Desirable Properties for Diversity and Truncated Effectiveness Metrics

Ameer Albahem
RMIT University
Melbourne, Australia
ameer.albahem@rmit.edu.au

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

Falk Scholer
RMIT University
Melbourne, Australia
falk.scholer@rmit.edu.au

Alistair Moffat
The University of Melbourne
Melbourne, Australia
ammoffat@unimelb.edu.au

Lawrence Cavedon
RMIT University
Melbourne, Australia
lawrence.cavedon@rmit.edu.au

ABSTRACT

A wide range of evaluation metrics have been proposed to measure the quality of search results, including in the presence of diversification. Some of these metrics have been adapted for use in search tasks with different complexities, such as where the search system returns lists of different lengths. Given the range of requirements, it can be difficult to compare the behavior of these metrics. In this work, we examine effectiveness metrics using a simple property-based approach. In particular, we present a case-analysis framework to define and study fundamental properties that seem integral to any evaluation metric. An example of a simple property is that a ranking with only one non-relevant document should never score lower than a ranking with two non-relevant documents. The framework facilitates quantifying the ability of metrics to satisfy properties, both separately and simultaneously, and to identify those cases where properties are violated. Our analysis shows that the Average Cube Test and Intent-Aware Average Precision are two metrics which fail to satisfy the desirable properties, and hence should be used with caution.

CCS CONCEPTS

• **Information systems → Evaluation of retrieval results**; **Retrieval effectiveness**;

KEYWORDS

Evaluation, Search result diversification, Axiomatic analysis

ACM Reference Format:
Ameer Albahem, Damiano Spina, Falk Scholer, Alistair Moffat, and Lawrence Cavedon. 2018. Desirable Properties for Diversity and Truncated Effectiveness Metrics. In *23rd Australasian Document Computing Symposium (ADCS '18), December 11–12, 2018, Dunedin, New Zealand.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3291992.3291996

1 INTRODUCTION

A wide range of evaluation metrics have been proposed, all with the goal of comparing and benchmarking IR systems. Given a search task, understanding the behavior of metrics is critical to a decision as to which of them is suitable to the task. Metrics can usually be categorized in terms of assumptions about users and systems. For instance, a common practice when defining a metric is to calculate the system performance at some cutoff k in the ranking, assuming that the user would have stopped inspecting documents at or before position k. However, the ranking returned by the system might be *truncated*, that is, of length $l < k$, perhaps as a result of the system finding fewer than k matching documents, or the collection having less than k relevant documents, or perhaps even as a requirement of the search task.

Ideally, a metric aims to measure a set of desirable properties associated with a given search task, and the metric score should reflect the level of attainment reached by the system in respect to those properties, based on the system's outputs. In this work, we introduce a framework that allows us to analyze metrics that may be applied to search tasks with two characteristics: diversified (or faceted) queries, and truncated rankings. Search result diversification aims to tackle query ambiguity and under-specified information needs. For ambiguous queries, we need to diversify results to satisfy the different possible interpretations or intents of users (for example, apple as fruit versus apple as a company), and for under-specified queries we may wish to provide shallow coverage of all of the possible interpretations. Consequently, many evaluation metrics have been proposed to benchmark diversification algorithms, for the most part as extended versions of metrics used for ad hoc evaluation. Indeed, some evaluation frameworks such as the Intent-Aware approach [1] and the D# mechanism [23] instantiate a diversity metric from any uni-dimensional ad hoc retrieval metric.

Evaluation metrics are typically studied using empirical approaches. One such option is to determine how closely a metric matches the behavior or preferences of users across a set of search systems [26]. However, this type of study is time and resource consuming. Another option is to use discrimination power analysis – performing statistical significance tests for a large number of system pairs, and calculating the percentage that result in (according to that particular metric) statistically significant differences [22]. Metrics with high discriminative power are considered superior,

ADCS '18, December 11–12, 2018, Dunedin, New Zealand

ACM ISBN 978-1-4503-6549-9/18/12…$15.00
https://doi.org/10.1145/3291992.3291996

on the purely pragmatic basis that they are more likely to lead to a reportable system comparison. Generally, both approaches are somewhat inscrutable, and do not tell us much about the particular behavior of the different metrics being considered. In addition, the findings of these approaches are dependent on the ground truths and search systems (runs) used in the experimental setup.

A recent and complementary proposal is the use of axiomatic analysis, where a set of fundamental and desirable properties are defined, and metrics are categorized in terms of whether they satisfy the properties or not [3, 4]. For instance, the replacement property of Ferrante et al. [13] states that if a less relevant document is replaced by a more relevant document, then the metric score should not decrease. Different axiomatic frameworks have been proposed in connection with retrieval models [12, 14], and to evaluate evaluation metrics [2, 13]. Moffat [18] provides an alternative taxonomy, based on numerical properties, and shows that metrics can be categorized according to which of the properties (in this work, a set of seven) they satisfy. Amigó et al. [5] proposed several formal constraints (or desirable properties) for diversity metrics. In this work, we also employ the idea of *desirable property* to study metrics designed to evaluate diversified and truncated rankings.

In contrast to prior research, our focus is to instantiate desirable properties of metrics in a case-analysis framework that allows researchers to quantify the ability of metrics to satisfy a certain property (or multiple properties simultaneously).

The remainder of the paper is organized as follows. Section 2 introduces the most commonly used evaluation metrics, and then Section 3 describes the proposed framework. We firstly define desirable properties of evaluation metrics and then explore their relationship with a range of actual metrics. In Section 4, we instantiate our framework on three sets of evaluation metrics that have been used in both ad hoc and diversified retrieval benchmarks, such as the TREC's Dynamic Domain [28] and Web [10] tracks. We discuss the implications and limitations of our findings in Section 5. Finally, Section 6 concludes the work.

2 BACKGROUND

We now describe the most commonly used evaluation metrics for ad hoc, diversified, and truncated rankings.

Ad Hoc Search. Early in the development of information retrieval, the focus of evaluation was on the system's ability to find available relevant information, and its accuracy; the set-based Recall (R) and Precision (P) metrics quantify these two aspects, respectively. For ranked output lists, P and R can be parameterized with a depth k, resulting in P@k and R@k. However, P@k and R@k do not consider the position of relevant documents within the first k ranks. Average Precision (AP) builds on these metrics by averaging P@i at each rank position i where a relevant document is returned, and normalizing by the total number of available relevant documents; overall this reflects the system's ability to retrieve relevant documents towards the top of the ranking. Another approach that also does this is Reciprocal Rank (RR), which can be regarded as modeling the actions of a user who is fully satisfied as soon as any single relevant document is found.

None of these metrics accommodate the degree to which a document is relevant to an information need. To tackle this, gain-based *graded relevance* approaches incorporate the system's ability to place highly relevant documents at or near the top of the ranking. In particular, Normalized Discounted Cumulative Gain (NDCG) [15] is now in wide use. Chapelle et al. [6] extended RR into Expected Reciprocal Rank (ERR) to handle graded relevance, via the assumption of a probabilistic user, following the example of Moffat and Zobel's Rank-Biased Precision (RBP) metric [19].

A shared aspect of many previous metrics is that document rankings are evaluated at a fixed depth k, and different values of k of necessity lead to different and incomparable metrics. Instead of using a cutoff k, Rank-Biased Precision (RBP) [19] uses a converging sequence of weights, and in doing so allows "evaluation at depth k" to be a useful (and converging) approximation of "evaluation at an arbitrary depth". In particular, RBP models a user who proceeds from each document in the SERP to the next with a probability ϕ that reflects the level of "patience" that they bring to this particular search context.

Diversification. A range of metrics have been proposed to evaluate search result diversification. Many are extended versions of traditional IR metrics, while others are based on specific meaphors for diversity. Given a query q and a metric cutoff k, a diversity metric m measures the extent to which the top retrieved documents cover the query aspects $A_q = a_1, ..., a_m$ where each aspect a_j is associated with a value p_j that represents its importance or popularity. Let $\mathsf{rel}(d_i, a)$ denote the relevance of the i th-ranked document to aspect a, and $\mathsf{nrel}(d_{i-1}, a)$ denote the number of seen documents relevant to aspect a in ranks 1 through $i-1$.

Subtopic Recall srec@k [30] measures the range of query aspects that have been covered up to depth k, computed as the ratio of the number of aspects covered to the total number of aspects. The Intent-Aware (IA) framework [1] provides a method to extend any information retrieval metric to evaluate a diversified ranking by using a linear combination of the metric scores of individual rankings, computed by treating each aspect in turn as the only relevant aspect to the user while assuming the rest are non-relevant:

$$\mathsf{m\text{-}IA}@k = \sum_{a \in A_q} p_a \cdot \mathsf{m}_a@k \quad (1)$$

where m can be any metric such as AP, RR or NDCG, and where $\mathsf{m}_a@k$ is the calculation of that metric treating a as the only desired aspect of the set A_q.

Many researchers have extended metrics based on the cascade model of user behavior [11, 19], which assumes that a probabilistic user examines documents by reading a ranked results list from the top, and stopping when they are satisfied (or frustrated with) the search results. In general, these metrics follow the formula:

$$\mathsf{m}@k = \frac{1}{\mathcal{N}} \sum_{i=1}^{k} \frac{1}{\mathsf{discount}(i)} \left[\sum_{a \in A_q} p_a \cdot \mathsf{rel}(d_i, a) \cdot \mathsf{novelty}(d \mid d_1, d_{i-1}) \right] \quad (2)$$

where $\mathcal{N}$ is a normalization factor to map scores to the range of $[0, 1]$; p_a is the importance of the query aspect; $\mathsf{discount}(i)$ is a discount factor that emphasizes top ranked documents; $\mathsf{rel}(d_i, a)$ is as already defined; and $\mathsf{novelty}(d \mid d_1, d_{i-1})$ is the metric's approach to measuring novelty. Some metrics model novelty indirectly, by penalizing documents that are returned for aspects that have been covered before rank i [7, 8]. Chapelle et al. [6] model novelty by how unlikely it is that the previous documents have covered the aspect.

The widely used α-NDCG metric [7] scores documents by how well they cover the different aspects of queries. In addition, it assumes that the relevance of a document to a query is proportional to the number of aspects, or *nuggets*, that it contains. Together with ideas from NDCG, the quality of the ranking is measured as:

$$\alpha\text{-NDCG@}k = \frac{1}{\alpha\text{-DCG}^{\text{Ideal}}} \sum_{i=1}^{k} \frac{1}{\log_2(i+1)} \left[\sum_{a \in A_q} \text{rel}(d_i, a)(1-\alpha)^{\text{nrel}(d_{i-1}, a)} \right] \quad (3)$$

where p_a is not modeled; and $\alpha \in [0, 1]$ controls the user tolerance for redundancy. Novelty and Rank Biased Precision (NRBP) [8] combines ideas from RBP and α-NDCG as follows:

$$\text{NRBP} = \frac{1-(1-\alpha)\beta}{|A_q|} \cdot \sum_{i=1}^{\infty} \beta^{i-1} \sum_{a \in A_q} \text{rel}(d_i, a)(1-\alpha)^{\text{nrel}(d_{i-1}, a)} \quad (4)$$

where β denotes the patience of the user; α denotes the user's tolerance in regard to redundant content; $\text{rel}(d_i, a)$ is a binary relevance; and p_a is not modeled. Note that Equation 4 was derived after applying a collection-independent normalization. There are two ways to normalize a diversified ranking score [9]: collection-dependent and collection-independent. *Collection-dependent* normalization uses an approximate ideal ranking based on ground truth, and will not be ideal if some aspects have no relevant documents. Normalized NRBP (nNRBP) [9] uses collection-dependent normalization. In contrast, *collection-independent* normalization is based on the highest possible score achievable from a perfect collection [8] and constructs a ranking of documents relevant to all aspects. Alongside ERR, Chapelle et al. [6] propose a diversity-aware metric as follows:

$$\text{ERR@}k = \sum_{i=1}^{k} \frac{1}{i} \sum_{a \in A_q} p_a \cdot p_i^a \cdot \prod_{j=1}^{i-1} (1 - p_j^a) \quad (5)$$

where p_i^a is the probability that a document at rank i is relevant to aspect a, and where $p_i^a = (2^{\text{rel}(d_i, a)} - 1)/2^{\text{relmax}}$, with relmax being the maximum relevance grade.

Both NRBP and α-NDCG ignore the graded relevance of documents with respect to the query aspects, do not model aspect importance, and do not normalize scores using approximate ideal rankings. The D#-measure framework introduced by Sakai and Song [23] combines ideas from intent-aware, subtopic-recall, and the cascade model to address several components: normalization, graded relevance, aspect importance, discounted gain and coverage of aspects. D#-m boosts aspect recall and rewards documents that are highly relevant to more popular aspects of the query as follows:

$$\text{D\#-m@}k = \gamma \cdot \text{srec@}k + (1-\gamma) \cdot \text{D-m@}k \quad (6)$$

where D-m@k is calculated assuming a global gain of relevance [24].

Most of the previous measures have an explicit or implicit assumption that the user gains utility from seeing relevant documents; Smucker and Clarke [25] explicitly model the amount of time that users need when viewing document summaries and full documents, and incorporate this into a metric called Time-Biased Gain.

Recently, the Cube Test (CT) [17] measure has been proposed, based on a metaphor of filling a cube with "document water", as a way of modeling the speed at which the user information need is met. The gain of each document is judged with consideration of both the query aspects it addresses, and the extent to which those aspects have already been addressed by previous documents. In addition, it assumes a model where the user can stop gaining benefit from some aspect – when that prism within the cube has become full. This model is motivated by professional search scenarios such as patent search, in which a patent officer does not need all relevant documents to invalidate certain claims in a patent application. The Cube Test can be represented as follows:

$$\text{CT}(q, D) = \frac{1}{\text{Time}(D)} \sum_{i=1}^{|D|} \sum_{a \in A_q} p_a \cdot \text{rel}(d_i, a) \cdot \gamma^{\text{nrel}(d_{i-1}, a)} \cdot \tau_a, \quad (7)$$

where D is the set of retrieved documents, $\text{rel}(d_i, a)$ is normalized to $[0, 1]$; and γ is a discounting factor that controls the trade-off between relevance and novelty. In addition, MH determines the cube height; τ_a is an indicator function such that, if $\sum_{j=1}^{i-1} \text{rel}(d_j, a) < \text{MH}$, then $\tau_a = 1$, and $\tau_a = 0$ otherwise; and $\text{Time}(D)$ is the time spent examining documents in D. The "Average Cube Test" [27] is then defined as:

$$\text{ACT}(q, D_q) = \frac{1}{|D_q|} \sum_{i=1}^{|D_q|} \text{CT}(q, D^i), \quad (8)$$

in which D^i is the set of documents in D_q form the first to the i th inclusive. In the 2015 TREC Dynamic Domain (DD) track [28], MH was set to 5 and Time is the number of iterations it takes a user to reach the i th document.

Truncated Rankings. There are many search scenarios in which it is desirable for the search system to truncate the ranked list of documents, rather that supplying documents that it has no confidence in, or that will not be of interest to the presumed user. Examples of these cases include question answering, and search via mobile devices. To measure this aspect, Liu et al. [16] propose appending a nominal *terminal document* to rankings, to indicate that there are no more documents to be considered. Liu et al. then extend a range of previous ad hoc retrieval metrics, defining a gain value for the terminal document, and then using that in the calculation of the metric.

3 FRAMEWORK

As an illustrative example, consider the space of all rankings of at most two documents, including truncated rankings, shown in Figure 1. We assume that a document is relevant to a single aspect from query aspects $\mathcal{A}_q = \{\mathfrak{a}, \mathfrak{b}\}$. Each ranking is represented by a sequence of relevance labels, where: a means a document is relevant to aspect $\mathfrak{a}$; b indicates a document is relevant to aspect $\mathfrak{b}$; x means a document is not relevant to any aspect; and $\emptyset$ indicates an empty ranking.

Intuitively, there are some rankings that are better – or at least no worse – than others. For instance, the ranking $\overrightarrow{aa}$ is not worse than the ranking $\overrightarrow{a}$, as the former has more relevant documents. From an evaluation perspective, we thus expect a sound evaluation metric to give the first of those two rankings a numerical score that is not less than the score assigned to the second ranking.

Following this reasoning, we define a set of desirable properties that metrics for evaluating diversified and truncated rankings should satisfy.

$$\begin{array}{cccc} \overrightarrow{a} & \overrightarrow{aa} & \overrightarrow{ab} & \overrightarrow{ax} \\ \overrightarrow{b} & \overrightarrow{bb} & \overrightarrow{ba} & \overrightarrow{bx} \\ \overrightarrow{x} & \overrightarrow{xa} & \overrightarrow{xb} & \overrightarrow{xx} \\ \emptyset & & & \end{array}$$

Figure 1: List of all possible diversified rankings of at most length two, including truncated rankings, with two aspects.

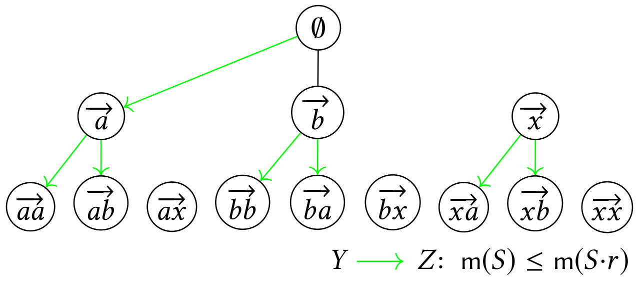

Figure 2: Tree of relations between rankings induced using the Relevance Monotonicity property.

3.1 Properties

The properties defined below are not intended to be comprehensive, but rather informative, and to create a better understanding of the behavior of metrics in complex search tasks involving diversification and truncation.

Relevance Monotonicity. Many pairs of rankings in Figure 1, such as $\overrightarrow{a}$ and $\overrightarrow{aa}$, or $\overrightarrow{b}$ and $\overrightarrow{bb}$, or $\overrightarrow{a}$ and $\overrightarrow{ab}$, are instantiations of extending a sequence by adding a relevant document. To describe the relation between these rankings, we suggest that if a ranking S is extended by a single relevant document r, then the computed effectiveness score should not decrease, but might remain the same. That is, given a metric m, we argue for the requirement

$$\mathsf{m}(S) \leq \mathsf{m}(S{\cdot}r)\,. \tag{9}$$

Figure 2 shows the graph of relations between rankings induced using the Relevance Monotonicity property. Note that this property is different from the monotonicity property defined by Moffat [18]. Moffat's property focuses on extending a ranking by a document whether it is relevant or not, which might be desirable if it is required to have a lower bound of performance. In our case, we believe that relevance monotonicity is desirable for all metrics.

It might also be argued that if the user has seen a relevant document in S that satisfied their information need, then adding another relevant document r with content already covered by S should be actively penalized, in that it somehow detracts from the search experience. There are two cases where this might occur. The first is when we add a duplicate of a seen relevant document. In this case, we note that a second occurrence of *any* document in a ranking should be regarded as being non-relevant; and in any case, a duplicate detection process if present would prevent that from happening. The second case where viewing more relevant documents might be detrimental is when the environment requires high effort from the user. For instance, Ong et al. [20] found that users do not necessarily click more when more relevant documents are retrieved in mobile search, an effect possibly caused by the high cost of inspecting documents on a small screen. Whether that difference in click rate should be regarded as evidence of dissatisfaction with the Search Engine Result Page (SERP) presented is a question that is interesting to consider, but beyond the scope of this current work.

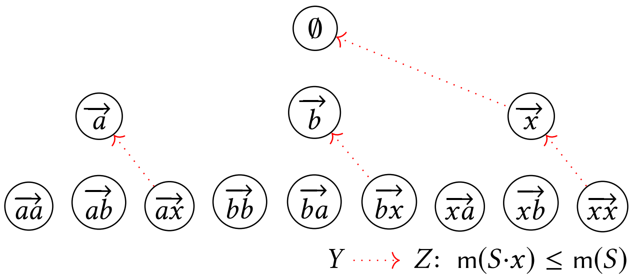

Figure 3: Tree of relations between rankings induced using the Irrelevance Monotonicity property.

Irrelevance Monotonicity. We further suggest that if a ranking S is extended by a single non-relevant document x, then the computed effectiveness score should not increase – Figure 3 shows the Irrelevance Monotonicity property using the illustrative rankings. That is, given a metric m:

$$\mathsf{m}(S{\cdot}x) \leq \mathsf{m}(S)\,. \tag{10}$$

We argue for this inequality on the grounds that the reward for adding a non-relevant document x to S should never be more than the reward of maintaining the original (even if truncated) ranking S. This is especially important in scenarios such as question answering, in which a desirable property is the ability of the search system to avoid giving a wrong, misleading, or offensive answer [21].

Again, a counter-view might be argued in certain cases – for example, that seeing one or more non-relevant documents might help the user to understand what they are actually searching for, and to help them clarify their intentions as they formulate further queries in the same search session. In this study, our primary focus is on a single search interaction, and therefore we assume that non-relevant documents lead to no gain. Nevertheless, measuring the system's ability to learn from user interactions over a sequence of queries is an interesting property to model in a broader whole-of-session evaluation context.

Redundancy. Suppose that the search system needs to diversify a Search Engine Result Page (SERP) to cover two or more distinct aspects associated with the query. Now the question becomes whether providing two (or multiple) documents for the same aspect is better or worse than providing documents relevant to different aspects of the query. Factors such as relevance grade and aspect importance are some of the variables that determine the answer. To contain the space of possibilities, in the first instance we make the following assumptions:

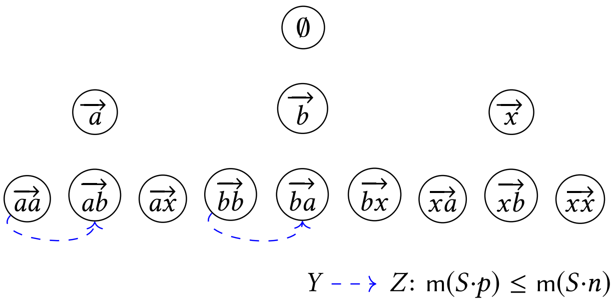

Figure 4: Tree of relations between rankings induced using the Redundancy property.

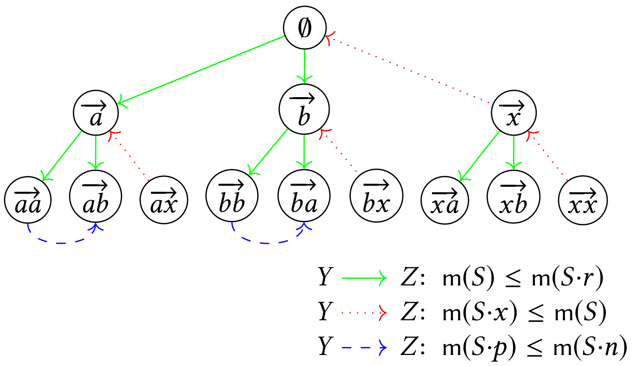

Figure 5: Tree of relations between rankings induced using multiple properties.

(1) *Uniform distribution of aspect probability.* All aspects have the same importance or weight $p_i = 1/m$ for a query q with m aspects $A_q = a_1, \ldots, a_m$, that is, there are no preferences among the aspects.
(2) *Binary relevance judgments to aspects.* A document is either relevant to an aspect or not relevant.
(3) *Single aspect document relevance.* If a document is relevant, it is relevant to only one aspect.

Under these assumptions, a diversification metric should not favor a run that adds a document relevant to an already-seen aspect over a run that adds a relevant document to an unseen, novel aspect [5]. That is, if p denotes a document relevant to an aspect already covered by S, and n denotes a document relevant to an aspect that has no coverage in S, then we suggest that:

$$\mathsf{m}(S{\cdot}p) \leq \mathsf{m}(S{\cdot}n) \qquad (11)$$

One might argue that, if a metric measures diversity under the three assumptions mentioned above, then the relationship should be one of strict inequality. However the relaxed version of this property is needed to properly handle metrics that have fixed cutoffs. For example, assuming two aspects a_1 and a_2, if $S = a_1{\cdot}a_1{\cdot}a_1$ and subtopic recall to depth three is being measured, then $\mathsf{srec@3}(S{\cdot}a_1) = \mathsf{srec@3}\ (S{\cdot}a_2) = \frac{1}{2}$, and equality is required. Figure 4 shows the relationships that arise from the Redundancy property.

3.2 Induction

The properties defined above do not cover all possible relations between rankings. However, as shown in Figure 5, many pairs of rankings have a relation between them induced using combination of properties. For instance, the $\overrightarrow{bx}$ ranking is not better than the $\overrightarrow{bb}$ ranking via the Irrelevance Monotonicity and Relevance Monotonicity properties.

3.3 Case Analysis

Given a set of properties and a set of metrics, there are different approaches that can be employed to determine whether the metrics satisfy the properties. One method consists of proving the satisfaction of a certain property mathematically, by examining the metric formulation and, for example, constructing an inductive proof. A second option, and the one employed here, is to enumerate all possible input rankings up to some maximum evaluation depth k, and note any exceptions to the properties that arise. This is useful in that it allows us to quantify the behavior of the metric. It also provides the cases where the metrics do not satisfy the property; which might be instrumental in understanding their behavior and suitability to a given search task.

In this framework, we perform the following three steps:

(1) Generate all possible instances of a truncated and diversified ranking. This is equivalent to building an $m + 1$-ary tree, where $m + 1$ is the number of relevance choices – either relevant for one of the m different aspects for this query, or completely non-relevant – of each document that a system can add to the ranking at a given position of some depth h, the length of the rankings being generated. A path from the root node to any leaf node represents a *full-length* ranking, and a path from the root to an internal node represents a *truncated* ranking that should also be assigned a score by each of the metrics. The total number of rankings of length h is given by

$$\text{Total number of cases} = \frac{(m+1)^{h+1} - 1}{m}.$$

(2) Evaluate all generated rankings using the metrics $\mathsf{m}(\cdot)$ being evaluated, that is, generate metric scores for each of the rankings.
(3) For every metric, use the computed scores and the known tree-based relationships of the generated runs to determine whether the metric violates any of the properties proposed by Equations 9, 10, and 11.

Note that this approach provides certainty of outcome if violations are detected; and, provided h is moderately large, confidence (but not certainty) of outcome if no violations are detected.

4 EVALUATION

We now employ this approach to study some of the metrics that have been used to evaluate search tasks.[1] In particular, we consider the metrics used in the TREC Web and Dynamic Domain tracks, and use each of them to score all possible rankings of length up to 10, and assuming that two different aspects of relevance occur.

[1] The code is available at https://github.com/aalbahem/ir-eval-meta-analysis.

Table 1: Violation summary of properties and metrics when considering all possible truncated rankings of up to 10 documents, with ✗ and ✓ indicating whether a property is violated or not.

Metric	Properties			
	$\mathrm{m}(S) \leq \mathrm{m}(S{\cdot}r)$	$\mathrm{m}(S{\cdot}x) \leq \mathrm{m}(S)$	$\mathrm{m}(S{\cdot}p) \leq \mathrm{m}(S{\cdot}n)$	Induction
ACT	✓	✗	✓	✗
AP-IA	✓	✓	✗	✗

4.1 Metrics

We use standard evaluation tools to perform the measurements. Ad hoc retrieval metrics RR, P@5,P@10 NDCG@5, NDCG@10 and AP were computed using `trec_eval`.[2]

Diversification metrics were computed using the official evaluation script for the TREC Web Diversity task using `ndeval`.[3] In particular, we considered Subtopic Recall [30], Intent-Aware Average Precision (AP-IA) and Precision (P-IA) [1], and several Cascade-based metrics: diversity-aware Expected Reciprocal Rank (ERR) [6]; α-NDCG [7]; and Novelty and Rank Biased Precision (NRBP) [8].

We also included the metrics used in the TREC Dynamic Domain (DD) Track [28].[4] TREC DD examines search scenarios that involve truncated rankings and diversification. The metrics used to evaluate systems are the Cube Test (CT) [17], Normalized Cube Test (nCT) [29], and the Average Cube Test (ACT) [28]. In our analysis, we generated runs according to the TREC DD run format, assuming a single iteration. As these metrics were used in TREC DD to evaluate the truncation and diversification dimensions of rankings, we believe they are good candidates for which to instantiate our framework.

4.2 Results

Table 1 shows the properties identified in our framework, and lists the metrics that were found to have violated them. ACT violates the Irrelevance Monotonicity property, and AP-IA violates the Redundancy property. ACT violates almost 99% of the cases, 29,496 violations out of 29,523 cases where Irrelevance Monotonicity is applicable; whereas AP-IA violates 100% of 2,026 cases in which Redundancy is applicable. In addition, both metrics violate relationships induced using Induction. None of the other metrics included in the evaluation violated any of the properties.

4.3 Violation Analysis

To understand the basis of these violations, we examined cases in detail. In the case of ACT, we found that a violation occurs because the metric calculates the score of the ranking by first calculating CT at each position in the ranking, and then averaging these scores. For instance, assuming a query has two aspects a and b, if $S = a{\cdot}b$ and covers the two aspects, then

$$\mathrm{ACT}(S) = \frac{\mathrm{CT}(a) + \mathrm{CT}(a{\cdot}b)}{2} = 0.0750$$

and

$$\mathrm{ACT}(S{\cdot}x) = \frac{\mathrm{CT}(a) + \mathrm{CT}(a{\cdot}b) + \mathrm{CT}(a{\cdot}b{\cdot}x)}{3} = 0.08330.$$

That is, the ACT score of $S{\cdot}x$ is greater than the ACT score for the shorter ranking S.

In the Intent-Aware metrics, the score of a ranking is the weighted sum of the scores of multiple rankings, considering the facets independently [1]. A violation can occur in AP-IA when a ranking contains documents relevant to a single aspect only. For example, given $S = a{\cdot}x{\cdot}x$, then $S{\cdot}p = a{\cdot}x{\cdot}x{\cdot}a$ and $S{\cdot}n = a{\cdot}x{\cdot}x{\cdot}b$. Suppose further that there are five relevant documents for facet a, and five relevant documents for facet b, and hence $R = 10$ relevant documents in total. When these two rankings are scored, we get $\mathrm{AP\text{-}IA}(S{\cdot}p) = 0.150$, with the AP score for facet a being computed as $(1/10) \cdot (1/1 + 2/4)$ and AP for facet b remaining zero; and $\mathrm{AP\text{-}IA}(S{\cdot}n) = 0.125$, computed as the sum of $(1/10) \cdot (1/1)$ (the AP score for facet a) and $(1/10) \cdot (1/4)$ (the AP score for facet b). The two separate AP scores in the second case do not outweigh the single larger one in the first case, giving rise to the violation.

5 DISCUSSION

These results have two implications. First, with respect to ACT, it raises a concern about its ability to evaluate the truncation dimension of a ranking system. However, ACT satisfies our Redundancy property. With respect to the AP-IA, it confirms what has been reported in previous work [8, 24]: the intent-aware evaluation framework [1] can produce counter-intuitive results when some aspects are less frequent in the ranking than others.

Therefore, it confirms the concern about their ability to measure the diversification aspect of rankings.

Metric Behavior in Truncated Rankings. As discussed previously, many metrics assume that the cutoff k is smaller than the length of the ranking l. However, the assumption does not hold if truncated rankings are permitted, that is, when $k < l$. When faced with a truncated ranking, the computation associated with each metric needs to be adapted. We studied the implementations in the various metric evaluation scripts to determine how they handle such situations.

In the case where k is part of the metric calculation (for example, P@10 and P-IA@10), `trec_eval` and `ndeval` assume the insertion of non-relevant documents in positions $l + 1$ to k. In addition, `trec_eval` has another implementation of P@k that divides by l instead of k. Similarly, metrics with an "ideal" normalization component (α-NDCG as implemented in `ndeval` and NDCG as implemented in `trec_eval`) still calculate the score of an ideal ranking of depth k (rather than to the shallower depth l) and use it to normalize ranking scores, thereby again interpolating non-relevant documents to "make up the numbers".

Evaluation of Single-Aspect Rankings. We also studied rankings with only a single aspect, a situation for which many of the evaluation metrics fall back to traditional ad hoc retrieval metrics. In this special case where $m = 1$, ACT still does not satisfy the Irrelevance Monotonicity property of Equation 10. For AP-IA, there is a single ranking, and it is equivalent to AP. As there are no other aspects,

[2] https://github.com/usnistgov/trec_eval
[3] https://github.com/trec-web/trec-web-2014
[4] https://github.com/trec-dd/trec-dd-jig

the third property is not applicable since there is no notion of an *unseen* aspect.

6 CONCLUSIONS AND FUTURE WORK

Evaluation is a cornerstone of information retrieval research, and a great number of metrics have been proposed to evaluate search systems. Therefore, it is vital that researchers understand and use evaluation metrics appropriately. In this work, we proposed a novel analysis framework that supports the quantification of the extent to which evaluation metrics satisfy a set of desirable properties. We instantiated the framework based on three desirable properties that metrics used to evaluate diversified and truncated rankings should satisfy: *Relevance Monotonicity*, *Irrelevance Monotonicity* and *Redundancy*. Using the framework, existing diversification metrics were analyzed, finding that the Average Cube Test (ACT) might be unpredictable when evaluating truncated rankings, as is the case for the TREC DD Track; and that Intent-Aware Average Precision (AP-IA) provides counter-intuitive results in some situations. This latter result is in line with previous studies on intent-aware metrics [9, 24].

In future work, we will investigate extensions and modifications to the framework, such as relaxing some of the assumptions that were made when defining the current properties, for example the requirement that the aspects have equal weights. In addition, we intend to study further diversification metrics such as the recently-proposed Rank-Biased Utility (RBU) [5], and those arising from the D-# evaluation framework [23].

ACKNOWLEDGMENT

This research was partially supported by Australian Research Council Project LP130100563 and Real Thing Entertainment Pty Ltd.

REFERENCES

[1] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *Proc. WSDM*, pages 5–14, 2009.
[2] Enrique Amigó, Julio Gonzalo, and Felisa Verdejo. A general evaluation measure for document organization tasks. In *Proc. SIGIR*, pages 643–652, 2013.
[3] Enrique Amigó, Hui Fang, Stefano Mizzaro, and ChengXiang Zhai. Axiomatic thinking for information retrieval: And related tasks. In *Proc. SIGIR*, pages 1419–1420, 2017.
[4] Enrique Amigó, Hui Fang, Stefano Mizzaro, and ChengXiang Zhai. Are we on the right track?: An examination of information retrieval methodologies. In *Proc. SIGIR*, pages 997–1000, 2018.
[5] Enrique Amigó, Damiano Spina, and Jorge Carrillo-de Albornoz. An Axiomatic Analysis of Diversity Evaluation Metrics: Introducing the Rank-Biased Utility Metric. In *Proc. SIGIR*, pages 625–634, 2018.
[6] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *Proc. CIKM*, pages 621–630, 2009.
[7] Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proc. SIGIR*, pages 659–666, 2008.
[8] Charles L. A. Clarke, Maheedhar Kolla, and Olga Vechtomova. An effectiveness measure for ambiguous and underspecified queries. In *Proc. ICTIR*, pages 188–199, 2009.
[9] Charles L. A. Clarke, Nick Craswell, Ian Soboroff, and Azin Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *Proc. WSDM*, pages 75–84, 2011.
[10] Charles L. A. Clarke, Nick Craswell, and Ellen M. Voorhees. Overview of the TREC 2012 Web Track. In *Proc. TREC*, 2012.
[11] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In *Proc. WSDM*, pages 87–94, 2008.
[12] Hui Fang and ChengXiang Zhai. An exploration of axiomatic approaches to information retrieval. In *Proc. SIGIR*, pages 480–487, 2005.
[13] Marco Ferrante, Nicola Ferro, and Maria Maistro. Towards a formal framework for utility-oriented measurements of retrieval effectiveness. In *Proc. ICTIR*, pages 21–30, 2015.
[14] Sreenivas Gollapudi and Aneesh Sharma. An axiomatic approach for result diversification. In *Proc. WWW*, pages 381–390, 2009.
[15] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Sys.*, 20(4):422–446, 2002.
[16] Fei Liu, Alistair Moffat, Timothy Baldwin, and Xiuzhen Zhang. Quit while ahead: Evaluating truncated rankings. In *Proc. SIGIR*, pages 953–956, 2016.
[17] Jiyun Luo, Christopher Wing, Hui Yang, and Marti Hearst. The water filling model and the cube test: Multi-dimensional evaluation for professional search. In *Proc. CIKM*, pages 709–714, 2013.
[18] Alistair Moffat. Seven numeric properties of effectiveness metrics. In *Proc. Asia Info. Retri. Soc. Conf.*, pages 1–12, 2013.
[19] Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Sys.*, 27(1):2:1–2:27, 2008.
[20] Kevin Ong, Kalervo Järvelin, Mark Sanderson, and Falk Scholer. Using information scent to understand mobile and desktop web search behavior. In *Proc. SIGIR*, pages 295–304, 2017.
[21] Anselmo Peñas and Álvaro Rodrigo. A simple measure to assess non-response. In *Proc. ACL*, pages 1415–1424, 2011.
[22] Tetsuya Sakai. Evaluating evaluation metrics based on the bootstrap. In *Proc. SIGIR*, pages 525–532, 2006.
[23] Tetsuya Sakai and Ruihua Song. Evaluating diversified search results using per-intent graded relevance. In *Proc. SIGIR*, pages 1043–1052, 2011.
[24] Tetsuya Sakai, Nick Craswell, Ruihua Song, Stephen Robertson, Zhicheng Dou, and Chin-Yew Lin. Simple evaluation metrics for diversified search results. In *Proc. EVIA*, pages 42–50, 2010.
[25] M. D. Smucker and C. L. A. Clarke. Time-based calibration of effectiveness measures. In *Proc. SIGIR*, pages 95–104, 2012.
[26] Andrew Turpin and Falk Scholer. User performance versus precision measures for simple web search tasks. In *Proc. SIGIR*, pages 11–18, 2006.
[27] Hui Yang and Ian Soboroff. TREC 2016 Dynamic Domain Track overview. In *Proc. TREC*, 2016.
[28] Hui Yang, John Frank, and Ian Soboroff. TREC 2015 Dynamic Domain Track overview. In *Proc. TREC*, 2015.
[29] Hui Yang, Zhiwen Tang, and Ian Soboroff. TREC 2017 Dynamic Domain Track Overview. In *Proc. TREC*, 2017.
[30] Cheng Xiang Zhai, William W. Cohen, and John Lafferty. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proc. SIGIR*, pages 10–17, 2003.