Evaluating Relevance Judgments with Pairwise Discriminative Power

Zhumin Chu[1], Jiaxin Mao[2], Fan Zhang[1], Yiqun Liu[1⋆], Tetsuya Sakai[3], Min Zhang[1], Shaoping Ma[1]
1 Department of Computer Science and Technology, Institute for Artificial Intelligence, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China
2 Beijing Key Laboratory of Big Data Management and Analysis Methods, Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
3 Department of Computer Science and Engineering, Waseda University, Tokyo, Japan
yiqunliu@tsinghua.edu.cn

ABSTRACT

Relevance judgments play an essential role in the evaluation of information retrieval systems. As many different relevance judgment settings have been proposed in recent years, an evaluation metric to compare relevance judgments in different annotation settings has become a necessity. Traditional metrics, such as κ, Krippendorff's α and Φ have mainly focused on the inter-assessor consistency to evaluate the quality of relevance judgments. They encounter "reliable but useless" problem when employed to compare different annotation settings (e.g. binary judgment v.s. 4-grade judgment). Meanwhile, other existing popular metrics such as discriminative power (DP) are not designed to compare relevance judgments across different annotation settings, they therefore suffer from limitations, such as the requirement of result ranking lists from different systems. Therefore, how to design an evaluation metric to compare relevance judgments under different grade settings needs further investigation. In this work, we propose a novel metric named pairwise discriminative power (PDP) to evaluate the quality of relevance judgment collections. By leveraging a small amount of document-level preference tests, PDP estimates the discriminative ability of relevance judgments on separating ranking lists with various qualities. With comprehensive experiments on both synthetic and real-world datasets, we show that PDP maintains a high degree of consistency with annotation quality in various grade settings. Compared with existing metrics (e.g., Krippendorff's α, Φ, DP, etc), it provides reliable evaluation results with affordable additional annotation efforts.

CCS CONCEPTS

• **Information systems** → **Relevance assessment**;

KEYWORDS

Relevance judgment, Preference test, Evaluation metric

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3482428

ACM Reference Format:
Zhumin Chu[1], Jiaxin Mao[2], Fan Zhang[1], Yiqun Liu[1⋆], Tetsuya Sakai[3], Min Zhang[1], Shaoping Ma[1]. 2021. Evaluating Relevance Judgments with Pairwise Discriminative Power. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3459637.3482428

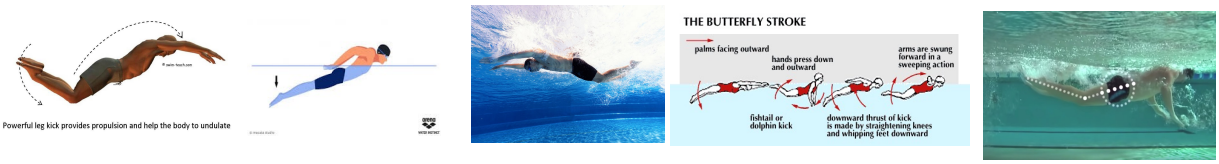
1 INTRODUCTION

Relevance judgments play a vital role in the evaluation of information retrieval systems and the optimization of machine learning based ranking models. Either in the practical application of large-scale commercial search engine or in the construction of various benchmarks (e.g., TREC [41], CLEF [19], NTCIR [31]), relevance judgments are still a necessary component. As different multi-grade settings have been proposed in recent years, such as 3-grade [11], 4-grade [25], 6-grade [17], 100-grade [27], or even magnitude estimation [38], an evaluation metric to compare relevance judgments in different annotation settings, especially in different grade settings, has become a necessity.

Traditional metrics, such as κ [16, 20, 39], Krippendorff's α [24] and Φ [10], adopt inter-assessor consistency to evaluate the quality of relevance judgments. However, these metrics are incompetent to compare relevance judgments collected in different settings because of the "reliable but useless" problem. To better illustrate this problem, we show an example of the relevance judgments collected in the image search scenario in Figure 1. This instance presents 7-grade relevance judgment results, as well as the 4-grade results. Although the annotation consistency calculated with Krippendorff's α (ordinal version) in the 7-grade (0.57) setting is lower than that of the 4-grade (0.68) setting, we can obtain more valuable information from the 7-grade relevance judgments. For example, the fourth image is more useful for users to understand the complete movement flow of butterfly stroke legs than others. The 7-grade judgment results can notice this difference, but not for the 4-grade results.

Another way to measure the quality of the relevance judgments and the relevance-based evaluation metrics is to test whether they can be used to reliably evaluate and compare different retrieval systems. Discriminative power (DP) [30], as a representative, can be applied to the evaluation of IR effective measures as well as different relevance judgment settings. However, evaluating relevance judgment collections with DP requires a number of ranking lists from different retrieval systems, which cannot be collected easily in some cases.

⋆Corresponding author



4-grade	3, 3, 3	2, 3, 3	2, 2, 2	3, 3, 3	3, 3, 3
7-grade	5, 6, 5	4, 5, 5	3, 5, 3	6, 6, 6	5, 5, 4

Figure 1: An example of relevance judgment with three assessors under different grade settings in the image search scenario (query content: butterfly stroke leg movement)

To better compare different relevance judgment settings, in this paper, we propose a novel evaluation metric for relevance judgments: pairwise discriminative power (PDP). Based on an affordable amount of additional user preference annotations [43], PDP estimates the discrimination ability of the relevance judgments on potential ranking lists to evaluate the quality of relevance judgments. Given the estimation of PDP, researchers can evaluate the annotation methods of relevance judgments and design a more reasonable annotation framework that can leverage both evaluation credibility and ranking algorithms' training effectiveness.

Inspired by the concept of informational entropy [35], we define PDP as the uncertainty of ranking list under topics (in Sec 3.3). To obtain PDP, we propose two methods to utilize the preference test results and then regard them as the gold standard to train for the permutation probability of the ranking list (in Sec 3.4). We summarize our main contributions as follows:

- We propose a novel metric to evaluate the discriminative ability of relevance judgment collections under different annotation settings.
- We present a unified framework for generating synthetic relevance judgment collections under multi-grade settings and then generate a series of synthetic datasets on different grade settings.
- We conduct extensive experiments on both synthetic and real-world datasets to verify the feasibility of our novel metric. The results show that our metric performs competitively compared to existing metrics.

2 RELATED WORK

Currently, Cranfield-like approach [15] is widely used to evaluate information retrieval systems. In this paradigm, researchers need to construct representative samples of topic and document collections, then conduct relevance assessments for the results returned by the retrieval system, and finally use evaluation metrics [9, 23, 26, 32, 34] to measure the goodness of the retrieved ranking list.

Many existing works have focused on the design of relevance assessments scale and setting [36]. Historically, binary scale (relevant or not) was used to conduct relevance assessments [40]. In recent years, a series of multi-grade relevance judgment settings have been proposed, including 3-grade scale used in TREC Terabyte Track [11], 4-grade scale in NTCIR WWW task [25], 6-grade scale in TREC Web Track [17], and so on. Tang et al. [37] compared the relevance scales ranging from 2 to 11 points, and found that the maximum confidence can be achieved in 7-grade scale. Turpin et al. [38] introduced the unbounded scale method, magnitude estimation, into document relevance judgments and observed a considerable consistency with traditional ordinal judgment settings. In [27], Roitero et al. tried a 100-grade relevance scale (S100) and showed the effectiveness and robustness of S100.

Another research direction attempts to validate the feasibility of preference judgment. Carterette et al. [6] first proposed to evaluate search engines using preference judgments rather than absolute judgments; they also designed the interface to conduct preference judgments. Chandar et al. [7, 8] introduced preference judgments to novelty and diversity search tasks, and proposed preference based metrics to evaluate it. Yang et al. [42] compared preference judgments with ordinal relevance judgments and found that preference judgments seem to be the same or even more reliable. Recently, Sakai et al. [34] proposed two novel types of preference-based measures, and also released a large-scale document preference judgments dataset. Clarke et al. [12–14] focused on evaluating top retrieved documents, and utilized preference judgments to estimate the maximum similarity between actual ranking and ideal ranking list.

Because of the importance of relevance judgments in the cranfield-like approach, it is critical to measure and control its quality. Traditionally, the quality of relevance judgments has been measured using the consistency of annotations between assessors, such as the widely used metrics Cohen's κ [16], Fleiss' κ [20], Weighted κ [20] and Krippendorff's α [24]. Based on the probabilistic parameter estimation, Checco et al. proposed Φ [10] to overcome the limitations existed in κ and α. Most of the consistency metrics are easy and convenient to calculate without any additional prior knowledge . However, they suffer from the phenomenon "reliable but useless", and perform poorly in the cross-scale relevance judgment comparison tasks, as we have shown in Figure 1. Sakai [30] proposed discriminative power (DP) measure to evaluate the sensitivity of evaluation metrics. DP can also be extended to compare the discriminative ability of different relevance judgment collections. However, DP is calculated based on a number of ranking lists from retrieval systems, which is difficult to obtain in some cases. Another way to compare relevance judgments collected in different relevance scales is to conduct scale transformation. Han et al [22] described several strategies to transform scale from fine-grained to coarse-grained, their experimental results illustrate that scale transformation strategies strongly affect the results of evaluation experiments. Also, Bailey et al. [3] categorized relevance judgments into three categories: "gold/silver/bronze/ standard" judges, and found a low agreement in relevance judgements between these three groups. In this work, we mainly deal with "silver standard" judges.

3 MODELS

3.1 Evaluation Metric for Relevance Judgments

We first establish a unified evaluation framework for relevance judgment collections. We assume $\mathcal{Q}$ is the total topic space. For each topic $q \in \mathcal{Q}$, the ranking system retrieves a series of corresponding documents, which are denoted as a set $\mathcal{D}_q$.

To construct a relevance judgment collection $\mathcal{R}$, we only sample a representative topic set from $\mathcal{Q}$, denoted as $\mathcal{Q}_\mathcal{R}$. Empirically, we often only select the top rank documents to conduct relevance judgments for each topic. We assume the set of top-K rank documents in topic q is the set $\mathcal{D}_{q;\leq K}$. In most scenarios, we hire the same batch of assessors to conduct relevance judgments within each topic. We can denote the assessor set under topic q as $\mathcal{U}_q$. In practice, the collection $\mathcal{R}$ contains a series of relevance judgment records. Just as Eq 1 shows, each record is identified by a tuple (topic q, document d, assessor u), denoted as $r_{d,q;u}$.

$$\mathcal{R} = \{r_{d,q;u} | q \in \mathcal{Q}_\mathcal{R}, d \in \mathcal{D}_{q;\leq K}, u \in \mathcal{U}_q\} \tag{1}$$

The evaluation metric for relevance judgments is then a mapping function from the collection $\mathcal{R}$ to a real number. Eq 2 shows the specific calculation formula:

$$\text{metric} = f(\mathcal{R}|O) = \frac{1}{|\mathcal{Q}_\mathcal{R}|} \sum_{q \in \mathcal{Q}_\mathcal{R}} f(\mathcal{R}_q|O_q). \tag{2}$$

In practice, the original metric is calculated at the topic-level, so we take the mean of metrics upon all the sampled topic set $\mathcal{Q}_\mathcal{R}$ as our collection-level evaluation metric. Sometimes, we need to introduce some empirical knowledge (O) to enhance the metric's performance. For example, in discriminative power (DP) [30] measure, the set O includes a series of collected ranking runs. In kappa [39] measure, O is just an empty set. In our PDP measure, the set O contains some preference judgment results.

3.2 The Architecture to Evaluate Relevance Judgments Based on Our Metric

When collecting a series of trial relevance judgment results under different settings, one might wonder which setting is better and deserves further annotation on a larger scale of data. In this case, our metric is a good choice to help evaluate relevance judgment results. Following shows the procedures to evaluate relevance judgment collections based on our metric:

(1) Determine the scale of preference tests, sample the preference test data, and then conduct preference tests.
(2) Estimate the document-level preference matrix under each topic based on the preference test results.
(3) Obtain the value of our metric on the basis of the preference matrix and relevance judgment results.

3.3 Pairwise Discriminative Power

Given the above framework, now we define our pairwise discriminative power (PDP) metric. With the relevance judgment results, we naturally obtain the ideal ranking results [12–14], that is, just ranking the documents based on the decreasing order of the documents' true relevance. The certainty of the ideal ranking results reflects the discriminative ability of relevance judgments. Inspired by informational entropy [35], we define PDP as the uncertainty of the ideal ranking list[1] , just as Eq 3 shows.

$$\text{PDP}(q) = \sum_{\pi \in \Pi_q} -p(\pi|q) \log p(\pi|q), \tag{3}$$

where π is a potential ranking list in topic q, Π_q is the set of all possible π. The probability $p(\pi|q)$ represents the permutation probability in which the ranking list under q is just the same as π. It is worth noting that Amigó et al. [2] also proposed an entropy-based topic-level metric, and perceived each preference relationship between documents as an uncertain unit and evaluated the observational uncertainty of topic based on the collected ranking lists. In contrast, PDP regards each potential ranking list as an uncertain unit, and then adopts preference information to estimate it.

Eq 4 shows our method to calculate the permutation probability $p(\pi|q)$, which is inspired by Plackett-Luce model [21]:

$$p(\pi|q) = \prod_{i=1}^{K} p(\pi_i = \pi(i)|\pi_{<i}, q) = \prod_{i=1}^{K} \frac{\exp(s^*_{\pi(i)|q})}{\sum_{j=i}^{K} \exp(s^*_{\pi(j)|q})}. \tag{4}$$

In Eq 4, the score vector s^* indicates the relevance information of documents in topic q, and it is also the parameter we need to train. Some previous works [6, 34] have demonstrated that preference judgments have advantages over absolute relevance judgments. Thus, in our setting, we assume the preference judgment results as the gold standard. And our training goal is to make s^* match the preference judgment results as much as possible. Here, we use the cross entropy [18] to evaluate the similarity between the preference judgment result and the score vector s^*. Eq 5 shows the optimization function to obtain the vector s^*, where $p(\pi(i) \succ \pi(j)|q)$ is the document-level preference probability, which will be discussed in Sec 3.4 . Specifically, we adopt stochastic gradient descent (SGD) [4] algorithm to optimize this loss function.

$$s^*_{\cdot|q} = \min_{s_{\cdot|q}} - \sum_{i \neq j} p\left(\pi(i) \succ \pi(j)|q\right) \log \frac{e^{s_{\pi(i)|q}}}{e^{s_{\pi(i)|q}} + e^{s_{\pi(j)|q}}} \tag{5}$$

In practice, to calculate PDP defined by Eq 3, the potential ranking list set might be large-scale, so we adopt the Monte Carlo simulation [29] strategy. In each simulation, we generate the sampling ranking list based on the permutation probability shown in Eq 4. Eq 6 shows the practical calculation formula of PDP, where the parameter T is the sampling scale:

$$\text{PDP}(q) = -\frac{1}{T} \sum_{t=1}^{T} \log \prod_{i=1}^{K} \frac{\exp(s^*_{\pi_t(i)|q})}{\sum_{j=i}^{K} \exp(s^*_{\pi_t(j)|q})}. \tag{6}$$

Note that we can generate the samples step by step. In each step, we use the probability $p(\pi_i = \pi(i)|\pi_{<i}, q)$ to determine the document selected at the i-th position. With this trick, the total simulation computational complexity is $O(TK^2)$ rather than $O(K! + TK \log K)$.

3.4 Estimate for Preference Matrix

Given the above introduction of PDP, there still exists an unclear question in the process of its calculation, that is, how to estimate the document-level preference probability $p(\pi(i) \succ \pi(j)|q)$?

[1] Analogous to informational entropy, the value of PDP ranges from 0 to $\log K!$, where K is the scale of ranking list π.

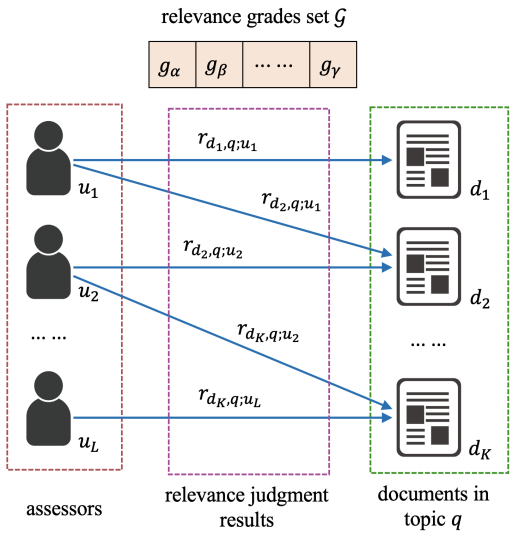


Figure 2: Description of notations used in Sec 3.4

Before introducing the specific methods, we first present the notation used in the subsequent introduction, just as Figure 2 shows. The symbols u_i, d_j, g_α represent an assessor, a document, and a relevance grade, respectively. The relevance grades set $\mathcal{G}$ contains a series of valid relevance grades. For example, in 4-grade setting, $\mathcal{G} = \{0, 1, 2, 3\}$, while in magnitude estimation setting, $\mathcal{G} = \mathbb{R}_+$. $r_{d_j,q;u_i}$ means that the assessor u_i annotates the document d_j under topic q with relevance level $r_{d_j,q;u_i}$.

3.4.1 Estimation of Document-Level Preference Matrix. In the subsequent content, we just take $p(d_1 \succ d_2|q)$ as a representative to show how to estimate for document-level preference matrix. In our setting, we assume that the relevance grade is the key factor to affect the preference probability between documents. For example, if the assessor u annotates both document d_1 and d_1' as grade g_α under topic q, we perceive that they have the same degree of preference compared with document d_2 in the view of u. The preference probabilities $p(d_1 \succ d_2|q)$ and $p(d_1' \succ d_2|q)$ are both controlled by the grade-level preference probability $p(g_\alpha \succ g_\beta)$, where g_β is the relevance score that assessor u annotates document d_2 as. To establish the connection between document-level preference probability $p(d_1 \succ d_2|q)$ and grade-level preference probability $p(g_\alpha \succ g_\beta)$, we propose two kinds of strategies: individual mode and aggregate mode.

In individual mode, we perceive that the relevance judgment result of each assessor as an independent sample. The document-level preference probability of each query-document-document tuple (q, d_1, d_2) is the mean value of all the assessors' preference probability. Eq 7 shows the calculation formula for document-level preference probability in individual mode:

$$p(d_1 \succ d_2|q) = \frac{1}{|\mathcal{U}_q|} \sum_{u \in \mathcal{U}_q} p(g_\alpha \succ g_\beta | g_\alpha = r_{d_1,q;u}, g_\beta = r_{d_2,q;u}). \tag{7}$$

In aggregate mode, we perceive that the aggregation result of all the relevance judgments in a query-document pair is a single sample. The document-level preference probability just equals to the preference probability between the aggregation relevance scores of these two documents, as Eq 8 shows. Eq 9 shows the definition of the symbol $\hat{r}_{d,q}$, where $Agg(\cdot)$ is denoted as the aggregate function to combine all the relevance judgment results under a particular document into a single relevance signal. We choose the median function as our aggregate function in the subsequent sections.

$$p(d_1 \succ d_2|q) = p\left(g_\alpha \succ g_\beta | g_\alpha = \hat{r}_{d_1,q}, g_\beta = \hat{r}_{d_2,q}\right) \tag{8}$$

$$\hat{r}_{d,q} = Agg\left(\{r_{d,q;u} | u \in \mathcal{U}_q\}\right) \tag{9}$$

3.4.2 The Requirement of Preference Tests. To estimate the grade-level preference probability, we need to introduce the preference judgments. In each time of preference judgment process, we show a topic q, and two related document d_1, d_2 to assessors. The assessors need to determine which document better satisfies the search needs. Finally, we combine all the assessors' preference judgments into a single signal $pf_{d_1 \succ d_2|q}$, to judge whether document d_1 is better than d_2 in topic q. When $pf_{d_1 \succ d_2} > (<) 0$, it means that document d_1 is (not) better than d_2. It is worth noting that we do not need to conduct preference test for all the preference tuples (q, d_1, d_2), we just need to sample some representative tuples for preference annotation instead. The sampling scale would be discussed in Sec 3.4.4.

3.4.3 Estimation of Grade-Level Preference Matrix. We estimate the grade-level preference matrix based on the collected preference test results.

In individual mode, we adopt all the independent tuple samples (e.g. tuple $(q, r_{d_1,q;u}, r_{d_2,q;u})$) to estimate for grade-level preference matrix $\{p(g_\alpha \succ g_\beta)\}|_{\alpha,\beta \in \mathcal{G}}$. Eq 10 shows the specific calculation formula, where the set S contains all the preference test tuples (q, d_1, d_2). The function $I(x)$ represents indicative function. When event x occurs, $I(x)$ equals to 1, otherwise it equals to 0.

$$p(g_\alpha \succ g_\beta) = \frac{\sum_{(q,d_1,d_2) \in S} \sum_{u \in \mathcal{U}_q} I\left(g_\alpha = r_{d_1,q;u}, g_\beta = r_{d_2,q;u}, pf_{d_1 \succ d_2|q} > 0\right)}{\sum_{(q,d_1,d_2) \in S} \sum_{u \in \mathcal{U}_q} I\left(g_\alpha = r_{d_1,q;u}, g_\beta = r_{d_2,q;u}\right)} \tag{10}$$

In aggregate mode, we need to aggregate the relevance judgment result and then estimate the grade-level preference probability. Eq 11 shows the specific calculation formula for grade-level preference probability:

$$p(g_\alpha \succ g_\beta) = \frac{\sum_{(q,d_1,d_2) \in S} I\left(g_\alpha = \hat{r}_{d_1,q}, g_\beta = \hat{r}_{d_2,q}, pf_{d_1 \succ d_2|q} > 0\right)}{\sum_{(q,d_1,d_2) \in S} I\left(g_\alpha = \hat{r}_{d_1,q}, g_\beta = \hat{r}_{d_2,q}\right)}. \tag{11}$$

3.4.4 The Scale of Preference Judgments. One more problem we are concerned with is how to determine the scale of preference judgments. When we conducted more preference judgments, the estimate for preference matrix could be more reliable, but at the same time we need to spend more funds. We need to find a balanced position between gain and cost.

Here, we will derive the relationship between the estimation reliability of the preference matrix and the scale of preference results. Due to lack of space, we only show the derivation process in aggregate mode. The process in individual mode is almost the same. Suppose we need to estimate for the probability $p(g_\alpha \succ g_\beta)$, its ideal value is p. For each related preference test sample, we denote

it as tuple (q, d_1, d_2), where $\hat{r}_{d_1,q} = g_\alpha$, $\hat{r}_{d_2,q} = g_\beta$. The random variable $I(pf_{d_1 \succ d_2|q} > 0)$ is sampled from Bernoulli distribution $B(p)$. Therefore, in Eq 11, our estimated probability $|S|\hat{p}(g_\alpha \succ g_\beta)$ obeys binomial distribution $B(|S|, p)$. Based on central limit theorem [28], when the random samples size $|S|$ is large enough, we can perceive that the random variable $\frac{|S|(\hat{p}(g_\alpha \succ g_\beta) - p)}{\sqrt{|S|p(1-p)}} \sim N(0,1)$. Eq 12 shows the relationship between $|S|$ and parameters δ. It illustrates that we can obtain a reliable preference probability estimate $\hat{p}(g_\alpha \succ g_\beta)$ under any parameter δ, as long as the scale $|S|$ is large enough.

$$|S| = \frac{p(1-p)}{\delta^2}\Phi^{-1}\left[\frac{1}{2}P\left(\hat{p}(g_\alpha \succ g_\beta) \in (p-\delta, p+\delta)\right) + \frac{1}{2}\right] \quad (12)$$

3.5 An Example to Calculate PDP

In this subsection, we will show the calculation method of PDP measure[2] through the example in Figure 1.

Following the procedures introduced in Sec 3.2, first we need to determine the details of preference tests, and then conduct preference tests. For simplicity, here we directly set the grade-level preference matrix in 4-grade and 7-grade setting as shown in Eq 13 and Eq 14 respectively. For example, in Eq 13, $P_{4g}(3,1) = 0.9$[3] , which means the probability that the user prefers a 3-grade document to a 1-grade document is 0.9 in 4-grade setting. Noting that in the real situation, the grade-level preference matrix in individual mode and aggregate mode are different, here we predefine them as the same value for simplicity.

$$P_{4g} = \begin{bmatrix} 0.50 & 0.15 & 0.10 & 0.05 \\ 0.85 & 0.50 & 0.15 & 0.10 \\ 0.90 & 0.85 & 0.50 & 0.15 \\ 0.95 & 0.90 & 0.85 & 0.50 \end{bmatrix} \quad (13)$$

$$P_{7g} = \begin{bmatrix} 0.50 & 0.20 & 0.15 & 0.10 & 0.06 & 0.03 & 0.01 \\ 0.80 & 0.50 & 0.20 & 0.15 & 0.10 & 0.06 & 0.03 \\ 0.85 & 0.80 & 0.50 & 0.20 & 0.15 & 0.10 & 0.06 \\ 0.90 & 0.85 & 0.80 & 0.50 & 0.20 & 0.15 & 0.10 \\ 0.94 & 0.90 & 0.85 & 0.80 & 0.50 & 0.20 & 0.15 \\ 0.97 & 0.94 & 0.90 & 0.85 & 0.80 & 0.50 & 0.20 \\ 0.99 & 0.97 & 0.94 & 0.90 & 0.85 & 0.80 & 0.50 \end{bmatrix} \quad (14)$$

Next, we need to estimate the document-level preference matrix in each setting. Here, we take the first image d_1 and the second image d_2 as an example. In 7-grade setting, using Eq 7, we can get that in individual mode $p(d_1 \succ d_2|q) = \frac{1}{3} \times \left[P_{7g}(5,4) + P_{7g}(6,5) + P_{7g}(5,5)\right] = 0.70$. In aggregate mode, based on Eq 8, when we use median function as the aggregate function, then in 7-grade setting, $\hat{r}_{d_1,q} = 5$, $\hat{r}_{d_2,q} = 5$, thus, $p(d_1 \succ d_2|q) = P_{7g}(5,5) = 0.50$. Using these methods, ultimately we can obtain the document-level preference matrices.

Finally, we calculate PDP based on the document-level preference matrices. We need to first optimize score vector s^* based on Eq 5, and then conduct simulation using Eq 6. As a result of this run[4] , in 4-grade setting, PDP equals to 4.2142 and 4.1705 in individual and aggregate modes respectively, while in 7-grade setting the values of PDP in individual and aggregate modes are 4.0249 and 3.8180 respectively. As expected, the PDP in 7-grade setting is lower than in 4-grade setting, which indicates that the 7-grade data is much more discriminative than the 4-grade data.

[2]The code can be found in https://github.com/chuzhumin98/PDP

[3]Here we build the index of the matrix starting from 0.

[4]Because of the randomness in the process of simulation, the values of PDP are slightly different between runs.

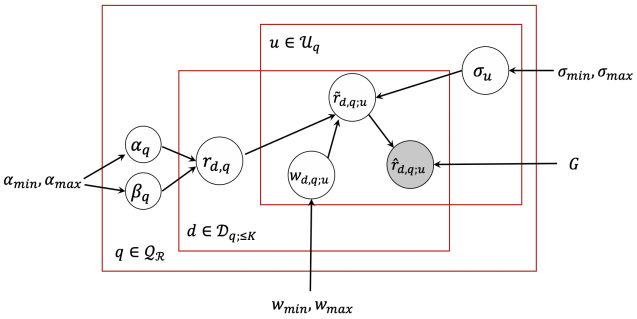


Figure 3: Diagram of synthetic dataset generation architecture

In the subsequent sections, we aim to answer the following three research questions:

- **RQ1**: What impacts do the quality and relevance scale of relevance judgments have on PDP? Can we use PDP to evaluate the quality of relevance judgments?
- **RQ2**: How many preference test results do we need to obtain a reliable PDP value?
- **RQ3**: Can we use PDP to evaluate the existing relevance judgment collections effectively?

In Sec 4, we conduct extensive experiments on synthetic datasets to answer **RQ1** and **RQ2**. To answer **RQ3**, we further conduct experiments on real-world datasets in Sec 5.

4 EXPERIMENTS ON SYNTHETIC DATASETS

4.1 Experimental Setting

To answer **RQ1**, we require a relevance judgment dataset with gold standard for measuring its quality. Also, this dataset needs to contain relevance judgment results under different grade settings, so that we can test the influence of relevance scale on PDP. These requirements are difficult to satisfy in real-world datasets, so we generate a large-scale synthetic dataset to answer **RQ1**. Figure 3 shows the synthetic dataset generation architecture.

In our setting, we perceive the true relevance of each query-document pair $r_{d,q}$ is a real number within the interval $[0, 1]$. The larger value of $r_{d,q}$ indicates the higher relevance between topic q and document d. To generate $r_{d,q}$, we introduced two topic-level variables α_q and β_q. We assume that $r_{d,q}$ obeys a beta distribution with parameter α_q and β_q, i.e., $r_{d,q} \sim Beta(\alpha_q, \beta_q)$. As for α_q, β_q under each topic q, they are all sampled from the uniform distribution $U(\alpha_{min}, \alpha_{max})$.

With the true relevance set $\{r_{d,q}|d \in \mathcal{D}_{q;\leq K}\}$, next, we need to generate the corresponding relevance judgment result $\hat{r}_{d,q;u}$. In general, the value of $\hat{r}_{d,q;u}$ is limited to a finite set. For example, in G-grade setting, usually $\hat{r}_{d,q;u} \in \{0, 1, \cdots, G-1\}$. To bridge the gap from $r_{d,q}$ and $\hat{r}_{d,q;u}$, we introduce intermediate hidden variables $\tilde{r}_{d,q;u}$ ($\in [0,1]$). We assume that the assessor's perceived relevance $\tilde{r}_{d,q;u}$ is determined by true relevance $r_{d,q}$, assessor-level bias factor σ_u and annotation-level bias factor $w_{d,q;u}$. The factors σ_u and $w_{d,q;u}$ are sampled from the uniform distribution $U(\sigma_{min}, \sigma_{max})$

and $U(w_{min}, w_{max})$, respectively. Eq 15 shows the specific calculation formula for assessor's perceived relevance, where ϵ_1, ϵ_2 are two random factors independently sampled from standard normal distribution.

$$\tilde{r}_{d,q;u} = r_{d,q} + \sigma_u \epsilon_1 + \sigma_u e^{w_{d,q;u}} \epsilon_2 \tag{15}$$

To obtain $\hat{r}_{d,q;u}$, we uniformly split the interval $[0, 1]$ into G slices: $[0, 1/G), [1/G, 2/G), \cdots, [(G-1)/G, 1]$. We just need to check which interval $\tilde{r}_{d,q;u}$ belongs to, then we can set $\hat{r}_{d,q;u}$ to be $0, 1, \cdots, G-1$, respectively.

To reduce the variance of metrics, for every Y topics, we choose the same batch of "assessors" to conduct relevance judgments. Also, to better evaluate the quality of relevance judgments under these Y topics with a single parameter, we set the assessor-level bias factors σ_u of the same batch assessors as the same.

In our experiments, we set $|Q_{\mathcal{R}}| = 10,000$, $Y = 100$, $K = 5$, $|\mathcal{U}_q| = 15$. We change the relevance scale $|G|$ ranging from 2 to 30. Also, we set $\alpha_{min}, \alpha_{max}, \sigma_{min}, \sigma_{max}, w_{min}, w_{max}$ as 2, 5, 0.001, 0.3, −0.5, 0.5, respectively.

At the same time, we need to generate synthetic preference judgment results in the process of PDP calculation. To answer **RQ1**, for the sake of the PDP value stability, we use the preference results of all the preference tuples (q, d_1, d_2) to estimate document-level preference matrix. To answer **RQ2**, we sample B preference tuples under each grade pair (g_α, g_β) using the sampling with replacement approach. We choose the parameter B from $\{10, 20, 40, 80, 160, 320\}$. Also, due to the high computational cost, we only select three representative batches of topics to conduct experiments in **RQ2**. The σ_u of these three batches are closest to 0.05, 0.15 and 0.25, which represent high-quality, medium-quality and low-quality of relevance judgments respectively.

Meanwhile, since we consider the preference judgments as the gold standard, all these preference results are generated without noise based on the true labels. In other words, we set $pf_{d_1 \succ d_2|q} > 0$ if and only if $r_{d_1,q} > r_{d_2,q}$.

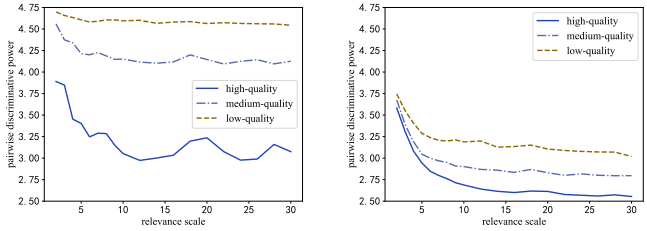
4.2 Experimental Results

4.2.1 Influence of the Quality and Relevance Scale of Relevance Judgments on PDP. To answer **RQ1**, we first investigate how PDP varies with different relevance scales and different annotation qualities. We select the relevance scale G ranging from 2 to 30. Based on σ_u, we divide the relevance judgment data into three categories: high quality ($0.001 \leq \sigma_u \leq 0.1$), medium quality ($0.1 < \sigma_u \leq 0.2$), and low quality ($0.2 < \sigma_u \leq 0.3$).

Figure 4 shows how relevance scale and annotation quality affect the mean of PDP. We can find that the mean value of PDP decreases when the relevance scale G increases. It indicates that if we can maintain the same annotation quality while the relevance scale increases, we will get a more discriminative relevance collection. However, in practice, that is not easy to achieve. A higher relevance scale setting requires more granular relevance perception. Assessors are more likely to be confused in the annotation process.

Considering the relationship between annotation quality and mean value of PDP, we observe that higher annotation qualities bring lower PDP values, when controlling the relevance scale as the same. That is what we expect.

Results also show that the mean value of PDP in aggregate mode is lower than the one in individual mode, which illustrates that in most cases the aggregation relevance scores could be more discriminative than directly using the original judgment scores. More specifically, we find that PDP in aggregate mode can extract discriminative information even in the low-quality collection, but not for PDP in individual mode. The mean of individual-mode PDP in the low-quality collection is about 4.6, which is only slightly better than the random annotation case (4.7875).



(a) Individual mode **(b) Aggregate mode**

Figure 4: PDP varies with different grade settings and annotation qualities in synthetic dataset, subfigures (a) and (b) represent PDP in individual mode and aggregate mode respectively

4.2.2 The Consistency Between Evaluation Metrics and Annotation Quality. To answer **RQ1**, we also analyze the consistency between evaluation metrics and annotation quality. Here, we use Fleiss' κ [20], Weighted κ[20], Krippendorff's α [24] and Φ [10] as comparisons. From Eq 15, we know that larger σ_u indicates that assessor u is more confused with the given grade setting to make more mistakes in the judgment process. In our experiments, we set the parameters σ_u of assessors annotated for each Y topics (i.e., a batch) as the same. Therefore, we can take the parameter σ_u to evaluate the mean annotation quality of the judgment results under these Y topics.

Figure 5 shows the consistency between these evaluation metrics and σ_u, where we use Spearman's rank correlation coefficient [1] to evaluate for the consistency. We can find that when the relevance scale increases, individual mode PDP, Krippendorff's α, weighted κ and Φ keep high consistency with annotation quality, aggregate mode PDP becomes more reliable, while kappa performs worse. This phenomenon is in line with our expectations. When relevance scale increases, even excellent assessors can hardly guarantee that the judgment results are identical. Hence, in high-grade settings, Fleiss' κ fails to distinguish the better one from given relevance collections. In contrast, the calculation of PDP is based on the pairwise comparison. When the relevance scale increases, the impact of annotation quality on PDP value is even more significant. High-quality assessors are more likely to bring consistent preference information compared with assessors in low quality. As for Krippendorff's α and Weighted κ, they consider the ordinal difference to design penalty weights of inconsistency. That seems to be reliable even when the relevance grade becomes large. Another consistency metric, Φ, maps different relevance grades into different real numbers within the interval $[0, 1]$. Higher relevance scale has no harm to estimate the posterior distribution $P(\vec{\mu}, \Phi|X)$. Thus, Φ still keeps high agreement with σ_u when relevance scale increases.

Figure 5 also depicts the differences between two modes of PDP. We observe that PDP in individual mode holds the highest consistency with annotation quality in all these metrics, no matter what

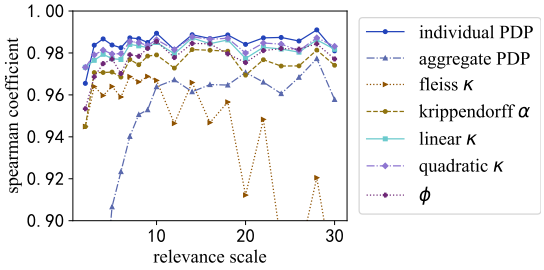


Figure 5: The consistency between evaluation metrics and annotation quality in different grade settings

the relevance scale is. PDP in aggregate mode performs not well, especially when the relevance scale is small. We assume that is because aggregate mode PDP might devote more attention to the intrinsic distinguishability characteristics among the documents, rather than the relevance judgments.

Thus, if one only focuses on the quality of relevance judgments, we recommend to employ individual mode PDP. If one expects to take the documents themselves into account, aggregate mode PDP might be a better choice.

4.2.3 Effect of the Preference Judgment Scale on the Reliability of PDP. To answer **RQ2**, we choose different scales of preference tuples to conduct experiments. Figure 6 shows the experimental results. We choose the relevance scale four, seven, and sixteen as representatives. The results of other relevance scales are similar to them. For each scale of preference test samples, we repeat the experiments 100 times independently to obtain each box-and-whisker in the graph.

From Figure 6, we observe a significant improvement of PDP's reliability when the scale of preference test samples become larger. When the relevance scale becomes larger, we can see a slight reliability improvement when keeping the same scale of preference test samples in each grade pair. However, the total required scale of preference test samples still becomes larger when the relevance scale increases, because the number of grade pair shows a quadratic relationship with the relevance scale. Compared with aggregate mode PDP, individual mode PDP shows more stable performance even on the small preference test samples and low-quality annotations condition. We assume that is because the actual samples size in individual mode is $|\mathcal{U}_q|$ larger than the size in aggregate mode, as the difference demonstrated by Eq 10 and Eq 11. Another interesting phenomenon occurs in Figure 6d: Aggregate mode PDP draws the wrong conclusion when comparing low-quality and medium-quality relevance judgment data in 4-grade setting. This phenomenon coincides with our findings in Sec 4.2.2: aggregate mode PDP exhibits lower consistency with relevance annotation quality when the relevance scale is not too large.

To answer the specific quantity required to obtain a reliable PDP value, we recommend to use individual mode PDP when the budget is not sufficient enough. In this case, no more than 40 preference test samples for each grade pair are required to evaluate relevance judgments with PDP reliably. The aggregate mode PDP can also be applied when we can collect no less than 160 preference test samples for each grade pair.

Table 1: Statistics of Original NTCIR-15 WWW-3 Data

subtask	Chinese	English
#topics	80	160
#assessors/topic	3	8
pool depth	30	15
total #docs pooled	11, 172	32, 375
relevance scale	4	3
submitted #runs	11	37

5 EXPERIMENTS ON REAL-WORLD DATASETS

5.1 Experimental Setting

We select two datasets to test the performance of our metric: NTCIR-15 WWW-3 Chinese and English subtask [33] datasets. Table 1 summarizes the statistics of the original NTCIR-15 WWW-3 data. Next, we will show more details about these two datasets.

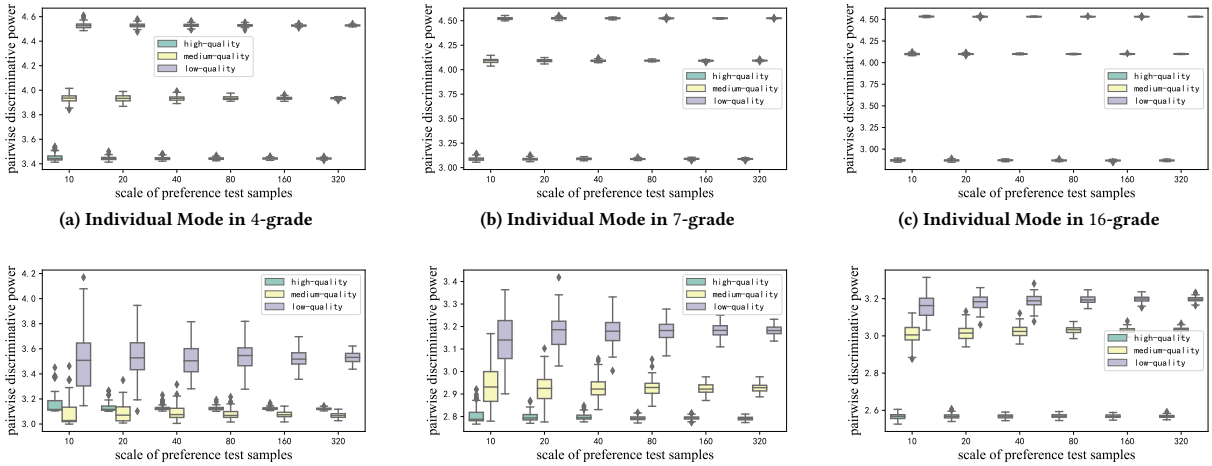
5.1.1 Chinese Subtask. In the Chinese subtask data, the original relevance judgments were conducted in a 4-grade setting. To obtain more types of relevance judgments data, we additionally conduct 7-grade relevance judgments. In our setting, we present a statement to assessors: the search need can be satisfied with document d under topic q. Assessors need to annotate the degree of agreement from 0 (strongly disagree), 1 (disagree), 2(a little disagree), 3 (neutral), 4 (a little agree), 5 (agree), and 6 (strongly agree). We conduct 7-grade relevance assessments in the top-10 documents pooling set, totaling of 5, 098 query-document pairs. For each query-document pair, we ask three assessors to annotate independently.

Based on a set of 4-grade relevance judgment results, we perform *relevance grade reduction* to obtain binary relevance assessments by treating all assessments with grade $M(<=3)$ or above as relevant and others as nonrelevant. We denote the result by *4to2* ($\geq M$). For example, "*4to2* (≥ 1)" means relevance grades 3,2,1 are treated as relevant and 0 as nonrelevant. In our experiments, we try all three reduction strategies (*4to2* (≥ 1), *4to2* (≥ 2), and *4to2* (≥ 3)) to obtain binary relevance collections from original 4-grade data.

As the requirement of preference judgment results in the process of PDP calculation, we also conduct preference tests to estimate the document-level preference matrix. To ensure the accuracy of the estimates, we randomly sample no less than 100 preference tuples (q, d_1, d_2) for each different grade pair (g_α, g_β) in original 4-grade and 7-grade settings, totaling 2, 244 preference tuples. We conducted preference tests under strict mode[5] [6], i.e., each assessor needs to choose from the preference levels $-2/-1/1/2$, which indicate the document d_1 is substantially-better/better/worse/substantially-worse than document d_2 under topic q, respectively. For each preference tuple, we asked three assessors to annotate. As the preference judgment results are perceived as ground truth in our experiments, these assessors are permitted to discuss the contradictory results to improve the annotation quality further after judgments. The final preference signal $pf_{d_1 \succ d_2|q}$ is set as the median of the assessors' judgment results.

5.1.2 English Subtask. In the English subtask data, due to the limited resources, we only make use of its original relevance judgments. In the WWW-3 task's overview paper [33], Sakai et al.

[5]The calculation method of PDP with weak mode preference tests is almost the same. Here we only conducted preference judgments under strict mode due to the limited resources.



(a) Individual Mode in 4-grade (b) Individual Mode in 7-grade (c) Individual Mode in 16-grade

(d) Aggregate Mode in 4-grade (e) Aggregate Mode in 7-grade (f) Aggregate Mode in 16-grade

Figure 6: The reliability of PDP varies with different preference judgments scale in different grade settings and different modes. Under each samples scale of each graph, the left, middle and right boxes represent the high quality, medium quality and low quality annotation data respectively.

adopted another interesting method to aggregate the judgment results. They obtained a 5-grade relevance score by taking the integer part of $\log_2(S+1)$, where S is the sum of the raw labels.

In our experiments, we adopt these 5-grade aggregation results, as well as the original 3-grade setting and two kinds of reduced 2-grade settings (*3to2* (≥ 1) and *3to2* (≥ 2)).

As for preference judgments, we use the same strategies as the Chinese subtask to conduct. In total, we collect 300 preference tuples for the follow-up experiments.

5.2 Experimental Results

Due to the effectiveness of Krippendorff's α and Φ on synthetic experiments, we continue to use them as comparisons with PDP on real-world data. Discriminative power (DP) [30], as an existing evaluation metric which is able to evaluate the discriminative power of relevance judgment collections, is also calculated.

Table 2 summarizes the experimental results on real-world dataset. We calculate all the evaluation metrics on both top-5 and top-10 relevance judgment collections. As for the top-K ($K = 5, 10$) collection, PDP is calculated based on the top-K ranking list of each collected run. The value of PDP shown in Table 2 is the mean value of PDP upon all the topics and all the submitted runs. Based on [30], we calculate the values of DP on a series of evaluation measures, including linear nDCG, exponential nDCG [5, 23] , Q-measure [32], nERR [9] and RBP [26]. As for consistency metrics, Krippendorff's α and Φ are calculated with top-K pooling document set.

5.2.1 Overall Analysis in Chinese Subtask. Table 2 shows the experimental results in the Chinese subtask. The value of Krippendorff's α decays dramatically when relevance scale increases. However, we cannot draw any conclusions from this decay because the annotation consistency between assessors would naturally decay as the relevance scale increases. Therefore, Krippendorff's α is not an appropriate metric to evaluate annotation quality in cross-grade scenarios.

As for another consistency metric Φ, when focusing on the three *4to2* collections, it draws an opposite conclusion compared with DP and PDP: Φ points that *4to2*(≥ 3) performs best, while both DP and PDP assume that *4to2*(≥ 1) performs best. This phenomenon reflects the contradiction between high consistency (Φ) and high distinction (DP and PDP). We perceive that the conclusions drawn by DP and PDP are a bit more plausible because of the existing "reliable but useless" phenomenon.

Two kinds of PDP measures tell us completely opposite conclusions when comparing collections in 4-grade and 7-grade settings. We conduct significance tests, and find that the differences of all kinds of PDP between these two settings are not significant even when we relax the p-value to 0.1. DP measure also has a mix results on the comparison of these two datasets. Some indicate 4-grade setting is better, while other stand for 7-grade setting. This phenomenon indicates that the discriminative ability of these two datasets is relatively close.

When compared the original 4-grade collection with the reduced binary collections, we find that the reduction strategies harm the quality of relevance judgment data. Among these reduction strategies, DP and PDP show that *4to2* (≥ 1) performs best while *4to2* (≥ 3) performs worst. It illustrates that the difference between grade 0 and higher grades contains most information in original 4-grade relevance judgment data, while the difference between grade 3 and lower grades makes the least contribution in original data. These conclusions are consistent with the findings in [36].

5.2.2 Overall Analysis in English Subtask. In English subtask, an interesting finding is that the aggregation 5-grade setting used in WWW-3 overview paper [33] performs best. We assume that is because the document distribution under the original 3-grade is unbalanced. The conversion from 3-grade into 5-grade makes data distribution more even.

The reduction strategies in English subtask also harm the annotation quality. It is worth noting that the *3to2* (≥ 1) collection has a competitive performance against with original 3-grade relevance collection. This phenomenon indicates that the difference between grade 1 and grade 2 in original relevance judgment data is not reliable. The data confirm our conjecture: the grade-level preference probability $p(2 > 1)$ is only 0.73 in aggregate mode.

In the English subtask, we can also observe the high consistency between PDP and DP. The low values of Krippendorff's α and Φ

Table 2: The mean values of Krippendorff's α, Φ, DP and PDP(ours) on both WWW-3 Chinese and English subtask datasets. The boldface highlights the best-performing annotation setting.

Subtask			Chinese Subtask					English Subtask			
#grade			4to2(≥ 1)	4to2(≥ 2)	4to2(≥ 3)	4(orig)	7(orig)	3to2(≥ 1)	3to2(≥ 2)	3(orig)	3to5(logS)
α[24]	top-5		0.6256	**0.6779**	0.4799	0.6598	0.3848	0.1213	0.0850	**0.1312**	/
	top-10		0.6048	**0.6782**	0.4715	0.6468	0.4037	0.1198	0.0833	**0.1291**	/
Φ[10]	top-5		0.7806	0.8743	0.8885	**0.9811**	0.9552	-0.2202	0.1062	**0.1501**	/
	top-10		0.7781	0.8866	0.9123	**0.9851**	0.9595	-0.2214	0.0822	**0.1313**	/
DP[30]	top-5	linear_nDCG	**69.09%**	65.45%	47.27%	**69.09%**	**69.09%**	69.52%	56.61%	72.82%	**74.32%**
		exp_nDCG	**67.27%**	65.45%	49.09%	**67.27%**	65.45%	69.07%	54.80%	70.57%	**73.57%**
		Q-measure	72.73%	69.09%	45.45%	**78.18%**	67.27%	69.37%	53.30%	72.07%	**73.57%**
		nERR	61.82%	56.36%	47.27%	**63.64%**	**63.64%**	65.62%	54.80%	68.17%	**69.52%**
		RBP	69.09%	67.27%	52.73%	**74.55%**	69.09%	70.27%	58.26%	72.07%	**75.38%**
	top-10	linear_nDCG	72.73%	72.73%	50.91%	72.73%	**80.00%**	75.53%	63.96%	76.28%	**78.83%**
		exp_nDCG	**72.73%**	70.91%	50.91%	69.09%	70.91%	75.08%	62.31%	75.68%	**79.43%**
		Q-measure	72.73%	70.91%	45.45%	72.73%	**80.00%**	73.87%	58.11%	74.77%	**78.38%**
		nERR	**63.64%**	60.00%	47.27%	61.82%	**63.64%**	67.72%	60.66%	69.67%	**68.92%**
		RBP	74.55%	72.73%	56.36%	**80.00%**	**80.00%**	76.43%	66.52%	77.33%	**80.48%**
PDP(ours)	top-5	aggregate	4.2442	4.2810	4.5682	4.1060	**3.9864**	4.4386	4.6871	4.4885	**4.1821**
		individual	4.3652	4.4013	4.6317	**4.2289**	4.3290	4.6931	4.7343	**4.6790**	/
	top-10	aggregate	13.5859	13.7294	14.5224	13.2123	**12.8050**	14.1281	14.8191	14.2650	**13.4380**
		individual	13.9136	14.0529	14.6942	**13.5511**	13.8157	14.8135	14.9304	**14.7831**	/

in all the collections indicates the low reliability of the original relevance judgment results.

5.2.3 The Relevance Grade Extension Experiments. In Chinese subtask, we observe the close performance between the original 4-grade and 7-grade data. We wonder if we just extend the original 4-grade data into 7-grade to generate pseudo 7-grade assessments, whether the pseudo 7-grade collection can have a similar or even better performance.

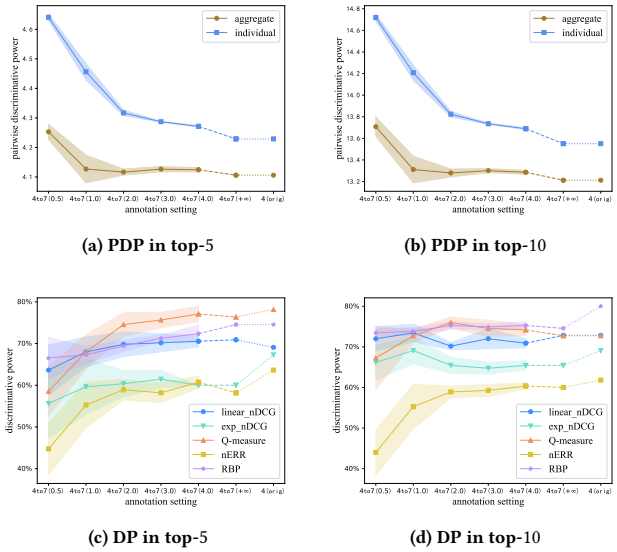
In the extension process to obtain pseudo 7-grade relevance judgment collection, we introduced parameter w to depict the degree of uncertainty. For simplicity, we shall refer to the resultant data as $4to7\ (w)$, e.g. "$4to7\ (1.0)$" (4-grade relevance grades extended to 7-grade data with $w = 1.0$).

Eq 16 shows the specific transition probability. The degree of uncertainty becomes weaker when the parameter w becomes larger. In particular, when w approaches positive infinity, the extension results become completely certain. The grade 0/2/4/6 in $4to7\ (+\infty)$ setting exactly corresponds to grade 0/1/2/3 in original 4-grade setting. In our experiments, we choose parameter w from the set $\{0.5, 1.0, 2.0, 3.0, 4.0, +\infty\}$. For each w, we randomly generated 5 slices of extension 7-grade collections to obtain a more stable values of evaluation metrics.

$$P(g_{\text{7-grade}} = g^{(7)} | g_{\text{4-grade}} = g^{(4)}) = \frac{e^{-w|2g^{(4)} - g^{(7)}|}}{\sum_{g \in \mathcal{G}^{(7)}} e^{-w|2g^{(4)} - g|}} \quad (16)$$

Figure 7 shows the the mean and standard derivation of PDP and DP under different parameters w. Almost all the evaluation metrics indicate that the collection discrimination becomes stronger when the w becomes larger. This phenomenon demonstrates that introducing randomness in the extension process has a side-effect on the annotation quality to some extents.

Another interesting phenomenon occurs on the data $4to7\ (+\infty)$. When comparing it with original 4-grade collection, we expect the values of evaluation metric in these two datasets would be the same due to their equivalence. As a result, we observe that PDP on $4to7\ (+\infty)$ is just the same as the 4-grade setting due to their identical



(a) PDP in top-5 **(b) PDP in top-10**

(c) DP in top-5 **(d) DP in top-10**

Figure 7: The mean and standard derivation of PDP and DP in top-5 and top-10 ranking lists under different extension parameters w

document-level preference matrices. However, on DP measures, there exists a significant shift between two collections. The DP values of most metrics on $4to7\ (+\infty)$ setting are lower than those on the 4-grade setting. We assume that this phenomenon occurs because many ranking evaluation measures are not scalable. The linear or exponential gain of nDCG and Q-measure, the exponential stop probability of nERR cause the variability of measures when the relevance scale changes. This phenomenon is not a good signal. It might lead us to get wrong conclusions when using DP measures to compare annotation quality in cross-grade scenarios.

6 CONCLUSIONS

In this paper, we propose a novel metric PDP to evaluate the discriminative ability of relevance judgment collections. Unlike the

DP measure proposed by Sakai et al. [30], PDP does not need to be calculated based on a series of ranking lists from different retrieval systems but only requires introducing affordable amount of additional preference tests to evaluate the relevance judgment collections. We propose a unified framework for generating relevance judgment collections under multi-grade setting and then generate a series of synthetic data sets on different grade settings. We conduct a series of experiments on both synthetic and real-world datasets. Experimental results confirm that PDP, especially the individual mode version, can characterize the annotation quality to some extent and show competitive performance compared with κ, Krippendorff's α, Φ and DP. In our experiments, we also observe that consistency metrics is not appropriate to compare the annotation quality in cross-grade scenarios. PDP and DP can satisfy this evaluation need, but DP suffers from the value shift between different grades. We recommend the follow-up researchers adopt PDP metric to compare relevance judgments collected on different annotation settings. Then they can decide which setting to be employed for larger-scale annotation experiments.

7 ACKNOWLEDGEMENTS

This work is supported by the National Key Research and Development Program of China (2018YFC0831700), Natural Science Foundation of China (Grant No. 61732008, 61532011, 61902209, U2001212), Beijing Academy of Articial Intelligence (BAAI), Tsinghua University Guoqiang Research Institute, Beijing Outstanding Young Scientist Program (NO. BJJWZYJH012019100020098) and Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China.

REFERENCES

[1] Haldun Akoglu. 2018. User's guide to correlation coefficients. *Turkish journal of emergency medicine* 18, 3 (2018), 91–93.
[2] Enrique Amigó, Fernando Giner, Stefano Mizzaro, and Damiano Spina. 2018. A Formal Account of Effectiveness Evaluation and Ranking Fusion. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval*. 123–130.
[3] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P de Vries, and Emine Yilmaz. 2008. Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 667–674.
[4] Léon Bottou. 2012. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*. Springer, 421–436.
[5] Christopher Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Gregory N Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine learning (ICML-05)*. 89–96.
[6] Ben Carterette, Paul N Bennett, David Maxwell Chickering, and Susan T Dumais. 2008. Here or there. In *European Conference on Information Retrieval*. Springer, 16–27.
[7] Praveen Chandar and Ben Carterette. 2012. Using preference judgments for novel document retrieval. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 861–870.
[8] Praveen Chandar and Ben Carterette. 2013. Preference based evaluation measures for novelty and diversity. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 413–422.
[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*. 621–630.
[10] Alessandro Checco, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. 2017. Let's agree to disagree: Fixing agreement measures for crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 5.
[11] Charles LA Clarke, Nick Craswell, and Ian Soboroff. 2004. Overview of the TREC 2004 Terabyte Track.. In *TREC*, Vol. 4. 74.
[12] Charles LA Clarke, Mark D Smucker, and Alexandra Vtyurina. 2020. Offline Evaluation by Maximum Similarity to an Ideal Ranking. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 225–234.
[13] Charles LA Clarke, Alexandra Vtyurina, and Mark D Smucker. 2020. Assessing top-k preferences. *arXiv preprint arXiv:2007.11682* (2020).
[14] Charles LA Clarke, Alexandra Vtyurina, and Mark D Smucker. 2020. Offline evaluation without gain. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*. 185–192.
[15] Cyril Cleverdon and EM Keen. 1966. Aslib–Cranfield research project. *Factors determining the performance of indexing systems* 1 (1966).
[16] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.
[17] Kevyn Collins-Thompson, Craig Macdonald, Paul Bennett, Fernando Diaz, and Ellen M Voorhees. 2015. *TREC 2014 web track overview*. Technical Report. MICHIGAN UNIV ANN ARBOR.
[18] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. 2005. A tutorial on the cross-entropy method. *Annals of operations research* 134, 1 (2005), 19–67.
[19] Nicola Ferro and Carol Peters. 2019. *Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF*. Vol. 41. Springer.
[20] Joseph L Fleiss, Jacob Cohen, and Brian S Everitt. 1969. Large sample standard errors of kappa and weighted kappa. *Psychological bulletin* 72, 5 (1969), 323.
[21] John Guiver and Edward Snelson. 2009. Bayesian inference for Plackett-Luce ranking models. In *proceedings of the 26th annual international conference on machine learning*. 377–384.
[22] Lei Han, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. 2019. On transforming relevance scales. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 39–48.
[23] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.
[24] Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. (2011).
[25] Cheng Luo, Tetsuya Sakai, Yiqun Liu, Zhicheng Dou, Chenyan Xiong, and Jingfang Xu. 2017. Overview of the ntcir-13 we want web task. *Proc. NTCIR-13* (2017).
[26] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)* 27, 1 (2008), 1–27.
[27] Kevin Roitero, Eddy Maddalena, Gianluca Demartini, and Stefano Mizzaro. 2018. On fine-grained relevance scales. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 675–684.
[28] Murray Rosenblatt. 1956. A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences of the United States of America* 42, 1 (1956), 43.
[29] Reuven Y Rubinstein and Dirk P Kroese. 2016. *Simulation and the Monte Carlo method*. Vol. 10. John Wiley & Sons.
[30] 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*. 525–532.
[31] Tetsuya Sakai, Douglas W Oard, and Noriko Kando. [n.d.]. *Evaluating Information Retrieval and Access Tasks: NTCIR's Legacy of Research Impact*. Springer Nature.
[32] 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*. 1043–1052.
[33] Tetsuya Sakai, Sijie Tao, Zhaohao Zeng, Yukun Zheng, Jiaxin Mao, Zhumin Chu, Yiqun Liu, Maria Maistro, Zhicheng Dou, Nicola Ferro, et al. 2020. Overview of the NTCIR-15 We Want Web with CENTRE (WWW-3) Task. *Proceedings of NTCIR-15. to appear* (2020).
[34] Tetsuya Sakai and Zhaohao Zeng. 2020. Good evaluation measures based on document preferences. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 359–368.
[35] Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal* 27, 3 (1948), 379–423.
[36] Eero Sormunen. 2002. Liberal relevance criteria of TREC- counting on negligible documents?. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. 324–330.
[37] Rong Tang, William M Shaw Jr, and Jack L Vevea. 1999. Towards the identification of the optimal number of relevance categories. *Journal of the American Society for Information Science* 50, 3 (1999), 254–264.
[38] Andrew Turpin, Falk Scholer, Stefano Mizzaro, and Eddy Maddalena. 2015. The benefits of magnitude estimation relevance assessments for information retrieval evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 565–574.
[39] Anthony J Viera, Joanne M Garrett, et al. 2005. Understanding interobserver agreement: the kappa statistic. *Fam med* 37, 5 (2005), 360–363.
[40] Ellen M Voorhees and Donna Harman. 2002. Overview of TREC 2002.. In *Trec*.
[41] Ellen M Voorhees, Donna K Harman, et al. 2005. *TREC: Experiment and evaluation in information retrieval*. Vol. 63. MIT press Cambridge.
[42] Ziying Yang, Alistair Moffat, and Andrew Turpin. 2018. Pairwise crowd judgments: Preference, absolute, and ratio. In *Proceedings of the 23rd Australasian Document Computing Symposium*. 1–8.
[43] Dongqing Zhu and Ben Carterette. 2010. An analysis of assessor behavior in crowdsourced preference judgments. In *SIGIR 2010 workshop on crowdsourcing for search evaluation*. 17–20.