Too complicated for a metric

Towards Unified Metrics for Accuracy and Diversity for Recommender Systems

Javier Parapar*
javier.parapar@udc.es
Universidade da Coruña
A Coruña, Spain

Filip Radlinski
filiprad@google.comm
Google
London, United Kingdom

ABSTRACT

Recommender systems evaluation has evolved rapidly in recent years. However, for offline evaluation, accuracy is the *de facto* standard for assessing the superiority of one method over another, with most research comparisons focused on tasks ranging from rating prediction to ranking metrics for top-n recommendation. Simultaneously, recommendation diversity and novelty have become recognized as critical to users' perceived utility, with several new metrics recently proposed for evaluating these aspects of recommendation lists. Consequently, the accuracy-diversity dilemma frequently shows up as a choice to make when creating new recommendation algorithms.

We propose a novel adaptation of a unified metric, derived from one commonly used for search system evaluation, to Recommender Systems. The proposed metric combines topical diversity and accuracy, and we show it to satisfy a set of desired properties that we formulate axiomatically. These axioms are defined as fundamental constraints that a good unified metric should always satisfy. Moreover, beyond the axiomatic analysis, we present an experimental evaluation of the metric with collaborative filtering data. Our analysis shows that the metric respects the desired theoretical constraints and behaves as expected when performing offline evaluation.

CCS CONCEPTS

• **Information systems** → **Recommender systems**.

KEYWORDS

diversity, recommender systems, offline evaluation, metrics

ACM Reference Format:
Javier Parapar and Filip Radlinski. 2021. Towards Unified Metrics for Accuracy and Diversity for Recommender Systems. In *Fifteenth ACM Conference on Recommender Systems (RecSys '21), September 27-October 1, 2021, Amsterdam, Netherlands.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3460231.3474234

*Work carried out as Visiting Faculty Researcher at Google.

RecSys '21, September 27-October 1, 2021, Amsterdam, Netherlands

ACM ISBN 978-1-4503-8458-2/21/09…$15.00
https://doi.org/10.1145/3460231.3474234

1 INTRODUCTION

Evaluation of Recommender Systems (RSs) is a research topic attracting significant attention (e.g. [8, 18, 43]). When evaluating, a sequence of decisions must be taken to determine the metrics. The first decision is between online and offline evaluation. Online evaluation is the gold standard, as it directly assesses user satisfaction. However, online evaluation of RSs imposes challenges compared to the more common offline evaluation [25]. For instance, the personalized nature of the RSs implies the need for more, many times scarce, resources. Moreover, how the system presents recommendations has been demonstrated to have a considerable effect [35]. Therefore, offline evaluation is the most common evaluation framework for RS research. When using offline evaluation, we have to make a second decision: evaluate with error metrics for the rating prediction task [22] or evaluate with ranking metrics for top-n recommendation [14]. Lately, different works have pointed out the advantages of ranking metrics over error metrics regarding the measure of user satisfaction [5, 22, 28].

The final decision when evaluating the top-n recommendation task on offline collaborative filtering data is to choose what to measure. Here, the most common objective is the accuracy of the recommendations. For that, classical ranking metrics are typically used [43]. More recently, the need for aspects beyond pure recommendation precision has been recognized [22]. Highly related properties such as serendipity, novelty, and diversity have been shown to be crucial in determining user engagement with systems [10]. Therefore new metrics for evaluating recommendation list diversity and novelty have been proposed [8]. This dichotomy forces the researcher to choose which metrics to optimize.

In this paper, we try to shed some light on this dilemma by advancing a unified evaluation of RSs. Inspired by extensive work carried out on search results diversification [41], we adapt a unified metric for evaluating top-n recommendations. Our metric considers in its formulation topical (aspect) redundancy and both personalized item and topical relevance. For validating this proposal, we extend work on axiomatic analysis of the properties of different metrics [16], and present experiments on collaborative filtering data. The analysis shows that our metric, $\alpha\beta$-nDCG, satisfies the desired axioms and that it also behaves well under all the experimental scenarios.

In the next section, we place our work in the context of related work, and provide background. Section 3 presents our proposal for a unified metric. Then, we perform the axiomatic analysis on Section 4 and the experimental validation on Section 5. Finally, we conclude in Section 6.

2 RELATED WORK

The need of offline evaluation metrics beyond accuracy has been pointed out repeatedly by the Recommender Systems community [4, 15, 19, 26]. Several metrics exist for evaluating either accuracy or diversity in RSs. On the one hand, Valcarce et al. [43] recently reviewed the most commonly used accuracy metrics for offline evaluation, showing how precision and nDCG behave best. On the other hand, Castells et al. [8] review the importance of novelty and diversity in RSs and the different existing metrics for those properties. However, limited effort has been put into unifying these, with Vargas et al.'s work probably the most relevant. Vargas and Castells [48] present a framework for analysing metrics regarding item choice, discovery, and relevance. Vargas' thesis [47] adapted some metrics from search result diversification [41] for RSs diversity evaluation.

In this paper, we focus on topical or aspect based diversity [51]. This concept refers to recommended items showing diverse aspects, e.g., different types of products or different genres of movies. This scenario matches the classical search results diversification task [41], where documents may present various aspects (*nuggets*) for a given information need. In that field, authors carried out extensive work in the joint evaluation of relevance and diversity.

In search settings, Clarke et al. [12] introduced the α-nDCG metric to evaluate aspect diversity. This metric adapts normalized Discounted Cumulative Gain (nDCG) [24], with different aspects (*nuggets*) also considered. Agrawal et al. [1] proposed a simple intent-aware adaptation of ranking metrics as a weighted average of isolated local aspect evaluation. Following this, Clarke et al. [13] presented NRBP as an extension of RBP for the diversification scenario. NRBP combines the strengths of the intent-aware metrics [1], RBP [31] and α-nDCG [12]. In document ranking, some authors also argue that the appearance of non-relevant documents should penalize the metric value, as for instance with Expected Utility (EU) [49]. Finally, we note work by Amigó et al. [3] on analyzing an extensive set of metrics for the search results diversification task. In particular, they present Rank-Biased Utility (RBU), a metric informed by a set of desired axioms that the authors defined for the search diversification scenario.

For evaluation of our proposed metric, we rely on two different strategies. First, we use axiomatics [16] to analyze the properties of the metric we present. The theoretical characterization of different research pieces has been extensively used in the past, both in document search and RSs. For instance, axioms have been used for the characterization of ranking models [32, 36, 42] or smoothing methods [21, 45]. Moreover, they were also used to characterize user expectations [29] and, therefore, metric behaviour [2, 3, 17]. Second, regarding our experimental part, we will evaluate mainly three different aspects of the metric (1) rank correlation (2) discriminative power and (3) robustness to incompleteness. These have been widely used both in RSs and search evaluation. For instance, Valcarce et al. [44] study the discriminative power and robustness to incompleteness in the RSs scenario, and Sakai and Song [39] demonstrate the importance of analyzing the discriminative power of metrics in search result diversification.

3 A UNIFIED METRIC FOR DIVERSITY AND ACCURACY

The α-nDCG metric was formulated in the Information Retrieval community and further adapted by Vargas [47] for RSs. We now reformulate α-nDCG to consider both topical diversity and accuracy specific to top-n recommendation. Formally, a top-n recommender produces a list of ranked item suggestions $\vec{i} = (i_1, \ldots, i_n)$ selected from the whole set of items $\mathcal{I}$ not belonging to the user profile $\mathcal{I}_u$, which is the list of items with preferences already expressed by the user u.

For item diversification, we contemplate a set of item aspects $\mathcal{A} = \{a_1, \ldots, a_c\}$. Item aspects could be any categorical classification of items, for instance, a movie may present different genres $a_{action}, a_{comedy}, a_{drama}, \ldots, a_{romance}$. An item may exhibit one or more aspects, e.g., $i_{titanic}$ exhibiting a_{drama} and $a_{romance}$. We will use the notation $a_\phi \in i$ to indicate that the item i exhibits aspect a_ϕ. In terms of relevance, $r_{u,i}$ represents the graded relevance of item i for user u, which corresponds with the rating that the user has given to the item. Typically, ratings are on a 5-point Likert scale: $r_{u,i} \in \{1 \ldots 5\}$. We assume that aspect dependant relevance for an item $r_{u,i,\phi} = r_{u,i}$ when i exhibits a_ϕ and 0 otherwise.

3.1 α-nDCG

The rationale for α-nDCG is that relevance of each document cannot be judged in isolation from the rest of the documents ranked. Thus α-nDCG considers dependencies among *nuggets* for computing an estimation of both relevance and diversity [12]. Although Clarke et al. proposed this metric for the search task, Vargas [46] shows how it can be used to evaluate recommendations. Vargas' approach assumes that there is a binary relevance judgment for an item i given a user u and an aspect a_ϕ. Following that assumption, the probability of relevance of an item for a user is:

$$P(R = 1, u, i) = P(\exists a_\phi : a_\phi \in u \cap i) \tag{1}$$

where a_ϕ is an aspect of interest to the user. Rewriting this in a probabilistic formulation, we obtain:

$$P(R = 1, u, i) = 1 - \prod_{\phi=1}^{c} 1 - P(a_\phi \in u) \times P(a_\phi \in i) \tag{2}$$

where c is the number of different aspects considered in the collection. If we consider that aspects are static and known, $P(a_\phi \in i)$ can be estimated as follows:

$$P(a_\phi \in i) = \begin{cases} 0 & a_\phi \notin i \\ 1 & \text{otherwise} \end{cases} \tag{3}$$

Otherwise, if we consider that items are not unequivocally assigned to categories but that there is some degree of uncertainty about the assignment, then we could assign an α to that probability when $a_\phi \in i$. With this choice, we can compute a diversity metric on the item ranking for a user.

3.2 $\alpha\beta$-nDCG

Two important aspects of α-nDCG that need to be reexamined when adapted to the recommendation task:

(1) The α parameter accounts for the possibility of the *assessor* being wrong in his or her judgement. The definition assumes that assessors' positive judgements may be erroneous, but negative judgements are always correct. In the item recommendation task, the situation is quite the contrary. When using explicit feedback for evaluation, the user produces positive ratings over items. In that situation, the certainty of the user judgement is relatively high. On the other hand, items not judged by the user are assumed to be non-relevant in offline evaluations. This assumption is quite strong, and there is significant evidence that it dramatically affects outcomes [7]. Specifically, user preferences are incomplete, so there are many missing item preferences in the test data. So, in the case of the recommendation, we should assume that positive judgements are mainly correct, while accounting for the possibility that assumed negative judgments could be wrong because of the missing ratings in offline data.

(2) In the case of retrieval, the assessors' judgements consider both relevance and topicality. That is, nuggets are facets of relevance to an information need. In RSs, item–aspect relations are influenced by the user's tastes. For example, a user may not like all horror movies the same amount. Therefore, item aspects should be considered conditioned on the user, instead of simply static aspect–item relations [1]. We will have liked-aspects of items for each user, e.g. *"horror movies that user u likes"*. In this sense, the set of horror movies that are relevant for a user could be {*"The Texas Chain Saw Massacre"*,*"Friday the 13th"*, *"Halloween"* } while for another it could be {*"The Exorcist"*, *"Poltergeist"*, *"The Omen"*}.

With this in mind, we can rewrite Equation 2 as:

$$P(R=1,u,i) = 1 - \prod_{\phi=1}^{c} 1 - P(a_\phi|u,i) \times P(a_\phi|u) \tag{4}$$

where the probability of item i contributing to satisfying the user's interest in a_ϕ is

$$P(a_\phi|u,i) = \begin{cases} 0 & a_\phi \notin i \\ \alpha(u,i) & \nexists\, r_{u,i} \text{ and } a_\phi \in i \\ \beta(u,r_{u,i}) & \exists\, r_{u,i} \text{ and } a_\phi \in i. \end{cases} \tag{5}$$

and $\beta(u,r_{u,i})$ is the confidence in the user's judgement value (we will assume discrete rating values, $r_{u,i} \in \{r_{min} \dots r_{max}\}$). This function can be defined in several ways. We propose a simple definition where the normalized rating value for the item is smoothed by a β factor accounting for user rating uncertainty (Equation 6). We leave for future work how to further personalise this factor to correct for user rating bias, rating scale or rating inconsistency [37]. Analogously, $\alpha(u,i)$ represents the weight of the item i with missing rating for the user. This factor may be personalized using different indicators, e.g., the user profile size, the item's freshness in the catalogue, the average rating of i, etc. Again, we will just test here with a small constant value α leaving personalized formulation for future work.

$$\beta(u,r_{u,i}) = \frac{r_{u,i}}{r_{max}} * \beta \tag{6}$$

Next, we introduce redundancy and novelty by estimating whether or not the user is (still) interested at position k in more items capturing a given aspect after having been shown earlier items in the ranking $S = \vec{i}[0,\dots,k-1]$:

$$P(a_\phi|u,S) = P(a_\phi|u) \prod_{i \in S} 1 - P(a_\phi|u,i) \tag{7}$$

By replacing $P(a_\phi|u)$ in Equation 4 by its redundancy aware variant (Equation 7), we obtain:

$$P(R_k = 1,u,i,S) = 1 - \prod_{\phi=1}^{c} 1 - P(a_\phi|u,i) \times P(a_\phi|u,S) \tag{8}$$

Now, the β parameter is responsible for a secondary role beyond its influence on the user-rating confidence. It also models the user's eagerness to look at items lower in the ranking. That is, the higher β, the more the relevant items contribute to satisfying the user's interests, and the fewer items are needed to exhaust the user's interest in an aspect.

3.3 Estimation

Clarke et al. [12] cannot estimate $P(a_\phi|u)$ because there is an absence of user preference data in traditional document retrieval offline evaluation. Instead, that work assumes that topics are independent and equally likely to be relevant to each user: $P(a_\phi|u) = \gamma$. We argue that this may be an oversimplification in the case of recommendation: When working with explicit user preferences over items, estimates can be computed about the degree of relevance of an aspect to a user. The maximum likelihood estimate is the simplest option:

$$P(a_\phi|u) \doteq \frac{\sum_{i \in \mathcal{I}_u | a_\phi \in i} r_{u,i}}{\sum_\varphi \sum_{i \in \mathcal{I}_u | a_\varphi \in i} r_{u,i}} \equiv \gamma_\phi^u \tag{9}$$

Generally speaking, the probability of relevance of an aspect to a user can be estimated on the prevalence of that aspect among past positive preferences from the user. For brevity, we will refer to that probability estimate as γ_ϕ^u. Following [12], we also define the number of ranked items up to position $k-1$ judged by the user as relevant and exhibiting aspect ϕ:

$$\rho_{u,r,\phi,k-1} = \sum_{j=1}^{k-1} \begin{cases} 1 & a_\phi \in i_j \text{ and } r_{u,i_j} = r \\ 0 & otherwise \end{cases} \tag{10}$$

$$\tau_{u,\phi,k-1} = \sum_{j=1}^{k-1} \begin{cases} 1 & a_\phi \in i_j \text{ and } \nexists\, r_{u,i_j} \\ 0 & otherwise \end{cases} \tag{11}$$

$\rho_{u,r,\phi,k-1}$ is the number of items ranked up to position $k-1$ judged by the user with rating r and showing aspect ϕ. $\tau_{u,\phi,k-1}$ is the number of items up to position $k-1$ belonging to aspect ϕ without information about the user's preference over them. Then, the inner term of Equation 7 becomes:

$$\prod_{j=1}^{k-1} 1 - P(a_\phi|u,i_j) = (1-\alpha)^{\tau_{u,\phi,k-1}} \prod_{r=r_{min}}^{r_{max}} (1-\beta(u,r))^{\rho_{u,r,\phi,k-1}} \tag{12}$$

We can now rewrite Equation 8 as:

$$P(R=1,u,i_1,\ldots,i_k) = 1 - \prod_{\phi=1}^{c}(1 - P(a_\phi|u,i_k) \times \gamma_\phi^u \times (1-\alpha)^{\tau_{\phi,k-1}} \times \prod_{r=r_{min}}^{r_{max}} (1-\beta(u,r))^{\rho_{u,r,\phi,k-1}}) \quad (13)$$

This probability is the gain value at the position k: $G[k]$. For computing $\alpha\beta$-nDCG@k, first, we have to compute the cumulative gain at the position k as in Equation 14:

$$CG[k] = \sum_{j=1}^{k} G[j] \quad (14)$$

The original nDCG applies a discount factor to penalize documents lower in the ranking. We propose to use the typical search ranking logarithmic discount function. More elaborate discounting functions reflecting the particular behaviour of recommender system users are left for future work:

$$DCG[k] = \sum_{j=1}^{k} G[j]/\log_2(1+j) \quad (15)$$

Finally, we normalize the Discounted Cumulative Gain by the Ideal Discounted Cumulative Gain (IDCG). Computing the Ideal Gain is NP-complete. As explained in [12], a good enough approximation can be computed greedily [50], based on selecting at each position the item with the highest gain, breaking ties arbitrarily.

Given all this, we use IDCG to normalize the DCG, resulting in the following definition for $\alpha\beta$-nDCG@k:

$$\alpha\beta\text{-nDCG@}k = \frac{DCG[k]}{IDCG[k]} \quad (16)$$

In summary, the changes of $\alpha\beta$-nDCG@k over α-nDCG@k are: (1) the role of the α parameter now accounts for the missing rating effect instead of the confidence in the assessor's decision, (2) the formulation of $P(a_\phi|u)$ allows us to estimate the aspect relevance for each user given his or her historical data rather than assuming all aspects are equally relevant, and (3) the added β parameter accounts for both the confidence in the rating that the user assigned to each item, and how fast the user is satisfied with relevant items while exploring the ranking.

4 AXIOMATIC ANALYSIS

We now derive axioms that our metric should satisfy, following the methodology of Amigó et al. [3]. First, we introduce common notation assumptions for the top-n recommendation case. Let $\vec{i}$ be a ranking of items; we use the notation $\vec{i}_{p\leftrightarrow q}$ for referring to a ranking where items in positions p and q are swapped. Analogously, we use $\vec{i}_{j\leftrightarrow j'}$ for denoting a ranking where the item j is swapped in $\vec{i}$ with the item j'. In both swaps, the item on the left was initially higher on the ranking than the one on the right. $w(\vec{i}, a_\phi)$ represents the weight of aspect a_ϕ in the list items $\vec{i}$. Analogously, the profile of user u may show interest in the different aspects with a weight $w(\mathcal{I}_u, a_\phi)$ that, for the sake of legibility, we refer as $w(u, a_\phi)$. The sum of all aspect weights for the user u add up 1 ($\sum_\phi w(u,a_\phi) = 1$). A user is satisfied with relevant items from particular aspect a_ϕ in the ranking until position p with value $0 \leq s(\vec{i}[0,\ldots,p], a_\phi) \leq 1$. We say that an aspect is saturated when the user is fully satisfied: $s(\vec{i}[0,\ldots,p], a_\phi) = 1$.

When judging an item ranking, we assume the standard behaviour of exploration of the ranked list [9], i.e., the user inspects the items in the recommendation list sequentially, from top to bottom. The user will stop inspecting the ranking when (a) his or her information need is satisfied (success) or (b) the user stops looking after some effort without finding any good item recommendations (fail). According to the Expected Reciprocal Rank user model presented by Chapelle et al. [9], the higher the relevance of the observed items in the ranking, the lower the need of the user to explore more documents.

Finally, we denote the quality score given by the metric to the top-n recommendation as $Q(\vec{i})$.

4.1 Axioms for Relevant and Diverse Item Rankings

Inspired by Amigó et al. [3], we adapt the desirable axioms for a recommendation ranking. As Sakai and Zeng [40] pointed out, Amigó et al. [3] axioms were defined in two isolated blocks (relevance and redundancy constraints). In turn, we consider both relevance and topical diversity jointly in the definition of the axioms:

Axiom 1 (Priority inside Aspect, Pri). *Swapping two items showing exclusively one aspect a_n in concordance with their relevance value increases the ranking quality score. Given $k > 0$:*

$$r_{u,i_{p+k}} > r_{u,i_p}, a_n \in i_p, i_{p+k} \wedge \forall_{\phi \neq n}\, a_\phi \notin i_p, i_{p+k} \implies Q(\vec{i}_{p\leftrightarrow p+k}) > Q(\vec{i}) \quad (17)$$

This axiom expresses the preference for the best-rated item when there is no difference in two items' aspects.

Axiom 2 (Deepness inside Aspect, Deep). *Correctly swapping items showing exclusively one aspect a_n in concordance with their relevance value has a bigger effect at earlier positions in the ranking. Given $k > 0$:*

$$p < q, r_{u,i_p} = r_{u,i_q} < r_{u,i_{p+1}} r_{u,i_{q+1}}, a_n \in i_p, i_q, i_{p+1}, i_{q+1} \wedge \forall_{\phi \neq n}\, a_\phi \notin i_p, i_q, i_{p+1}, i_{q+1} \implies Q(\vec{i}_{p\leftrightarrow p+1}) > Q(\vec{i}_{q\leftrightarrow q+1}) \quad (18)$$

This constraint is based on the concept of top-heaviness. That is, we prefer more relevance density at the beginning of the ranking than at the bottom.

Axiom 3 (Non Priority on Saturated Aspects, NonPriSatAsp). *There is a high enough aspect satisfaction of a user over an aspect $a_{n'}$ and a small enough positive difference r_Δ between two items relevance such that swapping those items showing exclusively one aspect each, a_n and $a_{n'}$, in concordance with their relevance value, decreases the ranking quality score. Given $k > 0$:*

$$\exists r_\Delta, s_\Delta \in \mathbb{R}^+ | r_{u,i_{p+k}} - r_{u,i_p} = r_\Delta, a_n \in i_p, a_{n'} \in i_{p+k}, s(\vec{i}[0,\ldots,p], a_{n'}) - s(\vec{i}[0,\ldots,p], a_n) = s_\Delta \implies Q(\vec{i}_{p\leftrightarrow p+k}) < Q(\vec{i}) \quad (19)$$

An example of this axioms is the case where the user is saturated with one aspect. When offered two items to be consumed, he prefers the one from a non-saturated aspect even when that item has a lower rating.

Axiom 4 (Top Heaviness Threshold, TopHeav). *There exists a value n large enough such that retrieving only one relevant item at the top of the ranking yields a higher quality score than retrieving n relevant items with the same graded relevance r after n non-relevant documents.*

$$\exists n \in \mathbb{N}^+ | Q(i^r_1, i^0_2 \dots i^0_{2n}) > Q(i^0_1, \dots, i^0_n, i^r_{n+1}, \dots, i^r_{2n}), r > 0 \quad (20)$$

This constraint builds upon the Pri axiom. It says that the user prefers to see only one relevant item at the top rather than having to go deep on the ranking for many relevant items. Informally, this axiom models a desire for efficiency when exploring ranked results.

Axiom 5 (Top Heaviness Threshold Complementary, TopHeavComp). *There exists a value m small enough such that retrieving only one relevant item at the top of the ranking yields a lower quality score than retrieving m relevant items with the same graded relevance r after m non-relevant documents.*

$$\exists m \in \mathbb{N}^+ | Q(i^r_1, i^0_2 \dots i^0_{2m}) < Q(i^0_1, \dots, i^0_m, i^r_{m+1}, \dots, i^r_{2m}), r > 0 \quad (21)$$

Complementarily to the previous axiom, this axiom refers to the existence of a lower bound for the user's effort for ranking exploration.

Axiom 6 (Aspect Relevance, AspRel). *Given two equally relevant items j and j′ showing exclusively and respectively two aspects a_n and $a_{n'}$, where neither aspect has been observed earlier in the ranking ($s(\vec{i}, a_n) = s(\vec{i}, a_{n'}) = 0$), then the item exhibiting the aspect with higher user weight $w(u, a_\phi)$ yields higher quality score.*

$$\begin{aligned} r_{u,j} = r_{u,j'} > 0, a_n \in j, a_{n'} \in j', \\ \forall_{\phi \neq n}\, a_\phi \notin j, \forall_{\phi \neq n'}\, a_\phi \notin j', w(u, a_n) > w(u, a'_n) \\ \implies Q(\vec{i}_{j \leftrightarrow j'}) < Q(\vec{i}) \end{aligned} \quad (22)$$

This axiom reflects the importance of one aspect over another. The user favours the item exhibiting the preferred aspect given two equally liked items.

Axiom 7 (Prefer More Aspect Contribution, MoreAsp). *Given two equally relevant items for the user both showing non-saturated aspects, $s(u, a_\phi) < 1$, of interest to the user, $w(u, a_\phi) > 0$, then the presence of the item with higher remaining interest from non-saturated aspects yields higher quality score than the other one.*

$$\begin{aligned} r_{u,j} = r_{u,j'}, \sum_{a_\phi \in j} w(u, a_\phi) - w(\vec{i}, a_\phi) < \sum_{a_\phi \in j'} w(u, a_\phi) - w(\vec{i}, a_\phi) \\ \implies Q(\vec{i}_{j \leftrightarrow j'}) < Q(\vec{i}) \end{aligned} \quad (23)$$

Given two equally relevant items, the user tends to favour the one showing aspects that are still of interest. In other words, the item with less aspect-level redundancy will get a higher score.

Axiom 8 (Missing over Non-Relevant, MissOverNon). *Given two items j and j′, where $r(u, j) = 0$ but the user's rating over j′ is unknown yet j′ exhibits a non-saturated aspect, then swapping j with j′ yields a higher quality score.*

$$r_{u,j} = 0, \nexists r_{u,j'}, \exists a_\phi \in j' | s(u, a_\phi) < 1 \implies Q(\vec{i}_{j \leftrightarrow j'}) > Q(\vec{i}) \quad (24)$$

This constraint models the missing rating effect: Intuitively the user would favour an unknown item rather than an item that he or she is known to dislike.

4.2 Metric Analysis

This section analyzes whether our proposed metric, or previous diversity aware metrics for document ranking, satisfy the desired axioms. A summary of that analysis is presented in Table 1. Due to space limitations, we restrict the analysis to the translation of some of the diversity-aware metrics from [3] to the recommendation setting. In particular, we do not comment on pure accuracy metrics as they do not satisfy most of the axioms. We also note that although intent-aware (IA) counterparts [1] solve some of the limitations of pure accuracy metrics, those variants suffer from two important limitations. First, their maximum value is not 1: it is highly improbable that a single ranked list is ideal for every aspect. Second, and arguably more difficult to resolve, they tend to under-represent the performance for minor aspects [11]. Therefore we leave the analysis of those variants to future work.

As noted above, α-nDCG@k does not consider item aspects conditioned on the user [46]. If we adapt it to graded relevance (the rating value), it satisfies Pri, Deep, NonPriSatAsp, TopHeavComp and MoreAsp. Due to the redundancy factor, it also satisfies TopHeav. α-nDCG does not consider user overall interest on the different aspects, therefore, it does not satisfy AspRel or MoreAsp. As it does not consider the difference between a missing rating and an item rated with 0 by the user, it also does not satisfy MissOverNon.

Zhai et al. [50] presented sub-topic recall (S-Recall) and sub-topic reciprocal rank (S-RR). Both are aspect level variants of the original metrics. S-Recall@k only computes the coverage of each of the aspects in the first k positions, i.e., how many of the possible aspects the top k items show. In that way, S-Recall@k is agnostic to the user's interest both on item relevance and topic preferences, so it does not satisfy Pri, Deep, TopHeav, TopHeavComp, AspRel nor MissOverNon. It would only partially satisfy NonPriSatAsp, and MoreAsp in the case of binary user satisfaction over aspects and with no other items exhibiting those aspects in the top k. The value of S-RR@k is the inverse of the first position where every possible aspect appeared on the ranking if that position is higher than k, 0 otherwise. Consequently, S-RR@k behaves similarly to S-Recall@k in terms of the axiomatic properties.

Clarke et al. [13] presented Novelty- and Rank-Biased Precision (NRBP). Contrary to most of the metrics, it does not work with cut-offs. In practical terms, that means that we theoretically need judgments values over the entire collection for computing the ideal gain vector. NRBP is essentially a combination of α-nDCG@k and Rank Biased Precision (RBP) [31] where both α and β penalize the user interest as he or she goes down the ranking. In NRBP,

Table 1: Properties of accuracy-diversity metrics (●= axiom satisfied, ○= axiom not satisfied).

Metric	*Axioms*							
	Pri	Deep	NonPriSatAsp	TopHeav	TopHeavComp	AspRel	MoreAsp	MissOverNon
α-nDCG@k	●	●	●	●	●	○	○	○
S-Recall@k	○	○	○	○	○	○	○	○
S-RR@100%	○	○	○	○	○	○	○	○
NRBP	●	●	●	●	●	○	○	○
EU	●	●	●	●	●	●	●	○
RBU@k	●	●	●	●	●	●	●	○
αβ-nDCG@k	●	●	●	●	●	●	●	●

α accounts for how fast a user gets bored of items from one aspect, while β models how willing he or she is to look for relevant documents down in the ranking. Thanks to those two discounting factors, when considering probabilistic relevance, NRBP satisfies Pri, Deep, NonPriSatAsp, TopHeav and TopHeavComp. However, again this metric does not consider personalized user interests on aspects. Therefore, it does not satisfy AspRel or MoreAsp. The proposed probabilistic relevance model considers missing judgments as zero relevance, so MissOverNon is also not satisfied.

Amigó et al. [3] included Expected Utility (EU) [49] in their analysis as the only metric penalizing non-relevant documents at the end of the ranking. This factor enables the satisfaction of their Conf axiom [3]. This axiom is not desirable in top-n recommendation; many non judged documents are relevant but with missing preferences. This explains why EU does not satisfy MissOverNon. Apart from that, EU is quite similar to NRBP, satisfying Pri, Deep, NonPriSatAsp, TopHeavComp and MoreAsp. It also satisfies AspRel and MoreAsp as it considers the weights of aspects.

Amigó et al. [3] proposed a new metric informed by their axioms on document ranking. Rank-Biased Utility is, again, a combination of existing metrics. It combines the exploration model of RBP [31] with the redundancy penalization of the intent-aware version of Expected Reciprocal Rank [9] (ERR-IA) with the user effort factor from EU. As a combination of those three metrics, RBU satisfies Pri, Deep (with p<1), NonPriSatAsp, TopHeav TopHeavComp (with p<1), MoreAsp, AspRel, and MoreAsp (with e<1) but not MissOverNon.

Finally, $\alpha\beta$-nDCG@k satisfies every axiom, as it was designed with them in mind. It keeps the α-nDCG@k properties that allow it to satisfy Pri, Deep, NonPriSatAsp, TopHeavComp, MoreAsp andTopHeav. Due to the consideration of users overall interest in different aspects through the formulation of $P(a_\phi|u)$ (see Equation 9), it also satisfies AspRel or MoreAsp. The inclusion of the $\alpha(u, i)$ factor (see Equation 4) enables the satisfaction of MissOverNon.

5 EXPERIMENTS

In this section, we further evaluate our unified metric proposal. In Section 4, we defined a set of theoretical properties for the behaviour of the quality metrics. However, as commented by Sakai and Zeng [40], axiomatic analysis has its limitations. As an alternative, in online experiments, user preferences over rankings could be obtained and assessed as to how observations correlate with the defined metrics. That is quite challenging in recommender system setting: When evaluating documents under explicit information needs, relevance can be judged rapidly by assessors with strong agreement. However, for RSs, relevance is user-dependent, and the information need is not explicit. Rather, it is indirectly reflected by users' past preferences over items. Therefore, producing a rigorous user study for the different metrics would require a massive number of users and significant time for the users to consume and evaluate suggestions. We therefore present an intermediate approach and leave the online user validation for future work. For doing so, we use the collaborative filtering information from the Movielens 20M collection [20]. Specifically, we take the users preferences from a 20% random test split. The dataset categorizes movies among 19 genres (e.g. action, adventure, drama, etc.).

When defining the axioms, we had in mind a desirable item ranking order for the users. Now we can use this concept to define an ideal item ordering when considering both accuracy and diversity. We acknowledge that this simplifies the actual RS scenario: in RSs, different users, even with the same past preferences, may favour slightly different orderings. However, when limited to offline evaluation, this conceptualization of an ideal ordering may serve us to observe the behaviour of the metrics. Of course, alternative and more complex idealizations of the item ranking may be proposed. The following aims to reflect the same ideas as those reflected in the axioms.

5.1 Ideal Ordering

For computing the ideal ordering for a user $\vec{\imath^u}$, we use the actual user preferences (the ground truth for the user). The ideal ranking of size k is a particular ordering of the items judged by the user greedily obtained following these steps:

- We create as many ideal per-aspect rankings $\vec{\imath^u}_{a_\phi}$ as different items aspect a_ϕ in the ground truth for the user. Each of them contains all the items in the user's preferences exhibiting that aspect.
- We sort the items in each of the per-aspect rankings following two criteria: first, we sort the items by the user expressed rating, and secondly, for breaking ties, we sort items based on the number of aspects they exhibit.
- We compute the aspect weights for the user $w(u, a_\phi)$ using the maximum likelihood estimate over the user profile in the ground truth data (as in Equation 9).

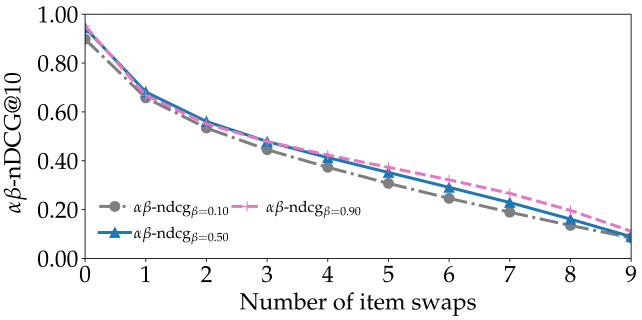

Figure 1: Values of $\alpha\beta$-nDCG for 10 cutoff for different number of bottom-to-top swaps.

- At each position we take the top item for the selected aspect a_{next} by popping $\vec{\iota^u}_{a_{\text{next}}}$
- The selection of a_{next} on the position k for the $\vec{\iota^u}$ is made iteratively. At each point, we choose the a_ϕ with the highest weight difference between the user aspect and the of the aspect in the ongoing ideal ranking.

$$\text{next} = \arg\max_{\phi} \left(w(u, a_\phi) - w(\vec{\iota^u}, a_\phi) \right)$$

5.2 Experiment 1: Rank Correlation with Ideal Ordering

A crucial aspect of a metric is how good it is in ranking systems according to the expected quality order. To assess that is not commonly possible in offline evaluation because the user's actual quality order is unknown. In practice, when possible, online evaluation is carried out [23]. Classical online experiments check if the ranking of systems produced by a metric correlates with user-assigned preferences over systems. However, given our notion of ideal item ordering, we may produce a similar experiment with offline data. In that way, we can assess which of the compared metrics produces a ranking of systems that correlates best with the actual order of systems under the ideal ordering assumption. For that purpose, we produce simulated systems as gradual perturbations of the ideal ordered system. The correct order of systems should be ranking them by how far they are from the ideal ranking (i.e. how much perturbation was applied). Here we consider three perturbation models that affect the three factors that we focus on: item relevance, diversity, and aspect relevance. Regarding the metric configurations, if not stated otherwise, we report the following configurations $\alpha = 0.005$ and $\beta = 0.5$ for $\alpha\beta$-nDCG, $p = 0.99$ $\beta = 0.9$ and $\alpha = 0.25$ for NRBP, $e = 0.05$ and $\alpha = 0.25$ for EU and $p = 0.99$ and $e = 0.05$ for RBU, following the best values reported in [3].

5.2.1 Swapping items bottom to top. Any swap of an item from the bottom of the ideal ranking to the top of it should yield a worse ranking. If we incrementally increase the number of swaps, we may create incrementally worse item rankings. A good metric should observe those differences and produce incrementally lower values. Formally we produce 1-perturbed bottom-top for an ideal ranking $\vec{\iota^u}$ of size k, as $\vec{\iota^u}_{1\leftrightarrow k}$ and an s-perturbed as, $\vec{\iota^u}_{1\leftrightarrow k, 2\leftrightarrow k-1, \ldots, s\leftrightarrow k-s}$. This perturbation should produce a gradual reduction in the system performance as shown in Figure 1 for different configurations on the $\alpha\beta$-nDCG@10. For this experiment, we generated 50 gradually perturbed systems.

5.2.2 Swapping items from redundant aspect. We produce perturbations of the ideal ranking for tackling aspect redundancy. We want to observe if the metrics react adequately to an excess of redundancy for an aspect in the ranking. Departing from the ideal ranking, if we add an item from the same category when the algorithm for computing the ideal ranking says to add an item from another category, it should produce an excess of redundancy on that part of the ranking. We can produce different variations by including not one but s redundant items in that category. Formally, let a_{current} and a_{next} be the aspect of the last item added to the ideal ranking and the aspect to pop an item from next respectively. Let p be the first position on the ranking where $a_{\text{current}} \neq a_{\text{next}}$. We produce 1-perturbed redundant aspect from an ideal ranking $\vec{\iota^u}$ at p as $\vec{\iota^u}_{p\leftrightarrow pop(\vec{\iota^u}_{a_{\text{current}}})}$. Analogously, the s-perturbed redundant aspect ranking would contain s additional items exhibiting aspect a_{current}. For this experiment, we also generated 50 gradually perturbed systems.

5.2.3 Swapping aspects. We produce perturbations of the ideal ranking for tackling aspect relevance. The objective is to check if the metrics can assess whether the ranking presents the aspects in the order preferred by the user. Departing from the ideal ranking, we select the first item for the user not from her most liked movie category but from the next category in order of preference, where the item from the favourite category is not present. We can produce an s-perturbed ranking by selecting from the $(s+1)^{th}$ most preferred category (if it exists). Formally, we produce 1-perturbed aspect relevance from an ideal ranking $\vec{\iota^u}$ where a_{most} is the aspect most liked by the user and a_{second} the second most liked aspect that $\vec{\iota^u_1}$ does not show as $\vec{\iota^u}_{1\leftrightarrow pop(\vec{\iota^u}_{a_{\text{second}}})}$. We only produced 1 to 10 perturbations, limited by the number of different genres that users tend to rate in the Movielens dataset.

5.2.4 Results. When analyzing the correlation of the system rankings of the different metrics with the actual system ranking (see Figure 2), there are various clear conclusions. First, as expected, subtopic metrics (S-Recall and S-RR) cannot produce a correct ranking of systems when systems diverge on relevance order (left). Instead, they produce a negatively correlated ranking. This negative correlation is observed because in swapping non-relevant items into the top of the ranking, we are speeding up sub-topic coverage.

Second, RBU, NRBP, EU, and our proposal are perfectly correlated on system ranking for the item relevance perturbed systems.

Third, when the system differences are due to an excess of aspect redundancy (centre), we can see how RBU and EU perform more poorly. They fail to produce the expected ranking of systems due to their inability to detect homogeneous excess of aspects' redundancy in parts of the ranking (all aspects have the same excess). This an example of the need for experimental evaluation beyond the idealized axiomatic analysis [40].

When considering the ability of the metrics to assess the correct order of aspect relevance (right), we see that in this case, NRBP,

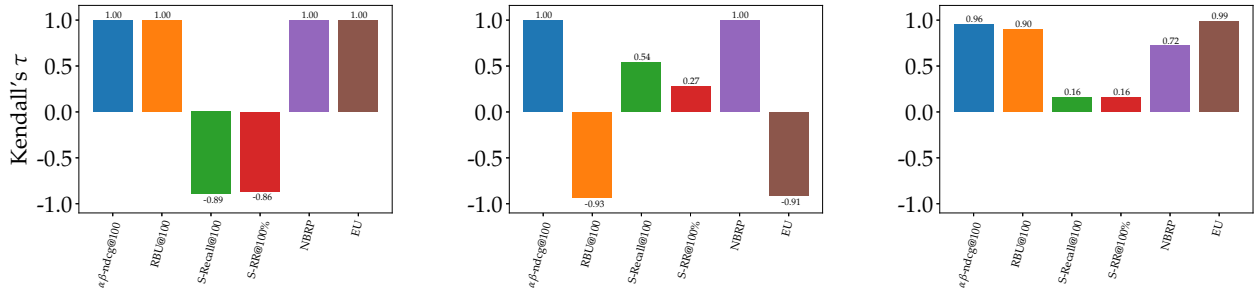

Figure 2: Experiment 1. Correlation between the gradually perturbed systems' ranking for the swapping items bottom to top (left), swapping items from redundant aspect (center), and swapping aspects (right) scenarios.

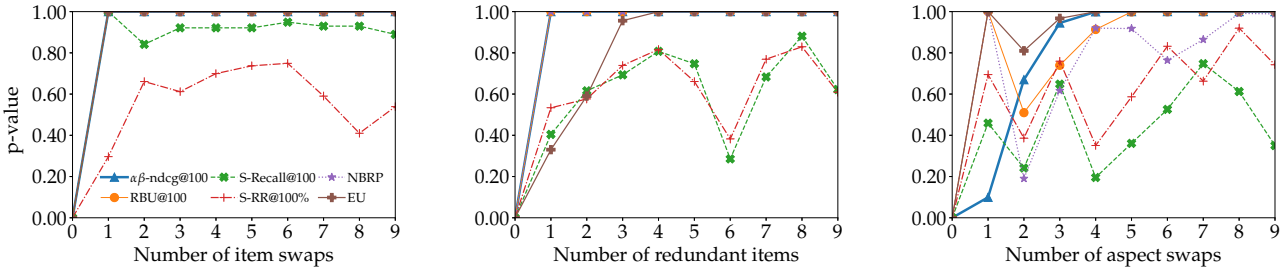

Figure 3: Experiment 2. P-values obtained by the significance test when comparing the performance of gradually worse systems for different metrics against the ideal system, under the three perturbation scenarios: swapping items bottom to top (left), swapping items from redundant aspect (center), and swapping aspects (right).

following the axiomatic analysis, behaves worse than the metrics that explicitly consider aspect weight in their formulation. We also note how other metrics do not have perfect correlation. This is because not all users rated items from 10 aspects, which introduces some noise in the experiment.

5.3 Experiment 2: Discriminative power

Apart from correctly ranking systems, researchers use metrics to assess whether their method is better than the state of the art. That improvement is usually considered by regarding statistical significance. By using statistical tests, researchers have a degree of confidence in the improvement, rather than that it was only due to chance on the data used in the offline evaluation. A good metric should allow us to spot a significant difference easily. To assess this property, we follow the approach that Sakai [38] presented for studying the p-value that the tests produce. However, instead of checking the p-value of all system pairs, we just compute the test p-value for different types of system improvements[1]. Then, we compare incrementally worse rankings (swapping items as in the three scenarios of Experiment 1) and check the p-value returned by the metric's statistical test. In that way, we may expect a curve similar to a power curve for a helpful metric.

In Figure 3, we observe the behaviour of the metrics on spotting statistical differences when the ideal order is altered. In general, the subtopic metrics behave worse than the other ones. Moreover, $\alpha\beta$-nDCG, RBU, NRBP, and EU perform quite well on the item swapping and the redundant items scenarios. The aspect swapping scenario seems to be the most challenging. Here, only $\alpha\beta$-nDCG performs optimally (monotonic fast increment); the other metrics show erratic behavior. This result points that for some users, the metrics cannot detect negative performance variations, making it harder to identify (non) significance.

5.4 Experiment 3: Robustness to Incompleteness

A particularly important property for RSs metrics is robustness to data incompleteness [27]. Missing ratings in offline evaluation have been found to affect the ranking of systems greatly [6]. That is one of the reasons why offline metrics often favour popularity biased methods: Popular items have a higher chance of having user ratings in offline data. The property of a metric to maintain the same system ranking in those situations is called robustness. Valcarce et al. [43] recently studied the robustness of different ranking metrics to missing preferences (sparsity bias). They explore how different metrics behave when gradually removing items from the test split. In this experiment, we follow a similar approach: We gradually remove items from the test split at random, and observe how the systems' ranking evolves. For doing so, we use 50 systems produced by random shuffles of the ideal ranking. When removing ratings, we use proportional random removal by category, i.e., we distribute the item removals among the categories proportionally to the category presence in the original ranking. We produce 50 different removals and report averaged values.

[1]We use a one-sided (null hypothesis corresponding with '*perturbed system performance' is lower than original one's performance*') Wilcoxon Signed-Rank Test on the paired per-user metric values [33, 34].

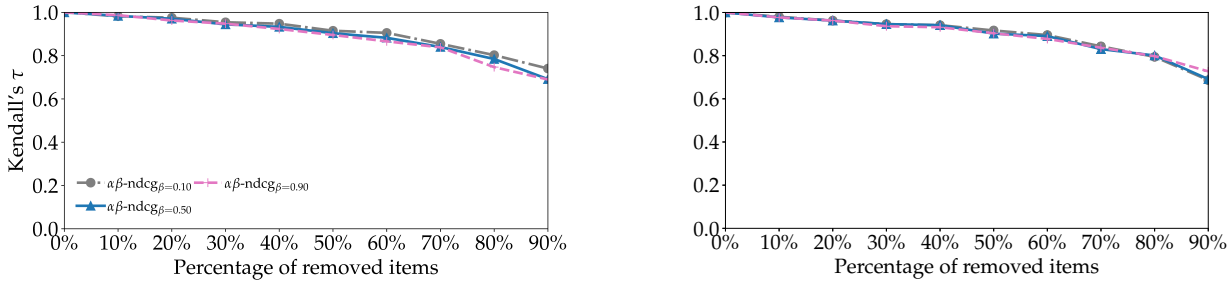

Figure 4: Experiment 3. Correlation between the systems' ranking (50 systems) computed with the whole test set and over gradually smaller test sets for $\alpha\beta$-nDCG with cutoffs 100 (left) and 10 (right), averaged over 50 different removals of ratings.

After Experiments 1 and 2, only our proposed metric behaves as desired for item and aspect relevance, redundancy, and discriminative power. Therefore, in Figure 4, we only show the results of how our metric correlates with itself (noting that a metric that produces no information, e.g. return always 1, would still correlate with itself).

We see that $\alpha\beta$-nDCG is robust to the missing rating effect, even accounting from both item and aspect relevance and aspect redundancy. It still shows a high correlation with itself when a significant proportion of preferences is missing for shallow and deeper cut-offs. This behaviour may be explained not only by the firm foundation of the normalized discounting model, as shown in [43], but also by explicitly considering the missing preferences in its formulation. As commented previously, we just used a small constant factor α for modelling the missing rating effect. In this experiment, we see the importance of that factor. With α being a constant, the lower the β parameter, the greater the influence of accounting for the missing rating. In the case of cutoff 100 (left), the lower the β, the higher the correlation of rankings. This observation suggests the need for further work on properly estimating the missing rating factor.

5.5 Discussion

The results of the experiments show that under our model for the ideal ranking, the adapted $\alpha\beta$-nDCG performs well in all the scenarios. The proposed perfect ordering follows the principles of ranking exploration models previously proposed for the search task [30]. The metric is good at detecting non-optimal item order, aspect distribution and ranking, and topical redundancy accumulations. Moreover, it also behaves well in terms of discriminative power and robustness to incompleteness.

We showed the merits and drawbacks of other metrics designed for search results diversification. In particular, when used in the RSs field, those metrics' performance is affected by how they model aspect weights over user preferences. The results also point out the importance of jointly considering item and topical relevance together with topical redundancy. Metrics that fail to do so tend to underperform in the tested scenarios.

6 CONCLUSIONS

This paper has further explored the problem of finding a unified metric for item relevance and aspect redundancy. We have proposed an adaptation of the existing redundancy aware version of nDCG for the particular needs of evaluating RSs. Further, we have defined a set of axioms that a useful unified metric should follow and have shown how $\alpha\beta$-nDCG satisfies them while other adapted metrics do not. Moreover, we have analyzed the behaviour of the metrics over actual collaborative filtering data, complementing the theoretical analysis.

Our metric accounts for a particular user exploration model derived from the original nDCG. As future work, we will explore other discount approaches based on a better understanding of user behaviour on RSs. Moreover, we will also further study the capacity of the α parameter to reflect different scenarios of missing ratings, including factors such as how popular or how liked a given item is in the collection. Finally, the most challenging next step is to design and execute a detailed user study to validate our metric in terms of the user's perceived value.

ACKNOWLEDGMENTS

The first author wants to thank Google's Visiting Researcher Program and Google AI London team for their warm welcome.

REFERENCES

[1] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying Search Results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM '09)*. ACM, 5–14.

[2] 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)*. ACM, 643–652.

[3] Enrique Amigó, Damiano Spina, and Jorge Carrillo-de Albornoz. 2018. An Axiomatic Analysis of Diversity Evaluation Metrics: Introducing the Rank-Biased Utility Metric. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. ACM, 625 – 634.

[4] Joeran Beel, Marcel Genzmehr, Stefan Langer, Andreas Nürnberger, and Bela Gipp. 2013. A Comparative Analysis of Offline and Online Evaluations and Discussion of Research Paper Recommender System Evaluation. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation* (Hong Kong, China) *(RepSys '13)*. ACM, New York, NY, USA, 7–14.

[5] Alejandro Bellogin, Pablo Castells, and Ivan Cantador. 2011. Precision-Oriented Evaluation of Recommender Systems: An Algorithmic Comparison. In *Proceedings of the Fifth ACM Conference on Recommender Systems* (Chicago, Illinois, USA) *(RecSys '11)*. ACM, New York, NY, USA, 333–336.

[6] Alejandro Bellogín, Pablo Castells, and Iván Cantador. 2017. Statistical biases in Information Retrieval metrics for recommender systems. *Information Retrieval Journal* 20, 6 (2017), 606–634.

[7] Rocío Cañamares, Pablo Castells, and Alistair Moffat. 2020. Offline evaluation options for recommender systems. *Information Retrieval Journal* 23 (2020), 387–410.

[8] Pablo Castells, Neil J. Hurley, and Saul Vargas. 2015. *Novelty and Diversity in Recommender Systems.* Springer, Boston, MA, 881–918.

[9] 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).* ACM, 621–630.

[10] Li Chen, Yonghua Yang, Ningxia Wang, Keping Yang, and Quan Yuan. 2019. How Serendipity Improves User Satisfaction with Recommendations? A Large-Scale User Evaluation. In *Proceedings of The Web Conference 2019 (WWW '19).* ACM, 240–250.

[11] Charles L.A. Clarke, Nick Craswell, Ian Soboroff, and Azin Ashkan. 2011. A Comparative Analysis of Cascade Measures for Novelty and Diversity. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM '11).* ACM, 75–84.

[12] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and Diversity in Information Retrieval Evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08).* ACM, 659–666.

[13] Charles L. Clarke, Maheedhar Kolla, and Olga Vechtomova. 2009. An Effectiveness Measure for Ambiguous and Underspecified Queries. In *Proceedings of the 2nd International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory (ICTIR '09).* Springer, 188–199.

[14] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. 2010. Performance of Recommender Algorithms on Top-n Recommendation Tasks. In *Proceedings of the Fourth ACM Conference on Recommender Systems (RecSys '10).* ACM, 39–46.

[15] Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. 2014. User Perception of Differences in Recommender Algorithms. In *Proceedings of the 8th ACM Conference on Recommender Systems* (Foster City, Silicon Valley, California, USA) *(RecSys '14).* ACM, New York, NY, USA, 161–168.

[16] Hui Fang and ChengXiang Zhai. 2005. An Exploration of Axiomatic Approaches to Information Retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05).* ACM, 480–487.

[17] 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).* ACM, 21–30.

[18] Nicola Ferro, Norbert Fuhr, Gregory Grefenstette, Joseph A. Konstan, Pablo Castells, Elizabeth M. Daly, Thierry Declerck, Michael D. Ekstrand, Werner Geyer, Julio Gonzalo, Tsvi Kuflik, Krister Lindn, Bernardo Magnini, Jian-Yun Nie, Raffaele Perego, Bracha Shapira, Ian Soboroff, Nava Tintarev, Karin Verspoor, Martijn C. Willemsen, and Justin Zobel. 2018. The Dagstuhl Perspectives Workshop on Performance Modeling and Prediction. *SIGIR Forum* 52, 1 (Aug. 2018), 91–101.

[19] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. 2010. Beyond Accuracy: Evaluating Recommender Systems by Coverage and Serendipity. In *Proceedings of the Fourth ACM Conference on Recommender Systems* (Barcelona, Spain) *(RecSys '10).* ACM, New York, NY, USA, 257–260.

[20] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 19 (Dec. 2015), 19 pages.

[21] Hussein Hazimeh and ChengXiang Zhai. 2015. Axiomatic Analysis of Smoothing Methods in Language Models for Pseudo-Relevance Feedback. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval (ICTIR '15).* ACM, 141–150.

[22] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. 2004. Evaluating Collaborative Filtering Recommender Systems. *ACM Trans. Inf. Syst.* 22, 1 (Jan. 2004), 5–53.

[23] Katja Hofmann, Lihong Li, and Filip Radlinski. 2016. Online Evaluation for Information Retrieval. *Found. Trends Inf. Retr.* 10, 1 (June 2016), 1–117.

[24] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (Oct. 2002), 422–446.

[25] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. 2009. Controlled Experiments on the Web: Survey and Practical Guide. *Data Min. Knowl. Discov.* 18, 1 (Feb. 2009), 140–181.

[26] Andrii Maksai, Florent Garcin, and Boi Faltings. 2015. Predicting Online Performance of News Recommender Systems Through Richer Evaluation Metrics. In *Proceedings of the 9th ACM Conference on Recommender Systems* (Vienna, Austria) *(RecSys '15).* ACM, New York, NY, USA, 179–186.

[27] Benjamin M. Marlin, Richard S. Zemel, Sam T. Roweis, and Malcolm Slaney. 2011. Recommender systems: Missing data and statistical model estimation. In *IJCAI International Joint Conference on Artificial Intelligence.*

[28] Sean M. McNee, John Riedl, and Joseph A. Konstan. 2006. Being Accurate is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems (CHI EA '06).* ACM, 1097–1101.

[29] Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. 2017. Incorporating User Expectations and Behavior into the Measurement of Search Effectiveness. *ACM Trans. Inf. Syst.* 35, 3, Article 24 (June 2017), 38 pages.

[30] Alistair Moffat, Paul Thomas, and Falk Scholer. 2013. Users versus Models: What Observation Tells Us about Effectiveness Metrics. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management* (San Francisco, California, USA) *(CIKM '13).* ACM, New York, NY, USA, 659–668.

[31] 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.

[32] Ali Montazeralghaem, Hamed Zamani, and Azadeh Shakery. 2016. Axiomatic Analysis for Improving the Log-Logistic Feedback Model. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16).* ACM, 765–768.

[33] Javier Parapar, David E. Losada, and Álvaro Barreiro. 2021. Testing the Tests: Simulation of Rankings to Compare Statistical Significance Tests in Information Retrieval Evaluation. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing* (Virtual Event, Republic of Korea) *(SAC '21).* ACM, New York, NY, USA, 655–664.

[34] Javier Parapar, David E. Losada, Manuel A. Presedo Quindimil, and Alvaro Barreiro. 2020. Using score distributions to compare statistical significance tests for information retrieval evaluation. *J. Assoc. Inf. Sci. Technol.* 71, 1 (2020), 98–113.

[35] Pearl Pu, Li Chen, and Rong Hu. 2011. A User-Centric Evaluation Framework for Recommender Systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys '11).* ACM, 157–164.

[36] Daniël Rennings, Felipe Moraes, and Claudia Hauff. 2019. An Axiomatic Approach to Diagnosing Neural IR Models. In *Advances in Information Retrieval*, Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra (Eds.). Springer, Cham, 489–503.

[37] Alan Said and Alejandro Bellogín. 2018. Coherence and inconsistencies in rating behavior: estimating the magic barrier of recommender systems. *User Modeling and User-Adapted Interaction* 28, 2 (2018), 97–125.

[38] Tetsuya Sakai. 2006. Evaluating Evaluation Metrics Based on the Bootstrap. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06).* ACM, 525–532.

[39] Tetsuya Sakai and Ruihua Song. 2011. Evaluating Diversified Search Results Using Per-Intent Graded Relevance. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Beijing, China) *(SIGIR '11).* ACM, 1043–1052.

[40] 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).* ACM, 595–604.

[41] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2015. Search Result Diversification. *Found. Trends Inf. Retr.* 9, 1 (March 2015), 1–90.

[42] Javier Sanz-Cruzado, Craig Macdonald, Iadh Ounis, and Pablo Castells. 2020. Axiomatic Analysis of Contact Recommendation Methods in Social Networks: An IR Perspective. In *Advances in Information Retrieval*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer, Cham, 175–190.

[43] Daniel Valcarce, Alejandro Bellogín, Javier Parapar, and Pablo Castells. 2018. On the Robustness and Discriminative Power of Information Retrieval Metrics for Top-N Recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18).* ACM, 260–268.

[44] Daniel Valcarce, Alejandro Bellogín, Javier Parapar, and Pablo Castells. 2020. Assessing ranking metrics in top-N recommendation. *Inf. Retr. J.* 23, 4 (2020), 411–448.

[45] Daniel Valcarce, Javier Parapar, and Álvaro Barreiro. 2017. Axiomatic analysis of language modelling of recommender systems. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 25, Suppl. 2 (2017), 113–127.

[46] Saúl Vargas. 2014. Novelty and Diversity Enhancement and Evaluation in Recommender Systems and Information Retrieval. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14).* ACM, 1281.

[47] Saúl Vargas. 2015. *Novelty and Diversity Evaluation and Enhancement in Recommender Systems.* Ph.D. Dissertation. Universidad Autonoma de Madrid.

[48] Saúl Vargas and Pablo Castells. 2011. Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems* (Chicago, Illinois, USA) *(RecSys '11).* ACM, New York, NY, USA, 109–116.

[49] Yiming Yang and Abhimanyu Lad. 2009. Modeling Expected Utility of Multi-Session Information Distillation. In *Proceedings of the 2nd International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory (ICTIR '09).* Springer, 164–175.

[50] ChengXiang Zhai, William W. Cohen, and John Lafferty. 2015. Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. *SIGIR Forum* 49, 1 (June 2015), 2–9.

[51] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. 2005. Improving Recommendation Lists through Topic Diversification. In *Proceedings of the 14th International Conference on World Wide Web* (Chiba, Japan) *(WWW '05).* ACM, New York, NY, USA, 22–32.