Towards Robust & Reusable Evaluation for Novelty & Diversity

Kai Hui
Max-Planck-Institut für Informatik
66123 Saarbrücken, Germany
khui@mpi-inf.mpg.de

ABSTRACT

Existing IR measures for offline evaluation directly bring in the labels into computation, where the labels are on the entire documents. This direct dependency makes the measure highly reliant on the completeness of the labels, consequently the measure values are sensitive towards missing labels, resulting in poor robustness and reusability. To mitigate this, we propose a novel evaluation approach, constructing an intermediate layer between the labels and the measure, improving the robustness and reusability by dampening the direct dependency, as well as considering the content of the document in the measure computation. In particular, we propose to estimate a language model based on a selected relevant document set to construct a ground truth, afterward using the divergence between the search result and this ground truth to compute measures. To further save labeling efforts and to improve efficiency, we select representative documents, query set and topic terms involved in the evaluation separately before computing the measure. Preliminary experiments on the diversity tasks of TREC Web Track 2009–2012, using ClueWeb09-A as a document collection, show that with as little as 30% of judgments our novel approach almost accurately reconstructs the original system rankings determined by α-nDCG, ERR-IA, and NRBP.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]

Keywords

Evaluation; Incomplete Judgments; Novelty; Diversity

1. INTRODUCTION

Evaluation in information retrieval is used to distinct the performance among different rivaling retrieval systems, evaluating how well the systems can satisfy users' information need given queries. Specifically, rivaling systems separately retrieve documents, subsequently being compared in terms of measure values. In this procedure, measures are computed based on click information, if applicable; otherwise, in existing offline evaluation, assessors need to label the retrieved documents, thereafter computing measures based on these labels. Nevertheless, the expensiveness of manual judgments makes the whole evaluation procedure costly and limits our ability in evaluation, consequently reliable measures requiring less human efforts in judgments are desirable.

In the pooling method employed in TREC, the cost of manual labels can be boiled down to two components, i.e., the returned document pool from all rivaling systems and the query set being used. Intuitively, to save the cost of the evaluation is to reduce the number of required labels, equivalently either to shrink the document pool to be labeled for unique query, or to reduce the size of test query set. Correspondingly, existing works for saving evaluation efforts were mainly on these two directions.

In current paradigm of offline evaluation, however, the computation of measures is merely based on the manual labels, not considering the document content. In another word, the document content is only used in assigning labels by assessors, after which the labels are introduced into measure computation [12]. Since these labels solely associate with entire document, the measures will lose reliability when evaluating systems containing unlabeled documents, even though the content of these unlabeled documents might be similar to labeled ones. Actually, there exist another option that we can reuse the historical labels in evaluation to save the labeling efforts. Nevertheless, due to the existence of the unlabeled documents, current measures for novelty and diversity are not reusable [26].

Consequently, the indirect consideration of document content in the measure computation, with assessors' judgment as intermediary, make the evaluation hardly extend to unlabeled document collection. Moreover, systems without being included in the judgment might be under bias evaluation, as results of falsely regarding the unlabeled relevant document as irrelevance. In short, the limit robustness and reusability of established measures is mainly due to the direct introduction of labels to the measure computation. Therefore, to improve the robustness and reusability, we need to revise the paradigm of the measures, introducing the document content into the measure computation, thereby breaking the direct dependency of the measure computation on the labels, consequently making the measures extendable and reusable in evaluating unlabeled data. To this end, we come up with a concrete preliminary study and a research schema for developing measures with better ro-

PIKM'14, November 3, 2014, Shanghai, China.

ACM 978-1-4503-1481-7/14/11 ...$15.00.
http://dx.doi.org/10.1145/2663714.2668045.

bustness and reusability, prompting novel measures which brings document content directly into measure computation, meanwhile we investigate the effects of representative document selection, query set selection and topical term selection for our novel evaluation approach.

There exists plenty of works on measures' robustness and reusability for adhoc measurement, where main ideas were either to recognize the crucial documents in evaluating different systems [2, 9, 27, 30] or to reduce the number of queries used [17, 19, 21, 25, 29]. The former type of works mainly employed the overlap of documents from search results of different rivaling systems to identify the discerning documents for manual labeling, like [9], or to assign pseudo relevance judgment as in [19]. Different from these works, we propose to bring in the document content in the measure computation, overcoming the heavy dependency of existing measures on the document labels. In query set selection works, the key problem is to find out a small set of queries on which the evaluation is reliable, given the facts that on different query sets the evaluation result may vary a lot [29]. It is obvious that reducing the size of test query set can significantly save the evaluation efforts, however most of existing works on this direction are retrospective, assuming available relevance judgment in the query selection. We propose to develop a practical query selection method, applying to our content-involved measures. Additionally, according to our knowledge, there exist no study regarding robustness and reusability for novelty and diversity measures, and our novel evaluation approach will mitigate this insufficiency.

In this work, we firstly investigate the manual efforts involved in evaluation with established novelty and diversity measures, i.e., α-nDCG [13], ERR-IA [11] and NRBP [14], by counting the number of unique labels in the TREC web track evaluation. What's more, we test the robustness of these measures in terms of their reliability under less labels. Additionally, following the study in [7], we further explore how the measures perform in evaluating unlabeled systems. Our data study shows that a significant effort is required for manual judgments in applying established measures and that the measures will lose reliability on incomplete judged document collection or for unlabeled systems. To mitigate this, we propose to construct novel measures for novelty and diversity, by introducing document content in the measure computation, dampening the dependency of measures on the labels. Moreover, we plan to further plug in the query selection, referring to [19], into our novel measures to reduce the label numbers. In particular, we introduce an intermediate layer between the measures and the manual labels, employing the known relevant documents as ground truth and compare the search results being evaluated with this ground truth. By understanding the reliability of our novel measures on different query sets, we plan to further propose a query selection method to shrink the query set required in reliable evaluation. Our preliminary experiments with the novel measures convinced that with as less as 30% labels available, the system rankings determined by our measures can approximate the original rankings with more than 0.8 Kendall τ correlation.

The rest of this paper is organized as follows. Section 2 gives some backgrounds. Section 3 displays the data study results. In Section 4, we investigate the related works, before describing our novel evaluation approach in Section 5. Finally, we summarize our work in Section 6.

2. BACKGROUND

In this section, we briefly describe the existing evaluation measures for novelty, namely α-nDCG, ERR-IA, and NRBP, which are widely used in recent years. Actually, all these measures are designed to evaluate both the diversity and novelty, which are highly mutually related, but still bear differences. Diversity mainly discusses how well the retrieval results cover the different potential user needs behind the given query, considering the relationship between ranking and the query. Meanwhile, the novelty is about how much novelty the user obtain when going through each document given what he already browsed in the ranking, mainly regarding the relationship in between the documents of ranking. Actually, they are the two sides of the same coin, focusing on different aspects, thus improvement on one also benefits the other. As indicated in Section 1, the evaluation on novelty is limited on the topics or facet levels, regarding the repetition of the subtopic or facet as the redundancy, meanwhile rewarding the documents with unseen topics. In this way, the measures expose to the risks that falsely penalizing the document with repeating subtopics but with novel content, and that over rewarding the document with novel subtopic but with redundant content, considering the subtopics are actually highly related with each other. The three representative measures for novelty evaluation are reviewed below. We can find out that the "novelty component" in these three measures are the counter of occurrence of the subtopics, plugging in a smaller factor to the evaluation score when with a large occurrence.

α-nDCG. Clarke et al. extended the traditional nDCG [20] to α-nDCG measure in evaluating diversity and novelty in search results [13]. α-nDCG scores a result set by rewarding results relevant to new subtopics and penalizing the ones relevant to redundant subtopics. Specifically, differently from nDCG, where the gain reflects the graded relevance value of the document to the query, α-nDCG uses a *novelty-biased gain*, which is defined as:

$$NG[r] = \sum_{i=1}^{m} J_i(r)(1-\alpha)^{C_i(r-1)} \tag{1}$$

The $J_i(r)$ is a flag which indicates if the document at rank r is relevant or not to the intent i. The $C_i(r-1)$ is the main "novelty component", it is the number of times the intent i covered by documents appearing before rank r, thereafter the factor multiplying a smaller factors to the evaluation score with larger C_i.

ERR-IA. The intent-aware version of the *Expected Reciprocal Rank* (ERR) [11]. It is defined as the weighted average of ERR computed separately for each query subtopic [10]. The ERR is based on "diminishing returns" for redundant documents , thus the contribution of each document is based on the relevance of documents ranked above it. The discount function is not just dependent on the rank but also on relevance of previously ranked documents:

$$ERR = \sum_{i=1}^{\infty} \frac{1}{i} \prod_{j=1}^{i-1} (1-R_j) \tag{2}$$

Where R_i is a function of the relevance grade of the document appearing at position i in the ranking, and it is commonly defined as $(2^g-1)/2^{g_{max}}$. In its intent aware version, i.e., ERR-IA, the novelty is rewarded by penalizing the repetition of the same subtopics with smaller R_i.

NRBP. The *Novelty- and Rank-Biased Precision* was proposed by Clarke et al. [14] to combine α-nDCG and RBP (*Rank Biased Precision*) [22]. It is computed as follows:

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

As we can see, NRBP uses two discount mechanisms: one is for the redundancy of documents and is based on the parameter α, the other one is based on the *persistence parameter* β , which is the probability that the user will go down in the ranked list of results. Similar to α-nDCG, the C_i is used as a "novelty component" here.

3. DATA STUDY

3.1 Dataset

We use ClueWeb09 (Cw) [1] as a document collection, which consists of one billion web pages (5 TB compressed, 25 TB uncompressed) in ten languages. In our experiments, we focus on the subset of more than 500 million English web pages, which are known as ClueWeb09 Category A (CwA). Queries & relevance judgments are taken 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, a.k.a., qrel. For our methods we convert graded labels into binary ones by treating labels minus two and zero as irrelevant and all other labels as relevant. Moreover, we 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. As standard in TREC, we consider top-20 query results, when comparing different systems.

3.2 Label Numbers and Pooling Depth

In this section, we analyze the pooling method employed in TREC evaluation, investigating the relationship between pooling depth and the number of documents to be labeled. What's more, we investigate the portion of the relevant documents with respect to the total labeled documents. In the pooling method, the top-k documents returned by different systems are collected, generating a pool of candidate documents for assessors to label. Intuitively, the pooling depth k determines the number of documents to be labeled, as well determines the depth of evaluation, thereby evaluation based on pooling method hardly evaluate deeper, especially for a large number of rivaling systems. However, since the rivaling systems process the same query on the same dataset in the evaluation, there exist overlap of the retrieved documents, making the estimation of the document number nontrivial. In addition, it is reasonable for us to assume pools with smaller k have larger portion of relevant documents, due to top ranked documents are more likely relevant. In particular, we gather in total 200 queries (50 queries per year) and 33251 relevant documents with at least one relevant subtopic or facet, assuming the deepest pooling is 20, thereafter converting all label numbers at other position as fraction to 20 in favor of our comparison. Figure 1 and Figure 2 summarize the results, where each curve represents the average percentage corresponding to one year's query set.

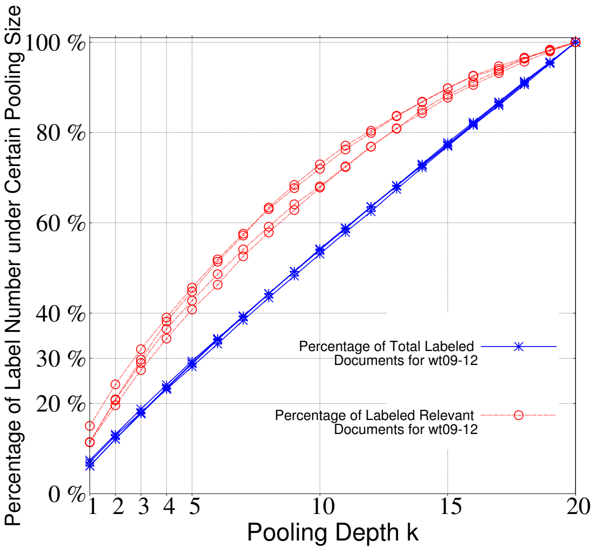

Figure 1: Pooling depth versus the percentage of total label number (blue curves), and relevant label number (red curves), all percentage are computed with respect to pooling depth 20. The x-axis is the pooling depth, and y-axis is the percentage of label number.

In Figure 1, the four diagonal curves (blue curves) represent the total number of unique document to be labeled, indicating an approximately linear increments of label number w.r.t. the pooling depth. Considering the expensiveness of the manual judgments, this trend consequently limits our ability in evaluating deeper results. Slightly different from our intuition, the overlap among returned document sets from different systems does not influence a lot, which may due to a large variety of the returned documents. The logarithm-shaped curves (red curves) along with the Figure 2 are about the relevant label numbers. In accordance with our assumption, there exist respectively more relevant labels in the pool with top ranked documents, given that the red curves are above the blue curves. Specifically, with depth 4 or 20% out of total labels, we can get more than 30% relevant labels.

3.3 Robustness with Less Labels

In this section, we further investigate the robustness of the existing measures with incomplete judgment. To this end, we inspect the correlation between system rankings determined on incomplete judgments and the ones determined on complete judgments. We follow the procedure employed in [5] and [6]. Given a query, we randomly shuffle the relevant documents in qrel, and pick up first $max(1, \lceil p\%|qrel|\rceil)$ relevant documents from qrel as known relevant label to construct qrel at $p\%$, meanwhile, under similar ideas, we pick up first $max(10, \lceil p\%|qrel|\rceil)$ documents from irrelevant documents in qrel. Note that, the qrel from 2010 did not contain irrelevant judgment for diversity task, thereby we only sample irrelevant documents for 2009, 2011 and 2012,

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

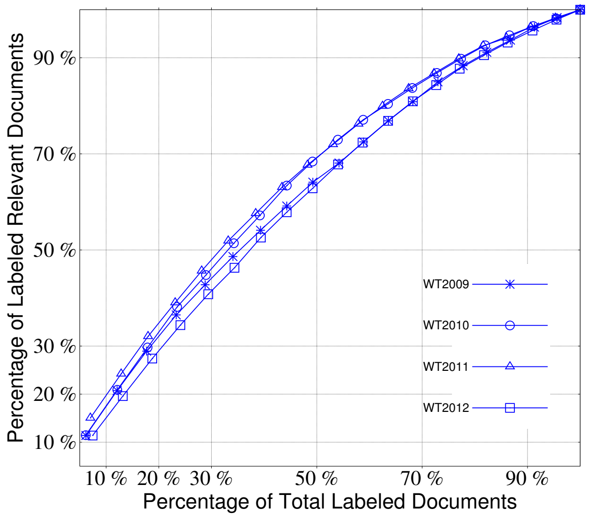

Figure 2: Percentage of relevant label versus the total label, with respect to pooling depth 20. The x-axis is the percentage of total label, and y-axis is the percentage of relevant label.

whereas regarding all unlabeled documents as irrelevance for 2010. On each sample rate p, we repeat 30 times to generate different incomplete qrel. In Figure 3, we summarize the robustness result for three existing measures ERR-IA, α-nDCG and NRBP, displaying the minimum and maximum correlation at each p with the dotted black curve, as well the average value with 95% confidence interval with the solid blue curve.

From the figures, we can observe significant decrease of the correlation when reducing the number of available labels. We claim that all these established measures are losing reliability with respect to less and less p value. When with 40% available labels, the evaluation from these measures becomes unreliable, where the correlation value is lower than 0.8. Moreover, for the minimum and maximum dotted curves, we find that the difference between them becomes larger with less p and that the maximum values are over 0.8 correlation even when $p = 20$, which may imply a potential to achieve better evaluation with very few labels. Similarly, in Figure 4, we further summarize the correlation trends with respect to available labels along with the pooling depth. With available labels from top-k documents of rivaling systems, we find that with pooling depth equal 3, we can achieve more than 0.8 correlation, indicating that with less labels the diversity measures could be robust as long as we remove labels along with the pooling depth. However, note that due to the existence of the direct dependency discussed in Section 1, the pooling depth qrel still can not fix the reusability problem.

3.4 Bias towards Unlabeled Systems

In this section, we investigate the evaluation bias when the established measures evaluate unlabeled systems. Similar to [7], we study the bias phenomenon by randomly removing documents from $p\%$ systems out of all systems, denoted as $(1-p\%)s\text{-}qrel$, subsequently comparing the system ranking determined by this incomplete qrel and by the full judgment qrel in terms of Kendall τ correlation. Different from [7], we repeat the comparison on a series of percentage for missing systems, varying $p\%$ between 5% and 99%. On each percentage, we randomly sample the systems for 30 times and compute the maximum, minimum and the 95% confidence interval near the average.

The results are summarized in Figure 5. The upper and lower dotted line indicate the maximum and minimum correlation among the 30 samples, meanwhile the solid line with confidence interval at each point presents the arithmetic mean of the correlation. From the figure we can conclude that the established measures evaluate the unlabeled systems biased, especially when over 50% of systems are missing, where the evaluation results are different from the full judgments significantly. Additionally, when removing more and more systems, the difference between the maximum and minimum correlation diverges, indicating the effects of systems in constituting the qrel are different.

4. RELATED WORK

Manual judgments incur expensive cost, therefore evaluation with less or no manual judgment has raised attentions in recent years. There exist several directions in the existing works such as the reduction of the unique document number to be labeled, including the automatic evaluation without manual judgment, the selection of the query subsets and the inference of the missing labels. These works either took advantages of different retrieved documents from different systems, recognizing key documents or generating pseudo labels, or of the different ability of different queries in comparing the retrieval systems. All these works dedicated to simulate the full evaluation, with complete judgment and query set, with less manual judgments, and evaluated their methods by comparing against the system rankings determined under full evaluation. Additionally, note that most of these methods depend on a specific measure, i.e., the reliability is highly related to the measures they used, and are not flexible in switching among different measures. We review works according to the ways they save the label cost, from query or document aspects. In addition to the low cost evaluation, we briefly review the works for identifying representative documents and terms in the end, enlightening our evaluation approach.

Sampling crucial documents with better distinguishability in judging the retrieval systems. Yilmaz and Aslam [30] as well as Aslam et al. [2] presented approaches for random sampling to estimate the actual values of average precision when relevance judgments are incomplete. Similarly, Sakai and Kando [27] applied traditional evaluation measures to "condensed" lists, which are ranked lists of documents obtained by removing all unlabeled documents. Carterette et al. [9] analyzed the distribution of the average precision over all possible assignments of relevance to all unlabeled documents and proposed a method to construct a test collection with minimal relevance assessments. All of these works focused on traditional effectiveness measures (e.g., average precision), whereas our focus in on more recent cascade measures for novelty and diversity. Moreover, different from these works, we use the contents of documents labeled as relevant when determining our measures.

Reducing the number of queries used in evaluation. Another option to save the cost of evaluation is to

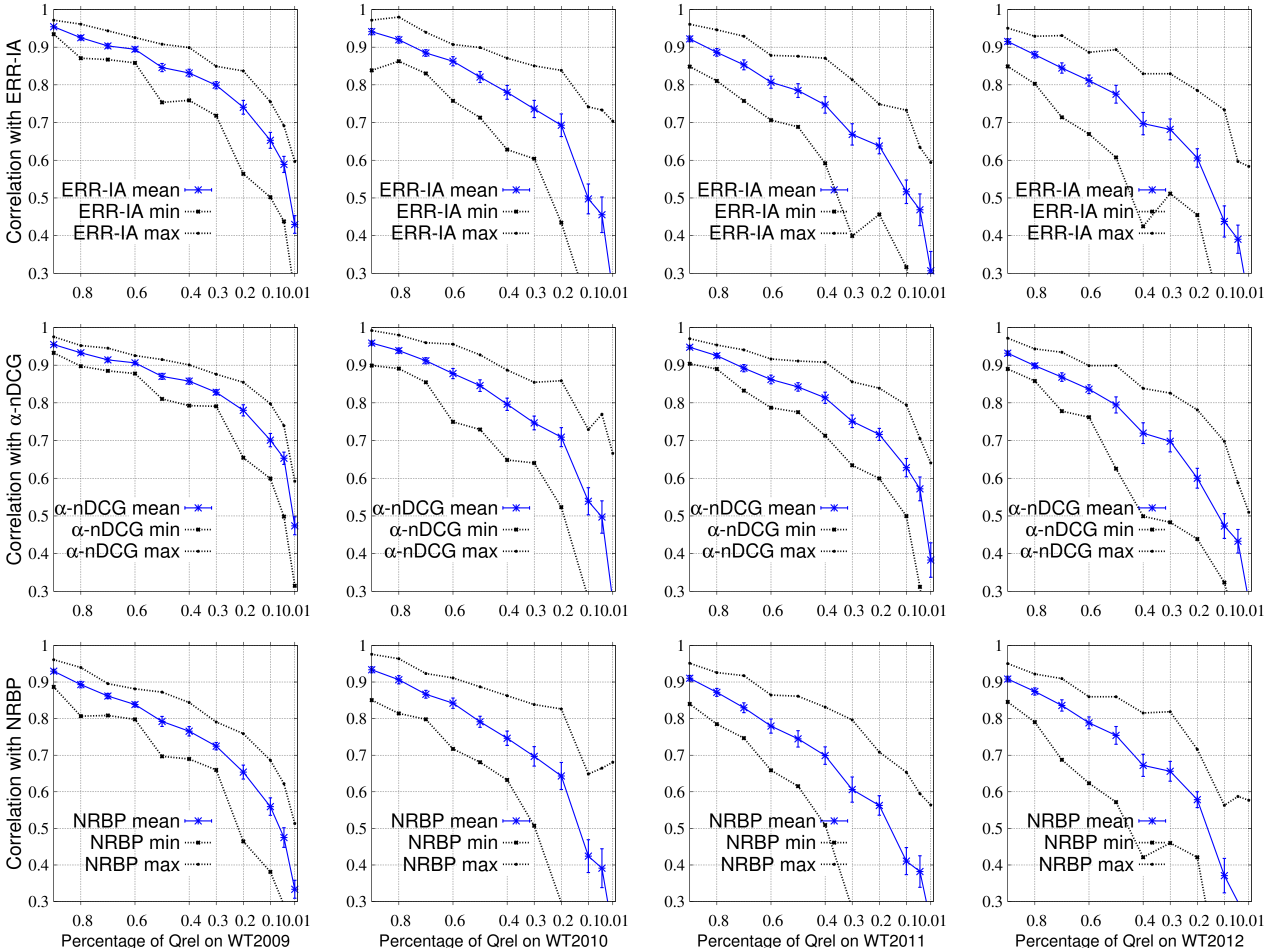

Figure 3: Robustness of the established measures, i.e., ERR-IA, α-nDCG and NRBP with randomly sampled partial qrel. The x-axis is the sampling rate, under each of them we evaluate the retrieval systems. The y-axis is the Kendall τ value. For each measure, we repeat the sampling 30 times, displaying mean value with 95% confident interval (solid blue curve), as well as the minimum and maximum value (dotted black curve).

reduce the size of query set being used. One early work by Voorhees et al. [29] demonstrated that different set of queries with same query number maybe highly varied in evaluation retrieval systems. In more recent work by Mizzaro & Robertson. [21], HITS analysis was used for the matrix between queries and retrieval systems, using hubness score as indicator of good topics in comparing systems. Subsequent work by Guiver et al. [17] analyzed how the evaluation with different subsets of queries could approximate the full query set and proposed a greedy algorithm to select query subset with best correlation. Robertson [25] compared methods in [21] and [17] with retrospective experiments and claimed that the generalsability of these methods were still not clear. Moreover, Hosseini et al. proposed a greedy algorithm, named Adaptive, in selecting the query subset in [19]. Different from prior works, it assumed no relevance judgment available in the query selection, thereafter are more practical. Our evaluation framework is more on the document and query parts, nevertheless, topic selection could be plugged into our framework given that the influence of the choices of query have been validated in all these aforementioned works.

Automatic evaluation with pseudo relevance judgments. The earliest work on this direction is by Soboroff et al. [28], randomly sampling pseudo relevant documents from a pool generated with documents returned by rivaling systems, like conducting majority vote by the frequency of occurrences of documents in different systems. Documents retrieved by more systems are more likely to be sampled, thereafter regarding as relevant, however, easily resulting in "tyranny of the masses" [3]. Nuray & Can [23] generated pseudo relevance documents on a p% system subset, finding that systems were different most from the average systems could get best performance. In this work, Nuray & Can actually took the rank position of documents into consideration apart from the occurrences of the documents, and achieved higher correlation than the method in [28]. Efron [16] further employed query aspects by generating the pseudo relevance with retrieval results from all these aspects. Diaz [15] proposed to evaluate the systems based on the intuition that better retrieval systems are more likely to fulfill the cluster hypothesis, i.e., similar documents should bear closing evaluating score for good retrieval systems. Diaz computed the correlation between the original ranking vector

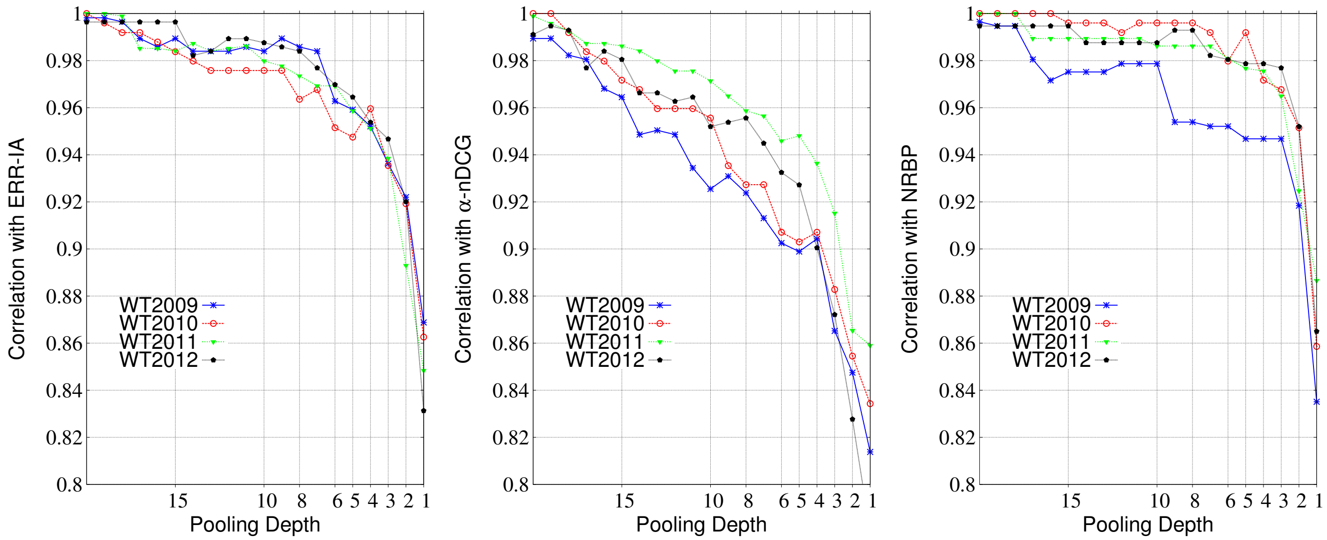

Figure 4: Robustness of the established measures, i.e., ERR-IA, α-nDCG and NRBP w.r.t. different pooling depth, under which we construct the partial qrel. The x-axis is the pooling depth. The y-axis is the Kendall τ value.

with document score as components and the ranking whose score are substituted with the average score of the corresponding similar documents, claiming that higher correlation could indicate better system performance. Moreover, Hauff et al. [18] further summarized several automatic evaluation methods and conducted experiments for comparing on sixteen TREC data sets. Similar to our setting, all these methods were evaluated by comparison against system rankings determined by some ground truth measures, e.g. MAP etc., however, our method mainly consider the documents within the same ranking and mainly concern the diversity measurement, whereas these methods concerned the interrelationship among documents from different rank lists to simulate adhoc measurement.

Inferring missing labels to make up the incomplete judgments. Carterette and Allan [8] as well as Büttcher et al. [7] tried to make up missing relevance judgments by predicting them using methods from machine learning. Moreover, Aslam & Yilmaz [4] proposed to infer the relevance judgments from rank lists given the associating average precision and the number of relevance documents. As a commonality to our work, both [8] and [7] made 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. Meanwhile [4] is different from us on the assumptions that the measure values, e.g., the AP, for the given rank lists are supposed available.

What's more: selecting representative documents and terms. In our evaluation framework, the selection of relevant documents in the construction of query language model might be a building block, i.e., either with all the available relevant documents or only small number of selected documents. Moreover, it remains questionable that do we need to use all the terms in the language model. For former question, works from the relevance feedback may give us hints. Raiber & Kurland [24] investigated different notions of representativeness for relevant documents in the similarity space, finding that the documents centrally located in the similarity space tend to have better representativeness. For the latter concern, an early work from Amitay et al. [1], used the relevant and irrelevant term sets to evaluate IR systems, enabling evaluation on dataset without labels by judging the documents with set of topic terms, possibly enlightening our work.

5. NOVEL EVALUATION APPROACH

5.1 Overview of the Approach

Our methods differ from established ones on the information used in the evaluation as described in Section 1. Apart from the labels, we take the document content into consideration, generating an intermediate layer between the labels and the measures, reducing the dependency between them, thereby improving the robustness and reusability. This intermediate layer is the key of our novel measures, combining the content of evaluated document and of the known relevant documents. Additionally, to further reduce the judgment efforts, we propose to select both documents to be labeled and the query set to be used. We also plan to promote the efficiency of the novel measures by picking up the representative terms for a given query, referring to the related works summarized in Section 4. In particular, in construction of the measures, we need to firstly select documents from search results to label, subsequently employing the labeled relevant documents to generate a ground truth language model. In evaluating the search results from a system, no matter whether it has been labeled, after constructing a language model on the search results, we subsequently compute the divergence between this constructed model and the ground truth. By converting the labeled documents into a language model, we avoid the direct dependency.

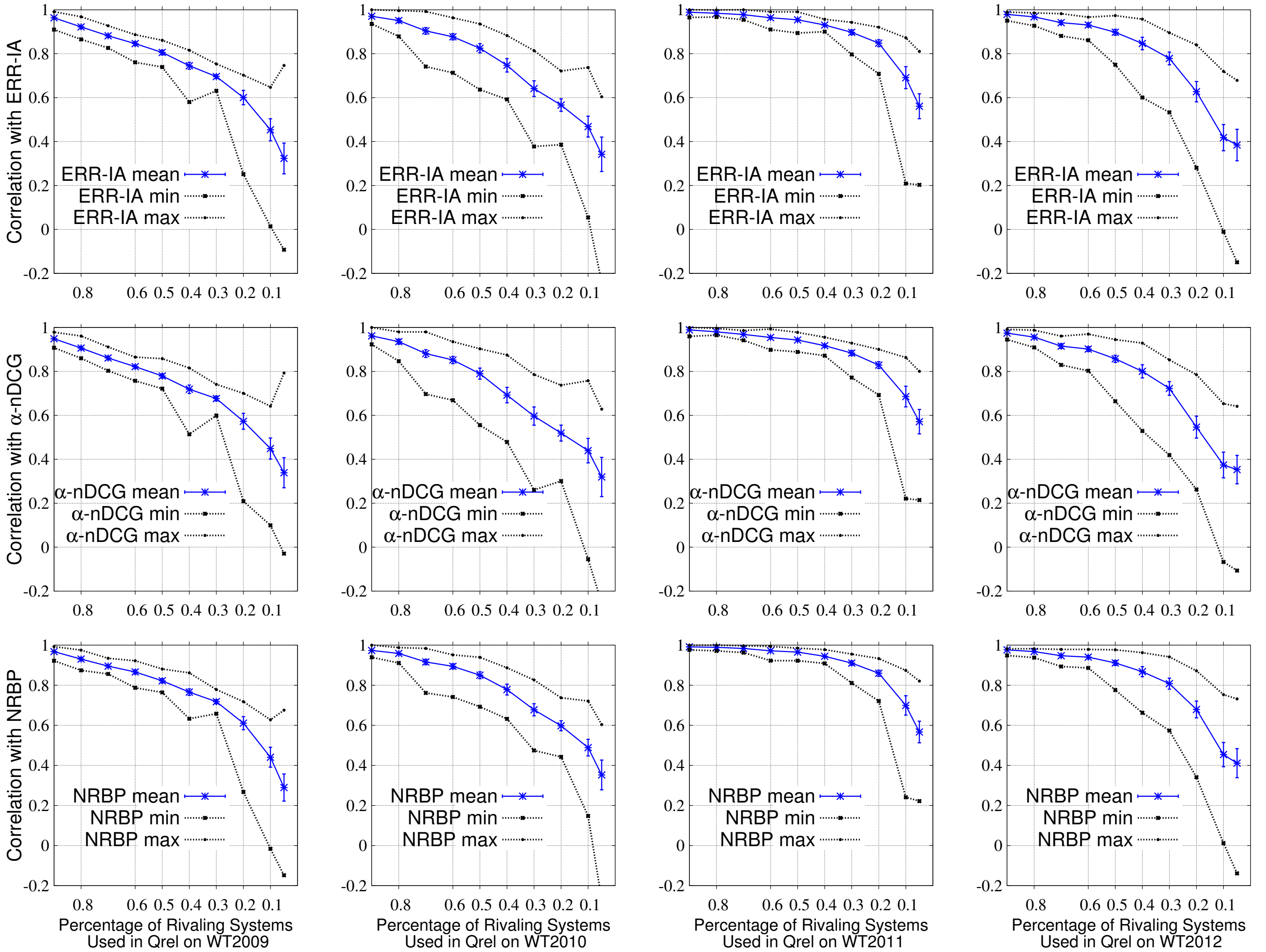

Figure 5: Evaluation bias towards unlabeled systems with the established measures, i.e., ERR-IA, α-nDCG and NRBP w.r.t. different percentage of labeled systems. The x-axis is the percentage of systems being judged. The y-axis is the Kendall τ value. On each percentage, we sample 30 times and display the maximum, minimum and the 95% confidence interval near the average.

5.2 Selecting Representative Documents

In Section 3, we convince that top ranked documents in rivaling systems are more likely to be relevant, thereby we collect top-k documents to generate the pool with a small k value. In our preliminary experiments, with $k = 3$ or $k = 4$ our measures can achieve acceptable reliability. Moreover, inspired by [24], we propose to take the divergence distribution among the documents into consideration to further filter the documents to be labeled. In [24], Raiber and Kurland proposed to employ the inter- and intra-document similarity to collect the representative documents, plugging them into the relevant feedback, finding that centrally located documents within the similarity space of the relevant documents tend to benefit the relevance feedback better. Our target is to pick a small set of representative documents to label meanwhile [24] targeted at relevance feedback, both are to select better documents in terms of indicator of other relevant documents. We can therefore compute the divergence distribution among the documents and select the centrally located documents. Note that, different from pooling method, we require relevant documents for every subtopic or facet beneath the given query, thereafter if relevant document is missing for one of the known subtopic or facet after judgment, we may need to perform the aforementioned selection process iteratively until every known subtopic or facet has at least one relevant document. According to our preliminary experiments, the document number to be labeled in this step is less than 20% comparing with the original pooling method with depth 20.

5.3 Constructing Measures

After the selection of documents, we evaluate the selected documents to label the relevance of document w.r.t. all known subtopics, subsequently generating ground truth language model for each subtopic respectively with the relevant documents. When evaluating search results $\mathcal{R}$ from a system, we compare the language model of documents in $\mathcal{R}$ and the ground truth model to compute the divergence in between, assigning search result with closer divergence higher measure score. Additionally, since we are dealing with novelty and diversity evaluation, for each subtopic or facet, we need to conduct the aforementioned comparison respectively. Alternatively, we can also predict the missing labels with similar ideas. Formally, our target can be summarized as a supervised classification or linear regression, where y is the label of each document, and x is a series of features of the document, including the above divergence feature. The value y could be concrete numbers, i.e., 0, 1, 2, $\cdots$, t, representing different subtopics, or be continuous when using linear regression. In both methods, we employ the divergence between evaluated documents and the ground truth model, thus converting the direct dependency on labels in established measures to an indirect dependency. The preliminary experiments convince the improvement of robustness and reliability of the proposed measures under incomplete judgment.

5.4 Selecting Topical Terms and Query Set

Apart from selecting the representative documents, the evaluation approach can be further optimized by limiting the vocabulary used in construction of the language model. This step may improve the efficiency of the algorithm considering the expensiveness of the language model computation. One early work on this direction by Amitay et al. [1] claimed that with a selected group of relevant terms in evaluating the retrieval results, instead using all vocabulary, the evaluation measures can retain reliable. In [1], Amitay et al. selected terms manually or with automatic methods. The kernel of our novel evaluation framework is based on the divergence between language models, which is also based on terms, thereby we plan to investigate whether we can prune the vocabulary and compute only with small portion of representative terms. Moreover, as indicated in [17] and [21], different query subsets are different in evaluating systems, meanwhile reducing the query numbers could significantly save the manual efforts. In our framework, we propose to select a set of queries before evaluation. In particular, we may firstly conduct retrospective experiments to select the topics on which our measures have higher correlation to the established measures, thereafter picking up the features which are able to distinct the different query set, furthermore, taking advantages of these features to select query set without labels available.

6. CONCLUSION

Our work investigates the reusability and robustness of the established cascade measures α-nDCG, ERR-IA, and NRBP. We find that their ability to rank systems reliably deteriorates quickly as we remove more and more relevance judgments, and the existence of the bias when evaluating systems containing unlabeled documents. To mitigate, we propose a novel evaluation framework by constructing an intermediate layer between the labels and measure to dampen the direct dependency in between. As future work, we plan to construct concrete measures following the framework proposed in this paper.

7. REFERENCES

[1] E. Amitay, D. Carmel, R. Lempel, and A. Soffer. Scaling ir-system evaluation using term relevance sets. In *SIGIR*, pages 10–17, 2004.

[2] 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.

[3] J. A. Aslam and R. Savell. On the effectiveness of evaluating retrieval systems in the absence of relevance judgments. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 361–362, New York, NY, USA, 2003. ACM.

[4] J. A. Aslam and E. Yilmaz. Inferring document relevance from incomplete information. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 633–642, New York, NY, USA, 2007. ACM.

[5] 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.

[6] 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.

[7] S. Büttcher, C. L. A. Clarke, P. C. K. Yeung, and I. Soboroff. Reliable information retrieval evaluation with incomplete and biased judgements. In *SIGIR*, pages 63–70, 2007.

[8] B. Carterette and J. Allan. Semiautomatic evaluation of retrieval systems using document similarities. In *CIKM*, pages 873–876, 2007.

[9] 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.

[10] 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.

[11] 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.

[12] C. L. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. 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, pages 75–84, New York, NY, USA, 2011. ACM.

[13] 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.

[14] 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.

[15] F. Diaz. Performance prediction using spatial autocorrelation. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 583–590, New York, NY, USA, 2007. ACM.

[16] M. Efron. Using multiple query aspects to build test collections without human relevance judgments. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ECIR '09, pages 276–287, Berlin, Heidelberg, 2009. Springer-Verlag.

[17] J. Guiver, S. Mizzaro, and S. Robertson. A few good topics: Experiments in topic set reduction for retrieval evaluation. *ACM Trans. Inf. Syst.*, 27(4):21:1–21:26, Nov. 2009.

[18] C. Hauff, D. Hiemstra, L. Azzopardi, and F. de Jong. A case for automatic system evaluation. In C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. Rüger, and K. van Rijsbergen, editors, *Advances in Information Retrieval*, volume 5993 of *Lecture Notes in Computer Science*, pages 153–165. Springer Berlin Heidelberg, 2010.

[19] M. Hosseini, I. J. Cox, N. Milic-Frayling, M. Shokouhi, and E. Yilmaz. An uncertainty-aware query selection model for evaluation of ir systems. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 901–910, New York, NY, USA, 2012. ACM.

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

[21] S. Mizzaro and S. Robertson. Hits hits trec: Exploring ir evaluation results with network analysis. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 479–486, New York, NY, USA, 2007. ACM.

[22] 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.

[23] R. Nuray and F. Can. Automatic ranking of information retrieval systems using data fusion. *Inf. Process. Manage.*, 42(3):595–614, 2006.

[24] F. Raiber and O. Kurland. On identifying representative relevant documents. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 99–108, New York, NY, USA, 2010. ACM.

[25] S. Robertson. On the contributions of topics to system evaluation. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 129–140, Berlin, Heidelberg, 2011. Springer-Verlag.

[26] T. Sakai. The unreusability of diversified search test collections. *Proceedings of EVIA 2013*, 2013.

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

[28] I. Soboroff, C. Nicholas, and P. Cahan. Ranking retrieval systems without relevance judgments. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 66–73, New York, NY, USA, 2001. ACM.

[29] E. M. Voorhees and C. Buckley. The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 316–323, New York, NY, USA, 2002. ACM.

[30] 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.