Dealing with Incomplete Judgments in Cascade Measures

Kai Hui
Max Planck Institute for Informatics
Saarbrücken Graduate School of
Computer Science
Saarbrücken, Germany
khui@mpi-inf.mpg.de

Klaus Berberich
htw saar
Max Planck Institute for Informatics
Saarbrücken, Germany
kberberi@mpi-inf.mpg.de

Ida Mele
Faculty of Informatics
Università della Svizzera italiana
Lugano, Switzerland
ida.mele@usi.ch

ABSTRACT

Cascade measures like α-nDCG, ERR-IA, and NRBP take into account novelty and diversity of query results and are computed using judgments provided by humans, which are costly to collect. These measures expect that all documents in the result list of a query are judged and cannot make use of judgments beyond the assigned labels. Existing work has demonstrated that condensing the query results by taking out documents without judgment can address this problem to some extent. However, how highly incomplete judgments can affect cascade measures and how to cope with such incompleteness have not been addressed yet. In this paper, we propose an approach which mitigates incomplete judgments by leveraging the content of documents relevant to the query's subtopics. These language models are estimated at each rank taking into account the document and the upper ranked ones. Then, our method determines gain values based on the Kullback-Leibler divergence between the language models. Experiments on the diversity tasks of the TREC Web Track 2009–2012 show that with only 15% of the judgments our method accurately reconstructs the original rankings determined by the established cascade measures.

CCS CONCEPTS

•**Information systems →Retrieval models and ranking;** *Web searching and information discovery;*

ACM Reference format:
Kai Hui, Klaus Berberich, and Ida Mele. 2017. Dealing with Incomplete Judgments in Cascade Measures. In *Proceedings of ICTIR '17, October 1–4, 2017, Amsterdam, The Netherlands, ,* 8 pages.
DOI: http://dx.doi.org/10.1145/3121050.3121064

1 INTRODUCTION

Information retrieval systems are evaluated based on their ability to return documents that are relevant to the query as well as on the novelty and diversity of the results. This is of valuable importance especially for faceted/ambiguous queries (e.g., java, jaguar, python) which have more than one possible interpretation. To avoid redundant information, the information retrieval system is supposed to show results that cover all possible different subtopics of the query (also known as *aspects* or *facets*). Recent years have seen an increasing interest in novelty and diversity as features complementary to relevance [13–15]. Cascade measures, such as α-nDCG, ERR-IA, and NRBP, have been proposed and adopted widely to evaluate novelty and diversity of a ranked list of query results. They reward the documents that diversify the result list in order to cover all possible information needs behind a faceted/ambiguous query. To quantify the novelty and diversity of the ranked list of query results, the cascade measures usually sum up the gain value of each result which is evaluated based on the relevance to the query and the novelty to the ranking. Such measures require manual judgments done by humans, which is costly to collect in terms of time and of money. An alternative is to partially evaluate the results which could lead to inaccurate computations. Hence, it is desirable to have a more effective use of the available judgments to better trade off the cost for evaluating results and the inaccurateness of the measure computations.

Several authors have investigated the reusability of test collections for diversification, comparing the influence of pooling depths and system bias [6, 20, 22]. Specifically, in Sakai et al. [22], the reusability is examined in terms of employing judgments collected for different pooling depths and of system bias in a leave-one-out experiment [6]. Moreover, a condensed-list method [20] was employed to improve the reusability by removing unjudged documents from the query results before the evaluation. It has been demonstrated that a condensed list can address the issue of incomplete judgments, but still a significant amount of judgments is needed (e.g., more than 50% of judgments [22]). A natural question that arises is how cascade measures and the condensed-list approach behave when substantially fewer judgments (e.g., less than 30%) are available. Subsequently, a second question worth exploring is how the very few judgments can be fully leveraged, namely, beyond a single label, to assess the effectiveness of the retrieval task.

To address these questions, we first investigate the behavior of cascade measures when judgments are highly incomplete and are only available for less than 30% of documents. In addition, inspired by the work on dealing with incomplete judgments in ad-hoc retrieval [6, 16, 18], we devise novel measures that use documents' contents to approximate the established cascade measures for novelty and diversity. The proposed measures are especially useful when very few judgments are available, by adequately employing the content of the judged documents in place of the labels.

Instead of directly using relevance judgments, our measures compare the language models estimated based on the contents of the returned documents against the language models estimated based on the contents of those documents that are judged as relevant to the different subtopics of the query. We estimate the gain values

ICTIR '17, October 1–4, 2017, Amsterdam, The Netherlands
© 2017 ACM. ISBN 978-1-4503-4490-6/17/10…$15.00
DOI: http://dx.doi.org/10.1145/3121050.3121064

of the documents from the language models' divergences which are then plugged into the established cascade measures. Intuitively, a system is rewarded, if it returns documents that are content-wise similar to documents considered relevant to the different subtopics of a query. If the system returns a non-relevant document, though, the language model estimated based on its result will diverge more from the subtopic language models.

We first examine the effectiveness of the established cascade measures over the raw list and a condensed list based on few judgments. Beyond that, we propose methods based on the observations that relevant documents for a subtopic tend to be homogeneous, i.e., relatively similar to each other. Irrelevant documents, in contrast, tend to exhibit widely different contents. To model the features that make the documents relevant to the possible subtopics of a query, our approach estimates a language model from the documents' contents. Likewise, our method estimates a language model at each rank of the returned results to characterize what the users can see when they go through the ranked list of the results of a query. Gain values are then determined based on the Kullback-Leibler divergence between the estimated language models. Exploring the design space, we consider different approaches with different ways to determine the gain values and aggregate them into a single measure. In total, we end up with four novel cascade measures, coined as AbsNb, AbsRb, DeltaNb, and DeltaRb. Differently from existing measures which explicitly penalize redundancy in the form of repetitions of the same label (on the same subtopic), our measures implicitly capture redundancy via the estimated language models.

The contribution of this paper is threefold:

- We study the robustness of the established cascade measures, i.e., α-nDCG, ERR-IA, and NRBP, as well as their condensed-list versions in presence of highly incomplete judgments;
- We propose four novel cascade measures that can robustly approximate α-nDCG, ERR-IA, and NRBP with very few judgments. Such novel measures allow reusing relevance judgments collected on one document collection to evaluate systems on another document collection;
- We performed a comprehensive experimental evaluation of our novel cascade measures based on ClueWeb09 datasets, using queries and runs from the TREC Web Track 2009–12.

Organization. The rest of this paper is organized as follows. We discuss related work in Section 2. Section 3 gives some background on effectiveness measures for novelty and diversity. Section 4 introduces our novel cascade measures. Our experimental evaluation is described in Section 5, before concluding in Section 6.

2 RELATED WORK

Measures for Novelty and Diversity. Standard effectiveness measures evaluate IR systems only in terms of the relevance of returned results, while other measures attempt to capture also their diversity and novelty. Zhai et al. [28] proposed the *subtopic recall* to measure the percentage of subtopics covered by a list of query results. Agrawal et al. [1] focused on ambiguous queries and presented intent-aware variants of established effectiveness measures. Differently from nDCG [17], which assumes independence of documents' relevance, Chapelle et al. [10] proposed the Expected Reciprocal Rank (ERR) measure, capturing the dependency among the documents by assuming a cascade-style user model. Extending ERR, Chapelle et al. [9] proposed ERR-IA to further measure the diversification of the ranking, following the general approach by Agrawal et al. [1]. Clarke et al. [11] considered underspecified queries, namely, queries with faceted interpretations. They presented α-nDCG which decomposes the information needs behind a query into so-called *information nuggets* and defines the utility of a document as the number of novel nuggets covered by it. In a follow-up work they proposed NRBP [12] which considers both ambiguous and underspecified (faceted) queries by combining α-nDCG and *Rank-Biased Precision* (RBP) proposed by Moffat and Zobel [19]. More recently, Want et al. [25] presented several hierarchical measures that consider the relationships among different subtopics. Our measures follow the cascade models, but differ from them since we implicitly capture the dependency between the returned results in the estimated query-result language model.

Dealing with Incomplete Judgments. To deal with the incompleteness of judgments, researchers have developed novel measures for robust evaluation when judgments are incomplete. Buckley and Voorhees [5] proposed *bpref* which uses binary relevance and completely ignores search results for which no relevance information is available. Yilmaz and Aslam [26] as well as Aslam et al. [3] presented approaches for random sampling to estimate the actual values of the average precision when relevance judgments are incomplete. Similarly, Sakai and Kando [23] applied traditional effectiveness measures to "condensed" lists which are ranked lists of documents obtained by removing all unjudged documents. Carterette et al. [8] analyzed the distribution of average precision over all possible assignments of relevance to all unjudged documents and proposed a method to construct a test collection with minimal relevance judgments. All of these works focused on traditional effectiveness measures (e.g., average precision), whereas we focus on more recent cascade measures for novelty and diversity. Moreover, differently from these works, we use the contents of documents judged as relevant when determining our measures. Beyond that, Amitay et al. [2] used the relevant and irrelevant term sets to evaluate IR systems, enabling evaluation on a dataset without judgments. Carterette and Allan [7] as well as Büttcher et al. [6] tried to make up for missing relevance judgments by predicting them using machine learning. Similarly to our work, they also make use of relevant documents' contents in their prediction models. Making use of machine learning and plugging predicted relevance judgments into cascade measures for novelty and diversity is an interesting direction for future research but orthogonal to our approach.

Finally, the works [21, 22] are the closest to ours in the sense that the incomplete judgments are considered in the context of diversification. In both papers, the situation of incomplete judgments from leave-one-out experiments and from expansion of judgment pooling are examined, but more than 50% of judgments are assumed to be available. Differently, in our work, we consider the situation when judgments are highly incomplete, that is, only between 1% and 50% are available, as shown in Section 5.2. Moreover, we also take into account the situation when no judgment is available at all by evaluating on disjoint document collections as shown in

Section 5.3. Hence, our work can be regarded as complementary to [21, 22] when established cascade measures fail to work.

3 BACKGROUND

In this section, we briefly describe existing evaluation measures for novelty and diversity, namely α-nDCG, ERR-IA, and NRBP.

α-nDCG. Clarke et al. [11] extended the traditional nDCG [17] to α-nDCG to capture novelty and diversity in query results. α-nDCG scores a query-result list by rewarding results relevant to new subtopics and penalizing the ones relevant to already covered subtopics. It balances relevance and diversity through the tuning of the α parameter. We used $\alpha = 0.5$ for our experiments to give equal importance to relevance and diversity, following the default setting in Trec[1]. α-nDCG includes a *novelty-biased gain* as indicated in Equation 1, where m is the number of query subtopics, k the number of results. $J_i(r)$ indicates the relevance of the document at rank r relative to the intent i, $C_i(r-1)$ is the number of times the subtopic i has been covered by documents appearing before rank r, and *IDCG* serves to normalize the measure.

$$\alpha\text{-}nDCG = \frac{\sum\limits_{r=1}^{k} \frac{\sum_{i=1}^{m} J_i(r)(1-\alpha)^{C_i(r-1)}}{log_2(1+r)}}{IDCG} \ . \qquad (1)$$

ERR-IA. The intent-aware version of the *Expected Reciprocal Rank* (ERR) has been proposed by Chapelle et al. [10], which is defined as the weighted average of ERR computed separately for each query subtopic [9] as summarized in Equation 2. $R_i(j)$ is a function of the relevance grade for subtopic i of the document at position j in the ranking. It is commonly defined as $(2^g - 1)/2^{g_{max}}$, where g is the grade given by the judges to the document. The scores are weighted by p_i which is the probability of the intent i.

$$ERR - IA = \sum_{i=1}^{m} p_i \sum_{r=1}^{k} \frac{1}{r} R_i(r) \prod_{j=1}^{r-1} (1 - R_i(j)) \ , \qquad (2)$$

NRBP. The *Novelty- and Rank-Biased Precision* was proposed by Clarke et al. [12] to combine α-nDCG and *Rank-Bias Precision* (RBP) as originally proposed by Moffat and Zobel [19]. It is defined as in Equation 3. This measure uses two discount mechanisms. Redundancy is penalized based on the parameter α, whereas the persistence by the parameter β. For our experiments we set α and β as 0.5, following the default configuration in Trec.

$$NRBP = \frac{1-(1-\alpha)\beta}{m} \sum_{r=1}^{\infty} \beta^{r-1} \sum_{i=1}^{m} J_i(r)(1-\alpha)^{C_i(r)} \ . \qquad (3)$$

4 DIVERGENCE-BASED MEASURES

In this section we introduce a family of novel measures. In contrast to existing ones, which only digest relevance judgments provided by humans, our measures operate on the content of the judged documents. Our approach indirectly provides robust measures which are able to deal with substantially incomplete relevance judgments, as we demonstrate in our experimental results.

4.1 Model

The document collection is denoted as $\mathcal{D}$. Each document is represented as a bag of words drawn from a vocabulary $\mathcal{V}$ consisting of indexed terms. For a term $v \in \mathcal{V}$ we use $tf(v,d)$ to denote its term frequency in document $d \in \mathcal{D}$ and let $|d| = \sum_{v \in \mathcal{V}} tf(v,d)$ be the document length. We refer to the subtopics of a query q as $\{q_1, \ldots, q_m\}$ with m subtopics, and let $r(q_i,d)$ be a predicate indicating the (binary) relevance of document d to subtopic q_i. Finally, we refer to a query result list as $R = \langle d_1, \ldots, d_{|R|} \rangle$ and as $R_k = \langle d_1, \ldots, d_k \rangle$ to its corresponding top-k results.

4.2 Statistical Language Models

Within the last two decades, statistical language models have also been successfully applied to Information Retrieval tasks [27]. In this work, language models serve two purposes. First, they can be used to model the characteristics that make a document relevant to a specific subtopic. Second, they capture what the users can see while sifting through the query's result list. While more advanced language models have been proposed (e.g., based on n-grams or allowing for term translations), for simplicity, we will focus on unigram language models with Dirichlet smoothing.

Top-k Query Result Language Model. From the top-k query result R_k we estimate a language model Θ_{R_k} as in Equation 5, where μ is a tunable parameter (set as μ = 2,500 [27]) which controls the influence of Dirichlet smoothing with the language model $\Theta_{\mathcal{D}}$ estimated from the document collection as in Equation 4.

$$P[v|\Theta_{\mathcal{D}}] = \frac{\sum_{d \in \mathcal{D}} tf(v,d)}{\sum_{d \in \mathcal{D}} |d|} \ . \qquad (4)$$

$$P[v|\Theta_{R_k}] = \frac{\sum_{d \in R_k} tf(v,d) \ + \ \mu}{\sum_{d \in R_k} |d| \ + \ \mu P[v|\Theta_{\mathcal{D}}]} \ . \qquad (5)$$

The language model Θ_{R_k} thus captures what users see when they inspect all the documents up to rank k in the query's result list. The smoothing with the document collection language model $\Theta_{\mathcal{D}}$ can be interpreted as their prior knowledge about general documents from the collection. By its definition, Θ_{R_k} captures the degree of diversity in the top-k query results. Intuitively, when homogeneous documents related to a single subtopic are returned, the estimated language model Θ_{R_k} will have lower entropy than in the case when heterogeneous documents related to various subtopics are returned. Moreover, Θ_{R_k} comes with an inherent bias against documents returned at lower ranks. When comparing Θ_{R_k} and $\Theta_{R_{k+1}}$ it is clear from the definition that the influence of the additional result on the estimate decreases as k increases.

Subtopic Language Models. Given a query q and its subtopics $\left\{q_1, \ldots, q_{|q|}\right\}$, we estimate a language model

$$P[v|\Theta_{q_i}] = \frac{\sum_{d \in \mathcal{D} \,:\, r(q_i,d)} tf(v,d) \ + \ \mu}{\sum_{d \in \mathcal{D} \,:\, r(q_i,d)} |d| \ + \ \mu P[v|\Theta_{\mathcal{D}}]} \qquad (6)$$

for each subtopic based on its relevant documents, again smoothed with the document collection language model $\Theta_{\mathcal{D}}$. The purpose of smoothing is twofold, namely, to avoid zero probabilities and to achieve a relative weighting of terms for the following divergence computation.

[1] http://trec.nist.gov/data/web2012.html

4.3 Divergence-Based Gain

We obtain the document gain values by comparing the language models estimated for subtopics and top-k results. In more details, let Θ_{q_i} be a subtopic language model and Θ_{R_k} be a top-k query result language model estimated as described above. For comparing the language models we can apply the Kullback-Leibler divergence

$$KLD(\Theta_{q_i} \parallel \Theta_{R_k}) = \sum_{v \in \mathcal{V}} P[v|\Theta_{q_i}] \log \left(\frac{P[v|\Theta_{q_i}]}{P[v|\Theta_{R_k}]} \right) , \qquad (7)$$

which ranges in $[0, \infty]$. We thus obtain high values of $KLD(\Theta_{q_i} \| \Theta_{R_k})$ when the top-k results in R_k are different from the documents relevant to the subtopic q_i, for instance, they use different terminology or key terms. Hence, we compute the per-subtopic gain value as

$$g(i,k) = max(0, 1 - \frac{KLD(\Theta_{q_i} \parallel \Theta_{R_k})}{KLD(\Theta_{q_i} \parallel \Theta_{\mathcal{D}})}) , \qquad (8)$$

which is normalized with the Kullback-Leibler divergence observed for the document collection language model $\Theta_{\mathcal{D}}$. According to our preliminary experiments, there always exist $KLD(\Theta_{q_i} \parallel \Theta_{R_k}) \leq KLD(\Theta_{q_i} \| \Theta_{\mathcal{D}})$ in our data, leading to $g(i,k) \in [0, 1]$. To turn these per-subtopic gain values $g(i,k)$ into per-rank gain values, which can then be aggregated, we consider two alternative formulations, coined as Abs and Delta.

Abs determines a per-rank gain value as

$$g(j) = \max_{1 \leq i \leq |q|} g(i,j) , \qquad (9)$$

thus rewarding query results whose top-j covers at least one of the subtopics well.

Delta derives per-rank gain values from the observed differences in per-subtopic gain values as

$$g(j) = \max \left(0, \max_{1 \leq i \leq |q|} (g(i,j) - g(i,j-1)) \right) . \qquad (10)$$

Note that the outer maximum function in Equation 10 guarantees that $g(j) \geq 0$. Given a query result, to obtain a high per-rank gain value under this formulation, its document at rank j must be closely related to a subtopic that has not been covered yet.

4.4 Position Bias

As a final step, we describe how the per-rank gain values $g(j)$ can be aggregated into a single measure reflecting the quality of a top-k result list. We propose Nb and Rb formulations.

Nb. By definition, as described above, in our approach the influence of documents at lower ranks is diminishing. Thus, we can simply sum up the per-rank gain values observed at ranks up to k as

$$\sum_{1 \leq j \leq k} g(j) . \qquad (11)$$

This formulation is referred to as Nb which stays for *no bias.*

Rb. We borrow the position-bias model from Rank-Biased Precision [19] and aggregate the per-rank gain values as

$$(1 - \theta) \cdot \sum_{1 \leq j \leq k} g(j) \cdot \theta^{j-1} . \qquad (12)$$

The parameter θ (set as $\theta = 0.8$ based on our pilot experiments to force a dramatic decay) models the user's persistence in sifting through the query result, or put differently, at each rank the user decides to stop inspecting query results with probability $(1 - \theta)$.

5 EXPERIMENTAL EVALUATION

In this section, we design experiments to investigate the reliability of established cascade measures and to examine our proposed measures under three different aspects: (i) the robustness when only few judgments available, (ii) how well they can reuse relevance judgments to evaluate systems on a previously unseen document collection, and (iii) their correlation with existing cascade measures.

5.1 Setup

Document Collection. We use ClueWeb09 [2] as a document collection. In our experiments, we focus on the subset of more than 500 million English web pages, which are known as ClueWeb09 Category A (CwA). For our robustness and reusability experiments, we also make use of ClueWeb09 Category B (CwB), as a well-defined subset of about 50 million English web pages. As a third subset of English web pages, called CwC, we consider all 450 million web pages that are part of CwA but not of CwB.

Queries & Relevance Judgments. We use data from the diversity track of the TREC Web Track 2009–2012. This leaves us with a total of 200 queries (50 per year) and their corresponding relevance judgments. For our methods we convert graded labels into binary ones by treating labels -2 (spam) and 0 (non-relevant) as irrelevant and all other labels as relevant. To compare our measures against existing cascade measures, we also obtained the runs submitted by participants of the TREC Web Track. There are 48 runs for 2009, 32 runs for 2010, 62 runs for 2011, and 48 runs for 2012. Some of the runs were produced considering only CwB; others considered the whole document collection CwA. As standard in TREC Web Track, we consider top-20 query results when comparing different systems.

Cascade Measures. As established cascade measures we consider α-nDCG, ERR-IA, and NRBP which are described in Section 3. Regarding our novel cascade measures, we combine the different design choices of per-rank gain and position bias and obtain four novel measures:

- AbsNb combining the absolute gain Abs from Eq. 9 with no position bias Nb as in Eq. 11;
- AbsRb combining the absolute gain Abs from Eq. 9 with the ranking bias Rb as in Eq. 12;
- DeltaNb combining the delta gain Delta from Eq. 10, with no position bias Nb as in Eq. 11;
- DeltaRb combining the delta gain Delta from Eq. 10, with the ranking bias Rb as in Eq. 12.

Rank Correlation. We use Kendall's τ as a correlation measure between system rankings determined by different cascade measures. Kendall's τ is the difference between concordant and discordant pairs divided by the total number of pairs. It ranges in $[-1, 1]$ with 1 (−1) indicating perfect agreement (disagreement). Vorhees [24] suggests 0.9 as a threshold to consider two rankings as equivalent, whereas a correlation below 0.8 reflects a significant difference. In this work, however, given the difficulty of our task, exploiting test

[2] http://www.lemurproject.org/clueweb09.php/

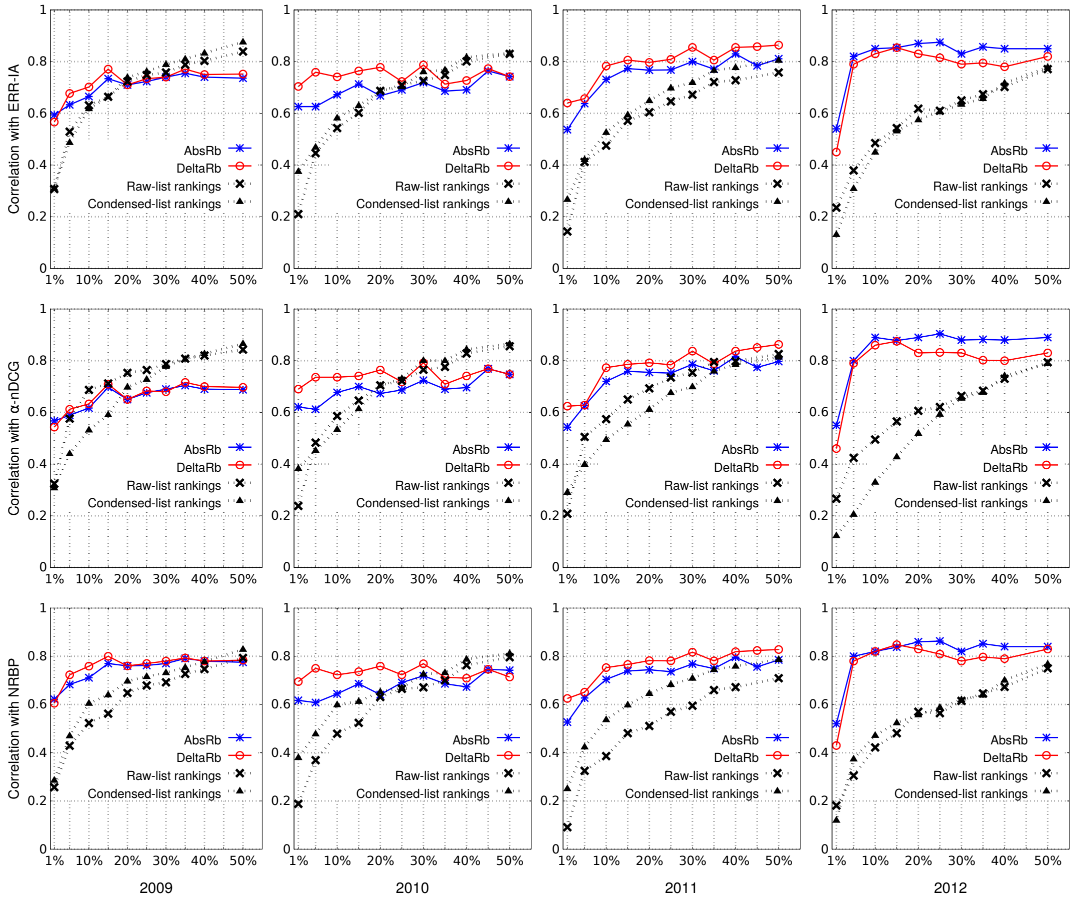

Figure 1: Effectiveness of cascade measures over the raw list and the condensed list. Rows correspond to ERR-IA, α-nDCG, and NRBP (top to bottom). Columns correspond to TREC Web Track 2009–2012 (left to right). In each figure, the x-axis indicates the sampling percentage $p\%$ and the y-axis indicates the Kendall's τ correlation.

collections with very few available judgments, namely, below 50%, we choose 0.8 as a threshold.

5.2 Robustness over Highly Incomplete Judgments

Firstly, we analytically investigate the effectiveness of different cascade measures when evaluating the raw list and the condensed list [20] of search results. Beyond that, we evaluate the proposed measures under the same set of judgments and make comparisons. In particular, we inspect the correlation between system rankings determined by different measures on incomplete judgments and the ones determined by established cascade measures over complete judgments. To do so, we follow Bompada et al. [4] and Buckley et al. [5] and denote the full relevance judgment document set as $qrel$. Given a query, we randomly shuffle the relevant documents in $qrel$ and pick up the first $max(1, \lceil p\%|qrel| \rceil)$ relevant documents from $qrel$ to build Θ_q. For each query we require at least one relevant document to construct our measures, and the relevance judgments are all that is required by our proposed measures. To compare with results based on established measures over condensed lists, we further sample $p\%$ non-relevant documents to construct an incomplete judgment set including only $p\%$ of the judgments. To remove the randomness of this sampling procedure, we report the average results based on 30 repetitions. Although this stratified random sampling is analytical, it can cover different situations through dozens of samplings.

The Kendall's τ correlations between our rankings and the ones under complete judgments with established measures are summarized in Figure 1. Each column corresponds to a query set (i.e., one year of TREC topics); each row represents one of the established cascade measures. The two dashed curves represent the system rankings determined by established measures, namely, ERR-IA, α-nDCG and NRBP, when measuring on raw lists and condensed lists, respectively. We only display results for AbsRb and DeltaRb, denoted as two solid curves, which look similar to AbsNb and DeltaNb, respectively. The x-axis indicates the sample percentage $p\%$, and the y-axis is the Kendall's τ correlation.

From Figure 1, it can be seen that the established measures require more than 40%-50% judgments to achieve 0.8 Kendall's τ correlation. Sakai et. al [22] demonstrated that the condensed-list

Table 1: Reusability of our measures. Relevance judgments collected on CwC (i.e., CwA-CwB) are used to evaluate systems on the disjoint document collection CwB. Kendall's τ above 0.8 are shown in bold.

		AbsNb	AbsRb	DeltaNb	DeltaRb
2009	α-nDCG	.72	.70	.73	.69
	ERR-IA	.78	.78	.76	.76
	NRBP	.74	.74	.70	.74
2010	α-nDCG	.72	.66	.74	.76
	ERR-IA	.70	.66	.72	.75
	NRBP	.68	.65	.71	.73
2011	α-nDCG	.71	.76	.79	**.81**
	ERR-IA	.67	.75	.73	**.81**
	NRBP	.64	.74	.69	.79
2012	α-nDCG	.23	.40	.31	.51
	ERR-IA	.26	.44	.31	.54
	NRBP	.26	.45	.31	.54

methods can address the incomplete-judgment issues in leave-one-out experiment [6]. However, it is clear from Figure 1 that with highly incomplete judgments (i.e., when less than 30% judgments are available), the correlation for condensed lists can be very low, e.g., lower than 0.4 when less than 1% judgments are available. This is not surprising since the highly incomplete judgments make the computation of the established measures highly depend on the very few documents that have been judged. Put differently, an unjudged document directly corresponds to a missing component in the formula of these measures. Our proposed measures behave much more smoothly, because they make use of all the judged documents instead of few documents labeled for a single query. Actually, even with only a single judged relevant documents, one can still estimate a reasonable language model Θ_q based on it, given that documents relevant to a query tend to have similar content. As a concrete example, the correlation numbers for the established measures vary a lot among different years, while the DeltaRb still has a Kendall's τ correlation above 0.8 for the year 2011 with as little as 15% of relevance judgments. Likewise, for the year 2012 we can observe a Kendall's τ correlation above 0.8 with as little as 5% relevance judgments. For the years 2009 and 2010, the proposed measures fail to get beyond 0.8 Kendall's τ, yet they achieve significantly higher correlation compared to the established measures when less than 15% judgments are available. Note that, differently from the established measures, both DeltaRb and AbsRb behave rather robustly when different amounts of judgments are available, and the correlation values do not increase monotonically. This is due to the fact that more judgments can only adjust Θ_q by including more observations of the distribution, which is fundamentally different from the way when computing established measures by taking individual relevant judgments into computation.

5.3 Reusability on Disjoint Document Collection

As a second aspect, we examine whether our measures are able to reuse relevance judgments collected on one document collection to evaluate systems on another (disjoint) document collection. Note that this setting is different from the one described in Section 5.2 with $p = 0\%$, which corresponds to having no relevance judgments available at all and is beyond hope for any measure. Instead, we estimate subtopic language models based on documents from CwC. Our objective is then to approximate the system rankings determined by ERR-IA and α-nDCG on CwB, which by construction is disjoint from CwC (we recall that CwC is the set of documents that appear in CwA but not in CwB). In this context, it is worth mentioning that CwB, despite of its smaller size, comes with 1.5× more relevance judgments than CwC, which is due to the facts that a lot of systems in TREC opted to work on CwB.

Table 1 reports the obtained Kendall's τ correlations. It can be seen that DeltaRb performs better among the proposed measures. Although only in 2011 over 0.8 correlation can be achieved, in other years the correlation is beyond 0.5, and in 2009-10 it is around 0.75. Note that the established cascade measures, in contrast, can not be employed in this setting due to the complete mismatch between relevance judgments and result documents. This actually highlights the advantages of our proposed measures in fully utilizing judged documents that do not appear in the evaluated query results.

In Table 1, we can observe relatively low values for the year 2012. Digging deeper we wanted to investigate the question to what extent the reusability depends on the document collection on which the relevance judgments were collected. Therefore, for the year 2012, we further employ all our document collections CwA, CwB, and CwC as a source of relevance judgments and study correlation with α-nDCG, ERR-IA, and NRBP on all these three document collections. Recall that CwC is disjoint from CwB, while both CwB and CwC are subsets of CwA. Table 2 shows Kendall's τ correlations for all combinations of document collections. From the table we can see that the choice of document collection on which relevance judgments are collected can have a significant impact. Thus, Kendall's τ correlations are generally higher for relevance judgments collected on CwA and CwB than on CwC. This is not completely surprising, given that many participants of TREC Web Track 2009-2012 initially focused on CwB, and CwC was constructed artificially. What is promising is that using relevance judgments from CwB to evaluate systems on the much larger CwA works fine when using our measures. It is confirmed by fact that the observed values of Kendall's τ decrease only slightly if at all. It is also worth mentioning that we performed analogous experiments for the years 2009–2011 with similar results which are omitted here for the space limitation.

5.4 Correlation

We now examine the correlation between our proposed measures and the established cascade measures. While our measures aim at addressing the cases when only highly incomplete judgments are available, one may desire to know the relationship between them and the established ones. To this end, we compute Kendall's τ between the system rankings determined by our measures and the

Table 2: Impact of the document collections used for collecting relevance judgments. The first row indicates the document collection on which relevance judgments were collected; the second row indicates the document collection on which query results were determined. Kendall's τ correlations above 0.8 are shown in bold.

Test collections / Pair of measures		CwA			CwB			CwC		
		CwA	CwB	CwC	CwA	CwB	CwC	CwA	CwB	CwC
α−nDCG	AbsNb	**.87**	**.89**	.71	.79	**.90**	.73	.30	.23	.71
	AbsRb	**.89**	**.89**	.71	**.84**	**.89**	.73	.43	.40	.71
	DeltaNb	.71	.79	.67	.61	**.81**	.67	.23	.31	.69
	DeltaRb	**.82**	**.88**	.69	.79	**.88**	.67	.41	.51	.70
ERR-IA	AbsNb	**.85**	**.87**	.65	**.81**	**.88**	.69	.29	.26	.67
	AbsRb	**.86**	**.88**	.69	**.85**	**.89**	.71	.43	.44	.68
	DeltaNb	.69	.75	.63	.60	.78	.60	.20	.31	.65
	DeltaRb	**.80**	**.86**	.65	**.80**	**.87**	.64	.41	.54	.66
NRBP	AbsNb	**.84**	**.83**	.60	**.82**	**.83**	.64	.26	.26	.63
	AbsRb	**.85**	**.84**	.64	**.86**	**.85**	.66	.40	.45	.63
	DeltaNb	.68	.70	.59	.61	.72	.55	.19	.31	.60
	DeltaRb	**.80**	**.82**	.60	**.82**	**.84**	.59	.39	.54	.61

ones determined by the established cascade measures with complete judgments. For comparison, Table 3 lists pairwise correlations between α-nDCG, ERR-IA, and NRBP in terms of their average Kendall's τ on TREC Web Track 2009–2012. As we can see, the established cascade measures are highly correlated.

Table 3: Average Kendall's τ between α-nDCG, ERR-IA, NRBP on TREC Web Track 2009–2012.

	α-nDCG	ERR-IA	NRBP
α-nDCG			
ERR-IA	.93		
NRBP	.88	.87	

Table 4 reports Kendall's τ between our four measures AbsNb, AbsRb, DeltaNb and DeltaRb, and the cascade measures α-nDCG, ERR-IA, and NRBP. For a different perspective, Figure 2 plots the system ranks assigned by our four methods against those assigned by ERR-IA on TREC Web Track 2009–2012. For a measure having perfect correlation with ERR-IA, the points in this plot would lie on the main diagonal $y = x$. Due to the space limitation, we only show plots against ERR-IA. From Table 4, we can see that the correlation between our measures and α-nDCG, ERR-IA, and NRBP varies across different query sets. The correlation is lowest for queries from the year 2010 and highest for queries from the year 2012. Comparing our methods, we observe a positive effect of the position bias with AbsRb and DeltaRb, consistently showing higher correlation than their non-biased counterparts. While our measures do not consistently achieve a Kendall's τ correlation above 0.8, we argue that the proposed measures are still useful since they can better deal with incomplete judgments and more effectively reuse relevance judgments, as we have discussed in Section 5.2 and 5.3. From Figure 2, it can be seen that all points distribute along the $y = x$, indicating that the system rankings determined by the proposed measures are close to the ones from ERR-IA.

Table 4: Correlations between our measures and the established cascade measures. Kendall's τ correlations above 0.8 are shown in bold.

		AbsNb	AbsRb	DeltaNb	DeltaRb
2009	α-nDCG	.78	.70	.78	.73
	ERR-IA	**.81**	.75	.79	.78
	NRBP	**.80**	.79	.73	.78
2010	α-nDCG	.70	.73	**.81**	.76
	ERR-IA	.68	.73	.77	.76
	NRBP	.65	.71	.75	.72
2011	α-nDCG	.74	**.81**	.76	**.85**
	ERR-IA	.74	**.81**	.75	**.85**
	NRBP	.71	.78	.70	**.81**
2012	α-nDCG	**.87**	**.89**	.71	**.82**
	ERR-IA	**.85**	**.86**	.69	**.80**
	NRBP	**.84**	**.85**	.68	**.80**

6 CONCLUSION

Our work investigates the performance of the established cascade measures (i.e., α-nDCG, ERR-IA, and NRBP) when less than 50% judgments are available. We found out that their ability to rank systems deteriorates quickly as we remove more and more relevance judgments. To mitigate this, we proposed novel cascade measures

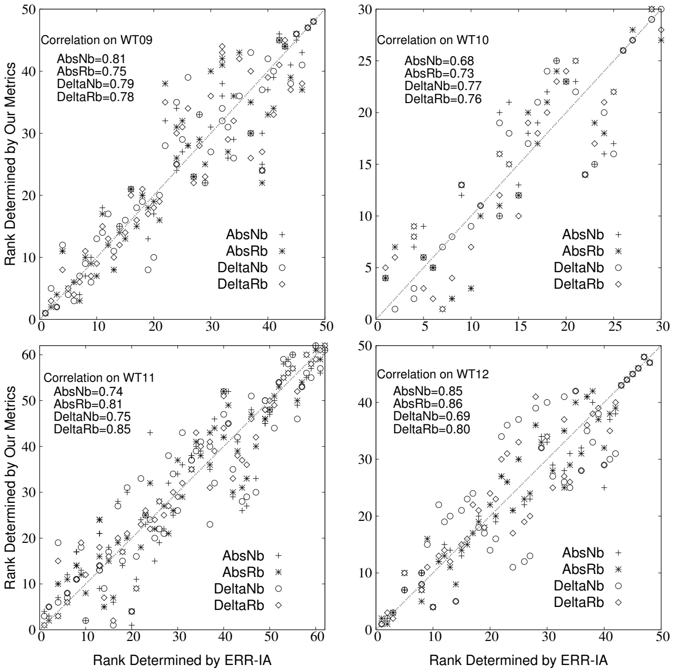

Figure 2: System rank assigned by ERR-IA vs. system rank assigned by our measures on TREC Web Track 2009–2012.

that are based on the Kullback-Leibler divergence between language models estimated for queries' subtopics and returned results. Our experiments showed that our novel measures correlate well with the established ones and, more importantly, are robust in the presence of highly incomplete judgments. Even with as little as 15% of relevance judgments, our cascade measures still get close to the established ones on complete judgments. Moreover, our measures can assess the retrieval performance of a system on unlabelled collections leveraging the relevance judgments gathered for a completely disjoint document collection.

REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5–14, New York, NY, USA, 2009. ACM.

[2] E. Amitay, D. Carmel, R. Lempel, and A. Soffer. Scaling IR-system Evaluation Using Term Relevance Sets. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 10–17, New York, NY, USA, 2004. ACM.

[3] J. A. Aslam, V. Pavlu, and E. Yilmaz. A statistical method for system evaluation using incomplete judgments. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 541–548, New York, NY, USA, 2006. ACM.

[4] T. Bompada, C.-C. Chang, J. Chen, R. Kumar, and R. Shenoy. On the robustness of relevance measures with incomplete judgments. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 359–366, New York, NY, USA, 2007. ACM.

[5] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 25–32, New York, NY, USA, 2004. ACM.

[6] S. Büttcher, C. L. A. Clarke, P. C. K. Yeung, and I. Soboroff. Reliable information retrieval evaluation with incomplete and biased judgements. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 63–70, New York, NY, USA, 2007. ACM.

[7] B. Carterette and J. Allan. Semiautomatic evaluation of retrieval systems using document similarities. In *Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 873–876, New York, NY, USA, 2007. ACM.

[8] B. Carterette, J. Allan, and R. Sitaraman. Minimal test collections for retrieval evaluation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 268–275, New York, NY, USA, 2006. ACM.

[9] O. Chapelle, S. Ji, C. Liao, E. Velipasaoglu, L. Lai, and S.-L. Wu. Intent-based diversification of web search results: metrics and algorithms. *Information Retrieval*, 14(6):572–592, 2011.

[10] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 621–630, New York, NY, USA, 2009. ACM.

[11] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. 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, pages 659–666, New York, NY, USA, 2008. ACM.

[12] C. L. Clarke, M. Kolla, and O. Vechtomova. 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, pages 188–199, Berlin, Heidelberg, 2009. Springer-Verlag.

[13] V. Dang and W. B. Croft. Diversity by proportionality: an election-based approach to search result diversification. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 65–74. ACM, 2012.

[14] V. Dang and W. B. Croft. Term level search result diversification. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 603–612. ACM, 2013.

[15] S. Hu, Z. Dou, X. Wang, T. Sakai, and J.-R. Wen. Search result diversification based on hierarchical intents. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 63–72. ACM, 2015.

[16] K. Hui and K. Berberich. Selective labeling and incomplete label mitigation for low-cost evaluation. In *International Symposium on String Processing and Information Retrieval*, pages 137–148. Springer International Publishing, 2015.

[17] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20:422–446, October 2002.

[18] G. K. Jayasinghe, W. Webber, M. Sanderson, and J. S. Culpepper. Improving test collection pools with machine learning. In *Proceedings of the 2014 Australasian Document Computing Symposium*, ADCS '14, pages 2:2–2:9, New York, NY, USA, 2014. ACM.

[19] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, 27(1):2:1–2:27, Dec. 2008.

[20] T. Sakai. Alternatives to bpref. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 71–78. ACM, 2007.

[21] T. Sakai. The unreusability of diversified search test collections. In *EVIA@ NTCIR*, 2013.

[22] T. Sakai, Z. Dou, R. Song, and N. Kando. The reusability of a diversified search test collection. In *Asia Information Retrieval Symposium*, pages 26–38. Springer, 2012.

[23] T. Sakai and N. Kando. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Inf. Retr.*, 11(5):447–470, 2008.

[24] E. M. Voorhees. Evaluation by highly relevant documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 74–82, New York, NY, USA, 2001. ACM.

[25] X. Wang, Z. Dou, T. Sakai, and J.-R. Wen. Evaluating search result diversity using intent hierarchies. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 415–424. ACM, 2016.

[26] E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, pages 102–111, New York, NY, USA, 2006. ACM.

[27] C. Zhai. Statistical language models for information retrieval a critical review. *Found. Trends Inf. Retr.*, 2:137–213, March 2008.

[28] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 10–17, New York, NY, USA, 2003. ACM.