Meta-Evaluation of Dynamic Search: How Do Metrics Capture Topical Relevance, Diversity and User Effort?

Ameer Albahem, Damiano Spina, Falk Scholer, and Lawrence Cavedon

RMIT University, Melbourne, Australia
{firstname.lastname}@rmit.edu.au

Abstract. Complex dynamic search tasks typically involve multi-aspect information needs and repeated interactions with an information retrieval system. Various metrics have been proposed to evaluate dynamic search systems, including the Cube Test, Expected Utility, and Session Discounted Cumulative Gain. While these complex metrics attempt to measure overall system "goodness" based on a combination of dimensions – such as topical relevance, novelty, or user effort – it remains an open question how well each of the competing evaluation dimensions is reflected in the final score. To investigate this, we adapt two meta-analysis frameworks: the Intuitiveness Test and Metric Unanimity. This study is the first to apply these frameworks to the analysis of dynamic search metrics and also to study how well these two approaches agree with each other. Our analysis shows that the complex metrics differ markedly in the extent to which they reflect these dimensions, and also demonstrates that the behaviors of the metrics change as a session progresses. Finally, our investigation of the two meta-analysis frameworks demonstrates a high level of agreement between the two approaches. Our findings can help to inform the choice and design of appropriate metrics for the evaluation of dynamic search systems.

Keywords: Evaluation, Dynamic Search, Intuitiveness Test, Metric Unanimity

1 Introduction

In many search scenarios, users interact with search systems multiple times to find documents relevant to a complex information need. During the search session, users might submit multiple queries [5], paginate [12] or provide fine-grained relevance feedback [23]. In recent years, *dynamic search* systems that can learn from user feedback and adapt subsequent results have been developed [5, 12, 23]. As a result of the complex actions and processes that may be part of dynamic search, different dimensions such as topical relevance, novelty and the amount of user effort can all play a role in the overall user satisfaction with search results [11]. To evaluate such systems, several information retrieval effectiveness metrics have been proposed that attempt to model some or all of these dimensions in their formulations. As these metrics address multiple (and sometimes

competing) aspects of performance, it is unclear whether they overall behave as intended, by rewarding relevant and novel documents while minimizing user effort.

In this work, we methodically study which dynamic search metrics are better able to capture different dimensions of effectiveness. Since dynamic search includes multiple iterations of interactions between a user and a dynamic search system, we also investigate how the length of a search session impacts on the ability of complex metrics to model these dimensions.

Recently, the TREC Dynamic Domain (DD) Track [21–23] and the CLEF Dynamic Search Lab [13] adopted metrics such as the Cube Test (CT) [14] and Session Discounted Cumulative Gain (sDCG) to evaluate dynamic search in an interactive setup where systems are expected to learn from user feedback and adapt their outputs dynamically. Luo et al. [14] compared these metrics with other widely used effectiveness measures by considering their discriminative power – the ability of metrics to detect statistically significant differences in the retrieval results of different systems. Specifically, CT was compared with diversity metrics (Alpha Normalized Discounted Cumulative Gain (α-nDCG) [9] and Intent-Aware Expected Reciprocal Rank (ERR-IA) [6]) and Time-Biased Gain [18]. However, this approach does not inform researchers about the behaviors of the metrics in terms of measuring the key effectiveness dimensions.

There are different approaches to study the behavior of evaluation metrics. One of such approaches consists of analyzing how closely a metric matches the behavior or preferences of users across a set of search systems [20]. Another option is axiomatic analysis, which consists of defining formal properties that metrics may or may not satisfy [1–4, 10, 15]. A complementary proposal is the use of statistical analysis over metric scores – this is the approach followed in this work.

To study complex dynamic search metrics and how they reflect effectiveness dimensions, we apply two meta-analysis frameworks: the Intuitiveness Test proposed by Sakai [16], and the Metric Unanimity framework proposed by Amigó et al. [3]. The *Intuitiveness Test* measures the extent to which complex metrics are able to capture key properties that are important to measure in a search task evaluation. In previous work, the test was used to analyze the intuitiveness of a range of diversity metrics such as α-nDCG and ERR-IA over a range of tasks including search result diversification [16, 17] and aggregated search [24]. Chuklin et al. [7] applied the test to evaluate the intuitiveness of click models in aggregated search.

Metric Unanimity [3] relies on the intuition that, if a system is superior to another system in every key property of an evaluation, then this should be unanimously reflected by metrics that measure these properties. Amigó et al. [3] analyzed the Metric Unanimity of a wide range of relevance and diversity metrics. However, they did not study the impact of session length on metric behavior, which is one of the key dimensions in dynamic search evaluation.

Our work provides the following key contributions:

- An evaluation of effectiveness metrics for dynamic search,[1] a domain where new complex metrics have been proposed, but whose ability to reflect different dimensions is not clear;
- Proposing a new dimension – *User Effort* – which has not been investigated in prior research into metric intuitiveness and unanimity;
- Investigating the agreement between the two meta-analysis frameworks (Intuitiveness Test and Metric Unanimity) that have previously been used to study evaluation metrics separately.

The results of our analysis show that the Normlaized Cube Test [19] (nCT) is generally more intuitive than other metrics when measuring all dimensions simultaneously. However, dynamic search metrics and diversity metrics can provide complementary information, and should both be reported to provide better insights into search effectiveness. The results also shed light on how metric behavior varies with search session length, where metrics tend to agree with each other more at earlier iterations. Lastly, both frameworks show a high level of agreement.

2 Methodology

2.1 Multidimensional Intuitiveness Test

The Intuitiveness Test proposed by Sakai [16] measures the ability of complex metrics to capture different dimensions of search task effectiveness evaluation. For instance, the test can be used to measure the extent to which diversity metrics such as α-nDCG or ERR-IA can intuitively capture relevance, or coverage of subtopics, two potentially competing dimensions of system effectiveness for search result diversification. In this framework, evaluation metrics are divided into *simple* and *complex* metrics: the former model a single dimension of retrieval effectiveness (for instance, Topical Relevance), while the latter incorporate two or more dimensions. In our work, we adapt the Intuitiveness Test [16] to evaluate how *intuitive* dynamic search metrics are in measuring multiple effectiveness dimensions simultaneously.

Algorithm 1 describes the process for comparing two complex metrics, M_{c1} and M_{c2}, given a set of simple metrics $M_s \in \mathcal{M}_S$ where each embodies a particular effectiveness dimension. In the algorithm, $M_x(t, r)$ denotes the final effectiveness score that the complex metric M_x assigned for the output of a run r (typically a ranked list of retrieved documents) produced by a given system in response to a search topic t. Where the complex metrics M_{c1} and M_{c2} disagree with each other regarding which run is more effective (line 12), the simple metrics are used to judge which of the complex metrics is more closely aligned with the simple metrics $\mathcal{M}_S$, and therefore more strongly embodies the effectiveness dimension that each simple metric M_s in $\mathcal{M}_S$ represents. A key assumption underlying this test is that the chosen simple metrics appropriately represent

[1] Code is available at https://github.com/aalbahem/ir-eval-meta-analysis.

1 **Input:** Complex Metrics M_{c1} and M_{c2}; Simple Metrics $\mathcal{M}_S$; Pairs of runs $\langle r_1, r_2 \rangle \in \mathcal{R}$; Set of topics $t \in \mathcal{T}$;
2 **Output:** Intuitiveness of M_{c1} and M_{c2};
3 Disagreements $= 0$; Correct$_1 = 0$; Correct$_2 = 0$;
4 **foreach** *pair of runs* $\langle r_1, r_2 \rangle \in \mathcal{R}$ **do**
5 **foreach** *topic* $t \in \mathcal{T}$ **do**
6 $\delta M_{c1} = M_{c1}(t, r_1) - M_{c1}(t, r_2)$;
7 $\delta M_{c2} = M_{c2}(t, r_1) - M_{c2}(t, r_2)$;
8 $\delta \mathcal{M}_S = \{\}$;
9 **foreach** M_s *in* $\mathcal{M}_S$ **do**
10 $\delta \mathcal{M}_S$.add$(M_s(t, r_1) - M_s(t, r_2))$;
11 **end**
12 **if** $\delta M_{c1} \times \delta M_{c2} < 0$ **then**
13 Disagreements++;
14 **if** $\forall \delta M_s \in \delta \mathcal{M}_S,\ \delta M_{c1} \times \delta M_s > 0$ **then**
15 Correct$_1$++;
16 **end**
17 **if** $\forall \delta M_s \in \delta \mathcal{M}_S,\ \delta M_{c2} \times \delta M_s > 0$ **then**
18 Correct$_2$++;
19 **end**
20 **end**
21 **end**
22 **end**
23 $Intuitiveness(M_{c1}|M_{c2}, \mathcal{M}_S) =$ Correct$_1$/Disagreements;
24 $Intuitiveness(M_{c2}|M_1, \mathcal{M}_S) =$ Correct$_2$/Disagreements;

Algorithm 1: Multi-dimensional Intuitiveness Calculation, based on Sakai [16].

the effectiveness dimensions to be considered; the choice of the simple metrics is therefore crucial to the analysis. Note that the focus of the test is only on the cases where complex metrics disagree with each other, since complex metrics generally correlate with each other [8, 16].

Algorithm 1 differs from the original version described by Sakai [16] in two aspects. First, we amended the original algorithm to support comparing complex metrics based on two or more simple metrics.[2] In lines 14 and 17, an agreement occurs if the complex metric agrees with *all* of the simple metrics in the set $\mathcal{M}_S$. Second, we refine the condition for agreement between a simple and a complex metric, such that this is only met when both metrics have the same preferences for the pair of runs under consideration (lines 14 and 17); for example, given a topic, both the simple and complex metrics prefer run r_1, or they both prefer run r_2. We discard the cases where the simple metric gives both runs the same scores (tie), i.e. $\delta M_s = 0$. The original paper does not discard these cases when reporting results; however, in practice they are discarded when evaluating statistical significance using a sign test. In our experiments, we found the number of ties

[2] Sakai [16] evaluated metrics considering diversity and relevance simultaneously, but the procedure was not detailed.

Input: A complex metric M_c; Set of simple Metrics $\mathcal{M}_\mathcal{S}$; Pairs of runs $\langle r_1, r_2 \rangle \in \mathcal{R}$; Set of topics $t \in \mathcal{T}$;
Output: Metric Unanimity (MU) of M_c with the set of metrics $\mathcal{M}_\mathcal{S}$
$\Delta m_{i,j} = 0$ $\Delta \mathcal{M}_{\mathcal{S}i,j} = 0$ $\Delta m\mathcal{M}_{\mathcal{S}i,j} = 0$
foreach *pair of runs* $\langle r_1, r_2 \rangle \in \mathcal{R}$ **do**
  foreach *topic* $t \in \mathcal{T}$ **do**
    if $M_c(t, r_1) == M_c(t, r_2)$ **then**
      $\Delta m_{i,j} += 0.5$
    end
    else
      $\Delta m_{i,j} += 1$
    end
    if $(\forall m \in \mathcal{M}_\mathcal{S},\ m(t, r_1) \geq m(t, r_2))$ *Or* $(\forall m \in \mathcal{M}_\mathcal{S},\ m(t, r_1) \leq m(t, r_2))$ **then**
      $\Delta \mathcal{M}_{\mathcal{S}i,j}$++;
    end
    if $(\forall m \in \mathcal{M}_\mathcal{S} \cup \{M_c\},\ m(t, r_1) \geq m(t, r_2))$ *Or* $(\forall m \in \mathcal{M}_\mathcal{S} \cup \{M_c\},\ m(t, r_1) \leq m(t, r_2))$ **then**
      $\Delta m\mathcal{M}_{\mathcal{S}i,j}$++;
    end
  end
end
$\mathrm{MU}(M_c, \mathcal{M}_\mathcal{S}) = \mathrm{PMI}(\Delta m_{i,j}, \Delta \mathcal{M}_{\mathcal{S}i,j})$
$\mathrm{PMI}\left(\Delta m_{i,j}, \Delta \mathcal{M}_{\mathcal{S}i,j}\right) = \log\left(\frac{P(\Delta m_{i,j}, \Delta \mathcal{M}_{\mathcal{S}i,j})}{P(\Delta m_{i,j} \times P(\Delta \mathcal{M}_{\mathcal{S}i,j})}\right) = \log\left(\frac{\frac{\Delta m\mathcal{M}_\mathcal{S}}{|\mathcal{R}|}}{\frac{\Delta m_{i,j}}{|\mathcal{R}|} \times \frac{\Delta \mathcal{M}_{\mathcal{S}i,j}}{|\mathcal{R}|}}\right)$

Algorithm 2: Metric Unanimity calculation based on Amigó et al. [3]

for simple metrics are high, which could obfuscate the actual trends regarding which complex metric has greater intuitiveness.

When conducting the sign test for our experimental results, we use the number of times a complex metric agrees with the simple metrics as the number of successes, i.e., the final $Correct_1$ for M_{c1} and $Correct_2$ for M_{c2}; the number of trials is $Correct_1 + Correct_2$; and the hypothesized probability of success is 0.5.

2.2 Metric Unanimity

Amigó et al. [3] define Metric Unanimity as the Point-wise Mutual Information between improvement decisions of a metric M and improvements captured simultaneously by a set of metrics $\mathcal{M}'$. In their work, Amigó et al. [3] represent $\mathcal{M}'$ by a set of various metrics such as ad hoc and diversity metrics, which is a mixture of simple and complex metrics. In this study, which investigates to what extent complex metrics embody different effectiveness dimensions, we instead instantiate M to be a complex metric M_c, and $\mathcal{M}'$ to be a set of one or more simple metrics, as described below.

Let $\Delta m_{i,j}$ denote the event that a complex metric M_c captures improvements between a pair of systems (r_1, r_2); similarly, let $\Delta \mathcal{M}_{\mathcal{S}i,j}$ denote the event that

all metrics in $\mathcal{M}_S$ simultaneously capture the improvements between system pairs. Then we calculate Metric Unanimity (MU) of M_c using Algorithm 2. Line 21 shows how the point-wise mutual information between the two events is calculated. A higher value of MU for a complex metric implies that it more effectively measures the individual dimensions of interest.

2.3 Simple Metrics

Sakai [16] suggested two simple (gold) metrics to measure the intuitiveness of diversification metrics: (i) Precision (prec) to measure the ability of metrics to capture topical relevance; and (ii) Subtopic Recall[3] (st-rec) to measure the ability of metrics to capture diversity.

In addition to these simple metrics that measure topical relevance and diversity, we study a new dimension not previously explored: *user effort*. As a first attempt, we propose three simple metrics to define user effort in terms of the time spent by users when inspecting the ranking:

- Reciprocal Iteration (r-it), 1/number of iterations.
- Negative Iteration (neg-it), $-1 \times$ number of iterations.
- Total Time (tot-time), computed based on the user model presented as part of the Time-Biased Gain metric [18]: $\sum_{d \in D} 4.4 + r_d \times (0.018 l_d + 7.8)$, where $r_d = 0.64$ if d is a relevant document in the ranking D, otherwise it is 0.39 and l_d is the number of words in d. We use the document length statistics provided by the TREC Dynamic Domain 2016 track to compute this.

The first two approaches, r-it and neg-it, are straightforward estimates based simply on the number of iterations in which a user interacts with a dynamic search system; both assume that the amount of user effort is the same for all documents. The third approach, tot-time, uses richer information and instead estimates user effort based on parameters that include the time to scan search result snippets, and read document content. The r-it approach mirrors the CT metric effort estimation used in the TREC Dynamic Domain track [22]; we investigate neg-it to test whether using the same information differently changes metric behaviors. Note that *user effort* could also be measured by considering other factors such as cognitive effort; in this work, we primarily represent user effort by simple metrics based on the number of iterations or time spent in dynamic search tasks, and leave other factors for future research.

In our experiments, we use different combinations of the simple metrics that reflect Topical Relevance, Diversity, and User Effort, to define $\mathcal{M}_S$. The different sets of simple metrics are then used to measure the intuitiveness and unanimity of complex metrics proposed in the literature to evaluate dynamic search and diversity tasks, which are described below.

[3] Also known as Intent Recall [16].

2.4 Complex Metrics

In this work, we study metrics used in dynamic search evaluation campaigns such as the TREC Dynamic Domain Track [21–23] and the CLEF Dynamic Search Lab [13]. In particular, we consider the Average Cube Test (ACT) [22] and the normalized version of CT(nCT) [14], Expected Utility (nEU), and Session Discounted Cumulative Gain (nsDCG) [19], which capture three key dimensions of dynamic search: Topical Relevance, Diversity and User Effort. We also study the Rank-Biased Utility (RBU) metric, recently introduced by Amigó et al. [3], which models these three dimensions of dynamic search, and was designed by incorporating ideas from different ad hoc and diversity metrics.

Since Diversity – supporting the retrieval of documents for different subtopics of a complex task – is a key dimension of dynamic search systems, we also study two well-known metrics for search result diversification: Alpha Normalized Discounted Cumulative Gain (α-nDCG) and Intent-Aware Expected Reciprocal Rank (nERR-IA), calculated using collection-dependent normalization as described by Clarke et al. [8]. These metrics are also complex, in that they combine two dimensions: Topical Relevance and Diversity. Therefore, comparing them with the dynamic search metrics may provide additional insights in terms of the relation between dynamic search metrics and the Diversity and Topical Relevance dimensions.

2.5 TREC Dynamic Domain Collections

The TREC Dynamic Domain (DD) track ran for three years, from 2015 to 2017 [21–23]. This track models an interactive search setup, where systems receive aspect-level feedback repeatedly and need to dynamically find relevant and novel documents for the query subtopics using the least possible number of iterations.

We make use of these collections and conduct our analysis of complex metrics based on the formal runs submitted to these tracks. While the specifics of the tracks differed slightly across the years (e.g. the number of search topics were 118, 53 and 60 in 2015, 2016 and 2017, respectively) the search tasks being modelled remained consistent, and we therefore conduct our analysis of metric behaviour both for specific instances of the track, as well as aggregating across all three years.

A total of 32, 21 and 11 runs were submitted to TREC DD 2015, 2016 and 2017, respectively. The runs include different diversification algorithms, relevance feedback methods, and retrieval models. In evaluating the runs, we used the official track evaluation script. We also evaluated these runs using the standard web diversity metrics implemented by the Web Diversity evaluation script `ndeval`. As TREC DD considers passage-level relevance judgments, we generate the document-level relevance judgments by summing up the respected subtopic-passage judgments, the same approach that is followed by the official TREC DD evaluation script. For the TREC DD evaluation, the participating runs were required to return 5 documents per iteration, thus to evaluate a standard web

Table 1. Results of the multidimensional intuitiveness test for the TREC DD 2016 data. For each pair of metrics, the intuitiveness scores for (metric in row)/(metric in column) are reported; the fraction of disagreements is shown in parenthesis (the ratio of disagreements to the total number of cases). $^+$ indicates statistically significant differences according to sign test with $p < 0.05$.

		Topical Relevance (Rel, prec) and Diversity (Div, st-rec) and User Effort (Eff, tot-time)					
it		nCT	nEU	nsDCG	α-nDCG	nERR-IA	RBU
1	ACT	0.0084/0.0823$^+$ (4.27%)	0.0215/0.0135 (14.65%)	0.0128/0.0171 (4.23%)	0.0044/0.0558$^+$ (6.14%)	0.0212/0.0466$^+$ (6.38%)	0.0077/0.0694$^+$ (5.84%)
1	nCT	-	0.0307$^+$/0.0028 (15.86%)	0.0684$^+$/0.0143 (5.67%)	0.0284/0.0269 (6.04%)	0.0440$^+$/0.0214 (7.18%)	0.0049/0.0245 (1.84%)
1	nEU	-	-	0.0118/0.0206 (15.34%)	0.0056/0.0302$^+$ (17.63%)	0.0105/0.0268$^+$ (17.18%)	0.0031/0.0303$^+$ (17.53%)
1	nsDCG	-	-	-	0.0042/0.0506$^+$ (6.42%)	0.0192/0.0410$^+$ (6.59%)	0.0130/0.0746$^+$ (5.56%)
1	α-nDCG	-	-	-	-	0.1126$^+$/0.0000 (1.36%)	0.0195/0.0265 (6.47%)
1	nERR-IA	-	-	-	-	-	0.0155/0.0418$^+$ (7.55%)
10	ACT	0.0019/0.0405$^+$ (14.47%)	0.0315$^+$/0.0235 (36.90%)	0.0357/0.0279 (24.23%)	0.0255/0.0330 (20.51%)	0.0473$^+$/0.0193 (18.68%)	0.0173/0.0192 (29.64%)
10	nCT	-	0.0349$^+$/0.0114 (36.40%)	0.0355$^+$/0.0121 (31.98%)	0.0259$^+$/0.0126 (30.68%)	0.0424$^+$/0.0058 (29.54%)	0.0293$^+$/0.0015 (18.16%)
10	nEU	-	-	0.0210/0.0242 (34.70%)	0.0158/0.0279$^+$ (36.54%)	0.0306/0.0244 (36.27%)	0.0258/0.0356$^+$ (35.68%)
10	nsDCG	-	-	-	0.0226/0.0533$^+$ (11.16%)	0.0560$^+$/0.0330 (14.48%)	0.0189/0.0240 (48.10%)
10	α-nDCG	-	-	-	-	0.1291$^+$/0.0017 (5.24%)	0.0195/0.0174 (46.15%)
10	nERR-IA	-	-	-	-	-	0.0146/0.0278$^+$ (43.85%)

diversity metric at the n-th iteration, we calculated the scores at a cutoff of $5 \times n$.

3 Results and Analysis

3.1 Intuitiveness Test Analysis

We first report and analyze results of the Intuitiveness Test (see Algorithm 1) between different pairs of complex metrics, and at different iterations, when the three dimensions – Topical Relevance, Diversity and User Effort – are considered through the simple metrics prec, st-rec and tot-time, respectively.

Table 1 shows results for the TREC Dynamic Domain 2016 runs, for early and late iterations.[4] Consistent trends were observed in the data for the other years. The results show that the metric nCT – which was defined to cover the three dimensions – is more intuitive overall (usually significantly better, and never significantly worse) than metrics designed to only cover Topical Relevance

[4] Due to space limitations, in Table 1 we only show the results for the TREC DD 2016 runs, which is the second edition of the track and had almost as twice as many runs as the last edition.

and Diversity dimensions (e.g. α-nDCG or nERR-IA). With respect to metrics that model all dimensions (nCT and RBU), nCT is generally more intuitive at iteration 1 and statistically significantly more intuitive at iteration 10.

The results also demonstrate that the behavior of metrics in terms of intuitiveness is dependent on the iteration in the dynamic search session at which they are measured. In particular, metrics tend to disagree with each other in their preferences of runs more at iteration 10 than at the first iteration.

3.2 Ranking of Metrics Based on the Intuitiveness Test

To gain a broader understanding of intuitiveness, it is desirable to aggregate the low-level results of individual intuitiveness test evaluations, such as those that were presented in Table 1. Ideally, given pairwise comparisons between complex metrics such as from the previous section, a ranking of the intuitiveness of metrics can be induced. However, statistical significance is not transitive. As a result, in Table 2, we report a ranking of metrics, based on the number of times that a complex metric obtains a significantly higher intuitiveness score against the other metrics. Representative combinations of the dimensions (Topical Relevance (Rel), Diversity (Div), and User Effort (Eff)) and iterations (1, 3, and 10) are reported.[5]

The aggregation process involves summing the number of times that one complex metric obtained a significantly higher intuitiveness score than another, across the TREC Dynamic Domain tracks from 2015, 2016 and 2017, and converting this count into a ranking, such that a rank of 1 indicates that a metric obtained the highest count of significantly higher scores, while a rank of 7 indicates the lowest count. For example, for the topical relevance dimension (column *Rel*) at iteration 1 (sub-column *1*), nERR-IA has a rank of 5, indicating that it is more intuitive than two other metrics across different years.

In terms of Topical Relevance, ACT is more intuitive in late iterations, whereas nsDCG is more intuitive than other metrics in early iterations. For Diversity, α-nDCG is more intuitive than other metrics, regardless of the iterations. Here, nCT and RBU start more intuitive than other metrics but become less intuitive in later iterations.

In terms of User Effort, complex metrics that directly model this dimension (ACT, nCT, nEU and RBU) are, as may be hoped, more intuitive than other metrics.

When considering Topical Relevance and Diversity together, α-nDCG has higher intuitiveness scores than other metrics. In addition, nCT shows good performance in early iterations. We suspect this may be due to its emphasis on time, which becomes greater in later iterations.

[5] Other combinations and iterations are not reported due to lack of space, but overall trends were consistent with these settings. We also calculated the ranking of metrics based directly on their intuitiveness test relationship (i.e. without taking statistical significance into account); overall trends were again consistent with those presented here.

Table 2. Intuitiveness Test-based ranking of complex metrics using the number of times that a complex metric obtained a statistically significantly higher Intuitiveness Test score than other metrics, across all TREC Dynamic Domain years.

	Rel				Div				Eff				Rel and Div				Rel and Div and Eff			
Metric/Iteration	*1*	*3*	*10*	*All*	*1*	*3*	*10*	*All*	*1*	*3*	*10*	*All*	*1*	*3*	*10*	*All*	*1*	*3*	*10*	*All*
ACT	4	5	1	3	4	5	4	4	4	4	3	4	5	4	3	5	4	4	2	4
nCT	3	3	1	2	2	3	6	3	2	3	2	2	2	3	5	4	1	1	1	1
nEU	6	6	6	7	6	7	5	5	1	1	3	3	6	5	4	6	5	4	6	6
nsDCG	1	1	2	1	5	4	2	2	6	6	6	7	5	2	1	2	4	4	5	5
α-nDCG	1	2	3	4	1	1	1	1	5	6	5	6	1	1	1	1	1	3	3	2
nERR-IA	5	4	5	6	3	2	3	2	3	5	4	5	4	3	2	3	3	4	6	6
RBU	2	3	4	5	2	6	7	6	2	2	1	1	3	4	6	7	2	2	4	3

Modelling User Effort. We also experimented with two other simple metrics for modeling user effort, as described in Section 2.3: r-it and neg-it. Compared to tot-time, they agree with nCT more than nEU, whereas tot-time agrees with nEU more than nCT. This is likely because nCT uses the number of iterations to represent time, hence iteration-based simple metrics of User Effort may have biased the analysis toward nCT. On the other hand, nEU uses document length as an estimate of user effort, and since tot-time also uses document length as one part of its calculation, which may lead to better agreement for nEU. Of the three simple metrics that we explored to represent User Effort, we recommend tot-time as the most suitable, since it more closely models user behaviour when interacting with search results, taking the relevance of answers, and the amount of time required to process both document summaries and document content, into account.

3.3 Ranking Based on Metric Unanimity

For Metric Unanimity (Algorithm 2), the MU scores between complex metrics and different sets of simple metrics (to represent the effectiveness dimensions individually, in pairs, or all together) are calculated. A ranking of metrics was then induced, using the frequency with which one complex metric showed a higher unanimity than another complex metric – i.e. $\text{MU}(M_{c1}) > \text{MU}(M_{c2})$ – across all pairwise comparisons.[6]

Table 3 shows the ranking of the complex metrics, for representative sets of combinations of simple metrics (columns) and iterations (sub-columns); the displayed configurations are consistent with those shown for the Intuitiveness Test analysis.

In general, when considering all dimensions, nCT and RBU are ranked higher than other metrics for different iterations. For the different years, nCT unani-

[6] The Metric Unanimity framework differs from the Intuitiveness Test framework in that there is no equivalent concept of underlying "number of successes", therefore a significance test similar to the sign test in the Intuitiveness Test framework cannot be carried out.

Table 3. Metric Unanimity-based ranking of complex metrics using the number of times that a complex metric obtained a higher Metric Unanimity score than other metrics, across all TREC Dynamic Domain years.

	Rel				Div				Eff				Rel and Div				Rel and Div and Eff			
Metric/Iteration	*1*	*3*	*10*	*All*	*1*	*3*	*10*	*All*	*1*	*3*	*10*	*All*	*1*	*3*	*10*	*All*	*1*	*3*	*10*	*All*
ACT	5	5	2	4	5	3	5	5	2	3	4	4	6	5	2	5	6	4	3	3
nCT	1	1	1	1	1	1	6	3	1	2	2	2	1	1	1	1	1	1	1	1
nEU	6	4	5	7	7	5	4	6	1	1	3	3	7	4	4	6	5	3	4	4
nsDCG	4	2	4	2	6	4	2	4	4	5	7	7	5	2	3	3	7	6	7	7
α-nDCG	2	3	3	3	2	1	1	1	3	5	6	6	2	2	1	2	3	5	5	5
nERR-IA	3	4	4	5	4	2	3	2	2	4	5	5	3	3	3	4	4	7	6	6
RBU	5	4	6	6	3	5	7	7	1	2	1	1	4	4	5	7	2	2	2	2

mously agreed more with the simple metrics than other metrics. This mirrors its behavior in the Intuitiveness Test analysis. However, considering other dimension combinations, metrics might have different behavior than ones observed with the Intuitiveness Test. For instance, for the topical relevance dimension (column Rel), nCT consistently ranked first across iterations, while in the Intuitiveness Test analysis, it ranked better in late iterations.

We therefore formally analyze the correlation between the Metric Unanimity ranking and the ranking induced by the Intuitiveness Test, which we describe in the next section.

3.4 Comparing Intuitiveness Test and Metric Unanimity Rankings

The Spearman rank correlations between the rankings induced by the Intuitiveness Test and the Metric Unanimity analysis, for different combinations of dimensions and iterations, are shown in Table 4.

When considering the effectiveness dimensions, the correlations are generally high. In particular, for Diversity, User Effort, and the combination of all three dimensions, all correlation coefficients are strong at 0.6 or higher. Regarding iterations, for combinations of two or more dimensions, the correlations are generally higher in early iterations than later iterations. Similar patterns were observed for the individual years (again not included due to space constraints). However, statistically significant correlations were found more frequently for the TREC DD 2015 and 2016 than for the 2017 edition. This is likely due the lower number of submitted runs (11) in 2017, while 2015 and 2016 received 32 and 21 runs, respectively.

With regard to comparing the Intuitiveness Test and Metric Unanimity frameworks, both approaches are motivated differently. Moreover, both differ in their fundamental units of comparison (as noted previously, one practical implication of this is that an additional significance test can be applied as part of the Intuitiveness Test approach). Nevertheless, high correlations are observed between both meta-evaluation approaches. This provides strong evidence to indicate that the observed metric behavior is not due to either framework being

Table 4. Spearman rank-order correlation coefficient between the Intuitiveness Test ranking (Table 2) and the Metric Unanimity ranking (Table 3). * indicates a statistically significant correlation with $p < 0.05$.

Iteration	Rel	Div	Eff	Rel and Div	Rel and Div and Eff
1	0.44	0.94*	0.96*	0.92*	0.82*
3	0.71	0.89*	0.99*	0.75	0.71
10	0.80*	0.96*	0.99*	0.39	0.65
All	0.89*	0.94*	1.00*	0.79*	0.61

biased towards certain complex metrics, but instead is a reflection of the ability of complex metrics to capture the properties instantiated by the chosen simple metrics.

4 Conclusions

In this work, we investigated the ability of complex effectiveness metrics to cover three key dimensions for dynamic search tasks: Topical Relevance, Diversity, and User Effort.

Our analysis – using both the Intuitiveness Test proposed by Sakai [16] and the Metric Unanimity approach proposed by Amigó et al. [3] – showed that complex metrics can differ substantially in their ability to capture the various dimensions. Across iterations and datasets, nCT captures the key properties better than other metrics. However, the results also showed that α-nDCG can provide complementary information to nCT, and therefore we recommend that both should be reported when considering the effectiveness of dynamic search systems. In addition, the results showed that the behaviour of metrics can change as a search session progresses: metrics tend to disagree with each other more at later iterations. Thus, we also recommend reporting results at different iterations of a search session. Finally, our investigation demonstrated a high level of correlation between the Intuitiveness Test and Metric Unanimity. This provides a solid understanding of how complex effectiveness metrics agree or differ in relation to simple metrics that represent specific dimensions.

Future work includes the exploration of the impact of assigning different weights to the dimensions in the meta-analysis frameworks, which currently implicitly assume equal importance. We also intend to extend this analysis by considering other metrics for each of the dimensions, e.g. considering cognitive complexity for user effort. It will also be interesting to apply these frameworks to study the suitability of complex metrics that are used in domains that are typified by different search tasks, such as slow search, or high-recall search. Finally, we plan to study how the framework can be applied to assist in the construction of new metrics, to ensure that they sufficiently cover desired properties.

Acknowledgement. This research was partially supported by Australian Research Council (projects LP130100563 and LP150100252), and Real Thing Entertainment Pty Ltd.

References

1. Albahem, A., Spina, D., Scholer, F., Moffat, A., Cavedon, L.: Desirable Properties for Diversity and Truncated Effectiveness Metrics. In: Proc. Aust. Doc. Comp. Symp. pp. 9:1–9:7 (2018)
2. Amigó, E., Gonzalo, J., Verdejo, F.: A General Evaluation Measure for Document Organization Tasks. In: Proc. SIGIR. pp. 643–652 (2013)
3. Amigó, E., Spina, D., Carrillo-de Albornoz, J.: An Axiomatic Analysis of Diversity Evaluation Metrics: Introducing the Rank-Biased Utility Metric. In: Proc. SIGIR. pp. 625–634 (2018)
4. Busin, L., Mizzaro, S.: Axiometrics: An Axiomatic Approach to Information Retrieval Effectiveness Metrics. In: Proc. ICTIR. pp. 8:22–8:29 (2013)
5. Carterette, B., Kanoulas, E., Hall, M., Clough, P.: Overview of the TREC 2014 Session Track. In: Proc. TREC (2014)
6. Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P.: Expected Reciprocal Rank for Graded Relevance. In: Proc. CIKM. pp. 621–630 (2009)
7. Chuklin, A., Zhou, K., Schuth, A., Sietsma, F., de Rijke, M.: Evaluating Intuitiveness of Vertical-aware Click Models. In: Proc. SIGIR. pp. 1075–1078 (2014)
8. Clarke, C.L., Craswell, N., Soboroff, I., Ashkan, A.: A Comparative Analysis of Cascade Measures for Novelty and Diversity. In: Proc. WSDM. pp. 75–84 (2011)
9. Clarke, C.L., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and Diversity in Information Retrieval Evaluation. In: Proc. SIGIR. pp. 659–666 (2008)
10. Ferrante, M., Ferro, N., Maistro, M.: Towards a Formal Framework for Utility-oriented Measurements of Retrieval Effectiveness. In: Proc. ICTIR. pp. 21–30 (2015)
11. Jiang, J., He, D., Allan, J.: Comparing In Situ and Multidimensional Relevance Judgments. In: Proc. SIGIR. pp. 405–414 (2017)
12. Jin, X., Sloan, M., Wang, J.: Interactive Exploratory Search for Multi Page Search Results. In: Proc. WWW. pp. 655–666 (2013)
13. Kanoulas, E., Azzopardi, L., Yang, G.H.: Overview of the CLEF Dynamic Search Evaluation Lab 2018. In: Proc. CLEF. pp. 362–371 (2018)
14. Luo, J., Wing, C., Yang, H., Hearst, M.: The Water Filling Model and the Cube Test: Multi-dimensional Evaluation for Professional search. In: Proc. CIKM. pp. 709–714 (2013)
15. Moffat, A.: Seven Numeric Properties of Effectiveness Metrics. In: Proc. Asia Info. Retri. Soc. Conf. pp. 1–12 (2013)
16. Sakai, T.: Evaluation with Informational and Navigational Intents. In: Proc. WWW. pp. 499–508 (2012)

17. Sakai, T.: How Intuitive Are Diversified Search Metrics? Concordance Test Results for the Diversity U-Measures. In: Proc. Asia Info. Retri. Soc. Conf. pp. 13–24 (2013)
18. Smucker, M.D., Clarke, C.L.: Time-based Calibration of Effectiveness Measures. In: Proc. SIGIR. pp. 95–104 (2012)
19. Tang, Z., Yang, G.H.: Investigating Per Topic Upper Bound for Session Search Evaluation. In: Proc. ICTIR. pp. 185–192 (2017)
20. Turpin, A., Scholer, F.: User Performance Versus Precision Measures for Simple Web Search Tasks. In: Proc. SIGIR. pp. 11–18 (2006)
21. Yang, H., Frank, J., Soboroff, I.: TREC 2015 Dynamic Domain Track Overview. In: Proc. TREC (2015)
22. Yang, H., Soboroff, I.: TREC 2016 Dynamic Domain Track Overview. In: Proc. TREC (2016)
23. Yang, H., Tang, Z., Soboroff, I.: TREC 2017 Dynamic Domain Track Overview. In: Proc. TREC (2017)
24. Zhou, K., Lalmas, M., Sakai, T., Cummins, R., Jose, J.M.: On the Reliability and Intuitiveness of Aggregated Search Metrics. In: Proc. CIKM. pp. 689–698 (2013)