Data-Driven Evaluation Metrics for Heterogeneous Search Engine Result Pages

Leif Azzopardi
University of Strathclyde
Glasgow, Scotland
leifos@acm.org

Ryen W. White
Microsoft
Redmond, United States
ryenw@microsoft.com

Paul Thomas
Microsoft
Canberra, Australia
pathom@microsoft.com

Nick Craswell
Microsoft
Bellevue, United States
nickcr@microsoft.com

ABSTRACT

Evaluation metrics for search typically assume items are homogeneous. However, in the context of web search, this assumption does not hold. Modern search engine result pages (*SERP*s) are composed of a variety of item types (e.g., news, web, entity, etc.), and their influence on browsing behavior is largely unknown.

In this paper, we perform a large-scale empirical analysis of popular web search queries and investigate how different item types influence how people interact on SERPs. We then infer a *user browsing model* given people's interactions with SERP items – creating a data-driven metric based on item type. We show that the proposed metric leads to more accurate estimates of: (1) total gain, (2) total time spent, and (3) stopping depth – without requiring extensive parameter tuning or *a priori* relevance information. These results suggest that item heterogeneity should be accounted for when developing metrics for SERPs. While many open questions remain concerning the applicability and generalizability of data-driven metrics, they do serve as a formal mechanism to link observed user behaviors directly to how performance is measured. From this approach, we can draw new insights regarding the relationship between behavior and performance – and design data-driven metrics based on real user behavior rather than using metrics reliant on some hypothesized model of user browsing behavior.

CCS CONCEPTS

• **Information systems** → **Retrieval effectiveness**; • **Human-centered computing** → *User models.*

ACM Reference Format:
Leif Azzopardi, Ryen W. White, Paul Thomas, and Nick Craswell. 2020. Data-Driven Evaluation Metrics for Heterogeneous Search Engine Result Pages. In *2020 Conference on Human Information Interaction and Retrieval (CHIIR '20), March 14–18, 2020, Vancouver, BC, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3343413.3377959

CHIIR '20, March 14–18, 2020, Vancouver, BC, Canada

ACM ISBN 978-1-4503-6892-6/20/03…$15.00
https://doi.org/10.1145/3343413.3377959

1 INTRODUCTION

Evaluation has been a long standing problem in information retrieval (**IR**) [40, 43]. As such, offline or test collection metrics [40] have played a pivotal role in shaping the comparison and ranking of queries and systems. While initially, metrics were mainly system-focused (e.g., F1-measure, average precision (AP), etc.), newer metrics have become more user-focused. Such metrics can be described by their *user browsing model* [11], and such models are entirely defined by the probability of a user continuing to read items in the result list [35]. For example, in rank biased precision (RBP) [37] the patience parameter determines the likelihood of the user continuing to the next item, whereas for precision metrics, users are assumed to continue up until some fixed rank k. Rather than impose a static or fixed user browsing model, newer metrics have become adaptive – where the browsing behavior varies depending on factors such as the length or height of items [32, 39, 41], the relevance of items [26, 33, 36], and/or the rate of gain [6]. These metrics have largely been designed under the implicit assumption that the items are the same type.

However, modern search engine result pages (**SERP**s) are heterogeneous in nature, presenting an array of different item types (e.g., ads, news, web, etc.) [4, 5] presenting a significant challenge when trying to measure the whole page [8]. These item types not only require different amounts of effort to process, but also are provided to address the varying and differing needs and intents of web users [9, 13, 22]. For example, news items may be relevant for some cases [13], while other items like advertisements are often skipped by certain users, but consumed by others [17]. These differences in behavior are not reflected in most user browsing models. Therefore, current metrics may be less accurate and less able to deal with SERP heterogeneity.

In this work, we explore how item types influence browsing behaviour (compared to other factors such as query type, relevance, and position). We then directly infer user browsing models given the items, and their type, on the SERP creating a new data-driven metric (**DDM**) – where the probability of continuing depends on the type of item and its position on the SERP. We then perform a comprehensive analysis comparing our new approach against five commonly-used metrics [12, 23, 29, 38], four recently proposed adaptive metrics [6, 33, 41, 48], and two other data-driven approaches [26, 47].

2 BACKGROUND

Metric development has been central to evaluation in IR. Accurate measurements of the performance of queries and systems are needed to decide which queries to improve and which algorithms and systems to employ [40]. As our understanding of measurement has advanced, it has been shown that most metrics used to evaluate ranked lists of search results are fundamentally related [11, 24, 34] and are underpinned by different user browsing models [11]. Within the **C/W/L** framework, Moffat et al. [35] demonstrate that these user browsing models can be described entirely by the conditional probability of a user continuing to the $(i+1)th$ item, given they are already at the ith item in the result list. This framework provides a formal basis for metrics and measurement [6, 35, 44]. We use it here both to measure result lists with current metrics and to build a new metric that also conditions on the item type.

Figure 1 represents the **C/W/L** user browsing model, where once a query is issued, it is assumed that the user will inspect item i, accumulate some gain (at a cost) and then either continue to the next item with probability $C(i+1|i)$, or stop with probability $1-C(i+1|i)$. $C(i+1|i)$ is the conditional probability of continuing to examine $(i+1)$ given i is examined (we shall use C_i as shorthand). So, for example, Precision at rank k (P@k) is defined by the probability of continuing $C_i = 1$, when $i < k$, otherwise $C_i = 0$; whereas RBP with persistence θ is given by $C_i = \theta$.

Under the **C/W/L** framework, the vector of continuation probabilities (**C**), which we shall call the *continuation* or **C** *function*, can be converted to a weight vector **W**. This can be interpreted as the expected proportion of attention given to the item at rank i. **C** can be converted to **W** as follows:

$$W_i = \frac{\prod_{j=1}^{i-1} C_j}{\sum_k \prod_{j=1}^{k-1} C_j} \tag{1}$$

Given weight vector **W**, the *Expected Utility* (**EU**) of a result list is:

$$EU = \sum_{i=1\ldots\infty} W_i \, r_i \tag{2}$$

where **r** is the relevance (gain) vector for each rank i. The EU essentially quantifies the rate of gain per item. The *Expected Depth* (**ED**) i.e., number of items that a user examines, is given by [33]:

$$ED = \frac{1}{W_1} \tag{3}$$

Taken together, the *Expected Total Utility* (**ETU**) is the rate of gain multiplied by the expected number of items examined (e.g., $ETU = EU \times ED$). Further, the *Expected Total Cost* (**ETC**) can be computed in a similar manner to ETU, but using the expected cost (EC), by using a cost vector **k**, where each k_i represents the cost of each item i, instead of a gain vector **r** [6].

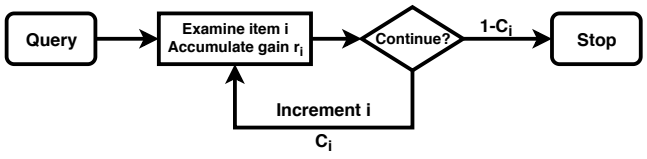

Figure 1: The user browsing model encoded in metrics is based on the conditional probability of continuing C_i.

While each metric encapsulates different models of browsing behavior, a key assumption is that a more realistic user browsing model will result in a more accurate estimate of performance. Under the **C/W/L** framework, we can determine which metric best approximates users' performance and behavior, in terms of: *(i)* how much gain users acquire from the SERP, *(ii)* how long users spend on the SERP, and *(iii)* how many items they examine on the SERP. The latter values (time and depth) are observable, while the former needs to be inferred from relevance labels. This means it is possible to directly compare the expected behavior given a metric against the behavior actually observed.

From Static to Adaptive C Functions. Evaluation measures have evolved from precision and recall-based to utility and cost-based – with more focus towards how the user interacts with the SERP based on the gains and costs (termed "adaptive" in Moffat et al. [34]). To date, the most commonly used metrics include P@k, Reciprocal Rank (RR), and AP [12, 28, 38]. Moving beyond precision based metrics, (normalized) discounted cumulative gain ((n)DCG) [23, 29] was proposed to discount the weight assigned to documents further down the ranking. However, Moffat and Zobel [37] argued that the log-based discount function of DCG was not grounded, and did not best characterize how people actually examine a ranked list. Instead, they proposed RBP, which assumes that the probability of continuing is a constant (θ). Rather than discount by rank, Smucker and Clarke [41] proposed Time Biased Gain (TBG), which discounts the weighting proportional to the amount of time spent reading each item. Sakai and Dou [39] proposed a similar metric which discounts based directly on the length of each item (in the context of news search), while Luo et al. [32] proposed a related metric, height biased gain (HBG) which discounts based on the height of each item (in the context of mobile search). These metrics all essentially assume that users are less likely to continue when items are longer, larger, or take longer to process. More recently, Jiang and Allan [24] propose a similar metric which considers the cost approximated by time, by calculating the ratio of gain to cost.

Rather than consider the cost, the goal-sensitive INST measure [33] assumes that the user has some idea of the number of documents that they want (T). As they examine documents, and find relevant documents, then the probability of continuing decreases. Conversely, the Bejeweled player model (BPM) [48] considers both how much gain the user wants (T) and how much they are willing to spend (K). If the user reaches T, or spends K, then the user stops (i.e., $C_i = 0$, otherwise $C_i = 1$). Under the dynamic version, if the user encounters relevant items, BPM assumes that they will increase their estimates of T and K, and *vice versa* for non-relevant items. In the IFT metric [6], it assumes that users would like T, but at a rate of gain of at least A. Rather than a discrete **C** function (like in BPM), C_i decreases as they approach T, and increases as A increases.

From Model based to Data-Driven C Functions. Many metrics encapsulate hypothetical user browsing models primarily based on intuitions of how users interact with ranked lists (e.g., model based). More recently, there have been efforts to ground the browsing model in empirical data. For example, INST was formulated based on the observation that the conditional probability of continuing increased with rank. Yilmaz et al. [47] proposed Expected

Browsing Utility (EBU), which directly estimated the proportion of users at each rank (i.e., the **W** function). They showed that this led to a better fit of the observed click-through distributions when compared to other parameterized metrics (e.g., RBP, DCG, etc.), but did not evaluate how well it predicted performance. Jiang and Allan [26] proposed a method to estimate the patience/persistence parameter of existing metrics by inferring the patience at each i based on the relevance of items, and observed user behaviour (i.e., whether or not people continued given the relevance of an item) to empirically set the patience parameters. Their adaptive relevance (AR) method led to better estimates of performance suggesting that parameter estimation should be informed by relevance data. More recently, Moffat and Wicaksono [36] observed that non-relevance also impacts behavior, and proposed a metric where the user is more likely to continue if the item is plausibly non-relevant rather than egregiously.

Our work continues on this line of developing adaptive and data-driven metrics. However, the key difference is that rather than using relevance information – which may not be available – our approach focuses on estimating continuation probabilities empirically given people's interactions with different item types (e.g., ads, news, etc.).

3 DATA-DRIVEN METRICS

Given the **C**/**W**/**L** Framework, we propose that the **C** function can be estimated directly from observed data – rather than relying on hypothesized models of user behaviour. The probability of a user continuing to item $i+1$ given they have examined item i is number of people who went to $i+1$ and beyond, divided by the number of people who went to i and beyond. Thus, the *Maximum Likelihood Estimate* of the conditional probability of continuing (from i to $i+1$) can be formally defined as:

$$C(i+1|i,x) = \frac{\sum_{j=i+1}^{j=\infty} n(u,j,x)}{\sum_{j'=i}^{j'=\infty} n(u,j',x)} \tag{4}$$

where $n(u,j,x)$ is the number of users who stop at position j given x i.e., the number of users who click on the jth result and did not return to the SERP given x. This general form of the user's continuation behavior means that, depending on the data available, an array of data-driven metrics (**DDM**) can be developed. For example, $\mathbf{C_i}$(**Position**) conditioned only on i gives Expected Browsing Utility (EBU) [47]; and $\mathbf{C_i}$(**Relevance**) conditioned on the relevance of the item at i, where x is the relevance label (highly relevant, relevant, etc.), is the empirical version of the Adaptive Relevance (AR) method [26]. Given the C_i values, a page specific **C** function can be constructed using the item information (x), which is then used within the **C**/**W**/**L** framework to predict EU, ETU, etc.

An Item-Type C Function. Prior metrics have typically assumed that the items presented are all of the same type. However, SERPs are heterogeneous in nature, composed of many different item types (e.g., web, ads, video, entity card, etc.). As we have noted, some metrics have tried to encode the differences between items implicitly or indirectly (for example, by assigning different costs); however, they have not explicitly conditioned on the item's type, which intuitively is likely to influence user browsing behavior. For example, users may be more likely to skip over ads, videos, and news, while they may be less likely to skip over web results, entity cards, and images. Therefore, in this work, we aim to explore the influence of item type on browsing behavior. Given Eq. 4, we can further define $\mathbf{C_i}$(**Type**), which is conditioned on the item type. We hypothesize that by accounting for heterogeneity of SERP elements, we can derive better performance estimates.

4 RESEARCH QUESTIONS AND METHOD

The aim of this study is to investigate the potential of data-driven metrics, and specifically, how they can account for the heterogeneity on SERPs. We seek to answer the following research questions:

RQ1 What is the influence of position, relevance, and item types on user browsing behavior (as described by **C**)?

RQ2 Which metric (given its **C** function) provides the most accurate prediction of performance (lowest error)?

RQ3 Which metric (given its **C** function) provides the most accurate ranking of queries (highest correlation)?

By performance, we mean how well the metrics predict the total amount of gain (i.e., ETU), the total amount of time they spent on the SERP (i.e., ETC), and the number of items that they examined (i.e., ED). We address these questions in the context of web search for frequently issued (popular) queries. This is because it: (1) represents a large proportion of search traffic that is very important to measure and improve, and (2) provides a large volume of search traffic to robustly estimate the data-driven **C** functions. While popular web queries tend to thought of as navigational [9], the intents (and subsequent behavior) can vary substantially from query to query [42]. For example, they can be *very focused*, whereby most users tend to select the same item. Or they can be *very diffuse*, whereby users tend to select a number of different items. For example, for queries such as "*facebook*" and "*youtube*," users would tend to select only one URL, the corresponding homepage, i.e., www.facebook.com, www.youtube.com, etc. where are for queries like "*music*" and "*cheap flights*" users tend to select from among a number of different URLs. Given that behavior may differ between these different query intents for popular queries, we also consider how the query type (focused to diffuse) also influences continuation behavior.

Data and Materials. For the purposes of this study, we created a data set containing 915 popular web search queries issued to a commercial search engine during October, 2017[1]. Over the course of two weeks, we collected a sub-sample of web traffic which contained approximately 16.4 million query impressions (where each of the 915 queries were submitted approximately 9,000 times per week, on average). To focus the analysis, we restricted the sample in a number of ways. The users were English speakers from the United States. The sample was from desktop users, as we hypothesized that different form factors will have different interaction costs, and therefore potentially different **C** functions. We also limited the sample to only include impressions where at least one click was observed. This is because we wanted predict the depth to which people went to on the SERPs – so that we could compare this against what the metrics predict. While this could have been potentially

[1]Note that we randomly selected 1000 popular queries, but in our sub-sampling, 85 queries were returned less than 1000 times, and so were excluded.

Table 1: The mean and standard error values for each of the features for the four levels of query focus. Gain, Time, and Depth values are reported relative to the overall mean for all 915 queries. ⋆ (†) indicates significance differences b/w all other groups (or marked groups). Less focused queries (MF) tend to be longer and more distinct URLs are displayed on their SERPs.

Focus	# Queries	# Impressions per Query	Query Length (Chars.)	# Distinct URLs Shown	Click Entropy	Click Ratio	Relative % Percentage Difference Gain	Time	Depth
HF	426	14557±3431	10.37±0.25†	19.77±1.73	0.34±0.01*	0.96±0.00*	24.6±2.7*	-14.4±0.8*	-21.5±1.0*
F	229	4809±574	11.17±0.37	26.80±3.25	1.03±0.02*	0.82±0.01*	0.7±3.8*	6.0±1.9*	3.1±2.7*
SF	157	4145±511	11.24±0.41	27.33±2.79	1.57±0.03*	0.61±0.01*	-32.3±3.6*	14.5±2.7*	19.3±4.3*
MF	103	2769±483	12.01±0.59†	56.16±12.3*	2.58±0.11*	0.40±0.01*	-54.2±3.3*	24.2±3.9*	52.5±6.6*

inferred from mouse/viewport information, this would have added more uncertainty in determining the last item viewed – as such, we leave the abandonment scenario where no clicks are observed for future work.

For each query impression, we then extracted the main result items shown on the SERP: "organic" web results (Web), advertisements (Ad), entity cards (Entity), News, Video, and Image blocks. All other items were categorized into "other," including query suggestions, disambiguation, stock quotes, etc. Each SERP had 12.8 result items on average. We recorded the total time spent on the SERP, along with which items were clicked and their location (core or right rail, and rank within the core/rail).

Query Intent and Navigational Focus. To determine how focused the navigational intents were, we employed two measures: *click entropy*, which provides an indication of spread in the selected results [42], and *click ratio*, which provides an indication of focus [16]. The measures were highly and significantly correlated ($r = -0.904, p < 0.001$), and so we used only the click ratio measure as it is simpler to interpret:

$$r(q) = \frac{\max_u c(u,q)}{\sum_{u'} c(u',q)} \tag{5}$$

where $c(u,q)$ is the number of times a URL u was clicked for the given query q. Intuitively, the more focused the navigational intent of the query (i.e., the more the user population only wanted one specific URL u) the higher the value (up to one, where one indicates that all users selected the same URL u for the given q). We partitioned the queries into different groups according to click ratio: (**HF**) Highly Focused, where $r(q) \geq 0.9$, (**F**) Focused, where $0.7 \geq r(q) > 0.9$, (**SF**) Somewhat Focused, where $0.5 \geq r(q) > 0.7$, and (**MF**) Marginally Focused, where $r(q) < 0.5$. Table 1 summarizes the breakdown over query focus levels.

To check that the groupings were informative, we performed significance testing using an ANOVA, with follow up multi-comparison tests using Tukey HSD to check whether there were differences between groups in terms of the features reported in Table 1. Given our breakdown, we observed significant differences across a number of aspects: the click entropy across groups ranged from 0.33 (HF) to 2.58 (MF) ($F(1,913) = 488.9, p < 0.001$), the relative depth, w.r.t. the average depth ranged from -21.5% (HF) to +52.5% (MF) ($F(1,913) = 330.8, p < 0.001$), the relative time, w.r.t. the average time spent ranged from -14.4% (HF) to + 24.2% (MF) ($F(1,913) = 250.5, p < 0.001$), and the relative gain accrued w.r.t. to the average gain accrued ranged from +24.6% (HF) to -54.2% (MF) ($F(1,913) = 260.0, p < 0.001$). These results provided confidence in our decision to break the queries into the four levels of focus. We also hypothesized that different model-based metrics would need to be tuned specifically to accurately estimate the performance within each group.

Relevance Labels. For each of the 915 queries, we randomly selected query impressions per query, and then obtained judgments for all the items on these SERPs. Similar to TREC and other evaluation forums, judges were given the query and item/URL, and asked to rate the relevance on a four-point scale. We used an in-house crowd-sourcing platform; to control quality, judges were experienced with this task and subject to random checks against "gold standard" labels. We collected approximately 43,000 judgments for 12,800 unique items, and the final label was decided by majority vote with extra judgments requested as needed to break ties. Of all the items displayed on the pages across all 16.4 million query impressions, 34.8% had corresponding labels - of which 7.3% of items were labeled not relevant (**NR**), 57.1% marginally relevant (**MR**), 12.3% relevant (**R**), and 23.1% highly relevant (**HR**). Following convention, the remaining unlabeled items were considered non-relevant. To attribute gain to relevance labels we used the following conversion: HR=1.0, R=0.4, MR=0.2, NR=0.0, which was shown to best correlate to user satisfaction [20].

Ordering of Items. Evaluation metrics require a ranking and assume a top-down browsing order. However, SERPs typically are two dimensional, and so a browsing order needs to be inferred. In prior work, it has been shown that users examine items based on an "F-shape pattern" (e.g., golden triangle) [10, 17] and that browsing order correlates with clicks [31]. Azzopardi et al. [6] showed that, on average, web users inspected two items from the core, then an item from the right rail, and so on. They showed that this was highly correlated with the click-through distribution. In this work, we used the same method. Note that we also tried different orderings, but these led to similar findings.

Cost of Items. To calculate the expected total cost (in seconds) i.e. time spend on page, and to instantiate TBG and IFT, we needed to estimate the average cost (in seconds) of assessing items of different types. Following Azzopardi et al. [6], we created a linear model to predict the time spent on the SERP. The input variables were the number of times each item type was observed on the SERP, up to and including the last item clicked, and the time spent on the SERP. We used all the impressions from week one to estimate the costs. The estimates (which have been normalized with respect to

t, the time taken to process one web result[2]) were as follows: Web (1.00t), Ad (1.90t), Ad Right (0.65t), News (5.53t), Image (2.2t), Video (4.06t), Entity (13.77t), Entity Right (0.83t) and Other (2.77t). Statistical testing indicated that the estimates were significant (ANOVA $F(9, 8238513) = 3.59e + 04$, $p \ll 0.0001$).

Metrics. The metrics in our analysis included standard and often-used metrics, such as AP, RR, and P@k. Since we have graded relevance judgments, we used the respective graded versions [12, 29, 38]. We note that while these metrics all make unrealistic or questionable assumptions regarding search behavior [19], we included them because they are commonly employed baseline metrics. We further included DCG and RBP since these two metrics have been widely used in IR research. We also employed INST [33], TBG [41] (which captures the same intent as the U-Measure and HBG), BPM [48] and IFT [6], as they each attempt to model different aspects of user browsing behavior and have shown to highly correlate with user satisfaction and/or performance – but have not been directly compared against each other. We then compared these metrics against the data-driven metrics (DDMs) estimated based on: (*a*) $\mathbf{C_i}$(**Position**) denoted as DDM-P, (*b*) $\mathbf{C_i}$(**Relevance**) denoted as DDM-R, and (*c*) $\mathbf{C_i}$(**Type**) denoted as DDM-T.

Parameters and Settings. For this work, we employ a train-and-test methodology using two-fold cross validation. The queries were divided into two groups (457 and 458) with approximately equal number of queries in each of the focus levels. We then performed a parameter sweep using the data from one group, and then undertook testing using the held out query group. Below we detail the parameters used. For P@k and DCG@k, we explored $k = 1, \ldots, 5, 10, 15, 20$. For RBP, we used: $\theta = 0.0001, 0.001, 0.01, 0.1, \ldots, 0.9$. For INST, T was set to: $0.2, 0.4, \ldots, 1.4, 2.0$. For TBG, the half-life (h) was set to: $0.0625, 0.125, 0.25, 0.75, 1.0, 1.5, \ldots, 4$. The time at C_i was the sum of the time associated with each of the items up to and including i as per the estimated times for items.

Both BPM and IFT have several parameters. It was not possible to do an exhaustive sweep of all the parameters. So we first grounded the sweep by selecting the suggested parameters reported for navigational, web queries [6, 48], and then explored around this space. For the BPM we used the dynamic version, since it was shown to provide a greater correlation with user satisfaction, T was set to: $0.8, 1.0, 1.2, 1.4$ and K was set to: $1, 2, 3, 4, 5, 10, 15$, while updating parameters were set to one. This led to 28 different combinations, i.e., $T \times K$. For IFT, we explored the following settings: $T = 0.2, 0.4, 0.8, 1.0, 1.2$, and for each T, varied $b_1 = b_2 = 0.15, 0.25, 0.35, 0.45$, and $A = 0.05, 0.1, 0.2$. The R_1 and R_2 parameters were fixed to ten, since higher values would lead to a step function like P@k and BPM, while lower values would lead to a flat continuation function like RBP. This resulted in sixty different combinations i.e., $T \times b \times A$. Note that it took over 8,000 hours of processor time to evaluate the combinations listed above (for testing and training). This highlights a growing problem with employing increasingly sophisticated metrics with many parameters: training and tuning them requires substantial computational resources.

For the data-driven metrics (DDM-P, DDM-R and DDM-T), no parameter tuning was required - instead the C_i was estimated from the training data using Eq. 4. For DDM-R, for relevance labels were used, such that $x \in \{\text{HR}, \text{R}, \text{MR}, \text{NR}\}$, and for DDM-T, item types were used, such that: $x \in \{\text{Ad}, \text{Web}, \text{Entity}, \text{Image}, \text{News}, \text{Video}, \text{Other}\}$. We also included a DDM variant based on the query focus (DDM-F), using $\mathbf{C_i}$(**Focus**) where: $x \in \{\text{HF}, \text{F}, \text{SF}, \text{MF}\}$. This was because we further hypothesized that the focus of the query would also impact browsing behavior.

Reproducibility. In order to promote reproducibility, we have developed an open source framework which is publicly available at: https://github.com/ireval/cwl and contains the code to calculate all of the above metrics [7]. This enables other researchers to reproduce similar experiments. While it is not possible to release the query sample used in this paper due to commercial sensitivity and privacy concerns, the core of this paper is about understanding how to best encode user browsing behavior, and what factors influence performance.

5 EMPIRICAL ANALYSIS OF C FUNCTIONS

To address RQ1 above, we plotted the mean average $\mathbf{C_i}$(**Position**), $\mathbf{C_i}$(**Focus**), $\mathbf{C_i}$(**Relevance**) and $\mathbf{C_i}$(**Type**) over all queries (see Figure 2[3]). Below, we discuss the resultant **C** functions – which express user browsing behavior as the conditional probability of continuing.

$\mathbf{C_i}$(**Position**). In Figure 2, the top-left plot shows the continuation behavior at each position over all queries. Clearly, most users stop after the first result. However, approximately 20% of users continue to position 2, and of these about 65% stop, while the remainder continue to position 3, and so forth. Interestingly, the aggregate **C** function shows that C_i increases as users go deeper, such that by position ten the probability of continuing to the next position flattens out to around 0.8 (i.e., 80% of users continue to the next rank, if at position 10 or later). The shape of this **C** function is similar to the **C** within DCG and INST [33] suggesting that these metrics maybe a good model of aggregated continuation behavior (which should, in turn, produce good performance estimates).

$\mathbf{C_i}$(**Focus**). However, when we break down the query set according to the different query focus levels, a different pattern emerges, and the **C** behavior of users for the different query levels changes. The top-right plot in Figure 2 shows that for the highly focused queries (i.e., those where users tend to select the same result, e.g., "*facebook*" → facebook.com), a similar trend is observed in the **C** function for position only (see left plot, in Fig. 2). This is because these queries dominate the estimate since they represent almost half of the queries in the set (426 out of 915 are highly focused). But, as the queries become less focused and more diffuse the continuation probabilities (especially early on) are higher. For example, for focused queries (F) the probability of containing to position 2 is around 39%, for somewhat focused queries (SF) it is 43%, while for the marginally focused queries (MF) it is 52%. This is in stark contrast to the highly focused queries (HF) where the probability of continuing to position two is 15% on average. For the marginally focused queries (MF) the probability of continuing increases from 52% up to around 80% by position ten – suggesting that for such

[2]The times were normalized due to their commercial sensitivity.

[3]The plots are up to position 10, as after position 10 we observed less than 10,000 impressions.

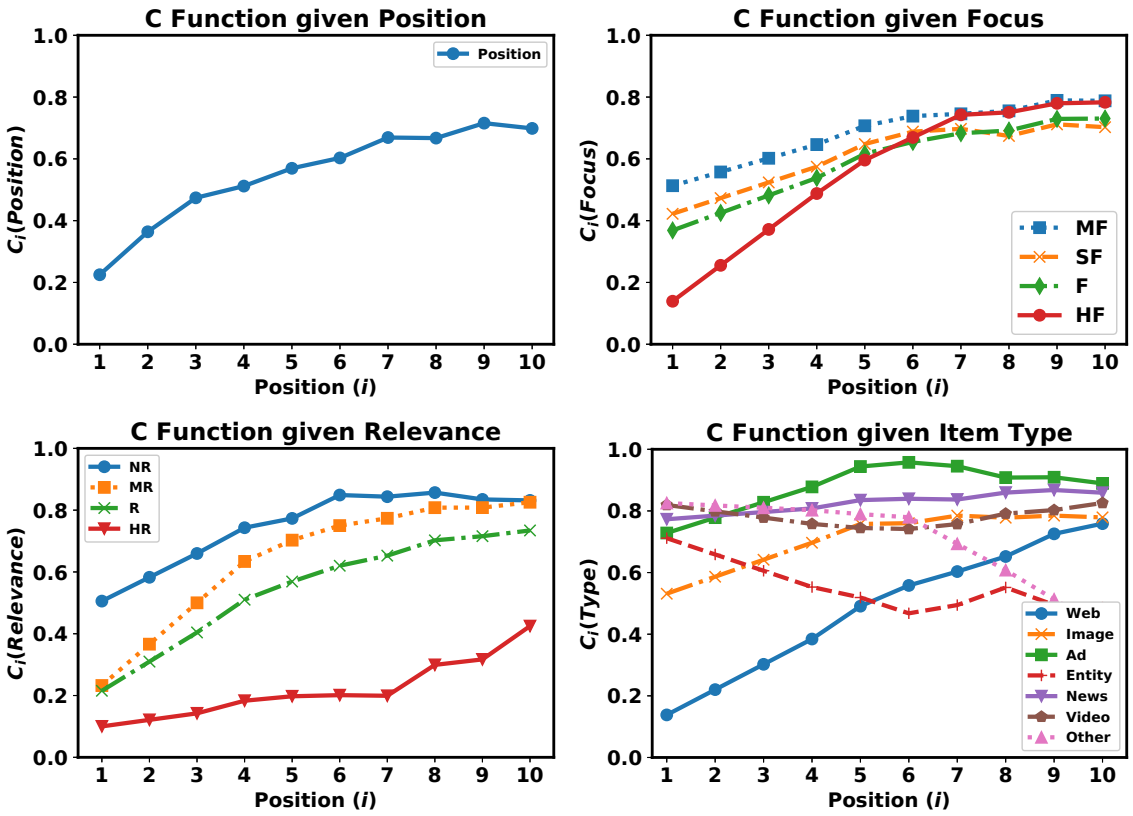

Figure 2: The mean ML estimate of the conditional probability of continuing to position $i+1$ given that i has been inspected given: position C_i(Position) (top-left), focus C_i(Focus) (top-right), relevance C_i(Relevance) (bottom-left) and item types C_i(Type) (bottom-right). The standard errors were less than 0.03 (not shown). For most factors C_i tends to increase with position. However, it is clear that focus, relevance, and items influence user browsing behavior in different ways.

queries, users are more persistent, or at least need to be more persistent, in order to find enough relevant information. Clearly, these plots show that people's continuation behavior changes given the query focus. This suggests that static metrics (such as P@k, DCG, and RBP) will be unlikely to fully capture the range of behaviors exhibited by users (unless explicitly tuned to different query types).

C_i(**Relevance**). In Figure 2, the bottom-left plot shows how the relevance label associated with the item relates to continuation. Firstly, users are more likely to stop when they encounter a highly relevant item (regardless of position). At position one, the probability of continuing to position two given a highly relevant item is around 10%, which is substantially lower than the aggregated C_1 value (i.e., 20%). Secondly, we can see that as the relevance of items changes from highly relevant to non-relevant, the probability of continuing increases across positions. Intuitively, this makes sense as for less relevant items users more likely would need to continue to find what they need. These plots confirm that relevance also has an impact on continuation behavior - and suggests that metrics which take relevance into account (i.e., INST, BPM, IFT, DDM-R), should lead to better estimates of performance.

C_i(**Type**). Finally, the bottom-right plot in Figure 2 shows the relationships between item types (Ad, Web, Entity cards, Image, News, Video, and Other), position and continuation behavior. Not too surprisingly, web results (which are the most common item type shown) follow a similar trend to the top-left plot, where the probability of continuing increases with position. For the other item types, we observe distinctly different patterns. For example, if an entity card is positioned first, then the continuation probability is very high (approximately 72%), but if positioned later on the page, the continuation probability drops down to around 40%. This shows that users are more likely to skip over entity cards if positioned too early suggesting that they either prefer other items, or want to inspect other possibilities on the SERP first. For Ads, we can see that the probability of continuing is relatively and consistently high (increasing to almost 100% around positions four to six). This suggests that users tend to continue past such items mid-SERP and that it might be better to put a different item type in their place. For the image items, the probability of continuing when placed at position one is much lower than these other items initially (56%) and then increases to be on par with video and news items (around 80%). Initially the lower C_i for images suggest that they are useful and well placed since the user's intent was to find images. However, when images are presented further down (perhaps to improve the appeal and diversity of the SERP), users are more likely to continue past them (as indicated by the higher C_i). While for video items

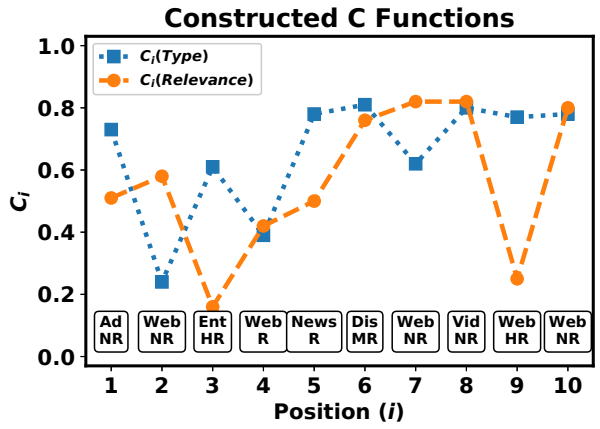

Figure 3: The resulting continuation functions for an example set of items on a SERP given their relevance or item type. The dashed line represents when C_i is conditioned on relevance, while the dotted line represent when C_i is conditioned on item type. Two different continuation functions are constructed as a result, capturing different aspects of user browsing behavior.

users, on average, are likely to continue past them about with high probability (80%) regardless of position. Finally, for web results we see that the probability of continuing is very low initially (around 20%) and then increases as i increases. Compared to the other items, this suggests that there is a trade-off between the item types and their positions on page, motivating further research into understanding the interactions between items on the SERP. Note that current metrics do not explicitly consider the item type directly, though metrics such as TBG, BPM, and IFT all consider the costs associated with items, and so this may reflect to some extent the influence of item type. However, it is an open question whether this is the case, and whether information about the item type is useful when measuring performance – we explore this next.

6 ERROR ANALYSIS

In this section, we focus on addressing our second research question, to determine how accurately the different metrics could predict the estimated performance. For each query, the estimated (predicted) total gain was computed for each metric (i.e., ETU in Sec. 2), and then this was compared to the inferred (observed) total gain. As done by Azzopardi et al. [6], we calculated this as the sum of the gain of the clicked items given the judgments. Table 2 reports the mean absolute error between the predicted and observed values for: *gain* i.e., expected total gain (ETU) vs. inferred, observed total gain, *time* i.e., expected total time (ETC) vs. actual time spent on SERP, and *depth* i.e., expected stopping depth (ED) vs. observed stopping depth (rank of the last item clicked). The metric/parameter setting that minimized the error in total gain was used. We report the average parameter value over all queries (parameter value shown), and when each query focus subset was used - where the best parameter found for each subset of queries (e.g. HF, F, SF & MF) was used (denoted by "-Best", e.g., RBP-Best)[4].

[4]For Precision, $k = 1$ was best for all subsets.

Table 2: The mean (and standard error) of the absolute error for gain, time and depth over both folds. Super/subscripts indicate which metrics (by row number) were significantly different when using ANOVA/Tukey HSD testing.

		Mean Absolute Error		
	Metric	Gain	Time	Depth
1	AP	0.49±0.02	1.40±0.07	1.58±0.07
2	RR	0.29 ± 0.01^{1}	1.21±0.06	10.74±3.17
3	P@1	$0.18\pm0.01^{1,2}$	1.28±0.07	$0.65\pm0.03^{1,2}$
4	DCG(k=2)	$0.17\pm0.01^{1,2}$	1.10 ± 0.06^{1}	$0.50\pm0.03^{1,2}$
5	DCG-Best	$0.17\pm0.01^{1,2}$	1.15 ± 0.06^{1}	$0.50\pm0.03^{1,2}$
6	RBP(θ=0.3)	$0.15\pm0.01^{1,2,8,9}_{12,13}$	$1.08\pm0.06^{1,3}_{8,9}$	$0.47\pm0.03^{1,2}$
7	RBP-Best	$0.15\pm0.01^{1,2,8,9}_{12,13}$	$1.07\pm0.06^{1,3}_{8,9}$	$0.45\pm0.03^{1,2}$
8	TBG(h=0.0625)	$0.18\pm0.01^{1,2}$	1.28±0.07	$0.65\pm0.03^{1,2}$
9	TBG-Best	$0.17\pm0.01^{1,2}$	1.18±0.06	$0.55\pm0.03^{1,2}$
10	INST(T=1.0)	$0.19\pm0.01^{1,2}$	1.19 ± 0.06^{1}	$0.94\pm0.03^{1,2}$
11	INST-Best	$0.19\pm0.01^{1,2}$	1.15 ± 0.06^{1}	$0.83\pm0.03^{1,2}$
12	BPM(T=0.8,K=2)	$0.18\pm0.01^{1,2}$	1.09 ± 0.06^{1}	$0.49\pm0.03^{1,2}$
13	BPM-Best	$0.18\pm0.01^{1,2}$	1.15 ± 0.06^{1}	$0.54\pm0.03^{1,2}$
14	IFT(T=0.2,A=0.02)	$\mathbf{0.11\pm0.01}^{1-13}_{16,17}$	1.11 ± 0.06^{1}	$0.41\pm0.03^{1,2}$
15	IFT-Best	$\mathbf{0.11\pm0.01}^{1-13}_{16,17}$	1.10 ± 0.06^{1}	$0.41\pm0.03^{1,2}$
16	DDM-P	$0.16\pm0.01^{1,2}_{8,11}$	1.12 ± 0.06^{1}	$0.49\pm0.03^{1,2}$
17	DDM-F	$0.15\pm0.01^{1,2}_{8,11}$	$1.06\pm0.06^{1,3}_{8,9}$	$0.48\pm0.03^{1,2}$
18	DDM-R	$0.12\pm0.01^{1-17}$	$1.03\pm0.06^{1,3}_{8,9}$	$0.38\pm0.02^{1,2}$
19	DDM-T	$0.12\pm0.01^{1,17}$	$\mathbf{1.02\pm0.05}^{1,3}_{8,9}$	$\mathbf{0.32\pm0.02}^{1,2}$

With respect to predicting total gain, AP and RR, as expected, were the least accurate, while the IFT, and DDM-R and DDM-T metrics were the most accurate. An ANOVA between metrics ($F(12, 11882) = 245.79, p < 0.001$) indicated a significant difference between gain error rates - and the follow-up Tukey HSD tests revealed numerous differences. In Table 2, we have denoted which metrics were significantly better by including a superscript next to the value. From the table, we can see that most metrics were significantly better than AP and RR, and that RBP was significantly better than INST and BPM. DDM-R, DDM-T were significantly better than all other metrics, while IFT was significantly better than all other metrics other than DDM-R and DDM-T.

In terms of total time (cost), an ANOVA also revealed differences between metrics ($F(12, 11882) = 6.99, p < 0.001$) but to a much lesser extent. Tukey HSD follow-up tests showed that AP was generally the worst. While the DDMs tended to perform the best, and DDM-T gave the lowest error overall, the differences were only significant against AP, P, and TBG. Finally, in terms of depth, the ANOVA ($F(12, 11882) = 20.23, p < 0.001$) reported that there were differences – but the follow-up tests showed only that AP and RR were significantly worse than all other metrics. However, from the table, we can see that the DDM-T metric resulted in the lowest mean absolute error for depth.

Taken together, these show that DDM-T tends to provide the most accurate estimates across gain, time and depth – without requiring relevance information, or requiring any extensive tuning.

6.1 Correlation Analysis

To answer our third research question, we also performed a correlation analysis. Here we wanted to rank the queries by observed gain, time and depth, and see how well each metric predicted the ranking of queries. This is so we can identify which queries perform poorly relative to other queries. If this can be done successfully, then attention and resources can be directed to improve these queries. Given the metrics employed, we calculated Pearson's correlation co-efficient between the rankings given by gain, time, and depth. Table 3 reports the Pearson's r co-efficient for each metric (along with the 95% confidence intervals)[5].

In terms of ranking by gain, we see that AP and RR are the poorest predictors (though RR is somewhat better than AP). On the other hand, the DDM metrics are consistently high with DDM-R and DDM-T providing the highest correlations followed very closely by IFT with r values of 0.895 to 0.9. The remaining metrics still provide moderate to high correlations, varying between 0.825 and 0.864. In terms of ranking by time spent, again we see the DDM metrics and IFT provided the highest correlations, though substantially lower (0.587–0.694), reflecting the difficulty in predicting how long users spend on SERPs. The other metrics show at best moderate correlations between 0.423 (for DCG) up to 0.582 (for RBP). Finally, in terms of ranking by predicted depth, we first note that for certain "static" metrics the expected depth is always the same, i.e., P@$k = 1$, DCG@$k = 2$, RBP and DDM-P. This is because W_1 is the same regardless of query or SERP, and so the expected depth is constant (see Eq. 3). This means there is no correlation (denoted by "-"). For the adaptive metrics we can see varying correlations, with AP and RR being surprisingly high, but not quite as high as the data-driven metrics. Interestingly, TBG and INST report negative correlations. This is because they tend to overestimate the depth for many queries, and underestimate the depth on a small handful (skewing the correlation to be negative). This shows that despite predicting gain quite well they were poor at predicting the depth.

To test whether the metric(s) with the highest correlation for gain (DDM-R, DDM-T), time (DDM-F) and depth (DDM-T) were significantly better than the other metrics, we performed Pearson and Filon's co-correlations test. We used the dependent groups comparison as the samples were both correlated against the actual values, and used $p < 0.05$ [15]. We found that for the gain correlations, both DDM-R and DDM-T were significantly better than all other metrics (denoted by the superscripts). For the time correlations, DDM-F was significantly better than all other metrics, except AP. For depth, DDM-T was significantly better than all other metrics except AP.

These results show that while some metrics correlate highly in terms of gain, they may be less accurate with respect to time and depth – which are other important factors which influence satisfaction [24, 25]. These findings taken together with the findings from the error analysis suggest the data-driven metrics, in particular DDM-R and DDM-T, do consistently well across gain, time, and depth prediction/correlations.

[5]Statistical testing revealed that all correlations were significantly different from zero, except for P, RR, RBP, INST, and TBG on depth.

Table 3: Pearson's correlation coefficients (and confidence intervals) for each metric when ranking by Gain, Time, or Depth. Significance is denoted by super/subscripts.

		Correlations		
	Metric	**Gain**	**Time**	**Depth**
1	AP	0.39±0.05	0.55±0.05	0.66±0.04
2	RR	0.67±0.04	0.49±0.05	0.07±0.06
3	P@$k = 1$	0.82±0.02	0.58±0.04	-
4	DCG@$k = 2$	0.86±0.02	0.42±0.05	-
5	DCG-Best	0.78±0.03	0.64±0.04	0.46±0.05
6	RBP($\theta = 0.3$)	0.86±0.02	0.56±0.04	-
7	RBP-Best	0.86±0.02	0.59±0.04	0.32±0.06
8	TBG(h=0.0625)	0.82±0.02	0.58±0.04	-0.52±0.05
9	TBG-Best	0.81±0.02	0.49±0.05	0.25±0.06
10	INST(T=1.0)	0.84±0.02	0.33±0.06	-0.46±0.05
11	INST-Best	0.83±0.02	0.39±0.05	-0.33±0.06
12	BPM(T=0.8,K=2.0)	0.83±0.02	0.49±0.05	0.28±0.06
13	BPM-Best	0.82±0.02	0.38±0.06	0.10±0.06
14	IFT(T=0.2,A=0.05)	0.89±0.01	0.61±0.04	0.54±0.05
15	IFT-Best	0.89±0.01	0.60±0.04	0.52±0.05
16	DDM-P	0.85±0.02	0.59±0.04	-
17	DDM-F	0.83±0.02	$\mathbf{0.69{\pm}0.03}^{2-16}_{18}$	0.54±0.05
18	DDM-R	$\mathbf{0.90{\pm}0.01}^{1-17}$	0.59±0.04	0.52±0.05
19	DDM-T	$\mathbf{0.90{\pm}0.01}^{1-17}$	0.60±0.04	$\mathbf{0.69{\pm}0.03}^{2-18}$

7 DISCUSSION AND FUTURE WORK

Traditionally, metrics have been model-based – making various assumptions about user browsing behavior encapsulated as rules or a function. In this paper, we proposed a data-driven approach to building the continuation functions that underlie metrics. While DDMs require a representative sample of user behavior data to be estimated, they require little or no parameter tuning, and make no assumptions regarding user behavior, *a priori*. To determine whether data-driven metrics can improve the accuracy of estimated performance, we performed a large-scale analysis in the context of web search for popular queries. Our results suggest that DDMs can provide some of the most accurate estimates of gain, time, and depth. On the other hand, the model-based metrics could, when tuned extensively, provide comparable estimates (IFT was the best model-based metric, but has six parameters). While model-based metrics provide a mechanism to simulate a variety of different possible user browsing behaviors – and thus let one explore how performance would vary under different user browsing models – data-driven metrics let us evaluate with the observed user behaviors and brings us a step closer to bridging the divide between online and offline metrics for evaluation [21].

As SERPs have become increasingly complex, composed of a variety of heterogeneous items, they have created new evaluation challenges [8]. In this work, we have explored the influence of item type on people's continuation behavior (and compared this to other features, e.g., position, query focus, and relevance). This led to new insights regarding user's browsing behavior. For example, we have shown that presenting highly relevant items decreases the probability of continuing, but for highly focused queries this probability is much lower. Intuitively, this make sense – if a user finds the highly relevant item then they will click and leave. Conversely, other types

of items result in different behaviors; most obviously, when adverts are presented there is a greater probability of continuing. Video items have a high but similar probability of continuing regardless of position, whereas for entities it decreases over position, and for web results and images the probability of continuing increases with position. These findings also suggest that there is an interaction between the type of items, their position in the SERP, and user behavior – which needs to be considered if results are to be ordered optimally [18]. While newer metrics that factor in the cost of items may implicitly cater for different types [6, 32, 39, 41], they also hide the influence that specific item types are having on behavior. Thus, they tend to be not only less accurate (unless extensively tuned), but also less informative for designers. In contrast, when item type information was used, DDM-T was much more accurate overall, providing very low (if not lowest) errors when predicting gain, time, and depth, while consistently providing very high correlations when ranking by gain, time, and depth. A key benefit of DDM-T is that no relevance information is required *a priori* to estimate the **C** function, which means cost and depth estimates can be obtained for pages lacking relevance judgments. Furthermore, these findings suggest that item types should be explicitly considered, especially if we are interested in measuring performance given the whole SERP (not just a subset/list of homogeneous items on the SERP). Given that relevance and query type also provided very good predictions of performance, it also suggests building estimates of the continuation probabilities given these dependencies.

With the present study, there are a number of limitations and issues that need to be considered when using/applying data-driven metrics. Firstly, our analysis has only been performed in a particular context (web search with popular queries) on a specific interface. Thus, it is an open question how well these continuation functions generalize to other domains. It is likely that interfaces with different items, layouts, etc. which are used for different tasks will result in different browsing behaviors. However, the proposed approach is designed such that a representative sample of user behavior is required to build a metric that is specific to the interface/task/etc. This brings up the next issue, which is, how much data is required to estimate the **C** function reliably and how representative does the sample need to be to produce robust estimates of performance? And, what types of smoothing functions many be applicable here in order to apply the method to less data rich contexts? Another limitation of this work is that our sample only contained SERPs where at least one click was associated with the query impression. Thus, it is an open question as to how these metrics can be applied more generally, and whether they can be used to also measure when there is good and bad abandonment? This might require a reformulation of the **C**/**W**/**L** framework, or require techniques to infer the gain from other signals (i.e., depth and time). Another interesting issue that arises from this work, which is more philosophical, and is regarding the nature of a metric – should they be data-driven (based directly on user behavior) or model-driven (based on some model of user behavior, whether that is hypothesized or inspired from observation)? On one hand, data-driven metrics provide an indication of the performance based on how people have behaved with the system. On the other hand, with model-driven metrics it is possible to explore how performance would vary given the different parameters that encode different ways in which the user can browse through results. So if we want to evaluate how good performance is if the user were to behave like the model suggests then model-driven metrics provide an appropriate mechanism – but if we want to evaluate how good performance is with the current user behavior then data-driven metrics may be more appropriate. However, we have only compared these metrics with respect to how well they predict observables and it would be interesting to explore whether the different metrics fair better across other properties (e.g., robustness, fidelity, correlation with satisfaction, etc.). These limitations motivate several lines of further research:

- estimating continuation probabilities from data using more sophisticated estimators and smoothing functions;
- approximating a functional form given the different factors to create grounded model-based metrics when data is not available, which generalize to other scenarios;
- examining other factors that may influence browsing behaviors such as layouts and devices (e.g., grids, regions, mobile, etc. [30, 32, 45]), other query types (e.g. torso, tail, informational, etc. [9]), and within different search tasks (e.g., novelty, diversity, sessions, etc. [2, 3, 14, 27, 46]);
- exploring a variety of other search contexts and scenarios in which heterogeneous search results and recommendations are dispatched (e.g., products, news, etc.);
- extending the **C**/**W**/**L** framework to include click behavior, as the current model (shown in Fig. 1) implicitly assumes that users consume each item that they visit (i.e., clicks and read) and so this would bring offline evaluation closer to online;
- further extending the **C**/**W**/**L** framework to include measurements when there are no clicks observed i.e., good/bad abandonment [36] (which may be inferred based on the difference in expected gain, cost and depth, from pages with clicks and without clicks), and;
- evaluating the metrics with different methods to determine whether more accurate estimates of gain and cost lead to higher correlations with user search satisfaction [1, 20].

In summary, we explored how well data-driven metrics can estimate the continuation functions given position, query focus, relevance, and item type for web search. We evaluated how well DDMs compared to existing metrics in terms of how accurately they can infer the gain users accrue, how long they will spend, and how many items they examine on the SERP. Our findings showed that the DDM metrics based on relevance and items resulted in the most consistent and most accurate predictions and correlations. A key benefit of DDMs is that they directly link how we are modeling user behavior with how we are measuring it – providing deeper insights into how people interact with the different interfaces, layouts, and page compositions, rather than making assumptions about how we expect or believe they will behave. This invariably leads to metrics that can more accurately evaluate the whole page. As a result, this study has opened up numerous directions and challenges in designing and developing future metrics and their analysis.

Acknowledgements. We would like to thank Susan Dumais and Paul Bennett for their helpful feedback and suggestions on this work. We would also like to thank the anonymous reviewers for their insights, suggestions, and comments.

REFERENCES

[1] Azzah Al-Maskari and Mark Sanderson. 2010. A Review of Factors Influencing User Satisfaction in Information Retrieval. *JASIST* 61, 5 (May 2010), 859–868.

[2] Ameer Albahem, Damiano Spina, Falk Scholer, Alistair Moffat, and Lawrence Cavedon. 2018. Desirable Properties for Diversity and Truncated Effectiveness Metrics. In *Proceedings of the 23rd Australasian Document Computing Symposium (ADCS '18)*. Article Article 9, 7 pages.

[3] Enrique Amigó, Damiano Spina, and Jorge Carrillo-de Albornoz. 2018. An Axiomatic Analysis of Diversity Evaluation Metrics: Introducing the Rank-Biased Utility Metric. In *Proc. of the 41st International ACM SIGIR Conference*. 625–634.

[4] Jaime Arguello, Robert Capra, and Wan Ching Wu. 2013. Factors affecting aggregated search coherence and search behavior. In *Proc. SIGIR*. 1989–1998.

[5] Jaime Arguello, Wan-Ching Wu, Diane Kelly, and Ashlee Edwards. 2012. Task Complexity, Vertical Display and User Interaction in Aggregated Search. In *Proc. of the 35th ACM SIGIR Conference*. 435–444.

[6] Leif Azzopardi, Paul Thomas, and Nick Craswell. 2018. Measuring the Utility of Search Engine Result Pages: An Information Foraging Based Measure. In *Proc. of the 41st International ACM SIGIR Conference*. 605–614.

[7] Leif Azzopardi, Paul Thomas, and Alistair Moffat. 2019. Cwl_eval: An Evaluation Tool for Information Retrieval. In *Proc. of the 42nd International ACM SIGIR Conference*. 1321–1324.

[8] Peter Bailey, Nick Craswell, Ryen W. White, Liwei Chen, Ashwin Satyanarayana, and S.M.M. Tahaghoghi. 2010. Evaluating Whole-page Relevance. In *Proc. of the 33rd International ACM SIGIR Conference*. 767–768.

[9] Andrei Broder. 2002. A taxonomy of web search. In *ACM Sigir forum*, Vol. 36. ACM, 3–10.

[10] Georg Buscher, Susan T. Dumais, and Edward Cutrell. 2010. The Good, the Bad, and the Random: An Eye-tracking Study of Ad Quality in Web Search. In *Proc. of the 33rd ACM SIGIR Conferenc*. 42–49.

[11] Ben Carterette. 2011. System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation. In *Proc. of the 34th ACM SIGIR Conference*. 903–912.

[12] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *Proc. of the 18th ACM CIKM Conference*. 621–630.

[13] Zhicong Cheng, Bin Gao, and Tie-Yan Liu. 2010. Actively Predicting Diverse Search Intent from User Browsing Behaviors. In *Proc. of the 19th International Conference on World Wide Web (WWW '10)*. 221–230.

[14] Charles LA Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proc. of the 31st ACM SIGIR Conference*. 659–666.

[15] Birk Diedenhofen and Jochen Musch. 2015. cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *PLOS ONE* 10, 4 (04 2015), 1–12. https://doi.org/10.1371/journal.pone.0121945

[16] Zhicheng Dou, Ruihua Song, and Ji-Rong Wen. 2007. A large-scale evaluation and analysis of personalized search strategies. In *Proc. of the 16th WWW*. 581–590.

[17] Susan T. Dumais, Georg Buscher, and Edward Cutrell. 2010. Individual Differences in Gaze Patterns for Web Search. In *Proc. of the 3rd IIiX Symposium*. 185–194.

[18] Norbert Fuhr. 2008. A probability ranking principle for IIR. *Information Retrieval* 11, 3 (2008), 251–265.

[19] Norbert Fuhr. 2017. Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. *SIGIR Forum* 52, 2 (2017), 32–41.

[20] Ahmed Hassan, Xiolin Shi, Nick Craswell, and Bill Ramsey. 2013. Beyond clicks: Query reformulation as a predictor of search satisfaction. In *Proc. of the ACM CIKM Conference*. 2019–2028.

[21] Katja Hofmann, Lihong Li, and Filip Radlinski. 2016. Online Evaluation for Information Retrieval. *Foundations and Trends in Information Retrieval* 10, 1 (2016), 1–117.

[22] Jim Jansen, Danielle L. Booth, and Amanda Spink. 2007. Determining the user intent of web search engine queries. *Proc. of the 16th International World Wide Web Conference*, 1149–1150.

[23] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (Oct. 2002), 422–446.

[24] Jiepu Jiang and James Allan. 2016. Adaptive Effort for Search Evaluation Metrics. In *Proc. of the 38th European Conference on IR Research*. 187–199.

[25] Jiepu Jiang and James Allan. 2016. Correlation Between System and User Metrics in a Session. In *Proc. of the ACM CHIIR Conference*. 285–288.

[26] Jiepu Jiang and James Allan. 2017. Adaptive Persistence for Search Effectiveness Measures. In *Proc. of the ACM CIKM Conference (CIKM '17)*. 747–756.

[27] Evangelos Kanoulas, Ben Carterette, Paul D. Clough, and Mark Sanderson. 2011. Evaluating Multi-query Sessions. In *Proc. of the 34th ACM SIGIR Conference*. 1053–1062.

[28] Paul Kantor and Ellen Voorhees. 2000. The TREC-5 Confusion Track. *Information Retrieval* 2, 2-3 (2000), 165–176.

[29] Jaana Kekäläinen and Kalervo Järvelin. 2002. Using graded relevance assessments in IR evaluation. *JASIST* 53, 13 (2002), 1120–1129.

[30] Fei Liu, Alistair Moffat, Timothy Baldwin, and Xiuzhen Zhang. 2016. Quit While Ahead: Evaluating Truncated Rankings. In *Proc. of the 39th International ACM SIGIR Conference (SIGIR '16)*. 953–956.

[31] Lori Lorigo, Maya Haridasan, Hrönn Brynjarsdóttir, Ling Xia, Thorsten Joachims, Geri Gay, Laura Granka, Fabio Pellacini, and Bing Pan. 2008. Eye Tracking and Online Search. *JASIST* 59, 7 (2008), 1041–1052.

[32] Cheng Luo, Yiqun Liu, Tetsuya Sakai, Fan Zhang, Min Zhang, and Shaoping Ma. 2017. Evaluating Mobile Search with Height-Biased Gain. In *Proc. of the 40th International ACM SIGIR Conference*. 435–444.

[33] Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. 2015. INST: An Adaptive Metric for Information Retrieval Evaluation. In *Proc. of the 20th ADCS Conference*. Article 5, 4 pages.

[34] Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. 2017. Incorporating User Expectations and Behavior into the Measurement of Search Effectiveness. *ACM Transactions on Information Systems* 35, 3, Article 24 (2017), 38 pages.

[35] Alistair Moffat, Paul Thomas, and Falk Scholer. 2013. Users Versus Models: What Observation Tells Us About Effectiveness Metrics. In *Proc. of the 22nd ACM CIKM Conference*. 659–668.

[36] Alistair Moffat and Alfan Farizki Wicaksono. 2018. Users, Adaptivity, and Bad Abandonment. In *Proc. of the 41st International ACM SIGIR Conference*. 897–900.

[37] Alistair Moffat and Justin Zobel. 2008. Rank-Biased Precision for Measurement of Retrieval Effectiveness. *ACM Trans. on Information Systems* 27, 1 (2008), 2:1–2:27.

[38] Stephen E. Robertson, Evangelos Kanoulas, and Emine Yilmaz. 2010. Extending Average Precision to Graded Relevance Judgments. In *Proc. of the 33rd International ACM SIGIR Conference*. 603–610.

[39] Tetsuya Sakai and Zhicheng Dou. 2013. Summaries, Ranked Retrieval and Sessions: A Unified Framework for Info. Access Evaluation. In *Proc. of the 36th ACM SIGIR Conference*. 473–482.

[40] Mark Sanderson. 2010. Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval* 4, 4 (2010), 247–375.

[41] Mark D. Smucker and Charles L.A. Clarke. 2012. Time-Based Calibration of Effectiveness Measures. In *Proc. of the 35th ACM SIGIR Conference*. 95–104.

[42] Jaime Teevan, Susan T Dumais, and Daniel J Liebling. 2008. To personalize or not to personalize: modeling queries with variation in user intent. In *Proc. of the 31st international ACM SIGIR Conference*. ACM, 163–170.

[43] Ellen Voorhees and Donna Harman. 2005. *TREC: Experiment and Evaluation in Information Retrieval*. The MIT press.

[44] Alfan Farizki Wicaksono and Alistair Moffat. 2018. Empirical Evidence for Search Effectiveness Models. In *Proc. of the 27th ACM CIKM Conference*. 1571–1574.

[45] Xiaohui Xie, Jiaxin Mao, Maarten de Rijke, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2018. Constructing an Interaction Behavior Model for Web Image Search. In *Proc. of the 41st International ACM SIGIR Conference*. 425–434.

[46] Grace Hui Yang, Xuchu Dong, Jiyun Luo, and Sicong Zhang. 2018. Session search modeling by partially observable Markov decision process. *Information Retrieval Journal* 21, 1 (2018), 56–80.

[47] Emine Yilmaz, Milad Shokouhi, Nick Craswell, and Stephen Robertson. 2010. Expected browsing utility for web search evaluation. In *Proceedings of the 19th ACM SIGIR Conference*. 1561–1564.

[48] Fan Zhang, Yiqun Liu, Xin Li, Min Zhang, Yinghui Xu, and Shaoping Ma. 2017. Evaluating Web Search with a Bejeweled Player Model. In *Proc. of the 40th ACM SIGIR Conference*. 425–434.