -\frac{1}{|\mathcal{D}|}\sum_{d\in\mathcal{D}}\log\left(\frac{|\{d' : \mathrm{Rank}(d') \leq \mathrm{Rank}(d)\}|}{|\mathcal{D}|}\right)$$

$$H(G) = -\frac{1}{|\mathcal{D}|}\sum_{d\in\mathcal{D}}\log\left(\frac{|\{d' : \mathtt{Rel}(d') \geq \mathtt{Rel}(d)\}|}{|\mathcal{D}|}\right)$$

$$H(S,G) = -\frac{1}{|\mathcal{D}|}\sum_{d\in\mathcal{D}}\log\left(\frac{\left|\left\{\begin{array}{c} d' : \mathrm{Rank}(d') \leq \mathrm{Rank}(d) \\ \wedge \mathtt{Rel}(d') \geq \mathtt{Rel}(d) \end{array}\right\}\right|}{|\mathcal{D}|}\right).$$

Note that, although the constant factor $\frac{1}{|\mathcal{D}|}$ can be removed, the collection size $|\mathcal{D}|$ is still a free parameter.

3.4.2 Observational Information Effectiveness (OIE). OIE [5] is based on the Information Contrast Model similarity measure [4]. The OIE of a system output random variable S is defined as a combination of the entropy in the system output S, the entropy in the ground truth G, and their joint observational entropy:

$$\mathrm{OIE}_G(S) = H(S) + H(G) - \beta H(S, G).$$

Note that the more S and G are dissimilar, the more $H(S, G)$ is large. Amigó et al. [5] prove that OIE satisfies the first five properties described in Section 2 – including Confidence – whenever $1 < \beta < \frac{2m-1}{m}$, being m the minimum amount of documents that are necessarily explored by the user as defined in the Shallowness Threshold property. In particular OIE penalizes non-truncated rankings since both $H(S)$ and $H(S, G)$ grow to the same extent with the amount of irrelevant documents at the end of the ranking. Therefore, as long as $\beta > 1$, OIE satisfies Confidence. In addition, both $H(G)$ and $H(S, G)$ grow to the same extent with the amount of non-retrieved relevant documents in the collection. Therefore, as long as $\beta > 1$, OIE satisfies Recall. Redundancy is also captured

since the logarithm component of OIQ makes it to decrease more slowly as the document occupies deeper positions in the ranking.

There are two parameters to set when using OIE in practice: $|\mathcal{D}|$ and β. According to our experiments, the document collection size $|\mathcal{D}|$ does not affect substantially OIE computation: we use in our experiments $|\mathcal{D}| = 20,000$. We set $\beta = 1.05$: this is the parameter value for which retrieving the unique relevant document in the collection at position 20 is equivalent to returning nothing. We leave for future work the calibration of β for different scenarios.

4 EXPERIMENTAL ANALYSES

We now analyze empirically the behavior of metrics. We emulate systems providing truncated rankings by truncating real system outputs in traditional datasets accordingly to each truncating ranking metric. In the first experiment, we study the distribution of optimal truncation positions. In the second experiment, we compare truncated ranking effectiveness metrics among themselves, as well as with traditional full ranking effectiveness metrics. In the third experiment, we study the ability of metrics to capture other quality aspects, such as Recall, Redundancy, Deepness Threshold, and Shallowness Threshold.

4.1 Experimental Setup

Metrics. We experiment with Flat Utility, RBU, ERRU, and RBPU with $e = \frac{1}{20}$. We also include the terminal document versions NCDGT, ERRT, and RBPT, the traditional metric F-measure, and OIE. For RBU and ERR derived metrics, we normalize the document relevance scale as indicated by Chapelle et al. [14]: $\texttt{Rel}'(d) = \frac{2^{\texttt{Rel}(d)}-1}{2^{\text{MaxRel}}}$.

Datasets. We selected as datasets to be used in our experiments the following nine ad hoc retrieval tracks from TREC (all of them can be downloaded from http://trec.nist.gov/): Robust'04 (R04), TREC 4, TREC 7, TREC 8, TREC 11, TREC 12, TREC 13, TREC 14, and TeraByte'06 (TB06). We made this choice to have diversified test collections: they vary in terms of the number of topics used in the track (from 50 to 249) and of the number of systems participating to the track (from 33 to 129); also, they span more than 10 years and are quite standard in the evaluation community. We keep the graded relevance of documents in the ground truth.

4.2 Optimal Truncation Points

Given a full ranking and a metric, we find the truncation point of the ranking that maximizes the value of the metric. This truncation point would be the best choice for a system being evaluated with such a metric. Then, we study for each metric the distribution of optimal truncation points. We use log binning and consider exponential growth intervals $[2^i + 1, 2^{i+1}]$ of optimal truncation points. That is, empty ranking, and then the intervals $[1, 4]$ (in which we collapse the two $[1, 2]$ and $[3, 4]$ intervals), $[5, 8]$, $[9, 16]$, etc. For each dataset, we study the percentage of cases for which the optimal ranking truncation point lies in each interval. The percentage results are averaged across the nine datasets.

In Figure 2, the horizontal axis shows the ranking position intervals, and the vertical axis shows the percentage of cases. A first interesting overall result that can be seen by looking at the three

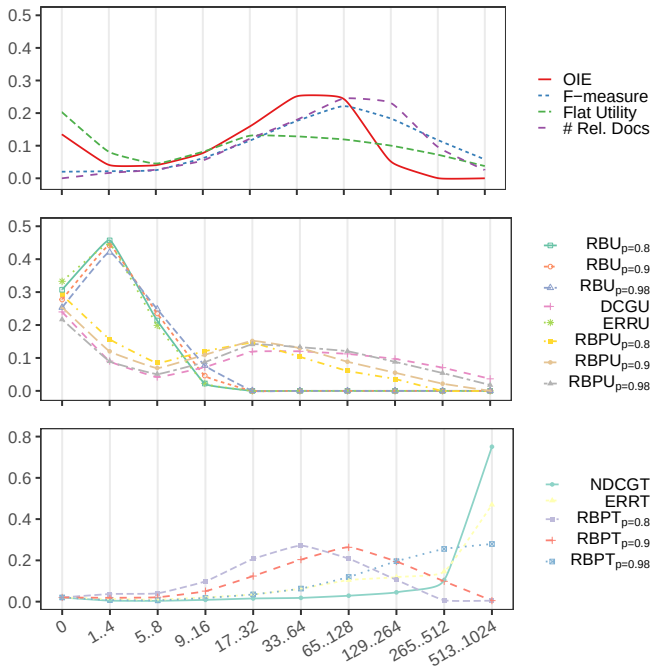

Figure 2: The distributions of optimal ranking lengths according to different truncation metrics. Results are averaged across the nine datasets.

plots is that for most metrics, with only three exceptions, the optimal truncation points tend to occur quite early in the rankings: for many cases (about 50%), the optimal truncation point is well before position 100. Also note that all the three exceptions are terminal document metrics that, as discussed above, do not modify the evaluation in a significant way, and lead to a weak truncation. In other terms, whatever truncation metric one uses, the systems should truncate their outputs in a rather substantial way. This is of course a very different scenario from the classical TREC-like setting, in which systems are required to retrieve a fixed amount of documents (usually 1,000).

Let us now analyze in how many cases returning nothing is the best option (i.e., the optimal ranking length is 0) according to each of the metrics. According to Flat Utility, and all the utility based metrics RBU, RBPU, DCGU, and ERRU (see the first two plots), this happens for a percentage between 20% and 30% of cases; this percentage changes depending on the value of the p parameter in RBPU and RBU. That is, if system outputs are evaluated by utility based metrics which accumulate user effort, then in many cases returning no documents is the best option, i.e., truncation is very strong. On the contrary, this is almost never the case for the terminal document based metrics (third plot). Indeed, in very few cases for these metrics the optimal ranking length is smaller than 10, i.e., truncation is much weaker. As discussed above, terminal document based metrics are not much sensitive to the effect of retrieving irrelevant documents. F-measure also never completely truncates rankings since zero recall would mean zero F-measure (which is the

harmonic mean of precision and recall). OIE has an intermediate behavior, completely leaving out about 10% of the rankings.

Let us analyze now how the optimal truncation is distributed across ranking positions. ERRU and RBU tend to reward strong ranking truncations (30% of cases between positions 1 and 4) due to the effect of Redundancy. This effect can be mitigated by smoothing the relevance probability scale $\left(\texttt{Rel}'(d) = \frac{2^{\texttt{Rel}(d)}-1}{2^{\text{MaxRel}}}\right)$. At the other extreme, the terminal document based metrics ERRT and NDCGT tend to reward weak truncations, or even no truncation at all. The optimal truncation of RBPT is highly sensitive to the p parameter. OIE concentrates most of optimal ranking lengths between positions 33 and 128, and F-measure between positions 33 and 264. On the other hand, Flat Utility and the RBPU utility based variants show a higher variability of optimal ranking lengths. A possible reason is due to both OIE and F-measure satisfying the Recall property. This explains why their distributions keep a certain correspondence with the distribution of the number of relevant documents per topic (see purple line in the first plot).

4.3 Truncated vs. Full Ranking Effectiveness

The second experiment aims to study to what extent the ranking truncation affects the relative score of systems. We firstly generate truncated rankings by finding the optimal length according to a truncated metric (see Section 4.2). By doing so, we are in practice using an oracle to simulate a situation in which a system participating in TREC has learned where to truncate each output in an optimal way. Then, we compare the truncated system effectiveness with the effectiveness of the original systems. We do so by computing Kendall's τ correlation between the effectiveness of the truncated rankings and the effectiveness of the original TREC systems, with full rankings. τ is computed for each dataset and then the values are averaged over the nine datasets. Intuitively, we aim to metrics to capture some additional signal after truncation – as they satisfy at least some of the truncated ranking formal properties discussed in Section 2 – without overly diverging from the scores obtained by the standard metrics for TREC, such as AP or NDCG. Generally speaking, we aim at a sort of sweet spot. On one hand, a too high / perfect correlation shows that the truncation has no effect in the ways the compared metrics are ranking the systems. On the other hand, a too low correlation means that effect is too strong (e.g., because the optimal truncated point is too aggressive).

Table 2 shows the results. The columns correspond with truncated ranking metrics, on which basis the TREC systems are truncated. The rows correspond with the corresponding full ranking metrics – plus some traditional metrics – applied to the full ranking. Note that terminal document and utility based truncated ranking metrics converge to the original metric when evaluating full rankings. For instance, RBPT and RBPU applied to full rankings are perfectly correlated with RBP, as well as RBU with iRBU. We include in the table 11 full ranking metrics and 14 truncated ranking metrics. The value for each metric pair is obtained by averaging 9 τ values, one for each dataset. Out of all of the $11 \times 14 \times 9 = 1,386$ τ values computed, 1,205 (87%) are statistically significant at the $p = .001$ level, and 1,250 (90%) at the $p = .01$ level.

Five main results can be highlighted. The first one, that corroborates the previous experiment, is that metrics based on terminal documents are not very sensitive to the ranking truncation effect: NDCGT, ERRT, and RBPT have a perfect correlation (in bold in the table) with the effectiveness of full ranking according to the corresponding traditional metric, i.e., NDCG, ERR, and RBP.

The second result, which also corroborates the previous experiment, is that truncating according to $\text{RBU}_{p=0.9}$ and $\text{RBU}_{p=0.98}$ somehow alters the relative scores of the systems: the truncated rankings have a slightly higher correlation with metrics such as ERR or $\text{iRBU}_{p=0.8}$, which are more top-heavy, than with the corresponding full ranking metrics $\text{iRBU}_{p=0.9}$ and $\text{iRBU}_{p=0.98}$. The reason is that, due to the effect of the Redundancy property, they tend to perform too strong truncations to the rankings (see also Figure 2).

The third result is that the effectiveness of systems truncated according to the utility based metrics (DCGU, ERRU, RBPU with $p = 0.8$, 0.9, and 0.98) somehow correlate with their measurement over full ranking. This suggests that they are capturing the effectiveness of the original full ranking but also the truncation effect.

The fourth result concerns OIE. In general, as we have seen above, the measurement of each metric over truncated rankings correlates with its full ranking counterpart. This does not hold for OIE, for which the relative effectiveness of systems changes substantially after truncation. We hypothesize that OIE effectiveness measurement is degraded when evaluating a full ranking due to the imbalance between the amount of information provided by the system and that provided by the ground truth. In fact, OIE over the full ranking is low correlated with all truncated metrics in general (see last row of the table).

Given that OIE is the only metric satisfying all theoretical properties, it deserves a more detailed analysis, which leads to the fifth result. Figure 3 shows some scatterplots that provide further insights. Each dot is a system participating in a TREC edition as indicated by the color code. The x-axis is effectiveness according to the indicated metric and the y-axis is effectiveness of the truncated ranking according to the indicated metric. The first scatterplot shows the low correlation between the truncated and full rankings according to OIE. The second scatterplot shows the relation between OIE over truncated rankings and full ranking $\text{RBP}_{p=0.98}$: the correlation is very high within each collection; the differences across collections are due to OIE being able to capture some collection features (such as the number of relevant documents in the ground truth). The third scatterplot compares instead two truncated ranking metrics and shows the high correlations of OIE and $\text{RBPU}_{p=0.98}$: both metrics measure in a similar way. Finally, the fourth scatterplot shows the high correlations of OIE over truncated rankings with full ranking AP. In fact, both $\text{RBPU}_{p=0.98}$ and, especially, OIE over truncated rankings have rather high correlations with the measurement of AP and NDCG over the full rankings (see the four values in bold on the top-right of the table: these correlation values are around, if not above, the 0.8 threshold that has traditionally been considered in metric comparison studies [38], although both $\text{RBPU}_{p=0.98}$ and OIE show a rather strong truncation in Figure 2. Even though OIE truncated does not present a high correlation with OIE over the full ranking, it seems anyway an adequate truncated ranking metric.

To summarize, these data confirm the weak effect of the terminal document based metrics; conversely, both OIE and $\text{RBPU}_{p=0.98}$ look to be adequate metrics for truncated ranking.

Table 2: Averages (across the nine datasets) of Kendall's τ correlation between the effectiveness of the full ranking according to the original metrics (rows) and the effectiveness of the rankings truncated at the optimal cutoff according to truncated ranking metrics (columns).

	Truncated Rankings													
Full	NDCGT	ERRT	RBPT			RBU			DCGU	ERRU	RBPU			OIE
Rankings			$p=0.8$	$p=0.9$	$p=0.98$	$p=0.8$	$p=0.9$	$p=0.98$			$p=0.8$	$p=0.9$	$p=0.98$	
AP	0.89	0.46	0.61	0.69	**0.85**	0.46	0.47	0.47	**0.80**	0.44	0.61	0.68	**0.79**	**0.87**
NDCG	**1.00**	0.49	0.62	0.68	**0.83**	0.48	0.48	0.49	**0.79**	0.46	0.61	0.66	**0.76**	**0.83**
DCG	0.92	0.48	0.62	0.69	0.84	0.48	0.48	0.49	**0.83**	0.46	0.61	0.67	0.77	0.80
ERR	0.48	**1.00**	0.78	0.70	0.53	**0.92**	**0.90**	**0.85**	0.58	**0.92**	0.77	0.69	0.56	0.52
$RBP_{p=0.8}$	0.62	0.78	**1.00**	0.89	0.69	0.80	0.80	0.79	0.73	0.76	**0.96**	0.89	0.73	0.67
$RBP_{p=0.9}$	0.68	0.69	0.89	**1.00**	0.78	0.71	0.72	0.72	0.80	0.66	0.87	**0.95**	0.83	0.77
$RBP_{p=0.98}$	0.83	0.53	0.69	0.78	**1.00**	0.52	0.53	0.55	**0.84**	0.50	0.68	0.76	**0.91**	**0.91**
$iRBU_{p=0.8}$	0.50	0.91	0.82	0.74	0.55	**0.91**	**0.94**	**0.92**	0.59	0.84	0.80	0.73	0.58	0.54
$iRBU_{p=0.9}$	0.51	0.84	0.79	0.74	0.57	0.83	0.87	0.89	0.59	0.76	0.76	0.72	0.59	0.56
$iRBU_{p=0.98}$	0.57	0.69	0.66	0.69	0.63	0.67	0.71	0.74	0.60	0.62	0.63	0.65	0.61	0.61
OIE	0.58	0.36	0.48	0.54	0.65	0.36	0.36	0.37	0.54	0.34	0.47	0.52	0.60	0.67

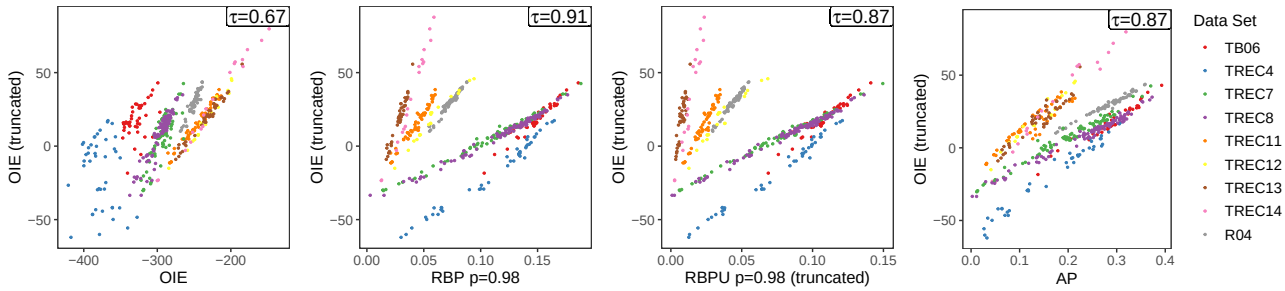

Figure 3: System effectiveness obtained according to evaluation metrics applied to full rankings (x-axis) and rankings truncated with the optimal truncation point (y-axis). The average of Kendall's τ values across collections is reported.

Table 3: $\text{Cov}_{\mathbb{M}}(\mathcal{M})$ for ranking effectiveness metrics with truncation at position 100 with respect to P@10, R@100, and RR.

OIE	NDCG	DCG	$RBP_{p=0.98}$	AP	F-measure	R@100	Flat Utility	$RBP_{p=0.9}$	P@10	$RBP_{p=0.8}$	$iRBU_{p=0.98}$	$iRBU_{p=0.9}$	$iRBU_{p=0.8}$	ERR	RR
0.921	0.918	0.911	0.911	0.901	0.896	0.891	0.891	0.890	0.867	0.846	0.832	0.810	0.782	0.761	0.249

4.4 Capturing Other Effectiveness Aspects

The next aspect to consider is the extent to which, under a fixed ranking length (e.g., effectiveness at 100), truncated ranking metrics capture different effectiveness features. To study this, we consider three traditional metrics that are not designed for truncated ranking evaluation and capture four basic and orthogonal ranking quality aspects: (i) P@10, which reflects the relevance of documents at the top of ranking (capturing Deepness Threshold and Shallowness Threshold properties); (ii) R@100 which captures the coverage of the relevant documents in the collection (Recall property); and (iii) RR, which is an extreme Redundancy avoiding based metric. Then, we study the ability of truncated ranking metrics to capture system improvements reported by these three metrics.

For this purpose, we apply the same method as Amigó et al. [7], which is based on the Unanimous Improvement Ratio (UIR) [6]. While other meta-evaluation metrics such as *robustness* [31] focus on consistency across datasets, UIR focuses on consistency across metrics. It essentially counts in how many test cases an improvement is captured by all metrics simultaneously. Being $\mathbb{M}$ a set of metrics, $\mathcal{T}$ a set of test cases or topics, and s_t a system output for the test case $t \in \mathcal{T}$, the Unanimous Improvement Ratio $\text{UIR}_{\mathbb{M}}(s, s')$ between two systems s and s' is defined as the difference of two probabilities:

$$\text{UIR}_{\mathbb{M}}(s, s') = P(s \geq_{\mathbb{M}} s') - P(s' \geq_{\mathbb{M}} s)$$

where $s \geq_{\mathbb{M}} s'$ represents that system s improves system s' unanimously for every metric. In this paper, we estimate these probabilities as follows. Being $\mathcal{T}^2$ a set of 1,000 test case (topic)

pairs randomly selected from a dataset, $\mathrm{UIR}_{\mathbb{M}}(s, s')$ is computed as:

$$\mathrm{UIR}_{\mathbb{M}}(s, s') \simeq \frac{\left|\left\{(t_1,t_2)\in\mathcal{T}^2 : s_{t_1} \geq_{\mathbb{M}} s'_{t_2}\right\}\right| - \left|\left\{(t_1,t_2)\in\mathcal{T}^2 : s'_{t_2} \geq_{\mathbb{M}} s_{t_1}\right\}\right|}{1000}.$$

UIR reflects to what extent a system outperforms another system for several metrics simultaneously. We compare rankings across different topics in order to capture quality aspects that are topic-dependent such as, for instance, Recall. The main strength of UIR is that it is non-parametric. That is, it does not assign any arbitrary relative weight to single metrics.

Then, we define our meta-evaluation measure *Quality Aspect Coverage* (Cov) for a single metric $\mathcal{M}$ as the Spearman's ρ correlation (over system output pairs s, s' in the set of system outputs) between differences in $\mathcal{M}$ and unanimous improvements over the reference metric set. Being $\mathbb{M}$ the reference metric set:[6]

$$\mathrm{Cov}_{\mathbb{M}}(\mathcal{M}) = \rho\left(\mathcal{M}(s) - \mathcal{M}(s'), \mathrm{UIR}_{\mathbb{M}}(s, s')\right).$$

The higher the coverage of a metric $\mathcal{M}$ w.r.t. a reference metric set $\mathbb{M}$, the more an improvement according to $\mathcal{M}$ reflects all quality aspects represented by $\mathcal{M}$.

Table 3 shows the Cov scores for truncated and non-truncated metrics. Recall-oriented metrics with top-heaviness such as OIE and NDCG achieve the highest Cov scores. This indicates that Recall is a particular quality aspect that is not captured by other metrics. On the contrary, metrics that penalize redundancy such as RBU, ERR, and RR achieve low Cov scores. This indicates that Redundancy is not an orthogonal quality aspect, being captured by the rest of metrics. In general, parameterizations oriented to deeper ranking positions (i.e., $p = 0.98$) cover more quality aspects than top-oriented metric variants (i.e., $p = 0.8$ or P@10).

5 A GUIDELINE FOR METRIC SELECTION

We have seen that different effectiveness evaluation rankings satisfy different formal properties (Section 3) and cover different quality aspects (Section 4.4). In order to help IR researchers and practitioners understand which metric to use – depending on the characteristics of the specific ranking problem they are aiming to study – we propose the guideline summarized in Figure 4.

The first question to ask is whether the system is required to choose the most appropriate ranking length. If not, the next question is whether the probability of the documents to be accessed by the user drops at lower positions in the ranking (Top-Weightedness). In the event that all displayed documents have the same probability of being inspected by the user, systems retrieve the same set of documents and only order matters: in that case, classical ranking correlation coefficients such as Spearman and Kendall are sufficient.

If we can assume that the user necessarily accesses the first N and no more, we can apply metrics such as P@N or R@N. Otherwise, it is possible that the user will only need to find one of the relevant documents to satisfy his/her information need, in which case an appropriate metric would be RR. Otherwise, we can distinguish between situations where the amount of relevant documents in the collection affects the ranking quality (Recall) or cases where each relevant document brings us gain regardless the amount of relevant

[6]Using the non-parametric coefficient Spearman instead of Pearson focuses the meta-evaluation on system score ordering rather than particular scale properties of metrics.

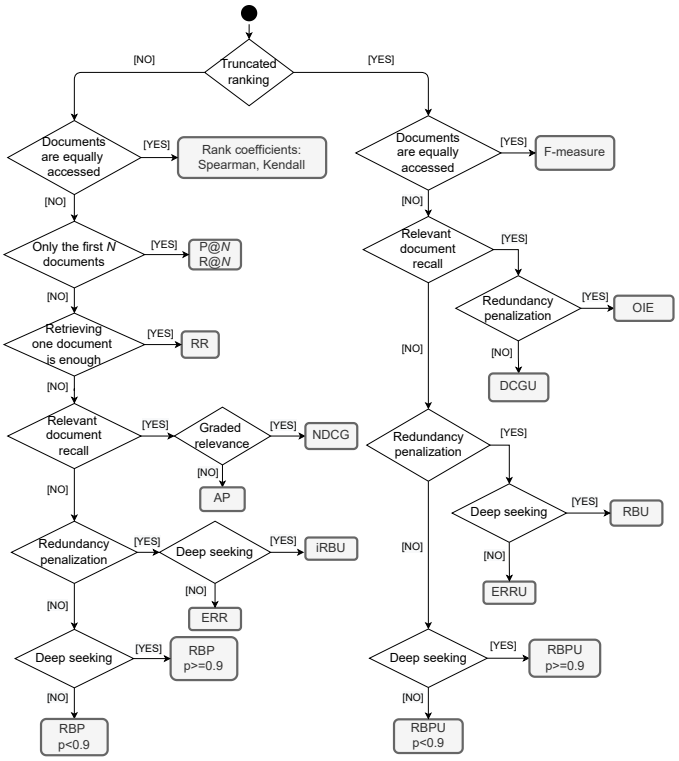

Figure 4: Workflow for metric selection in different ranking problems.

documents in the collection. In the first case we would use AP or NDCG, depending on whether the we deal with binary or graded relevance levels. In the second case, the next question is whether we want to consider the effect of Redundancy. That is, having previously found relevant documents reduces the gain obtained with newly found relevant documents. Appropriate metrics for this case would be ERR or iRBU, depending on the depth with which the user explores the ranking. Otherwise an appropriate metric would be RBP, with a parameter p depending on the expected ranking exploration depth.

Going back to the first question, suppose we are interested in evaluating the ranking length determined by the system (Confidence). In that case, if all ranking positions are accessed with the same probability, traditional classification metrics such as F-measure would be appropriate. Again, the questions to ask are whether the system should cover a representative set of relevant documents (recall-oriented) and whether we want to penalize the redundancy that relevant documents have with each other. RBPU does not consider Recall nor Redundancy, and its parameter will depend on the degree of depth of exploration. If we are not interested in Recall, but we are interested in Redundancy, then we would use ERRU or RBU depending on the expected user exploration depth. If we are interested in Recall but not Redundancy, we would then choose DCGU. Finally, if we want to capture the optimal cutoff (i.e., we want both Recall and Redundancy), then we can use OIE – a metric based on information theory.

6 CONCLUSIONS AND FUTURE WORK

We have addressed the issue of evaluating truncated rankings. In Section 2 we have proposed a set of formal properties that should be addressed by adequate effectiveness metrics. In our review of the state of the art and our theoretical study, we have grouped the truncated ranking metrics into three families: terminal document based, utility based, and Information Theory based. We have extended the notion of effort coefficient with ranking decay function applied in RBU to other standard metrics such as DCG, ERR, and RBP. Our theoretical analysis on seven formal properties shows that the Information Theory based OIE satisfies all of them simultaneously (Section 3). However, not all properties are necessary or desirable in any scenario. Therefore, we have included a general guideline on the basis of formal properties for metric selection (Section 5). The generalization of metrics for truncated rankings has allowed us to acquire a more general view in this sense.

As for our empirical results, we have observed that the optimal truncation point can vary depending on the chosen metric and parameters. In general, we have observed that the utility based metrics RBPU and DCGU and the information theory based metric OIE behave appropriately. On the other hand, our experiments have shown us that the OIE metric can degenerate when applied to full rankings, so it seems in principle only suitable for truncated rankings. To our knowledge, this is the first empirical study with this type of metrics.

Overall, this paper is a step forward towards a better understanding of truncated ranking metrics. In particular, we believe our theoretical and empirical analyses help identifying metrics that are useful for tasks where is important to assess the ability of systems to stop giving information to users – i.e., the way the system estimates at which point there is no more useful information to deliver to the user. This is the case of scenarios where inspecting irrelevant information is particularly costly – due to the nature of the communication channel or the task – as in mobile search on a small screen and with distracting surroundings [18], conversational search [10, 24, 36, 37, 41], or product search over voice assistants [12].

The limitations of this work still leave many facets to be understood about truncated rankings. In our experiments, we generated truncated rankings applying an oracle for each truncated metric. We plan to perform experiments with more realistic settings (e.g., dynamic search [1, 2]). In our analyses we focused on the most interesting correlations among metrics, but there are more to be studied. The effort factor e used by utility based metrics can be calibrated with user studies by following a similar methodology as proposed by Smucker and Clarke [34]. Finally, user studies would also shed light upon understanding the role of truncated rankings in the context of user's preferences and satisfaction.

ACKNOWLEDGMENTS

This work is partially supported by the Spanish Ministry of Economic Affairs and Digital Transformation (ref. C039/21-OT) and the Australian Research Council (DE200100064 and CE200100005).

REFERENCES

[1] Ameer Albahem, Damiano Spina, Falk Scholer, and Lawrence Cavedon. 2019. Meta-Evaluation of Dynamic Search: How Do Metrics Capture Topical Relevance, Diversity and User Effort?. In *Proceedings of the 41st European Conference on Information Retrieval (ECIR '19)*. Springer International Publishing, Cham, 607–620. https://doi.org/10.1007/978-3-030-15712-8_39

[2] Ameer Albahem, Damiano Spina, Falk Scholer, and Lawrence Cavedon. 2021. Component-Based Analysis of Dynamic Search Performance. *ACM Trans. Inf. Syst.* 40, 3, Article 61 (nov 2021), 47 pages. https://doi.org/10.1145/3483237

[3] Enrique Amigó, Hui Fang, Stefano Mizzaro, and ChengXiang Zhai. 2018. Are We on the Right Track? An Examination of Information Retrieval Methodologies. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 997–1000. https://doi.org/10.1145/3209978.3210131

[4] Enrique Amigó, Fernando Giner, Julio Gonzalo, and Felisa Verdejo. 2020. On the foundations of similarity in information access. *Inf. Retr. J.* 23, 3 (2020), 216–254. https://doi.org/10.1007/s10791-020-09375-z

[5] Enrique Amigó, Fernando Giner, Stefano Mizzaro, and Damiano Spina. 2018. A Formal Account of Effectiveness Evaluation and Ranking Fusion. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '18)*. Association for Computing Machinery, New York, NY, USA, 123–130. https://doi.org/10.1145/3234944.3234958

[6] Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2011. Combining Evaluation Metrics via the Unanimous Improvement Ratio and Its Application to Clustering Tasks. *J. Artif. Int. Res.* 42, 1 (sep 2011), 689–718. https://doi.org/10.5555/2208436.2208454

[7] Enrique Amigó, Julio Gonzalo, Stefano Mizzaro, and Jorge Carrillo-de Albornoz. 2020. An Effectiveness Metric for Ordinal Classification: Formal Properties and Experimental Results. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL '20)*. Association for Computational Linguistics, Online, 3938–3949. https://doi.org/10.18653/v1/2020.acl-main.363

[8] Enrique Amigó, Julio Gonzalo, and Felisa Verdejo. 2013. A General Evaluation Measure for Document Organization Tasks. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13)*. Association for Computing Machinery, New York, NY, USA, 643–652. https://doi.org/10.1145/2484028.2484081

[9] Enrique Amigó, Damiano Spina, and Jorge Carrillo-de Albornoz. 2018. An Axiomatic Analysis of Diversity Evaluation Metrics: Introducing the Rank-Biased Utility Metric. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 625–634. https://doi.org/10.1145/3209978.3210024

[10] Avishek Anand, Lawrence Cavedon, Hideo Joho, Mark Sanderson, and Benno Stein. 2020. Conversational Search (Dagstuhl Seminar 19461). *Dagstuhl Reports* 9, 11 (2020), 34–83. https://doi.org/10.4230/DagRep.9.11.34

[11] Chris Buckley and Ellen M. Voorhees. 2004. Retrieval Evaluation with Incomplete Information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04)*. Association for Computing Machinery, New York, NY, USA, 25–32. https://doi.org/10.1145/1008992.1009000

[12] David Carmel, Elad Haramaty, Arnon Lazerson, Liane Lewin-Eytan, and Yoelle Maarek. 2020. Why Do People Buy Seemingly Irrelevant Items in Voice Product Search? On the Relation between Product Relevance and Customer Satisfaction in ECommerce. In *Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM '20)*. Association for Computing Machinery, New York, NY, USA, 79–87. https://doi.org/10.1145/3336191.3371780

[13] Ben Carterette, Paul N. Bennett, David Maxwell Chickering, and Susan T. Dumais. 2008. Here or There: Preference Judgments for Relevance. In *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval (ECIR'08)*. Springer-Verlag, Berlin, Heidelberg, 16–27. https://doi.org/10.5555/1793274.1793281

[14] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09)*. Association for Computing Machinery, New York, NY, USA, 621–630. https://doi.org/10.1145/1645953.1646033

[15] Charles L.A. Clarke, Chengxi Luo, and Mark D. Smucker. 2021. Evaluation Measures Based on Preference Graphs. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 1534–1543. https://doi.org/10.1145/3404835.3462947

[16] Charles L.A. Clarke, Mark D. Smucker, and Alexandra Vtyurina. 2020. Offline Evaluation by Maximum Similarity to an Ideal Ranking. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 225–234. https://doi.org/10.1145/3340531.3411915

[17] Cyril Cleverdon. 1967. The Cranfield Tests on Index Language Devices. In *Aslib proceedings*, Vol. 19. MCB UP Ltd, Emerald Publishing Ltd, 173–194. https:

//doi.org/10.1108/eb050097

[18] Fabio Crestani, Stefano Mizzaro, and Ivan Scagnetto. 2017. *Mobile Information Retrieval*. Springer. https://doi.org/10.1007/978-3-319-60777-1

[19] Marco Ferrante, Nicola Ferro, and Norbert Fuhr. 2021. Towards Meaningful Statements in IR Evaluation: Mapping Evaluation Measures to Interval Scales. *IEEE Access* 9 (2021), 136182–136216. https://doi.org/10.1109/ACCESS.2021.3116857

[20] Marco Ferrante, Nicola Ferro, and Maria Maistro. 2015. Towards a Formal Framework for Utility-Oriented Measurements of Retrieval Effectiveness. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval (ICTIR '15)*. Association for Computing Machinery, New York, NY, USA, 21–30. https://doi.org/10.1145/2808194.2809452

[21] Hans P. Frei and Peter Schäuble. 1991. Determining the Effectiveness of Retrieval Algorithms. *Information Processing & Management* 27, 2 (1991), 153–164. https://doi.org/10.1016/0306-4573(91)90046-O

[22] Fernando Giner, Enrique Amigó, and Felisa Verdejo. 2020. Integrating Learned and Explicit Document Features for Reputation Monitoring in Social Media. *Knowl. Inf. Syst.* 62, 3 (2020), 951–985. https://doi.org/10.1007/s10115-019-01383-w

[23] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (2002), 422–446. https://doi.org/10.1145/582415.582418

[24] Johannes Kiesel, Damiano Spina, Henning Wachsmuth, and Benno Stein. 2021. The Meant, the Said, and the Understood: Conversational Argument Search and Cognitive Biases. In *Proceedings of the 3rd Conference on Conversational User Interfaces (CUI '21)*. Association for Computing Machinery, New York, NY, USA, Article 20, 5 pages. https://doi.org/10.1145/3469595.3469615

[25] Fei Liu, Alistair Moffat, Timothy Baldwin, and Xiuzhen Zhang. 2016. Quit While Ahead: Evaluating Truncated Rankings. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 953–956. https://doi.org/10.1145/2911451.2914737

[26] Simon Mason and N.E. Graham. 2002. Areas Beneath the Relative Operating Characteristics (ROC) and Relative Operating Levels (ROL) Curves: Statistical Significance and Interpretation. *Quarterly Journal of the Royal Meteorological Society* 128 (07 2002), 2145 – 2166. https://doi.org/10.1256/003590002320603584

[27] Alistair Moffat. 2013. Seven Numeric Properties of Effectiveness Metrics. In *Information Retrieval Technology (AIRS'13)*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1–12. https://doi.org/10.1007/978-3-642-45068-6_1

[28] Alistair Moffat and Justin Zobel. 2008. Rank-Biased Precision for Measurement of Retrieval Effectiveness. *ACM Trans. Inf. Syst.* 27, 1, Article 2 (dec 2008), 27 pages. https://doi.org/10.1145/1416950.1416952

[29] Anselmo Peñas and Álvaro Rodrigo. 2011. A Simple Measure to Assess Non-response. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL '11)*. Association for Computational Linguistics, Portland, Oregon, USA, 1415–1424. https://aclanthology.org/P11-1142

[30] Tetsuya Sakai. 2004. New Performance Metrics Based on Multigrade Relevance: Their Application to Question Answering. In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization (NTCIR-4)*. National Institute of Informatics (NII), Tokyo, Japan, 17 pages. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/OPEN/NTCIR4-OPEN-SakaiTrev.pdf

[31] Tetsuya Sakai. 2009. On the Robustness of Information Retrieval Metrics to Biased Relevance Assessments. *Journal of Information Processing* 17 (2009), 156–166. https://doi.org/10.2197/ipsjjip.17.156

[32] Tetsuya Sakai and Zhaohao Zeng. 2019. Which Diversity Evaluation Measures Are "Good"?. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 595–604. https://doi.org/10.1145/3331184.3331215

[33] Falk Scholer, Diane Kelly, and Ben Carterette. 2016. Information Retrieval Evaluation Using Test Collections. *Inf. Retr.* 19, 3 (2016), 225–229. https://doi.org/10.1007/s10791-016-9281-7

[34] Mark D. Smucker and Charles L.A. Clarke. 2012. Time-Based Calibration of Effectiveness Measures. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*. Association for Computing Machinery, New York, NY, USA, 95–104. https://doi.org/10.1145/2348283.2348300

[35] Karen Sparck Jones and Cornelis J. van Rijsbergen. 1976. Information Retrieval Test Collections. *J. Documentation* 32, 1 (1976), 59–75. https://doi.org/10.1108/eb026616

[36] Damiano Spina, Johanne R. Trippas, Paul Thomas, Hideo Joho, Katriina Byström, Leigh Clark, Nick Craswell, Mary Czerwinski, David Elsweiler, Alexander Frummet, Souvick Ghosh, Johannes Kiesel, Irene Lopatovska, Daniel McDuff, Selina Meyer, Ahmed Mourad, Paul Owoicho, Sachin Pathiyan Cherumanal, Daniel Russell, and Laurianne Sitbon. 2021. Report on the Future Conversations Workshop at CHIIR 2021. *SIGIR Forum* 55, 1, Article 6 (jul 2021), 22 pages. https://doi.org/10.1145/3476415.3476421

[37] Johanne R. Trippas, Damiano Spina, Paul Thomas, Mark Sanderson, Hideo Joho, and Lawrence Cavedon. 2020. Towards a Model for Spoken Conversational Search. *Information Processing & Management* 57, 2 (2020), 102162. https://doi.org/10.1016/j.ipm.2019.102162

[38] Ellen M. Voorhees. 2001. Evaluation by Highly Relevant Documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01)*. Association for Computing Machinery, New York, NY, USA, 74–82. https://doi.org/10.1145/383952.383963

[39] Ellen M. Voorhees. 2002. The Philosophy of Information Retrieval Evaluation. In *Evaluation of Cross-Language Information Retrieval Systems (CLEF '02)*, Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 355–370. https://doi.org/10.1007/3-540-45691-0_34

[40] Ellen M. Voorhees and Donna K. Harman. 2005. *TREC: Experiment and Evaluation in Information Retrieval*. Vol. 1. MIT Press Cambridge.

[41] Hamed Zamani, Johanne R Trippas, Jeff Dalton, and Filip Radlinski. 2022. Conversational Information Seeking. *arXiv preprint arXiv:2201.08808* (2022). https://doi.org/10.48550/arXiv.2201.08808