Revisiting the Evaluation of Diversified Search Evaluation Metrics with User Preferences

Fei Chen[1], Yiqun Liu[1], Zhicheng Dou[2,⋆], Keyang Xu[1], Yujie Cao[1], Min Zhang[1], and Shaoping Ma[1]

[1] Tsinghua University, Beijing, China
[2] Renmin University of China, Beijing, China
chenfei27@gmail.com

Abstract. To validate the credibility of diversity evaluation metrics, a number of methods that "evaluate evaluation metrics" are adopted in diversified search evaluation studies, such as Kendall's τ, Discriminative Power, and the Intuitiveness Test. These methods have been widely adopted and have aided us in gaining much insight into the effectiveness of evaluation metrics. However, they also follow certain types of user behaviors or statistical assumptions and do not take the information of users' actual search preferences into consideration. With multi-grade user preference judgments collected for diversified search result lists displayed parallel, we take user preferences as the ground truth to investigate the evaluation of diversity metrics. We find that user preference at the subtopic level gain similar results with those at the topic level, which means we can use user preference at the topic level with much less human efforts in future experiments. We further find that most existing evaluation metrics correlate with user preferences well for result lists with large performance differences, no matter the differences is detected by the metric or the users. According to these findings, we then propose a preference-weighted correlation, the *Multi-grade User Preference* (MUP) method, to evaluate the diversity metrics based on user preferences. The experimental results reveal that MUP evaluates diversity metrics from real users' perspective that may differ from other methods. In addition, we find the relevance of the search result is more important than the diversity of the search result in the diversified search evaluation of our experiments.

1 Introduction

Evaluation metrics have always been one of the most important and challenging topics in information retrieval research because of the part they play in tuning and optimizing retrieval systems [3]. For diversified search tasks, many evaluation methods, such as $\alpha - nDCG$ [7] and $D\# - measures$ [14], have been proposed. These metrics more or less simplify the assumptions about user behaviors. For example, with the assumption that users always view search results from top

⋆ The work was done when Zhicheng Dou was working at Microsoft Research.

A. Jaafar et al. (Eds.): AIRS 2014, LNCS 8870, pp. 48–59, 2014.

to bottom, most metrics leverage a ranking-based discount. These assumptions may help to simplify the evaluation process but also make the evaluation deviate from user's actual experience and satisfaction [17].

In this paper, we propose to take user preferences as the ground truth to evaluate diversity metrics. We first compare user preferences collected at the subtopic level, user preferences at the topic level, and the weighted user preferences with each other. Diversity metrics are then discussed in terms of the performance differences of run pairs detected by the metric, which is similar with Sanderson's work [16] (more recent than Turpin's work [18,19]) except that we involve more diversity metrics. Other differences between Sanderson's work and ours are that we collect user preferences in a graded strategy and leverage τ_b to evaluate the correlations between diversity metrics and the graded user preferences, whereas Sanderson *et al.* use agreement/disagreement between metrics and binary user preferences to discuss metrics' properties. And on the other hand, we further discuss the same metrics in terms of the performance differences of run pairs detected by the users. Based on the graded user preferences, we then propose a preference-weighted correlation, namely Multi-grade User Preference (MUP), to evaluate the diversity metrics. Finally, three widely-used methods for evaluating diversity metrics, namely Kendall's τ, Discriminative Power and the Intuitiveness Test are compared with MUP.

The major contributions of this paper are as follows:

1. We construct a test collection that contains 6,000 graded preferences collected at both the topic and subtopic levels (50 queries with 3 subtopics per query) on 10 run pairs. We investigate the consistency between the graded user preferences collected at the subtopic level and the preferences at the topic level for a better strategy to collect user preferences efficiently.
2. The correlations (τ_b) between a large number of diversity metrics and the graded user preferences are studied in two dimensions. The one is in terms of performance differences of run pairs detected by the metric and the other is in terms of the differences detected by the user.
3. We propose a preference-weighted correlation, namely Multi-grade User Preference (MUP), to evaluate the diversity metrics based on user preferences. Discussions between MUP and Kendall's τ, Discriminative Power and the Intuitiveness Test are performed in details.

The remainder of this paper is organized as follows. In Section 2, we review related work regarding the ways to evaluate diversity metrics. Section 3 compares the user preferences collected at the subtopic level with the user preferences at the topic level. Next, in Section 4 we compare the correlations between several widely-used diversity metrics and user preferences. Section 5 presents the proposed method to evaluate diversity metrics. In Section 6, we provide our experiments and corresponding analyses. Finally, Section 7 presents our conclusions and directions for future work.

2 Related Work

It is difficult to evaluate a diversity metric because different metrics make different assumptions to simplify the process of diversity evaluation. To present the possible effectiveness of a diversity metric, several methods have been developed.

Sakai *et al.* [15] propose to leverage Discriminative Power [11] to assess the effectiveness of diversity metrics. The method computes a significance test between every pair of the system runs and reports the percentage of pairs that a metric can significantly distinguish at some fixed significance level.

Kendall's τ [9] is another method used to compare different metrics. It is defined as the value proportional to the number of pairwise adjacent swaps needed to convert one ranking into the other ranking. Many previous works related to evaluation leverage it to compare their proposed metrics with other widely-used metrics [4,5].

The Intuitiveness Test [12] is developed by Sakai to compare the intuitiveness of different diversity metrics. In this method, a metric that is simple but can effectively represent the intuitiveness, e.g., the most important property that the diversity metrics should satisfy, is taken as the gold standard metric. The relative intuitiveness of two diversity metrics is measured in terms of preference agreement with its gold standard metric.

Moffat [10] proposes to characterize metrics by seven numeric properties, i.e. boundedness, monotonicity, convergence, top-weightedness, localization, completeness, and realizability. These properties are used to partition the evaluation metrics and help the metrics to be better understood.

Amigó *et al.* [2] propose reliability and sensitivity to compare evaluation metrics. Reliability is the probability of finding the document relationships of a system run in a chosen gold standard run, whereas sensitivity is the probability of finding the document relationships of a chosen gold standard run in a system run.

In general, these methods are lack of consideration about user preferences. After all, the ultimate aim of diversified search is satisfying the diverse information needs of users. In this paper, we highlight the possible effectiveness of user preferences in the evaluation of diversity metrics.

3 User Preferences Discussion

We first select 5 of the 12 runs created by different methods in NTCIR 10 Intent-2 task [13] (Table 1). Each two of the 5 runs are then presented to users in a paralleled way to collect user preference. To decrease the total work of preference collection, we only choose 50 of the 200 queries that contain as fewer (but at least 3) subtopics as possible. This is because in the experiments, only the top 3 subtopics ordered by weight are reserved for every query (for small workloads). We need to possibly decrease the bias of the subtopic reservation. The weights of the three reserved subtopics are then re-normalized by their sum.

3.1 Graded User Preferences Collection

Because in diversified search a query topic is considered to contain several subtopics, we collect user preferences at both the subtopic and the topic levels. We present a subtopic with search results from two different runs to collect user preference at the subtopic level, whereas simultaneously present all the three subtopics underlying a query with the search results to collect user preference at the topic level. The annotator is required to assess his preference by a number between 0 and 4. Because we have selected 50 queries with 150 subtopics, there are 200 presentations for each run pair. Considering 5 runs can generate 10 run pairs and each is confirmed to be presented to 3 annotators in the experiment, 110 annotators participate in this experiment with one person finishing 60 annotations. 10 annotators are filtered because the number of their decisions made within 10 seconds are larger than 30. Therefore, we collect 6,000 user preferences.

3.2 Graded User Preferences Comparisons

User preferences at the subtopic level are collected by presenting one subtopic at a time without giving the subtopic weight. However, the weights of subtopics underlying a query may always differ from each other. To compare the bias, we linearly combine the user preferences at the subtopic level with corresponding subtopic weights to form the weighted preferences at topic level. We can then compare the preferences at the topic level with the weighted preferences and even with the user preferences at the subtopic level.

We average the user preferences and present results in Table 1. The average preferences for each run pair at the subtopic level are computed based on the 150 subtopics, whereas the average weighted preferences and the average preferences at the topic level are based on the 50 queries. Table 1 shows that in our experiment no matter what type of user preferences is considered, the relative orders between each pair are identical. For example, users prefer BASELINE-D-C-1 to THUIR-D-C-1A in terms of all the three types of preferences according to the first row of Table 1.

Table 1 shows that the three types of user preferences have the same assess results for all the run pairs, although the significant results at the subtopic level contain the most items, which completely includes all of the significant results of the other two types of user preferences. This may be caused by a larger number of instances at the subtopic level considered in the significance test. Because user preferences at the subtopic level are collected without giving subtopic weights, it is more reasonable to consider the results of weighted preferences or the use preferences at the topic level. On the other hand, Table 1 shows that both the assess results and the significant results of weighted preferences are similar with the results of user preferences at the topic level. If we only collect user preferences at the topic level, 1,500 rather than 6,000 user preferences need to be collected. The corresponding cost would decrease to one-fourth. In the following sections of this paper, we only consider the user preferences at the topic level.

Table 1. User preference comparison. ">" in column PF indicates the left system is on average better than the right, whereas "<" indicates the right system is better than the left. The checkmarks in the columns SSL, SWS, and STL indicate the significance of user preferences at the subtopic level, the weighted level and the topic level, respectively.

Left	Right	PF	SSL	SWS	STL
BASELINE-D-C-1	THUIR-D-C-1A	<	✓	✓	✓
KECIR-D-C-3B	THUIR-D-C-1A	<	✓	✓	✓
KECIR-D-C-2B	THUIR-D-C-1A	<	✓	✓	✓
THUIR-D-C-1A	THUIR-D-C-4A	>			
BASELINE-D-C-1	THUIR-D-C-4A	<			
KECIR-D-C-3B	THUIR-D-C-4A	<	✓		
KECIR-D-C-2B	THUIR-D-C-4A	<	✓	✓	✓
BASELINE-D-C-1	KECIR-D-C-3B	>	✓		
KECIR-D-C-2B	KECIR-D-C-3B	<	✓	✓	✓
BASELINE-D-C-1	KECIR-D-C-2B	>	✓		✓

4 Correlation between Diversity Metrics and User Preferences

There are many works about methods to evaluate a diversified search result. These works have proposed diversity metrics such as $\alpha-nDCG$, $IA-measures$, $D\#-measures$. These metrics more or less simplify the assumptions about user behaviors, which prevent the metrics from reflecting aspects of the search processes that are experienced by the user. In this section, we take user preferences collected in Section 3 as the ground truth to present the user behavior related properties of diversity metrics in details. We consider a large range of diversity metrics such as $IA-nDCG$ [1], $IA-ERR$ [6], $\alpha-nDCG$ [7], $NRBP$ [8], $I-rec$ [11], $D-nDCG$, $D\#-nDCG$, $D-Q$, $D\#-Q$ [14], $DIN\#-nDCG$, $DIN\#-Q$, $P+Q$, $P+Q\#$ [12].

4.1 Comparing Correlations on Run Pairs with Different Performance

For a certain query, we leverage diversity metrics to evaluate the retrieval results of every two different runs. According to the evaluation score, we can obtains the run preferred by the metric within each pair. On the other hand, we have also collected user preferences on the same pairs. The correlation between the metric and user preferences can then be computed for these pairs. Since there may be a tie in both the evaluation scores and the user preferences, we compute the τ_b coefficient. τ_b is similar with the Kendall's τ [9] except that the former explicitly excludes the influences of the tie in the rankings.

We first demonstrate the changes of correlations from run pairs with small differences to run pairs with large differences by classifying the run pairs into two

bins. The run pairs whose difference in terms of a metric is greater than the average difference of all the pairs are assigned into a large Δ bin and the other pairs are into a small Δ bin. The correlations between metrics and user preferences are computed within each bin, respectively. From another dimension, we also classify the run pairs into different categories according to user preferences. As described in Section 3, we collect user preferences in a graded strategy(between 0 and 4). We equally split this range into 4 different subranges. Within each subrange, we compute the correlations between the metrics and user preferences.

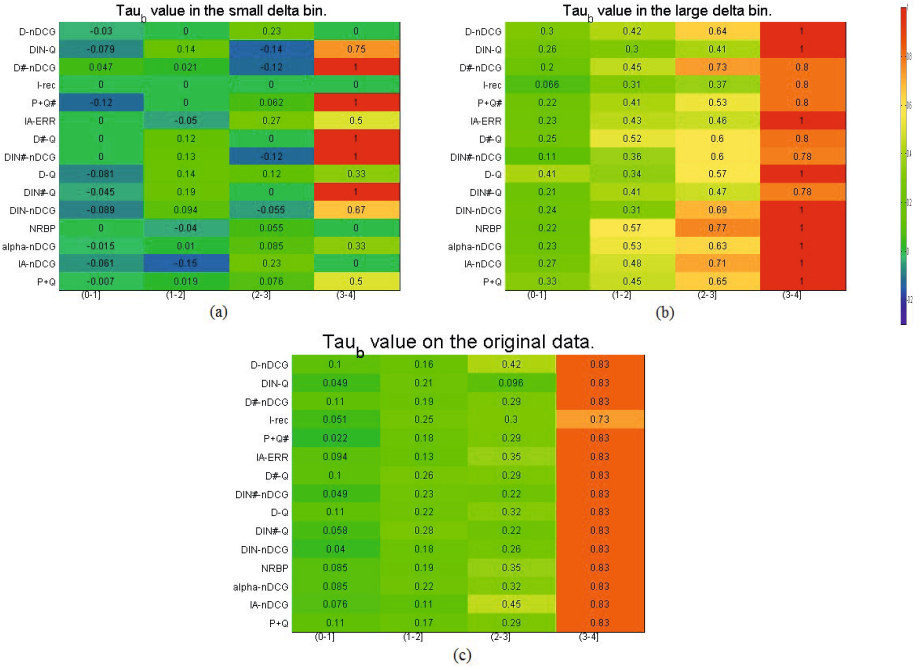

Fig. 1. The τ_b values between diversity metrics and user preferences. The x axis is the subranges of user preferences. Warmer color indicates stronger correlation with user preference. (a) presents the τ_b values computed on run pairs of the small Δ bin. (b) presents the τ_b values computed on run pairs of the large Δ bin. (c) presents the τ_b values on all the run pairs.

Fig. 1 presents all the results. In these heatmaps, a rectangle with color near the red indicates a strong positive correlation, whereas a rectangle with color near the blue indicates a strong negative correlation. From these heatmaps, we can find:

1. Comparing Fig. 1(a) with Fig. 1(b), we can find τ_b values in the large Δ bin are larger than the corresponding τ_b values in the small Δ bin. This means when the differences of two systems detected by metrics become larger, the metrics may agree with user preferences better.

2. From the dimension of user preferences (x axis), we also find that the metrics agree with user preferences better when the differences of run pairs detected by users are larger, whereas it is difficult for diversity metrics to agree with user preferences on the run pairs with small differences.
3. The τ_b values of $DIN\#-Q$, $DIN\#-nDCG$, $D\#-Q$, $P+Q\#$, and $D\#-nDCG$ in subrange (3-4] of the small Δ bin are larger than the corresponding τ_b values in the same subrange of the big Δ bin. By investigating the data, we find these metric only contain one system pair in subrange (3-4] of the small Δ bin, in which case the τ_b value includes much bias. In fact, for all the metrics in the small Δ bin there are few run pairs (on average, 0.9% of the total) in subrange (3-4]. This means if the users think the difference between two runs is small, then the metric is likely to think the difference is small as well.

The discussions above show that when the differences of run pairs are large, the metrics are more likely to achieve agreements with user preferences, whereas the agreements are difficult to achieve when the differences are small. This conclusion keeps true no matter the differences are detected by the metrics or the users. This inspires us to penalize the metric more in evaluating diversity metrics if it disagrees with user preferences on the run pairs with large differences. That is because the metric makes mistakes on run pairs where other metrics seldom make a mistake. The τ_b itself is not aware of this, although we have discussed the metrics based on the τ_b value.

5 Proposed Method to Evaluate Diversity Metrics

We first define some symbols in use. We denote a run pair as c. All of the 500 run pairs mentioned above compose a pair set C. Then, we define an indicator $J(c)$ satisfying $J(c) = 1$ if the metric agrees with the user preference on run pair c, whereas $J(c) = -1$ if the metric disagrees with the user preference. The user preference of c is denoted as u_c (where $0 \leq u_c \leq 4$). We propose the Multi-grade User Preference (MUP) to evaluate diversity metrics as follows:

$$MUP = \frac{sum_{c \in C}(u_c \times J(c))}{sum_{c \in C} u_c} \tag{1}$$

In Formula 1, if the metric agrees with the user preference on a pair c, then both the numerator and the denominator increase by u_c. However, if the metric disagrees with the user preference, then $J(c) = -1$ and the sum of the corresponding $u_c \times J(c)$ in the numerator is indeed equal to subtracting the user preference u_c from the sum. In contrast, the denominator always increases by u_c. This is taken as the penalization of the disagreement. If u_c is larger, MUP punishes the disagreement more. This is meaningful because the experiments and corresponding discussions in Section 4 show that the metrics achieve a better agreement with user preferences when the differences of run pairs are larger, whereas they perform worse on the run pairs with smaller differences. If a metric

makes a mistake (which means the metric disagrees with the user preference) on a run pair whose difference can be easily detected, it should be heavily penalized when we evaluate the metric. However, if the mistake is made on the run pair whose differences are small and difficult to detect, the corresponding punishment may be slight. Especially, MUP does not consider the run pairs with $u_c = 0$.

We can also discuss Formula 1 from the user's perspective. A large user preference means the user considers the difference between the run pair to be large. It is reasonable to consider that the user can, on average, detect a large difference with more confidence than when detecting a small difference. Therefore, if a metric makes a mistake on the run pair whose difference is detected by the user with much confidence, the metric should be penalized more. In contrast, a small u_c indicates a small difference detected by the user. We can also consider that the user is more confused when he is required to decide which one of two similar runs is better. Therefore, a smaller u_c would indicate less user confidence on the user's preference decision. If a metric makes a mistake on the run pair with a small u_c, the penalization may be slight.

5.1 Relationships between MUP and Kendall's τ

The MUP defined in Formula 1 is similar with the τ value. The difference is that MUP leverages the user preference u_c to weight the agreements and the disagreements considered in τ. As we discussed above, this weighted agreements (disagreements) would make MUP to penalize the mistakes more on the run pairs with large differences while to weaken the penalization to the mistakes on the run pairs with small differences.

We also can define MUP_b based on MUP, just like the extension from τ to τ_b.

$$MUP_b = sum_{c\in C}(u_c \times J(c)) \times \frac{1}{\sqrt{sum_{c\in C}(u_c + u_c \times T_0(c))}} \times \frac{1}{\sqrt{sum_{c\in C}(u_c + u_c \times T'_0(c))}} \quad (2)$$

Where $T_0(c)$ is an indicator satisfying $T_0(c) = 1$ when the run pair c is tied only in terms of the metric, otherwise $T_0(c) = 0$. $T'_0(c)$ is a similar indicator satisfying $T'_0(c) = 1$ when the run pair c is tied only in terms of user preferences, otherwise $T'_0(c) = 0$. Note that if $T'_0(c) = 1$, then we obtain $u_c = 0$ according to the definition of $T'_0(c)$. This means $u_c \times T'_0(c)$ is always equal to 0. Formula 2 can then be simplified as:

$$MUP_b = \frac{sum_{c\in C}(u_c \times J(c))}{\sqrt{sum_{c\in C}(u_c \times (1 + T_0(c))) \times sum_{c\in C} u_c}} \quad (3)$$

6 Experimental Comparisons

As discussed in Section 5, MUP (MUP_b) is a weighted τ (τ_b). In this section, we first construct experiments to discuss the consistency of them. As most of

the existing studies in the literature usually leverage Discriminative Power and the Intuitiveness Test to investigate the different aspects of the diversity metric, the discussions about them are also included in the experiments.

We compute the MUP, MUP_b, τ, τ_b, and Discriminative Power values on the selected 500 run pairs, respectively. We leverage the two-tailed paired bootstrap test with 1,000 bootstrap samples [11] for the Discriminative Power. The significance level in use is $\alpha = 0.05$. The results are presented in Fig. 2.

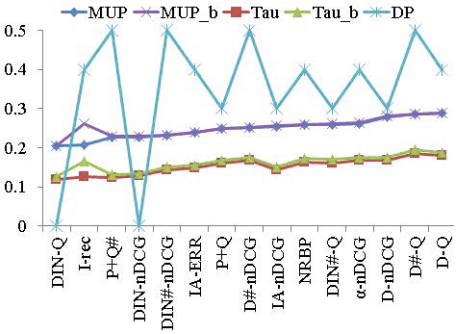

Fig. 2. The values of MUP, MUP_b, τ, τ_b and Discriminative Power. The metrics are ordered by their MUP values in an ascending order.

Fig. 2 shows:

1. D-Q gets higher MUP value than $\alpha-nDCG$, $NRBP$, and $IA-ERR$. This means D-Q simulates the user behaviors better in terms of the weighted correlation. From the definition of D-Q metric, we know that D-Q considers users browser search results from top to bottom until his information need is satisfied. the results indicate this assumption may be useful in the diversity evaluation. After adding other features, such as user information type or subtopic recall, $D\# - Q$, $DIN - Q$ and $DIN\# - Q$ become worse. This can also be observed from $D-nDCG$ to $D\#-nDCG$, $DIN-nDCG$, and $DIN\#-nDCG$.
2. The MUP_b values of most metrics are nearly the same with (indeed, slightly different from) the MUP values of the corresponding metrics, except for the values of $I-rec$. Note that there is an additional factor $\sqrt{1+T_0(c)}$ in the denominator of MUP_b comparing to MUP. According to the definitions of $J(c)$ and $T_0(c)$, the value of $\sqrt{1+T_0(c)}$ is equal to either 1 or 2 (its value equals to 2 when the run pair c is tied only in terms of the metric). The larger MUP_b value of $I-rec$ indicates there are a lot of run pairs on which the score of $I-rec$ is tied, whereas the user preference is not tied. This may result from the fact that $I-rec$ is a set-based metric, and only considers a binary relevance. In contrast, the slight difference between the MUP_b and MUP values of the other metrics indicates that there exist few run pairs which are tied only in terms of the corresponding metric.

3. The τ (τ_b) value decreases from $I-rec$ to $P+Q\#$ while The MUP (MUP_b) value increases, which shows the main difference between MUP (MUP_b) and τ (τ_b). This is caused by the weighted agreement/disagreement in MUP. Fig. 1(c) shows that in the subranges (0-1] and (1-2], the τ_b values of $I-rec$ are larger than the corresponding values of $P+Q\#$ while in the subrange (3-4], the τ_b value of $I-rec$ is smaller than the τ_b value of $P+Q\#$. Considering that MUP leverages the user preferences to weight the agreement/disagreement, the user preferences lying in subrange (3-4] cause the increment of MUP value. Similar reasons can be found in the decrements of $IA-nDCG$ and $D-Q$.
4. The Discriminative Power is not correlative with the MUP or MUP_b. Metrics with large discriminative power may not have large MUP or MUP_b values, whereas metrics with small discriminative power may not indicate small MUP or MUP_b values. This means the aspects evaluating by Discriminative Power may differ from those aspects evaluating by MUP or MUP_b. We only have 10 run pairs to compute the discriminative power here, which may cause some bias in the experiment.

Table 2. Intuitiveness based on preference agreements with the gold standard metric. The number of disagreements is shown in parentheses. We **highlight** the item if the relative order of the corresponding two metrics in this table agrees with their relative order in terms of MUP values.

The gold standard metric: $I-rec$							
Metric	$D-Q$	$D\#-Q$	$DIN\#-Q$	$DIN\#-nDCG$	$P+Q$	$P+Q\#$	$NRBP$
$\alpha-nDCG$	0.936/0.489 (94)	**0.618/0.971** (68)	0.573/0.987 (75)	0.545/1 (77)	**0.933/0.472** (89)	0.480/1.000 (75)	**0.741/0.667** (27)
$D-Q$	-	0.000/1.000 (66)	0.180/1.000 (89)	0.306/1.000 (111)	0.695/0.712 (59)	0.229/1.000 (105)	0.483/0.933 (89)
$D\#-Q$	-	-	0.696/1.000 (23)	0.745/0.979 (47)	**0.989/0.267** (90)	0.600/0.975 (40)	**0.970/0.582** (67)
$DIN\#-Q$	-	-	-	0.808/0.962 (26)	**0.989/0.215** (93)	0.690/0.966 (29)	**0.987/0.553** (76)
$DIN\#-nDCG$	-	-	-	-	0.991/0.287 (108)	0.861/0.972 (36)	0.988/0.537 (82)
$P+Q$	-	-	-	-	-	0.000/1 (80)	**0.500/0.906** (96)
$P+Q\#$	-	-	-	-	-	-	0.988/0.488 (82)
The gold standard metric: $Ef-P$							
$\alpha-nDCG$	**0.287/0.894** (94)	**0.441/0.765** (68)	0.427/0.827 (75)	0.506/0.740 (77)	0.438/0.775 (89)	0.547/0.680 (75)	0.630/0.630 (27)
$D-Q$	-	**0.864/0.333** (66)	**0.787/0.483** (89)	**0.811/0.459** (111)	**0.881/0.434** (59)	**0.848/0.400** (105)	**0.899/0.258** (89)
$D\#-Q$	-	-	0.565/0.913 (23)	**0.745/0.660** (47)	0.578/0.667 (90)	**0.800/0.500** (40)	**0.776/0.448** (67)
$DIN\#-Q$	-	-	-	**0.923/0.462** (26)	0.624/0.624 (93)	**1.000/0.310** (29)	**0.829/0.434** (76)
$DIN\#-nDCG$	-	-	-	-	**0.574/0.685** (108)	**0.833/0.611** (36)	0.732/0.512 (82)
$P+Q$	-	-	-	-	-	**0.713/0.463** (80)	0.760/0.448 (96)
$P+Q\#$	-	-	-	-	-	-	0.671/0.549 (82)

6.1 Comparison between MUP and the Intuitiveness Test

The Intuitiveness Test is proposed by Sakai [12] to quantify the intuitiveness of a metric. It requires a gold standard metric to represent the intuitiveness that the diversity metric should satisfy. Following the work of Sakai, we take $I-rec$ and $Ef-P$ as the gold standard metrics and list the intuitiveness computed

in Table 2. The metrics considered here are distributed in the top, middle and bottom positions of the ranking ordered by the MUP value.

From Table 2, we can find: When $Ef-P$ is taken as the gold standard, the Intuitiveness Test agrees better with MUP than it does when $I-rec$ is taken as the gold standard. The τ_b value between MUP and the Intuitiveness Test in the former case is 0.333, whereas the τ_b value in the latter case is -0.407. Since both $I-rec$ and $Ef-P$ are set-based metrics based on binary relevance, the possible bias of this type may be weakened. Considering that $I-rec$ is the gold standard of the diversity property and $Ef-P$ is the gold standard of the relevance property, the larger τ_b value in the former case indicates in the user's opinion, the relevance of the search result is more important than the diversity of the search result in the diversified search evaluation of our experiments. This result may direct the design of new diversity metrics and help us tune the trade-off parameters between relevance and diversity in diversity metrics such as $D\#-nDCG$. We will do a further research for this in future work.

7 Conclusions and Future Work

It is difficult to evaluate the effectiveness of diversity metrics. Most of the existing studies leverage Discriminative Power, Kendall's τ, or the Intuitiveness Test to evaluate the possible effectiveness of a diversity metric. However, they are lack consideration about the behaviors of user preferences. In this paper, we first collect 6,000 effective user preferences for 500 difference run pairs. A comparison between the weighted user preferences and the user preferences collected at the topic level shows they share similar characters, which means we only need to collect user preferences at the topic level with much less efforts. Then we investigate the correlations between the diversity metrics and user preferences. We find that diversity metrics agree better with user preferences when the difference of a run pair is larger, no matter the difference is detected in terms of the metric or user preferences. Based on these findings, we propose a preference-weighted correlation, namely MUP to evaluate diversity metrics. In the experiments, we first present the effort of the "preference-weighted" correlation by comparing MUP (MUP_b) with τ (τ_b). The results also show that MUP method evaluates diversity metrics from the aspects that may differ from Discriminative Power. In addition, we construct experiments to compare MUP with the Intuitiveness Test and find that when $Ef-P$ is taken as the gold standard of relevance evaluation, the Intuitiveness Test agrees better with MUP than it does when $I-rec$ is taken as the gold standard of diversity evaluation. Since the MUP method evaluates diversity metrics from real users' perspective, then this larger agreement reveals that in the user's opinion, the relevance of the search result is more important than the diversity of the search result in the diversified search evaluation of our experiments. This result may direct the design of new diversity metrics and help us tune the trade-off parameters between relevance and diversity in diversity metrics. In future work, we will base on the conclusions in this paper to develop new diversity metrics. We will also do a further research for tuning trade-off parameters in diversity metrics such as $D\#-nDCG$.

Acknowledgement. This work was supported by Natural Science Foundation (61073071) and a Research Fund FY14-RES-SPONSOR-111 from Microsoft Research Asia.

References

1. Agrawal, R., Gollapudi, S., Halverson, A., Leong, S.: Diversifying search results. In: Proc. of ACM WSDM 2009, pp. 1043–1052. ACM, Barcelona (2009)
2. Amigó, E., Gonzalo, J., Verdejo, F.: A general evaluation measure for document organization tasks. In: Proc. of SIGIR 2013, pp. 643–652. ACM, Ireland (2013)
3. Ashkan, A., Clarke, C.L.A.: On the informativeness of cascade and intent-aware effectiveness measures. In: Proc. of ACM, Hyderabad, India, pp. 407–416 (2011)
4. Aslam, J.A., Pavlu, V., Savell, R.: A unified model for metasearch, pooling, and system evaluation. In: Proc. of ACM CIKM 2003, pp. 484–491. ACM, New Orleans (2003)
5. Buckley, C., Voorhees, E.M.: Retrieval evaluation with incomplete information. In: Proc. of ACM SIGIR 2004, pp. 25–32. ACM, New York (2001)
6. Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: Proc. of ACM CIKM 2009, pp. 621–630. ACM, New York (2009)
7. Clarke, C.L.A., Kolla, M., Cormack, G.V., Vechtomova, O.: Novelty and diversity in information retrieval evaluation. In: Proc. of ACM SIGIR 2008, pp. 659–666. ACM, Singapore (2008)
8. Clarke, C.L.A., Kolla, M., Vechtomova, O.: An effectiveness measure for ambiguous and underspecified queries. In: Azzopardi, L., Kazai, G., Robertson, S., Rüger, S., Shokouhi, M., Song, D., Yilmaz, E. (eds.) ICTIR 2009. LNCS, vol. 5766, pp. 188–199. Springer, Heidelberg (2009)
9. Kendall, M.: A new measure of rank correlation. Biometrica 30, 81–89 (1938)
10. Moffat, A.: Seven numeric properties of effectiveness metrics. In: Banchs, R.E., Silvestri, F., Liu, T.-Y., Zhang, M., Gao, S., Lang, J. (eds.) AIRS 2013. LNCS, vol. 8281, pp. 1–12. Springer, Heidelberg (2013)
11. Sakai, T.: Evaluating evaluation metrics based on the bootstrap. In: Proc. of ACM SIGIR 2006, pp. 525–532. ACM, Seattle (2006)
12. Sakai, T.: Evaluation with informational and navigational intents. In: Proc.s of ACM WWW 2012, pp. 499–508. ACM, Lyon (2012)
13. Sakai, T., Dou, Z., Yamamoto, T., Liu, Y., Zhang, M., Song, R.: Overview of the ntcir-10 intent-2 task. In: Proc. of NTCIR 2010, Tokyo, Japan (2011)
14. Sakai, T., Song, R.: Evaluating diversified search results using per-intent graded relevance. In: Proc. of SIGIR 2011, pp. 1043–1052. ACM, Beijing (2011)
15. Sakai, T., Song, R.: Diversified search evaluation: Lessons from the ntcir-9 intent task. Journal of Information Retrieval 16, 504–529 (2013)
16. Sanderson, M., Paramita, M.L., Clough, P., Kanoulas, E.: Do user preferences and evaluation measures line up? In: Proc. of ACM SIGIR 2010, pp. 555–562. ACM, Geneva (2010)
17. Smucker, M.D., Clarke, C.L.A.: Time-based calibration of effectiveness measures. In: Proc. of ACM SIGIR 2012, pp. 95–104. ACM, Portland (2012)
18. Turpin, A., Scholer, F.: User performance versus precision measures for simple search tasks. In: Proc. of SIGIR 2006, pp. 11–18. ACM, Seattle (2006)
19. Turpin, A.H., Hersh, W.: Why batch and user evaluations do not give the same results. In: Proc. of SIGIR 2001, pp. 225–231. ACM, New Orleans (2001)