Adaptive Effort for Search Evaluation Metrics

Jiepu Jiang and James Allan

Center for Intelligent Information Retrieval
College of Information and Computer Sciences
University of Massachusetts Amherst
jpjiang@cs.umass.edu, allan@cs.umass.edu

Abstract. We explain a wide range of search evaluation metrics as the ratio of users' gain to effort for interacting with a ranked list of results. According to this explanation, many existing metrics measure users' effort as linear to the (expected) number of examined results. This implicitly assumes that users spend the same effort to examine different results. We adapt current metrics to account for different effort on relevant and non-relevant documents. Results show that such adaptive effort metrics better correlate with and predict user perceptions on search quality.

Keywords: evaluation metric; effort; cost; user model; adaptive model

1 Introduction

Searchers wish to find more but spend less. To accurately measure their search experience, we need to consider both the amount of relevant information they found (gain) and the effort they spent (cost). In this paper, we use *effort* and *cost* interchangeably because nowadays using search engines is mostly free of costs other than users' mental and physical effort (e.g., formulating queries, examining result snippets, and reading result web pages). Other costs may become relevant in certain scenarios – e.g., the price charged to search and access information in a paid database – but we only consider users' effort in this paper.

We show that a wide range of existing evaluation metrics can be summarized as some form of gain/effort ratio. These metrics focus on modeling users' gain [6,15] and interaction with a ranked list [2,6,13]—for example, nDCG [6], GAP [15], RBP [13], and ERR [2]. However, they use simple effort models, considering search effort as linear to the (expected) number of examined results. This implicitly assumes that users spend the same effort to examine every result. But evidence suggests that users usually invest greater effort on relevant results than on non-relevant ones, e.g., users are more likely to click on relevant entries [20], and they spend a longer time on relevant results [12].

To better model users' search experience, we adapt these metrics to account for different effort on results with different relevance grades. We examine two approaches: a parametric one that simply employs a parameter for the ratio of effort between relevant and non-relevant entries; and a time-based one that measures effort by the expected time to examine the results, which is similar to

time-biased gain [17]. Both approaches model users' effort adaptively according to the results in the ranked list. We evaluate the adaptive effort metrics by correlating with users' ratings on search quality. Results show that the adaptive effort metrics can better predict users' ratings compared with existing ones.

2 Existing IR Evaluation Metrics

Much previous work [1,17] summarized search evaluation metrics in the form of $\sum_{i=1}^{k} d(i)g(i)$, where $g(i)$ is the ith result's gain, and $d(i)$ is the discount on the ith result. This framework does not explicitly consider users' effort. Instead, we summarize existing metrics as the ratio of users' gain to effort on the ranked list. We categorize these metrics into two groups:

***M_1*: *E*(gain)/*E*(effort)**. M_1 metrics separately measure the expected total gain (E(gain)) and effort (E(effort)) on the ranked list. They evaluate a ranked list by the ratio of E(gain) to E(effort). Existing M_1 metrics are usually implemented as Equation 1: $P_{\text{examine}}(i)$ is the chance to examine the ith result; $g(i)$ and $e(i)$ are the gain and effort to examine the ith result. E(gain) and E(effort) simply sum up the expected gain and effort at each rank, until some cutoff k.

$$M_1 = \frac{E(\text{gain})}{E(\text{effort})} = \frac{\sum_{i=1}^{k} P_{\text{examine}}(i) \cdot g(i)}{\sum_{i=1}^{k} P_{\text{examine}}(i) \cdot e(i)} \tag{1}$$

***M_2*: *E*(gain/effort)**. M_2 metrics measure the expected ratio of gain to effort over different ways that users may interact with a ranked list. This is normally implemented by modeling the chances to stop at each rank when users examine results from top to bottom sequentially. M_2 metrics can be written as Equation 2, where $P_{\text{stop}}(j)$ is the probability to stop after examining the jth result and $\sum_j P_{\text{stop}}(j) = 1$. Users' gain and effort for stopping at rank j, $g_{\text{stop}}(j)$ and $e_{\text{stop}}(j)$, simply sum up the gain and effort for all examined results.

$$M_2 = E(\frac{\text{gain}}{\text{effort}}) = \sum_{j=1}^{k} P_{\text{stop}}(j) \cdot \frac{g_{\text{stop}}(j)}{e_{\text{stop}}(j)} = \sum_{j=1}^{k} P_{\text{stop}}(j) \cdot \frac{\sum_{i=1}^{j} g(i)}{\sum_{i=1}^{j} e(i)} \tag{2}$$

Table 1 lists components for the M_1 and M_2 metrics discussed in this paper. Note that this is only one possible way of explaining these metrics, while other interpretations may also be reasonable. We use $r(i)$ for the relevance of the ith result and $b(i)$ for its binary version, i.e., $b(i) = 1$ if $r(i) > 0$, otherwise $b(i) = 0$.

2.1 Precision, AP, GAP, and RBP

P@k can be considered as an M_1 metric where users always examine the top k results. Each examined result provides $b(i)$ gain, and costs 1 unit effort.

Following Robertson's work [14], we explain average precision (AP) as an M_2 metric in which users stop at each retrieved relevant result with an equal probability $1/N_r$. N_r is the total number of judged relevant results for the topic

Table 1: Components of existing M_1 and M_2 evaluation metrics.

Type	Metric	$P_{\text{examine}}(i)$ or $P_{\text{stop}}(j)$	$g(i)$	$e(i)$
M_1	P@k	1 if $i \leq k$, or 0	$b(i)$	1
	GP@k	1 if $i \leq k$, or 0	$\sum_{s=1}^{r(i)} g_s$	1
	DCG	$1/\log_2(i+1)$	$2^{r(i)} - 1$	$1/\sum_{i=1}^{k} \frac{1}{\log_2(i+1)}$
	RBP	$p^{(i-1)}$	$b(i)$	1
	GRBP	$p^{(i-1)}$	$\sum_{s=1}^{r(i)} g_s$	1
M_2	AP	$b(j)/N_r$	$b(i)$	1
	GAP *	$b(j)/N_r$	$\sum_{s=1}^{r(i)} g_s$	1
	RR	1 if $j = t$, or 0	1 if $i = j$, or 0	1
	ERR	$R(j)\prod_{m=1}^{j-1}(1 - R(m))$	1 if $i = j$, or 0	1

* GAP requires a normalization factor $N_r/E(N_r)$, as in Equation 3.

(or query). Therefore, the stopping probability at the jth result is $P_{\text{stop}}(j) = b(j)/N_r$. AP and P@$k$ share the same $g(i)$ and $e(i)$, as Table 1 shows.

Graded average precision (GAP) [15] generalizes AP to multi-level relevance judgments. It models that users may agree on a relevance threshold s with probability g_s—the probability that users only regard results with relevance grades $\geq s$ as relevant. Thus, the ith result has probability $\sum_{s=1}^{r(i)} g_s$ to be considered as relevant. The ith result's gain equates this probability, and $e(i) = 1$.

Similar to AP, we can explain GAP as an M_2 metric. Users stop at each retrieved relevant result with an equal probability $1/N_r$, regardless of the relevance grade. To obtain the original GAP, we need to further normalize the metric by $N_r/E(N_r)$, where $E(N_r)$ is the expected total number of results that users may consider as relevant, which takes into account the distribution of g_s. Equation 3 describes GAP, where: r_{max} is the highest possible relevance grade; N_m is the number of judged results with the relevance grade m.

$$\text{GAP} = \frac{N_r}{E(N_r)} \cdot \sum_{j=1}^{k} \frac{b(j)}{N_r} \cdot \frac{\sum_{i=1}^{j}\sum_{s=1}^{r(i)} g_s}{\sum_{i=1}^{j} 1},\ E(N_r) = \sum_{m=1}^{r_{max}} N_m \sum_{s=1}^{m} g_s \qquad (3)$$

Rank-biased precision (RBP) [13] models that after examining a result, users have probability p to examine the next result, and $1 - p$ to stop. Users always examine the first result. RBP is an M_1 metric. Users have p^{i-1} probability to examine the ith entry. RBP and P@k have the same gain and effort function. Note that the original RBP computes effort to an infinite rank ($E(\text{effort}) = \frac{1}{1-p}$). Here we measure both gain and effort to some cutoff k ($E(\text{effort}) = \frac{1-p^k}{1-p}$). This results in a slight numerical difference. But the two metrics are equivalent for evaluation purposes because they are proportional when p and k are predefined.

We also extend P@k and RBP to consider graded relevance judgments using the gain function in GAP ($g(i) = \sum_{s=1}^{r(i)} g_s$). We call the extensions graded P@k (GP@k) and graded RBP (GRBP).

2.2 Reciprocal Rank and Expected Reciprocal Rank

Reciprocal rank (RR) is an M_2 metric where users always and only stop at rank t (the rank of the first relevant result).

Expected reciprocal rank (ERR) [2] further models the chances that users stop at different ranks while sequentially examining a ranked list. ERR models that searchers, after examining the ith result, have probability $R(i)$ to stop, and $1 - R(i)$ to examine the next result. Chapelle et al. [2] define $R(i) = \frac{2^{r(i)}-1}{2^{r_{max}}}$. In order to stop at the jth result, users need to first have the chance to examine the jth result (they did not stop after examining results at higher ranks) and then stop after examining the jth result—$P_{\text{stop}}(j) = R(j)\prod_{m=1}^{j-1}(1 - R(m))$.

Both RR and ERR model that users always have 1 unit gain when they stop. They do not have an explicit gain function for individual results, but model stopping probability based on result relevance. To fit them into the M_2 framework, we define, when users stop at rank j, $g(i) = 1$ if $i = j$, otherwise $g(i) = 0$. For both metrics, $e(i) = 1$, such that stopping at rank j costs j unit effort.

$$\text{ERR@}k = \sum_{j=1}^{k} P_{\text{stop}}(j) \cdot \frac{1}{\sum_{i=1}^{j} 1},\ P_{\text{stop}}(j) = R(j)\prod_{m=1}^{j-1}(1 - R(m)) \tag{4}$$

2.3 Discounted Cumulative Gain Metrics

Discounted cumulative gain (DCG) [6] sums up each result's gain in a ranked list, with a discount factor $1/\log_2(i+1)$ on the ith result. It seems that DCG has no effort factor. But we can also consider DCG as a metric where a ranked list of length k always costs the user a constant effort 1. We can rewrite DCG as an M_1 metric as Equation 5. The log discount can be considered as the examination probability. Each examined result costs users $e(i)$ effort, such that $E(\text{effort})$ sums up to 1. $e(i)$ can be considered as a constant because it only depends on k.

$$\text{DCG@}k = \sum_{i=1}^{k} \frac{2^{r(i)} - 1}{\log_2(i+1)} = \frac{\sum_{i=1}^{k} \frac{2^{r(i)}-1}{\log_2(i+1)}}{\sum_{i=1}^{k} \frac{e(i)}{\log_2(i+1)}},\ e(i) = \frac{1}{\sum_{i=1}^{k} \frac{1}{\log_2(i+1)}} \tag{5}$$

The normalized DCG (nDCG) metric [6] is computed as the ratio of DCG to IDCG (the DCG of an ideal ranked list). For both DCG and IDCG, $E(\text{effort})$ equates 1, which can be ignored when computing nDCG. However, $E(\text{effort})$ for DCG and IDCG can be different if we set $e(i)$ adaptively for different results.

3 Adaptive Effort Metrics

3.1 Adaptive Effort Vector

Section 2 explained many current metrics as users' gain/effort ratio with a constant effort on different results. This is oversimplified. Instead, we assign different effort to results with different relevance grades. Let $0, 1, 2, ..., r_{max}$ be the possible relevance grades. We define an effort vector $[e_0, e_1, e_2, ..., e_{r_{max}}]$, where e_r is the effort to examine a result with the relevance grade r.

Table 2: Estimated time to examine results with each relevance grade.

Relevance (r)	t_{summary} [17]	$P_{\text{click}}(r)$	$t_{\text{click}}(r)$	$t(r)$
Non-relevant (0)	4.4 s	0.26	20.6 s	9.8 s
Relevant (1)	4.4 s	0.50	37.1 s	23.0 s
Highly Relevant (2)	4.4 s	0.55	60.3 s	37.6 s

We consider two ways to construct such effort vector in this paper. The first approach is to simply differentiate the effort on relevant and non-relevant results using a parameter $e_{r/nr}$. We set the effort to examine a relevant result to 1 unit. $e_{r/nr}$ is the ratio of effort on a relevant result to a non-relevant one—the effort to examine a non-relevant result is $\frac{1}{e_{r/nr}}$. For example, if we consider three relevance grades ($r = 0, 1, 2$), the effort vector is $[\frac{1}{e_{r/nr}}, 1, 1]$. Here we restrict $e_{r/nr} \geq 1$—relevant results cost more effort than non-relevant ones (because users are more likely to click on relevant results [20] and spend a longer time on them [12]).

The second approach estimates effort based on observed user interaction from search logs. Similar to time-biased gain [17], we measure effort by the amount of time required to examine a result. We assume that, when examining a result, users first examine its summary and make decisions on whether or not to click on its link. If users decide to click on the link, they further spend time reading its content. Equation 6 estimates $t(r)$, the expected time to examine a result with relevance r, where: t_{summary} is the time to examine a result summary; $t_{\text{click}}(r)$ is the time spent on a result with relevance r after opening its link; $P_{\text{click}}(r)$ is the chance to click on a result with relevance r after examining its summary.

$$t(r) = t_{\text{summary}} + P_{\text{click}}(r) \cdot t_{\text{click}}(r) \tag{6}$$

Table 2 shows the estimated time from a user study's search log [9]. We use this log to verify adaptive effort metrics. Details will be introduced in Section 4. The log does not provide t_{summary}. Thus, we use the reported value of t_{summary} in Smucker et al.'s article [17] (4.4 seconds). $P_{\text{click}}(r)$ and $t_{\text{click}}(r)$ are estimated from this log. The search log collected users' eye movement data such that we can estimate $P_{\text{click}}(r)$. The estimated time to examine *Highly Relevant*, *Relevant*, and *Non-relevant* results is 37.6, 23.0, and 9.8 seconds, respectively.

We set the effort to examine a result with the highest relevance grade (2 for this search log) to 1 unit. We set the effort on a result with the relevance grade r to $\frac{t(r)}{t(r_{max})}$. The effort vector for this log is $[9.8/37.6, 23.0/37.6, 1] = [0.26, 0.61, 1]$.

3.2 Computation

The adaptive effort metrics are simply variants of the metrics in Table 1 using the effort vectors introduced in Section 3.1—we replace $e(i)$ by $e(r(i))$, i.e., the effort to examine the ith result only depends on its relevance $r(i)$. For example, let a ranked list of five results have relevance $[0, 0, 1, 2, 0]$. Equation 7 computes adaptive P@k and RR using an effort vector $[\frac{1}{e_{r/nr}}, 1, 1]$.

$$\mathrm{P}_{\text{adaptive}} = \frac{2}{2 + 3 \times \frac{1}{e_{r/nr}}},\ \mathrm{RR}_{\text{adaptive}} = \frac{1}{\frac{1}{e_{r/nr}} + \frac{1}{e_{r/nr}} + 1} \tag{7}$$

When we set different effort to results with different relevance grades, users' effort is not linear to the (expected) number of examined results anymore, but further depends on results' relevance and positions in the ranked list. We look into the same example, and assume an ideal ranked list $[2, 2, 2, 1, 1]$. In such case, DCG has $E(\text{effort}) = \frac{1}{e_{r/nr}} + \frac{1}{e_{r/nr} \log_2 3} + \frac{1}{\log_2 4} + \frac{1}{\log_2 5} + \frac{1}{e_{r/nr} \log_2 6}$, but IDCG has $E(\text{effort}) = 1 + \frac{1}{\log_2 3} + \frac{1}{\log_2 4} + \frac{1}{\log_2 5} + \frac{1}{\log_2 6}$. The effort part is not trivial anymore when we normalize adaptive DCG using adaptive IDCG (adaptive nDCG).

Adaptive effort metrics have a prominent difference with static effort metrics. When we replace a non-relevant result with a relevant one, the gain of the ranked list does not increase for free in adaptive effort metrics. This is because the users' effort on the ranked list also increases (assuming relevant items cost more effort).

Equation 8 rewrites M_1 and M_2 metrics as $\sum_{i=1}^{k} g(i) \cdot d(i)$. It suggests that, when discounting a result's gain, adaptive effort metrics consider users' effort on each result in the ranked list. For M_1 metrics, $d(i) = \frac{P_{\text{examine}}(i)}{E(\text{effort})}$. Increasing the effort at any rank will increase $E(\text{effort})$, and penalize every result in the ranked list by a greater extent. For M_2 metrics, $d(i)$ depends on the effort to stop at rank i and each lower rank. Since $e_{\text{stop}}(j) = \sum_{m=1}^{j} e(m)$, $d(i)$ also depends on users' effort on each result in the ranked list. This makes the rank discounting mechanism in adaptive effort metrics more complex than conventional ones. We leave the analysis of such discounting mechanism for future work.

$$M_1 = \sum_{i=1}^{k} g(i) \cdot \frac{P_{\text{examine}}(i)}{E(\text{effort})},\ M_2 = \sum_{i=1}^{k} g(i) \cdot \sum_{j=i}^{k} \frac{P_{\text{stop}}(j)}{e_{\text{stop}}(j)} \tag{8}$$

3.3 Relation to Time-biased Gain and U-measure

The time-based effort vector looks similar to the time estimation in time-biased gain (TBG) [17]. But we estimate t_{click} based on result relevance, while TBG uses a linear model that depends on document length. We made this choice because the former better correlates with t_{click} in the dataset used for evaluation.

Despite their similarity in time estimation, adaptive effort metrics and TBG are motivated differently. TBG models "the possibility that the user stops at some point by a decay function $D(t)$, which indicates the probability that the user continues until time t" [17]. The longer (the more effort) it takes to reach a result, the less likely that users are persistent enough to examine the result. Thus, we can consider TBG as a metric that models users' examination behavior (P_{examine}) adaptively according to the effort spent *prior to* examining a result.

U-measure [16] is similar to TBG. But the discount function $d(i)$ is dependent on the cumulative length of the texts users read *after* examining the ith result. The more users have read (the more effort users have spent) when they finish examining a result, the less likely the result will be useful. This seems a reasonable heuristic, but it remains unclear what the discounting function models.

In contrast to TBG, adaptive effort metrics retain the original examination models (P_{examine} and P_{stop}) in existing metrics, but further discount the results' gain by the effort spent. The motivation is that for each unit of gain users acquire, we need to account for the cost (effort) to obtain that gain (and assess whether or not it is worthwhile). Comparing to U-measure, our metrics are different in that a result's gain is discounted based on not only what users examined prior to the result and for that result, but also those examined afterwards (as Equation 8 shows). The motivation is that user experience is derived from and measured for searchers' interaction with the ranked list as a whole—assuming a fixed contribution for an examined result regardless of what happened afterwards (such as in TBG and U-measure) seems oversimplified.

Therefore, we believe TBG, U-measure, and the proposed metrics all consider adaptive effort, but from different angles. This leaves room to combine them.

4 Evaluation

We evaluate a metric by how well it correlates with and predicts user perception on search quality. By modeling search effort adaptively, we expect the metrics can better indicate users' search experience. We use data from a user study [9] to examine adaptive effort metrics[1]. The user study asked participants to use search engines to work on some search tasks, and then rate their search experience and judge relevance of results. The dataset only collected user experience in a *search session*, so we must make some assumptions to verify metrics for a single query.

Relevance of results were judged at three levels: *Highly Relevant* (2), *Relevant* (1), or *Non-relevant* (0). Users rated search experience by answering: *how well do you think you performed in this task*? Options are: *very well* (5), *fairly well* (4), *average* (3), *rather badly* (2), and *very badly* (1). Users rated 22 sessions as *very well*, 27 as *fairly well*, 22 as *average*, 7 as *rather badly*, and 2 as *very badly*.

When evaluating a metric, we first use it to score each query in a session. We use the average score of queries as an indicator for the session's performance. We assess the metric by how well the average score of queries in a session correlates with and predicts users' ratings on search quality for that session. This assumes that average quality of queries in a session indicates that session's quality.

We measure correlations using Pearson's r and Spearman's ρ. In addition, we evaluate a metric by how well it predicts user-rated performance. This approach was previously used to evaluate user behavior metrics [8]. For each metric, we fit a linear regression model (with intercept). The dependent variable is user-rated search performance in a session. The independent variable is the average metric score of queries for that session. We measure the prediction performance by normalized root mean square error (NRMSE). We produce 10 random partitions of the dataset, and perform 10-fold cross validation on each partition. This yields prediction results on 100 test folds. We report the mean NRMSE values of the 100 folds and test statistical significance using a two-tail paired t-test.

[1] The dataset and source code for replicating our experiments can be accessed at `https://github.com/jiepujiang/ir_metrics/`

Table 3: Pearson's r and NRMSE for evaluation metrics.

		Pearson's r			NRMSE (smaller is better)					
	Metric	**static**	$\boldsymbol{e_{r/nr}=4}$	**time**	**static**	$\boldsymbol{e_{r/nr}=4}$		**time**		
A	P@k	**0.326**	0.295	0.228	**0.246**	0.249	↑↑	0.253	↑↑↑	**
	AP	**0.065**	0.062	0.054	**0.257**	0.257		0.257		
	RR	0.208	**0.236**	-0.052	0.253	**0.251**		0.256		*
B	GP@k	**0.371**	0.371	0.364	**0.241**	0.241		0.243	↑	
	GAP	**0.062**	0.061	0.055	0.257	**0.257**		0.257		
C	RBP, $p=0.8$	**0.331**	0.324	0.201	**0.245**	0.246		0.253	↑↑↑	***
	RBP, $p=0.6$	0.305	**0.335**	0.154	0.247	**0.245**		0.255	↑↑↑	***
D	GRBP, $p=0.8$	0.405	**0.440**	0.421	0.237	**0.233**	↓	0.236		*
	GRBP, $p=0.6$	0.402	**0.463**	0.444	0.238	**0.230**	↓↓↓	0.233	↓↓	*
	ERR	0.385	**0.427**	0.375	0.240	**0.236**	↓↓	0.242		***
	DCG	0.398	**0.424**	0.418	0.238	**0.235**		0.237		
	nDCG	0.352	0.398	**0.404**	0.243	0.238	↓↓↓	**0.238**	↓↓	
	TBG			**0.440**				**0.234**	† † †	
	U-measure			**0.445**				**0.233**	†	
S	sDCG [7]	0.009			0.258	† † †				
	nsDCG [10]	0.350			0.243	† † †				
	esNDCG [10]	0.355			0.244	†				

Light , medium , and dark shading indicate Pearson's r is significant at 0.05, 0.01, and 0.001 levels, respectively. Arrow indicates NRMSE value is significantly different from **static**. * indicates significant difference between $\boldsymbol{e_{r/nr}=4}$ and **time**. † indicates significant difference comparing to GRBP ($p=0.6$, $e_{r/nr}=4$). One, two, and three symbols indicate $p<0.05$, 0.01, and 0.001, respectively. **Bold font** and underline indicate the best value in its row and column, respectively.

5 Experiment

5.1 Parameters and Settings

For each metric in Table 1, we compare the metric using static effort with two adaptive versions using the parametric or time-based effort vector. We evaluate to a cutoff rank $k=9$, because the dataset shows only 9 results per page.

For GP@k, GAP, and GRBP, we set the distribution of g_s to $P(s=1)=0.4$ and $P(s=2)=0.6$. This parameter yields close to optimal correlations for most metrics. For RBP and GRBP, we examine $p=0.8$ (patient searcher) and $p=0.6$ (less patient searcher). We set $e_{r/nr}$ to 4 (the effort vector is thus $[0.25, 1, 1]$).

We compare with TBG [17] and U-measure [16]. The original TBG predicted time spent using document length. However, in our dataset, we did not find any significant correlation between the two ($r=0.02$). Instead, there is a weak but significant correlation between document relevance and time spent ($r=0.274$, $p<0.001$). We suspect this is because our dataset includes mostly web pages, while Smucker et al. [17] used a news corpus [18]. Web pages include many navigational texts, which makes it difficult to assess the size of the main content.

Thus, when computing TBG, we set document click probability and expected document examine time based on the estimation in Table 2. The dataset does not

provide document save probability. Thus, we set this probability and parameter h by a brute force scan to maximize Pearson's r in the dataset. The save probability is set to $P_{\text{save}}(r = 1) = 0.2$ and $P_{\text{save}}(r = 2) = 0.8$. h is set to 31. Note that this corresponds to a graded-relevance version of TBG well tuned on our dataset.

To be consistent with TBG, we also compute U-measure [16] based on time spent. We set $d(i) = \max(0, 1 - \frac{t(i)}{T})$, where $t(i)$ is the expected total time spent after users examined the ith result, and T is a parameter similar to L in the original U-measure. We set T to maximize Pearson's r. T is set to 99 seconds.

5.2 Results

Table 3 reports Pearson's r and NRMSE for metrics. We group results as follows:

- Block A: metrics that do not consider graded relevance and rank discount.
- Block B: metrics that consider graded relevance, but not rank discount.
- Block C: metrics that consider rank discount, but not graded relevance.
- Block D: metrics that consider both graded relevance and rank discount.
- Block S: session-level metrics (for reference only).

Following Kanoulas et al.'s work [10], we set $b = 2$ and $bq = 4$ in sDCG and nsDCG. For esNDCG, we set parameters to maximize Pearson's r—$P_{\text{down}} = 0.7$ and $P_{\text{reform}} = 0.8$. As results show, for most examined metrics (Blocks A, B, C, and D), their average query scores significantly correlate with users' ratings on search quality in a session. The correlations are similarly strong compared with the session-level metrics (Block S). This verifies that our evaluation approach is reasonable—average query quality in a session does indicate the session's quality.

Adaptive Effort vs. Static Effort As we report in Table 3 (Block D), using a parametric effort vector ($e_{r/nr} = 4$) in GBRP, ERR, DCG, and nDCG can improve metrics' correlations with user-rated performance. The improvements in Pearson's r range from about 0.03 to 0.06. The adaptive metrics with $e_{r/nr} = 4$ also yield lower NRMSEs in predicting users' ratings compared with the static effort ones. The differences are significant except DCG ($p = 0.078$).

Such improvements seem minor, but are in fact a meaningful progress. Block A stands for metrics typically used before 2000. Since 2000, we witness metrics on modeling graded relevance (e.g., nDCG, GAP, and ERR) and rank discount (e.g., nDCG, RBP, and ERR). These work improve Pearson's r from 0.326 (P@k, the best "static" in Block A) to 0.405 (GRBP, $p = 0.8$, the best "static" in Blocks B, C, and D). The proposed adaptive effort metrics further improve Pearson's r from 0.405 to 0.463 (GRBP, $p = 0.6$, the best in the table). The magnitude of improvements in correlating with user-rated performance, as examined in our dataset, are comparable to those achieved by modeling graded relevance and rank discount. We can draw similar conclusions by looking at NRMSE. Although it requires further verification using larger datasets and query-level user ratings, our results at least suggest that the improvements are not negligible.

The best performing metric in our evaluation is adaptive GRBP ($p = 0.6$) with $e_{r/nr} = 4$. It also outperforms well-tuned TBG and U-measure. All these

metrics consider adaptive search effort, but from different angles. The adaptive GRBP metric shows stronger Pearson's r with user-rated performance, and yields significantly lower NRMSE than TBG ($p < 0.001$) and U-measure ($p < 0.05$).

The preferable results of adaptive effort metrics confirms that users do not only care about how much relevant information they found, but are also concerned with the amount of effort they spent during search. By modeling search effort on relevant and non-relevant results, we can better measure users' search effort, which is the key to the improvements in correlating with user experience.

Parametric Effort Vector vs. Time-based One Comparing the two ways of constructing effort vectors, results suggest that the time-based effort vector is not as good as the simple parametric one ($e_{r/nr} = 4$). Compared with the time-based effort vector, metrics using the parametric effort vector almost always yield stronger correlations and lower NRMSE in prediction. Compared with metrics using static effort, the time-based effort vector can still improve GRBP, DCG, and nDCG's correlations, but it fails to help ERR (Table 3, Block D). In addition, it can only significantly reduce NRMSE for GRBP ($p = 0.6$) and nDCG.

We suspect this is because time is only one aspect of measuring search effort. It has an advantage—we can easily measure time-based effort from search logs—but it does not take into account other effort such as making decisions and cognitive burden. Whereas other types of effort are also difficult to determine and costly to measure. Thus, it seems more feasible to tune parameters in the effort vector to maximize correlations with user-rated performance (or minimize prediction errors), if such data is available. In the presented results, we only differentiate the effort on relevant and non-relevant results. We also experimented assigning different effort to each relevance grade. But this yields not much better performance, probably due to the limited size of the dataset (80 sessions).

Adaptive Effort, Graded Relevance, and Rank Discount Despite the success of adaptive effort in Block D, we did not observe consistent improvements in Blocks A, B, and C. The time-based effort vector even shows significantly worse prediction of user-rated performance for P@k, GP@k, and RBP. This suggests that adaptive effort needs to be applied together with graded relevance and rank discount. We suspect this is because: 1) when we apply adaptive effort vector to binary relevance metrics, they may prefer marginally relevant results over highly relevant ones in evaluation (for example, when using the time-based effort vector), which is problematic; 2) users are probably more concerned with search effort on top-ranked results, such that rank discount helps adaptive effort metrics. Although results in Blocks A, B, and C are negative, this is not a critical issue because all recent metrics (Block D) consider both factors. For these metrics, applying the adaptive effort vectors is consistently helpful.

Parameter Sensitivity Figure 1 plots metrics' correlations when $e_{r/nr}$ varies from 1 to 10. $e_{r/nr} = 1$ (the leftmost points) stands for metrics using a static effort vector. The figure shows that when we increase the value of $e_{r/nr}$ (decrease the effort on non-relevant results, $1/e_{r/nr}$), all these metrics consistently

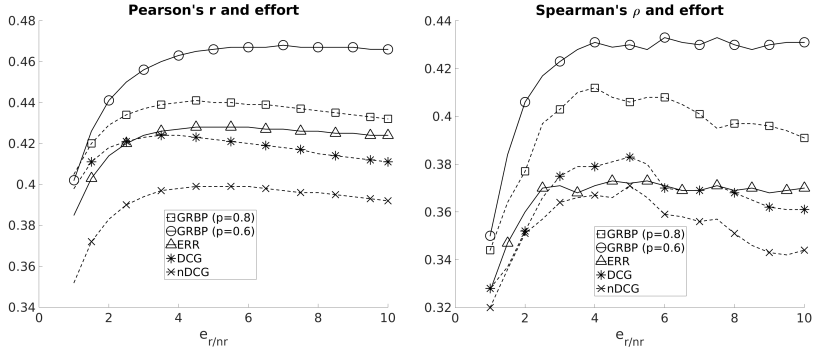

Fig. 1: Sensitivity of Pearson's r and Spearman's ρ to $e_{r/nr}$.

achieve better correlations with user-rated performance. The adaptive effort metrics achieve near-to-optimal correlations with user-rated performance when we set $e_{r/nr}$ to about 3 to 5. In addition, some metrics seem quite stable when we set $e_{r/nr}$ to values greater than 5, such as GRBP ($p = 0.6$) and ERR.

6 Discussion and Conclusion

Effort-oriented evaluation starts from expected search length [3], which measures the number of examined results to find a certain amount of relevant information. Dunlop [5] extended the metric to measure the expected time required to find a certain amount of relevant results. Kazai et al. [11] proposed effort-precision, the ratio of effort (the number of examined results) to find the same amount of relevant information in the ranked list compared with in an ideal list. But these works all assume that examining different results involves the same effort.

This paper presents a study on search effectiveness metrics using adaptive effort components. Previous work on this topic is limited. TBG [17] considered adaptive effort, but applies it to the discount function. U-measure [16] is similar to TBG, and possesses the flexibility of handling SERP elements other than documents (e.g., snippets, direct answers). De Vries et al. [4] modeled searchers' tolerance to effort spent on non-relevant information before stopping viewing an item. Villa et al. [19] found that relevant results cost assessors more effort to judge than highly relevant and non-relevant ones. Yilmaz et al. [21] examined differences between searchers' effort (dwell time) and assessors' effort (judging time) on results, and features predicting such effort. Our study shows that the adaptive effort metrics can better indicate users' search experience compared with conventional ones (with static effort).

The dataset for these experiences was based on session-level user ratings and required that we make assumptions to verify query-level metrics. One important area of future research is to extend this study to a broader set of queries of different types to better understand the applicability of this research. Another direction for future research is to explore the effect of different effort levels, for example, assigning different effort for *Relevant* and *Highly Relevant* results rather than just to distinguish relevant from non-relevant.

Acknowledgment This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant #IIS-0910884. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

1. Carterette, B.: System effectiveness, user models, and user utility: A conceptual framework for investigation. In: SIGIR '11. pp. 903–912 (2011)
2. Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: CIKM '09. pp. 621–630 (2009)
3. Cooper, W.S.: Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. American Documentation 19(1), 30–41 (1968)
4. De Vries, A.P., Kazai, G., Lalmas, M.: Tolerance to irrelevance: A user-effort oriented evaluation of retrieval systems without predefined retrieval unit. In: RIAO 2004. pp. 463–473 (2004)
5. Dunlop, M.D.: Time, relevance and interaction modelling for information retrieval. In: SIGIR '97. pp. 206–213 (1997)
6. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. 20(4), 422–446 (2002)
7. Järvelin, K., Price, S.L., Delcambre, L.M.L., Nielsen, M.L.: Discounted cumulated gain based evaluation of multiple-query IR sessions. In: ECIR'08. pp. 4–15 (2008)
8. Jiang, J., Hassan Awadallah, A., Shi, X., White, R.W.: Understanding and predicting graded search satisfaction. In: WSDM '15. pp. 57–66 (2015)
9. Jiang, J., He, D., Allan, J.: Searching, browsing, and clicking in a search session: Changes in user behavior by task and over time. In: SIGIR '14. pp. 607–616 (2014)
10. Kanoulas, E., Carterette, B., Clough, P.D., Sanderson, M.: Evaluating multi-query sessions. In: SIGIR '11. pp. 1053–1062 (2011)
11. Kazai, G., Lalmas, M.: extended cumulated gain measures for the evaluation of content-oriented xml retrieval. ACM Trans. Inf. Syst. 24(4), 503–542 (2006)
12. Kelly, D., Belkin, N.J.: Display time as implicit feedback: Understanding task effects. In: SIGIR '04. pp. 377–384 (2004)
13. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. Inf. Syst. 27(1), 2:1–2:27 (2008)
14. Robertson, S.E.: A new interpretation of average precision. In: SIGIR '08. pp. 689–690 (2008)
15. Robertson, S.E., Kanoulas, E., Yilmaz, E.: Extending average precision to graded relevance judgments. In: SIGIR '10. pp. 603–610 (2010)
16. Sakai, T., Dou, Z.: Summaries, ranked retrieval and sessions: A unified framework for information access evaluation. In: SIGIR '13. pp. 473–482 (2013)
17. Smucker, M.D., Clarke, C.L.: Time-based calibration of effectiveness measures. In: SIGIR '12. pp. 95–104 (2012)
18. Smucker, M.D., Jethani, C.P.: Human performance and retrieval precision revisited. In: SIGIR '10. pp. 595–602 (2010)
19. Villa, R., Halvey, M.: Is relevance hard work?: Evaluating the effort of making relevant assessments. In: SIGIR '13. pp. 765–768 (2013)
20. Yilmaz, E., Shokouhi, M., Craswell, N., Robertson, S.E.: Expected browsing utility for web search evaluation. In: CIKM '10. pp. 1561–1564 (2010)
21. Yilmaz, E., Verma, M., Craswell, N., Radlinski, F., Bailey, P.: Relevance and effort: An analysis of document utility. In: CIKM '14. pp. 91–100 (2014)