A Flexible Framework for Offline Effectiveness Metrics

Alistair Moffat
The University of Melbourne
Melbourne, Australia
ammoffat@unimelb.edu.au

Joel Mackenzie
The University of Queensland
Brisbane, Australia
joel.mackenzie@uq.edu.au

Paul Thomas
Microsoft
Canberra, Australia
pathom@microsoft.com

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

ABSTRACT

The use of offline effectiveness metrics is one of the cornerstones of evaluation in information retrieval. Static resources that include test collections and sets of topics, the corresponding relevance judgments connecting them, and metrics that map document rankings from a retrieval system to numeric scores have been used for multiple decades as an important way of comparing systems. The basis behind this experimental structure is that the metric score for a system can serve as a surrogate measurement for user satisfaction.

Here we introduce a user behavior framework that extends the C/W/L family. The essence of the new framework – which we call C/W/L/A – is that the user actions that are undertaken while reading the ranking can be considered separately from the benefit that each user will have derived as they exit the ranking. This split structure allows the great majority of current effectiveness metrics to be systematically categorized, and thus their relative properties and relationships to be better understood; and at the same time permits a wide range of novel combinations to be considered.

We then carry out experiments using relevance judgments, document rankings, and user satisfaction data from two distinct sources, comparing the patterns of metric scores generated, and showing that those metrics vary quite markedly in terms of their ability to predict user satisfaction.

CCS CONCEPTS

• **Human-centered computing** → *User models*; • **Information systems** → **Task models**; **Retrieval effectiveness**; **Presentation of retrieval results**.

KEYWORDS

User browsing model; effectiveness metric; offline evaluation

ACM Reference Format:
Alistair Moffat, Joel Mackenzie, Paul Thomas, and Leif Azzopardi. 2022. A Flexible Framework for Offline Effectiveness Metrics. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3477495.3531924

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531924

1 INTRODUCTION

The use of offline effectiveness metrics is one of the cornerstone approaches to evaluation in information retrieval [34], the other two being laboratory-based user studies with relatively small numbers of subjects [21], and large-scale in-situ observation of operational systems with relatively large numbers of subjects, and the ability to run A/B processing streams [19]. An offline evaluation makes use of a set of static resources: one or more test collections and corresponding topic sets (queries); sets of relevance judgments connecting the documents in each collection and the topics applicable to it; and one or more *effectiveness metrics*, each of which takes a document ranking (a *search engine result page*, or *SERP*) that has been generated by a retrieval system and the corresponding relevance judgments, and from them constructs a numeric score [18].

The offline experimental methodology has been used for more than fifty years as an important way of comparing systems [11]. The basis behind it is the assumption that the metric score (or scores, if more than one metric is applied) for a system can serve as a surrogate measurement for the underlying attribute of SERP *usefulness* as expressed via *user satisfaction*, the degree to which the SERP addresses the information need that prompted the user's query and thus helps them carry out some task. Consideration of a pool of representative topics then permits statistical inferences to be drawn in regard to the relative performance of systems, allowing the field to progress. Sanderson [34] provides an overview of offline experimental evaluation, and the resources that it employs.

The C/W/L framework of Moffat et al. [26, 27] (itself motivated in part by earlier work by Moffat and Zobel [25], Yilmaz et al. [42], Wang et al. [39], and Carterette et al. [7]) describes a class of effectiveness metrics – the C/W/L family – in terms of the aggregate behavior of a pool of users as they examine each SERP generated by the system. Each individual user is assumed to begin viewing the SERP at the top-ranked item, and to proceed sequentially from one item to the next, moving through the SERP until they discontinue their perusal. They then issue a reformulated query, change search mode (perhaps using a different service), or resume the task that prompted their information need. While this sequential user browsing model is only an approximation to true user behavior within a search interface – see, for example, Thomas et al. [38] – it nevertheless captures the essence of the users' interactions with

the ranked lists of documents that form the output from the core retrieval system, before they are intermingled with elements from other search verticals such as images, knowledge graph entries, "did you mean" suggestions, and so on [12].

The key element of the C/W/L framework is that it formalizes this user browsing model through the use of a *conditional continuation probability*, denoted $C(i)$, the probability that a user who has examined the item at rank i in the SERP will go on to also examine the item at rank $i+1$. From $C(i)$ the fraction of user attention $W(i)$ paid to each rank in the SERP can be calculated; and hence the expected aggregate per-user gain of utility derived from the SERP can be computed as the sum of that user's observed per-document gains once the SERP's corresponding relevance judgments are factored in. Details of this approach are provided in Section 2.

Section 3 then introduces a user behavior framework that extends the previous C/W/L approach. The new framework – which we call C/W/L/A – adds an *aggregation* function $A(i)$ to the computation, to represent the derived benefit that users receive when they exit the SERP at rank i and have thus viewed the first i items in the SERP. One obvious aggregation function is to sum the gain scores of the documents through to depth i; that approach yields all of the "pure C/W/L" metrics, as a proper subset of the C/W/L/A metrics. But other aggregation rules are also possible, and via them the new framework allows the great majority of current effectiveness metrics to be systematically categorized, and thus for their relative properties and relationships to be better understood.

Section 4 then uses the C/W/L/A framework to suggest new evaluation options, and considers the properties and applicability of those alternatives. We also carry out experiments using relevance judgments, document rankings, and user satisfaction data from two distinct sources including data derived from a large commercial search service. That data allows us to compare the scores generated from the very wide range of effectiveness metrics described by the C/W/L/A framework with whole-of-SERP quality assessments; with the results showing that a metric's ability to predict self-reported user-satisfaction and judge-assessed SERP quality values can vary quite markedly.

2 BACKGROUND

A range of mechanisms have been suggested to categorize offline effectiveness metrics [6, 31, 43]. In this section we present the C/W/L framework of Moffat et al. [26, 27], and provide examples showing its use and limitations. We also show how the framework encapsulates many existing effectiveness metrics, but fails to adequately describe other commonly used metrics, including ERR and Succ@k.

C/W/L and ERG Metrics. First, define $0 \le C(i) \le 1$ to be the *conditional continuation probability* of a user who has examined the item at rank i in a SERP also viewing the item at rank $i+1$ (rather than exiting the SERP) [25, 26]. Any particular specification of a function $C(i)$ defines a *user browsing model* – the manner in which the user consumes the items in the SERP. For now, we work with arbitrary functions $C(i)$ and place no constraints on their properties; factors that might influence $C(i)$ are discussed later.

From $C(i)$ we calculate $V(i)$, the fraction of users that view the document at rank i, starting with the assumption that $V(1) = 1$ and every user examines the first item in the ranking, and computed thereafter as the product of the fraction that view the $i-1$ st document and the fraction that continue from depth $i-1$ to depth i:

$$V(i) = \prod_{j=1}^{i-1} C(j)\,. \tag{1}$$

The expected number of items viewed per user is then given by:

$$V^+ = \sum_{i=1}^{\infty} V(i)\,. \tag{2}$$

Note that $C(\cdot)$ should be such that V^+ is finite and calculable for any particular ranking. This requirement is satisfied if $C(i) = 0$ at least once, and also if $\exists \epsilon > 0$ such that $\{i \mid C(i) \ge 1-\epsilon\}$ is finite.

Once V^+ is known, $W(i)$, the fraction of user attention paid to the i th item in the ranking, can be calculated:

$$W(i) = \frac{V(i)}{V^+}\,. \tag{3}$$

In the C/W/L framework [27], *expected rate of gain* (ERG) metrics have units of "expected utility accrued per item inspected" with the expectation relative to the attention fractions implied by $W(\cdot)$:

$$M_{\text{CWL-ERG}}(\mathbf{r}) = \sum_{i=1}^{\infty} W(i) \cdot r_i\,, \tag{4}$$

in which $0 \le r_i \le 1$ is the gain (or *utility*) associated with the i th item in the ranking, ranging from "not at all relevant" ($r_i = 0$) to "perfectly relevant" ($r_i = 1$). Table 1 provides a worked example, assuming a set of $C(i)$ values taken to be a model of user browsing behavior, and a set of r_i gains associated with some ranking of interest. Note that because $C(6)$ happens to be zero, $V(7)$ and beyond are also zero, and finite sums are sufficient to calculate V^+ and $M_{\text{CWL-ERG}}(\cdot)$ exactly. In more general cases in which $C(i)$ is always greater than zero, the sum V^+ is computed either via a closed form for the corresponding infinite sum, when $C(i)$ has an amenable form; or to some required degree of precision by summing a suitable number of terms when $C(i)$ does not. At the same time, if not all corresponding values of r_i are known, the weights $W(i)$ attached to the unknown ones can be accumulated as a *residual* [25], indicating the extent of the imprecision in the computed metric score.

Table 1: Example of C/W/L ERG computation. The columns headed r_i and $C(i)$ are arbitrary, everything else follows from those values. The first of the four sums is V^+, the last is the ERG metric value.

i	r_i	$C(i)$	$V(i)$	$L(i)$	$W(i)$	$W(i) \cdot r_i$
1	0.7	0.8	1.000	0.200	0.239	0.167
2	0.4	1.0	0.800	0.000	0.191	0.076
3	0.0	1.0	0.800	0.000	0.191	0.000
4	1.0	0.7	0.800	0.240	0.191	0.191
5	0.5	0.4	0.560	0.336	0.134	0.067
6	0.3	0.0	0.224	0.224	0.054	0.016
sums			4.184	1.000	1.000	0.518

C/W/L and ETG Metrics. The third component of the C/W/L functions, $L(i)$, is the probability that the item at rank i will be the *last* one viewed by any given user:

$$L(i) = V(i) - V(i+1) = V(i)\,(1 - C(i))\ . \tag{5}$$

Note that $\sum_{i=1}^{\infty} L(i) = 1$ and that $L(\cdot)$ is a probability distribution over a population of users, with their expected exit rank given by:

$$\sum_{i=1}^{\infty} i \cdot L(i) = \sum_{i=1}^{\infty} i \cdot (V(i) - V(i+1)) = \sum_{i=1}^{\infty} V(i) = V^{+}\,.$$

With $L(\cdot)$ so defined it becomes possible to compute the total utility gained by an average user given the model of browsing as the *expected total gain* (ETG) metric score:

$$M_{\text{CWL-ETG}}(\mathbf{r}) = \sum_{i=1}^{\infty} L(i) \cdot \left(\sum_{j=1}^{i} r_j \right). \tag{6}$$

Note that Equations 4 and 6 imply the relationship:

$$M_{\text{CWL-ETG}}(\mathbf{r}) = V^{+} \cdot M_{\text{CWL-ERG}}(\mathbf{r})\,,$$

with the expectation in ERG metrics being "*per item inspected, over all document views undertaken by all users*", and the expectation in ETG metrics being "*over all users*".

Defining Metrics. If $C(\cdot)$ is a conditional continuation function, then the C/W/L framework in essence provides two operators, $\text{CWL}_{\text{ERG}}(\cdot)$ and $\text{CWL}_{\text{ETG}}(\cdot)$, that map $C(\cdot)$ to two corresponding batch evaluation metrics. Conversely, many current metrics can be completely described via specification of their underlying $C(\cdot)$ function [26, 27]:

- Precision at depth k, Prec@k, is a CWL-ERG metric defined by

$$C_{\text{Prec@}k}(i) = \begin{cases} 1 & \text{if } i < k \\ 0 & \text{otherwise}\,. \end{cases} \tag{7}$$

  That is, Prec@$k \equiv \text{CWL}_{\text{ERG}}(C_{\text{Prec@}k})$. While Prec@$k$ is typically regarded as applying only to binary relevance grades $r_i \in \{0, 1\}$, this definition via $\text{CWL}_{\text{ERG}}(\cdot)$ also handles real-valued gains $0 \le r_i \le 1$ in a natural manner.

- Rank-biased precision with persistence parameter $0 \le \phi < 1$, RBP@ϕ [25], is also a CWL-ERG metric, and is defined by

$$C_{\text{RBP@}\phi}(i) = \phi\,. \tag{8}$$

  That is, RBP@$\phi \equiv \text{CWL}_{\text{ERG}}(C_{\text{RBP@}\phi})$.

- Discounted cumulative gain to depth k, DCG@k [20], is a CWL-ETG metric described by

$$C_{\text{DCG@}k}(i) = \begin{cases} \log_2(i+1)/\log_2(i+2) & \text{if } i < k \\ 0 & \text{otherwise}\,. \end{cases} \tag{9}$$

  As is also the case with RBP@ϕ, the gain values r_i are real-valued. The corresponding CWL-ERG metric normalizes by the DCG@k score of a ranking in which $r_i = 1$ for all $1 \le i \le k$, and is sometimes referred to as *scaled DCG*, SDCG@k, with metric values between zero and one; see, for example, Moffat et al. [26].

- Normalized discounted cumulative gain, NDCG@k [20], is computed as DCG@k normalized by a different constant (the ideal DCG@k score for this query at depth k, given knowledge of all relevant or partially relevant documents) and is thus also an ERG-like metric. Because the ideal DCG@k score used to derive NDCG@k from DCG@k is always less than the all-fully-relevant normalization used to compute SDCG@k, for any given topic NDCG@k is larger than SDCG@k by a fixed multiple.

- Average precision, AP [5], is a CWL-ERG metric for binary relevance grades $r_i \in \{0, 1\}$, defined by

$$C_{\text{AP1}}(i) = \frac{\sum_{j=i+1}^{\infty} (r_j / j)}{\sum_{j=i}^{\infty} (r_j / j)}\,. \tag{10}$$

  That is, AP $\equiv \text{CWL}_{\text{ERG}}(C_{\text{AP1}})$ describes the exact same metric as the standard definition of AP via the average of the precisions at the points in the ranking at which the relevant documents occur:

$$\text{AP}(\mathbf{r}) = \frac{1}{R} \sum_{i=1}^{\infty} \left(r_i \cdot \text{Prec@}i(\mathbf{r}) \right), \tag{11}$$

  in which $R = \sum_{i=1}^{\infty} r_i$ is the total number of relevant documents in the collection. As was the case with Prec@k, this definition includes a natural extension to real-valued gains $0 \le r_i \le 1$, an enhancement that has also been considered directly [13, 29]. (A second continuation function that describes AP is introduced in Section 3 and is denoted C_{AP2}.)

- Reciprocal Rank, RR, is a CWL-ERG metric defined for binary relevance grades by

$$C_{\text{RR}}(i) = 1 - r_i\,. \tag{12}$$

- INST is a CWL-ERG metric defined for graded relevance via the continuation function [27]:

$$C_{\text{INST@}T}(i) = \left(\frac{i + T + T_i - 1}{i + T + T_i} \right)^2 \tag{13}$$

  where T is a parameter governing the user's browsing behavior and corresponds to the total gain that they seek to reap as a result of their query, and where T_i is the amount of that nominal gain target that remains unfound after the ith entry in the SERP has been considered, that is, $T_i = T - \sum_{j=1}^{i} r_j$.

Note that INST@T is an *adaptive* metric, as is RR, in that the gain values associated with the items already viewed will influence the user's likelihood of proceeding to the next item in the SERP. On the other hand, the browsing models associated with Prec@k, RBP@ϕ, and DCG@k are *static*, and assume that the user's actions are completely unaffected by the quality of the documents that they observe as they proceed through the SERP. Item gain values r_i also influence the continuation conditional probabilities associated with C_{AP1}, but it is *future* gains that are taken into account, rather than observed gains. That is, AP doesn't assume users are adaptive, it assumes that they are *clairvoyant*.

A range of further metrics can also be described with the C/W/L framework [2] such as Time Biased Gain (TBG) TBG [36, 37]; the U-Measure[31]; the Bejewelled Player Model (BPM) [43]; Information Foraging Theory (IFT) [1]; along with data driven approaches to estimate the C function directly [3, 40, 45].

Expected Reciprocal Rank. On the other hand, there are also metrics that do *not* have corresponding explanations via the C/W/L framework. One such metric is expected reciprocal rank, ERR [8],

defined for real-valued gain values $0 \le r_i \le 1$ as

$$\mathrm{ERR}(\mathbf{r}) = \sum_{i=1}^{\infty} \left(\frac{r_i}{i} \cdot \prod_{j=1}^{i-1} (1 - r_j) \right). \tag{14}$$

When all of the r_i values are binary, $\mathrm{ERR}(\mathbf{r}) = \mathrm{RR}(\mathbf{r})$. But the added flexibility of ERR compared to RR in allowing real-valued gain values (for example, $r_i = 0.5$ as "partially relevant") breaks that connection, and means that ERR cannot be described via a $C(\cdot)$ function. Azzopardi et al. [4] demonstrate that fact, and consider a suite of C/W/L-compliant metrics that might be used to approximate (but not match) ERR's behavior. Similarly, the metric Succ@k – which measures the degree to which *any* document in the top k matches the query – instantiates a different gain accumulation assumption, and cannot be described by the C/W/L framework.

3 GENERALIZING C/W/L TO GET C/W/L/A

In this section we develop a more flexible approach – the C/W/L/A structure – that accommodates ERR and Succ@k directly. More importantly, that flexibility opens a path to a broad range of other batch evaluation metrics, many of which are considered in Section 4.

Proposed Generalization. The C/W/L expected total gain metrics, defined via Equation 6, are computed as an inner product of two components: (1) the fraction of users who exit the SERP at depth i, $L(i)$; and (2) the sum of the gains observed by a user who traverses ranks 1 through to i. That second quantity – $\sum_{j=1}^{i} r_j$ – represents that user's aggregate benefit, derived from their particular journey through the SERP. It has the role of quantifying how each user "feels" as they move on to their next task, using the information that they gleaned via their search; and hence if $C(\cdot)$ summarizes the universe of users' *actions* in regard to that query and that SERP, then $\sum_{j=1}^{i} r_j$ is an assessment of the universe of users' *reactions*. That is, the C/W/L framework assumes that each of the SERP elements that the user observes has equal and independent merit in determining their overall reaction, and that per-item relevance utilities should be summed to compute user reaction.

Our key observation in this paper is that the aggregation process need not be fixed, and can instead be given its own identity, with the summation $\sum_{j=1}^{i} r_j$ one option, but others also possible. That is, we now separate the "continuation" part of the C/W/L framework, the description of user *action*, from the derived benefit part, to independently account for user *reaction*.

Define $C(\cdot)$ and $L(\cdot)$ as above. Each user in the population is assumed to traverse the ranking from the top item though to their exit point, with fraction $L(i)$ of the user population exiting immediately after viewing the item at rank i. Those users will have viewed documents 1 to i inclusive, and are assumed to have constructed an *aggregate* gain value $A(\langle r_1, r_2, \ldots, r_i \rangle)$ from the corresponding item gains. Using the shorthand $A(i)$ for $A(\langle r_1, r_2, \ldots, r_i \rangle)$, a new family of metrics can then be specified via:

$$M_{\mathrm{CWLA}}(\mathbf{r}) = \sum_{i=1}^{\infty} L(i) \cdot A(i) \,. \tag{15}$$

We further define $\mathsf{CWLA}(\cdot, \cdot)$ to be an operator that uses Equations 1, 5, and 15, to combine a $C(\cdot)$ and an $A(\cdot)$ function to create an effectiveness metric in which the measured quantities are in units of "average aggregate gain per user".

Table 2: Example of generalized C/W/L/A computation using A_{avg} (Equation 19). The columns headed r_i and $C(i)$ are arbitrary (as in Table 1), the other values then follow.

i	r_i	$C(i)$	$V(i)$	$L(i)$	$A_{\mathrm{avg}}(i)$	$L(i) \cdot A(i)$
1	0.7	0.8	1.000	0.200	0.700	0.140
2	0.4	1.0	0.800	0.000	0.550	0.000
3	0.0	1.0	0.800	0.000	0.367	0.000
4	1.0	0.7	0.800	0.240	0.525	0.126
5	0.5	0.4	0.560	0.336	0.520	0.175
6	0.3	0.0	0.224	0.224	0.483	0.108
sums			4.184	1.000		0.549

To capture the aggregation function already referred to, the per-item gains can be summed:

$$A_{\mathrm{ETG}}(i) = \sum_{j=1}^{i} r_j \,. \tag{16}$$

With this formulation for $A(i)$ we then have

$$\mathsf{CWL}_{\mathrm{ETG}}(C) \equiv \mathsf{CWLA}(C, A_{\mathrm{ETG}}) \,,$$

that is, the C/W/L expected total gain (ETG) metrics are a subset of the C/W/L/A metrics. For example, DCG@$k \equiv \mathsf{CWLA}(C_{\mathrm{DCG@}k}, A_{\mathrm{ETG}})$.

Another viable option is to take:

$$A_{\mathrm{ERG}}(i) = \frac{1}{V^+} \sum_{j=1}^{i} r_j \,, \tag{17}$$

from which it is clear that:

$$\mathsf{CWL}_{\mathrm{ERG}}(C) \equiv \mathsf{CWLA}(C, A_{\mathrm{ERG}}) \,;$$

that is, the C/W/L expected rate of gain (ERG) family are also a subset of the new larger family: RBP@$\phi \equiv \mathsf{CWLA}(C_{\mathrm{RBP@}\phi}, A_{\mathrm{ERG}})$, for example.

The remainder of this section revisits a number of existing metrics using the C/W/L/A structure; considers several further $A(\cdot)$ functions that open up other possibilities; presents the broad range of metrics that can then be addressed; and then finally compares our taxonomy to other categorizations that have been proposed.

Expected Reciprocal Rank Revisited. As a first example, ERR is readily defined in the C/W/L/A framework. One part of it has already been provided as C_{ERR} (Equation 12). However, as shown by Azzopardi et al. [4], $\mathsf{CWL}(C_{\mathrm{RR}})$ does *not* define ERR. But an $A(\cdot)$ function that captures how the user "feels" when they leave the ranking after inspecting i items can fill that gap. In the case of ERR the required expression is:

$$A_{\mathrm{ERR}}(i) = 1/i \,, \tag{18}$$

indicating that as the user looks at more documents in the ranking they become increasingly dissatisfied with it. Note, however, that if they are seeing good quality documents, the low values generated by C_{RR} mean that they are likely to examine only a few documents, and will leave satisfied.

Whether this $A(\cdot)$ function – or any of the others – is a plausible model for user reactions as they view SERPs can then either be argued for in rhetorical terms, or evidence in support of the conjectured behaviors can be sought. The important thing provided by the C/W/L/A framework is the explanation for the derived score, given the assumptions about user behavior. It provides an equivalence between the mathematical computation associated with the metric and hypothesized user behaviors that may be observable in an experimental setting. Section 4 returns to this notion.

The browsing model associated with ERR assumes that each user scans the ranking until they encounter a document that, to their probabilistic perception, satisfies their information need, see Azzopardi et al. [4, Figure 1]. This means that even though A_{ERG} is not employed, ERR nevertheless has a sense of measuring SERP quality in units of expected gain per document. The "per" component is instead a result of A_{ERR}'s normalization by i, the number of documents inspected by *this* user. That is, ERR can be considered to have units of "expected perceived gain per document inspected", but calculated for each topic as a *macro average* across users rather than a *micro average* across documents viewed by users.[1] Similarly, in a "ETG-like" sense, ERR asserts that the user always exits the ranking with a total perceived gain of one.

Further Aggregation Functions. The shift in perspective from C/W/L to C/W/L/A means that a range of other aggregation functions can now be considered. Drawing on the macro- versus micro-averaging issue that makes ERR distinctive, one plausible option is to suppose that each user's perception of the ranking is determined by the average relevance that *they* observe:

$$A_{\text{avg}}(i) = \frac{1}{i}\sum_{j=1}^{i} r_j \,. \tag{19}$$

Another option is to argue that the user's opinion of the SERP will be completely dominated by the *best* element that they observed:

$$A_{\text{max}}(i) = \max_{j=1}^{i} r_j \,; \tag{20}$$

or by the *last* element that they observed:

$$A_{\text{fin}}(i) = r_i \,. \tag{21}$$

Linear combinations of Equations 20 and 21 with the intention of accommodating the well-known "peak end" rule [15] are also possible, and are considered shortly.

Average Precision Revisited. The flexibility of the new framework also allows for Robertson's [28] alternative explanation of average precision to be encoded in C/W/L/A.

[1] As a brief explanation of macro- and micro-averages: suppose that three rounds of observations are available: 3 successes out of 5 trials; 4 successes out of 7 trials; and 7 successes out of 9 trials. Then the *micro average* of the observations is a simple aggregation of successes relative to trials, without regard to the rounds:

$$\frac{(3+4+7)}{(5+7+9)} = \frac{14}{21} = 0.667$$

whereas the *macro average* averages the individual rounds' success rates:

$$\frac{1}{3}\left(\frac{3}{5}+\frac{4}{7}+\frac{7}{9}\right) = (0.600+0.571+0.777)/3 = 0.681 \,.$$

The computation in Table 1 employs micro-averaging, treating each item inspection as an observation, whereas Table 2 employs macro averaging, taking each user as a round of observations, to get a different score for the same SERP.

Again assuming binary relevance judgments, $r_i \in \{0, 1\}$, AP can be described via the alternative formulation:

$$C_{\text{AP2}}(i) = \frac{\sum_{j=i+1}^{\infty} r_j}{\sum_{j=i}^{\infty} r_j} \,, \tag{22}$$

with $C_{\text{AP2}}(i)$ defined to be zero once i is large enough that the denominator $\sum_{j=i}^{\infty} r_j$ reaches zero. Then AP $\equiv$ CWLA$(C_{\text{AP2}}, A_{\text{avg}})$: each user selects one of the relevant documents at random, and then asks what the precision is at that point. The "choose a relevant document at random" part is accounted for by the definition of $C_{\text{AP2}}(i)$, and then precision at that point by the corresponding $A_{\text{avg}}(i)$ formulation, noting that $L(i)$ will only be non-zero at the points in the ranking at which $r_i \neq 0$, and hence $A(i)$ can be allowed to be any value at the points at which $r_i = 0$ (and thus $C(i) = 1$ and $L(i) = 0$). Note that both user models for AP stipulate that a fraction of the user population pursues the ranking through until the deepest relevant document has been encountered – and that those users then somehow know that they have reached the last relevant document and that it is time to stop.

These two different formulations of AP are another "macro versus micro" situation, but here the two variants (averaging over observations, the C/W/L approach; and averaging over users, the new approach) can be arranged to give the same result. The user browsing models (the user "actions") yield different expected numbers of documents viewed and hence different $L(i)$ values, but the user "reaction" models compensate for that, making the two computed values the same. Note also that C_{AP2} cannot be usefully combined with either A_{ETG} or A_{fin}, because the resulting metrics have no top-weighted component, and all permutations of a set of documents receive the same metric score (the entries "XX" in Table 3).

The two C/W/L/A-based AP formulations need not be restricted to binary relevance judgments, and when non-binary relevance values $0 \le r_i \le 1$ are employed, give rise to a *graded average precision* (grAP) metric (still with CWLA$(C_{\text{AP1}}, A_{\text{ERG}})$ $\equiv$ CWL$(C_{\text{AP2}}, A_{\text{avg}})$) that complements the previous proposals of Dupret and Piwowarski [13] and Robertson et al. [29].

Rank-Biased Precision Revisited. Carterette [6] noted that there was also an alternative interpretation of RBP@ϕ, as the expected gain arising from the last SERP item viewed. That is,

$$\text{CWLA}(C_{\text{RBP@}\phi}, A_{\text{ERG}}) \equiv \text{RBP@}\phi \equiv \text{CWLA}(C_{\text{RBP@}\phi}, A_{\text{fin}}) \,,$$

a relationship that arises because of the particular properties of the geometric distribution, for which $W_{\text{RBP@}\phi}(i) = L_{\text{RBP@}\phi}(i)$. Carterette's work is discussed in more detail at the end of this section.

More Combinations. Table 3 summarizes the situation described to this point, placing all of the metrics discussed so far, and adding a number of further ones: RelRet@k, the total volume of relevance identified in the first k items of the SERP; Succ@k, an assessment of whether there are any useful items in the first k positions of the SERP; and, as already noted, a graded average precision option that has a clear interpretation in terms of user behavior.

Metrics can also be placed in all of the unlabeled positions in Table 3. For example, the combination CWLA$(C_{\text{RBP@}\phi}, A_{\text{max}})$ might be interesting for web search tasks, with a strongly top-weighted focus when ϕ is less than around 0.7, and an overall assessment via

Table 3: Combining $C(\cdot)$ and $A(\cdot)$ functions using the CWLA$(\cdot,\cdot)$ operator to describe existing metrics and other combinations that may be of interest. A superscript * indicates that the listed metric assumes binary relevance, $r_i \in \{0, 1\}$. An "X" indicates a combination where the relevance r_i used in neither the $C(\cdot)$ nor the $A(\cdot)$ component: these "metrics" are a constant. An "XX" indicates a combination where the metric score is determined entirely by set of gains $\mathbf{r}$, that is, without regard to the ordering of the documents within the SERP: these metrics are a constant for any set of gains, and cannot distinguish two rankings of the same documents. Three of the $A(\cdot)$ functions are non-decreasing; one is non-increasing; and the other two are not mono-directional.

$C(\cdot)$ function	Eq.	$A(i)$ function and equation number					
		A_{ETG}, 16 non-dec.	A_{ERG}, 17 non-dec.	A_{ERR}, 18 non-inc.	A_{avg}, 19 –	A_{max}, 20 non-dec.	A_{fin}, 21 –
$C_{\text{Prec@}k}$	7	RelRet@k	Prec@k	X	Prec@k	Succ@k	–
$C_{\text{RBP@}\phi}$	8	–	RBP@ϕ	X	–	–	RBP@ϕ
$C_{\text{DCG@}k}$	9	DCG@k	SDCG@k	X	–	–	–
C_{AP1}	10	–	AP*, grAP	–	–	–	–
C_{RR}	12	–	RR*	RR*, ERR	RR*	–	–
$C_{\text{INST@}T}$	13	–	INST@T	–	–	–	–
C_{AP2}	22	XX	–	–	AP*, grAP	–	XX

the best item that each user encountered in their perusal. The use of A_{max} rather than A_{ERG} also has implications for residual calculations [25]. For a given level of human judgment effort smaller residuals for high-scoring systems – the ones that tend to be of greatest interest in any experiment – makes measurement more precise. This $C_{\text{RBP@}\phi}$-based option, plus others amongst the combinations shown in Table 3, is considered further in Section 4.

Not all combinations of $C(\cdot)$ and $A(\cdot)$ are useful. The entries in Table 3 marked "X" do not use document relevance, r_i, in $C(\cdot)$ or $A(\cdot)$ at all: they are constants for all collections, queries, and rankings, with their exact values depending only on the choice of parameter. In addition, the entries marked "XX" do make use the r_i values, but not in a way that reflects the SERP document ordering. Any permutation of the same set of documents receives the same score. This might be useful for estimating topic difficulty, but not for measuring ranking quality, and nor for comparing two rankings for the same query and collection.

New Aggregation Functions. The flexibility of the C/W/L/A mechanism means that it is also possible to hypothesize further ways in which users might react to the set of i results that they observe, and develop new $A(\cdot)$ functions. We provide two examples that illustrate that ability.

First, consider a user who is more influenced by recent observations than ones from further back in time. As they proceed from the item at rank i of the SERP to the one at rank $i+1$ they "forget" some fraction of their previously acquired utility, but (perhaps) gather replacement utility from the $i+1$ st element. Define

$$A_{\text{fg},\delta}(1) = r_1 \text{ and } A_{\text{fg},\delta}(i+1) = \delta \cdot A_{\text{fg},\delta}(i) + r_{i+1}\,, \qquad (23)$$

so that the last viewed document is credited with its full gain, and previous viewed elements from the SERP have their value eroded via a geometric sequence based on δ. For example if $\delta = 0.5$, then $A(4) = r_1/8 + r_2/4 + r_3/2 + r_4$, and so on. In this formulation the measurement units are bounded above by the limit $A_{\text{fg},\delta}(i) \le 1/(1-\delta)$. That is, as i becomes large, the user assumed to forget old utility as fast as they can gather new utility – meaning that there is a limit to how beneficial any single search can be (in a total gain sense), without needing the somewhat artificial "@k" limit that is part of the definition of DCG@k. In the experiments reported in Section 4 we use a single value, and take $\delta = 0.8$, meaning that the most recently-viewed document has around twice the weight associated with it as the fourth most recent one does.

An interesting consequence of this definition is that $A_{\text{fin}}(i)$ is equivalent to $A_{\text{fg},}$ with $\delta = 0$, and $A_{\text{ETG}}(i)$ is equivalent to $A_{\text{fg},}$ with $\delta = 1$; that is the parameter δ creates an $A_{\text{fg},}$ spectrum with two already-discussed aggregation functions at its extremal points. We note that Wicaksono and Moffat [41] recently proposed a similar "forgetting" rule to handle users who undertake a sequence of queries, and see a sequence of SERPs, as part of a search session.

The second new aggregation function we propose is loosely based on the "peak-end" rule, which suggests that people judge an experience by its peaks (both high and low), and by what occurred most recently [15]. Here we take one peak, the best that has been observed in the ranking through to this point, and define

$$A_{\text{PE},\beta}(i) = \beta \cdot A_{\text{max}}(i) + (1-\beta) \cdot A_{\text{fin}}(i)\,, \qquad (24)$$

to capture this notion, where β is a blending ratio that again establishes a spectrum with two of the previous $A(\cdot)$ functions at its end points. To avoid experimental parameter explosion, the experiments in Section 4 consider just a single value, $\beta = 0.5$.

Related Work. As noted above, Carterette [6] has already commented on one of the dualities reported in Table 3 – that RBP@ϕ has two alternative interpretations, both equally valid, one derived using A_{ERG}, and one derived via A_{fin}. Indeed, although we have presented the C/W/L/A structure here as a development of the previous C/W/L approach, Carterette also anticipated much of its structure in his 2011 paper.

Carterette's interpretation of RBP@ϕ as the expected gain resulting from the last item viewed is an example of what he refers to as a "Model 1, Expected Utility" metric. More generally, Carterette's Model 1 metrics are the ones we categorize here as the CWLA$(\cdot, A_{\text{fin}})$ family, those for which the metric value is determined by the last item the user viewed. Carterette's $P(\cdot)$ function is thus the equivalent of the $L(\cdot)$ function defined by Equation 5. However Carterette

also assumes a condition that we do not include here, namely that $P(\cdot)$ is non-increasing, $P(1) \geq P(2) \geq P(3) \geq \cdots$.

Similarly, Carterette's "Model 2, Expected Total Utility" grouping corresponds to our CWLA$(\cdot, A_{\text{ETG}})$ family, with DCG@k presented as an exemplar. Carterette notes that $1/\log_2(i+1)$ has a limiting value of zero; nevertheless, as we have observed here, the problem remains that $A_{\text{ETG}}(i)$ can grow faster with i than $L_{\text{DCG}}(i) \propto (1/\log_2(i+1)) - (1/\log_2(i+2))$ decreases; hence the necessity of imposing a limiting depth k if discounted cumulative gain is to result in meaningful (and calculable) effectiveness scores.

Carterette's third grouping, the "Model 3, Expected Effort" class, seeks to capture the number of SERP items a user inspects to achieve a given level of utility. Carterette places ERR in this family, arguing that if a relevant document gives rise to (in our terms) $C(i) = 1 - \theta$, and a non-relevant one to $C(i) = 0$, then a weighting function of $1/i$ yields the expected reciprocal stopping rank. Finally, Carterette's fourth category, "Model 4, Expected Average Utility" corresponds to the one we have labeled here as the A_{avg} family, with AP provided as an example. Carterette goes on to describe two further (in his terms) $P(\cdot)$ functions, one that (in our terms) corresponds to the continuation functions $C(i) = i/(i+1)$, and one that is computed as $C(i) = 1$ when $r_i = 0$, and $C(i) = R_i/(R_i+1)$ when $r_i > 0$, where $R_i = \sum_{j=1}^{i} r_j$. As was also noted in connection with DCG@k, neither of these $C(\cdot)$ functions has the necessary convergence properties to be taken to arbitrary depths, and hence must be accompanied by an "@k" depth limit in order to be used. In other words, both of these $C(\cdot)$ functions imply that in the absence of a cutoff a non-negligible fraction of the population of users will proceed through to any arbitrary point in the document ranking.

Our development here builds on both its C/W/L origin and on Carterette's work. By retaining the user's actions as crystallized into a $C(\cdot)$ function, and placing constraints on the behavior of that function, we have the ability to develop models that can be checked against observable user behaviors (for example, via eye tracking or click models) that do not require "@k" cutoffs to be specified. In addition, by explicitly separating the user's set of possible reactions to the SERP into an $A(\cdot)$ function, we are able to develop mechanisms that reflect user satisfaction in regard to different types of search task and different aggregation outcomes.

Other related work is presented by Zhang et al. [44], and in the body of research undertaken those authors and a range of collaborators, including Mao et al. [23]; Zhang et al. [43]; Chen et al. [10]; Liu et al. [22]; and Zhang et al. [45]. Many of those papers also report experimental user studies in which user-SERP satisfaction ratings are requested; taken together, those datasets also form a valuable resource. Throughout this work Zhang et al.'s goal has been to develop metrics that reflect assumptions in regard to user behavior, and that yield metric scores that correlate with user satisfaction. Our work here fits that same context – our goal is to better fit the metric's evaluation to the users' perceptions of usefulness, so as to develop metrics that predict user satisfaction.

In particular, Zhang et al. [44] trace the development of user behavior modeling through the last two decades, starting with Järvelin and Kekäläinen's [20] DCG@k metric; and provide a comprehensive overview of the various options that have been proposed. They also tabulate more than twenty metrics (and hence user behavior models) against a wide range of criteria, including (as just one example) whether that model takes snippets as well as documents into account in user decision-making.

Finally, we note that Fuhr [17] and Ferrante et al. [14] have presented arguments in regard to properties that they believe must be required of all offline effectiveness metrics; and that those arguments are currently being debated by the IR community [30].

4 EXPERIMENTS

We now turn to experimentation, to explore the spectrum of metric possibilities suggested by the pairings of $C(\cdot)$ and $A(\cdot)$ that are evident in Table 3. Our goal is to establish whether there are previously-unconsidered C/W/L/A combinations that it might be worth elevating from being metrics that are merely "possible" to being metrics that are "possibly of interest".

Experiment 1: Tsinghua SERP Satisfaction. The first experiment takes the per-SERP satisfaction data from the Tsinghua *Q-Ref* collection[2] and correlates whole-of-SERP satisfaction scores with metric scores computed via a broad range of C/W/L/A combinations. The collection contains around 7,500 SERP-level impressions, each the result of an observed query. In this dataset individual document grades are coded in $[0, 3]$, with 0 indicating "*this result is useless*" and 3 corresponding to "*this result is serendipitous*"; these document grades are then used to compute SERP metric scores via a linear gain mapping to $r_i \in \{0, 1/3, 2/3, 1\}$. The same SERPs were also scored holisticly on a five-point scale $[0, 4]$, with a satisfaction value of 0 meaning *unsatisfied*, and a satisfaction score of 4 meaning *very satisfied*. In our analysis these five SERP satisfaction scores were treated as ordinal classes and not as numeric values.

Table 4 shows the results of this experiment. Each entry is a Kendall's τ_{b} correlation coefficient over the approximately 7,500 SERPs, comparing the metric scores and the global set of ordinal SERP satisfaction labels, with each row fixing one of the $C(\cdot)$ functions and each column fixing one of the $A(\cdot)$ functions. The higher the correlation coefficient, the more consistent that metric is as a predictor of user satisfaction. The maximum values in each row are highlighted in blue; corresponding metric names are (where labeled) shown in the corresponding positions in Table 3. All correlations are significantly different from zero, with $p \ll 10^{-100}$ in all cases except CWLA$(C_{\text{Prec@}k}, A_{\text{fin}})$, where $p < 0.5$, and CWLA$(C_{\text{AP1}}, A_{\text{ERR}})$, and CWLA$(C_{\text{AP2}}, A_{\text{ERR}})$ which are not significantly correlated with satisfaction at all. Table 4 also indicates, for each row ($C(\cdot)$ function), a canonical $A(\cdot)$ function. These are marked with "*a*" and are metrics listed in Table 3: for example, C_{RR} plus A_{ERR} is the ERR metric. Entries "*b*" represent statistically significant improvements over this baseline: for example, the metric CWLA$(C_{\text{RR}}, A_{\text{max}})$ correlates significantly better with satisfaction than ERR does.

The first three rows of Table 4 correspond to static continuation functions, ones that are usually coupled with A_{ERG} (or, equivalently, A_{ETG}). In all three cases the aggregation function A_{max} outperforms the usual A_{ERG}, and for this dataset it appears that for static metrics – ones in which $C(i)$ is not affected by r_i – the maximum r_i value seen is more predictive of satisfaction than is the total of the observed r_i values. The final r_i seen is also less important than

[2] See Chen et al. [9] and data available from http://www.thuir.cn/tiangong-qref/.

Table 4: Correlation between computed metric scores and user-reported whole of SERP satisfaction scores, computed as Kendall's τ_b coefficients derived from the Tsinghua dataset. Each entry in the table represents a distinct metric; values shown as "X" are constant functions for which it makes no sense to report a correlation, see Table 3. The parameters used in the various $C(\cdot)$ functions were $k = 5$, $\phi = 0.8$, and $T = 2.25$. The canonical metric for each $C(\cdot)$ function is marked with a superscript "a"; any $A(\cdot)$ functions giving significantly improved correlation relative to those ones (holding $C(\cdot)$ steady) are marked with "b". The largest values in each row are highlighted in blue.

	A_{ETG}	A_{ERG}	A_{ERR}	A_{avg}	A_{max}	A_{fin}	$A_{\text{fg},0.8}$	$A_{\text{PE},0.5}$
$C_{\text{Prec@}k}$	0.328	0.328^a	X	0.328	0.418^b	−0.024	0.285	0.393^b
$C_{\text{RBP@}\phi}$	0.326	0.326^a	X	0.334	0.398^b	0.326	0.326	0.366^b
$C_{\text{DCG@}k}$	0.334	0.334^a	X	0.323	0.390^b	0.324	0.332	0.362^b
C_{AP1}	0.369	0.388^a	0.001	0.384	0.435^b	0.446^b	0.384	0.446^b
C_{RR}	0.439^b	0.381^b	0.268^a	0.381^b	0.439^b	0.439^b	0.439^b	0.439^b
$C_{\text{INST@}T}$	0.335	0.335^a	0.321	0.330	0.365^b	0.341	0.333	0.357
C_{AP2}	0.351	0.394	0.007	0.388^a	0.435^b	0.447^b	0.371	0.446^b

the largest (column A_{fin}), even when blended with immediately preceding r_i values (column $A_{\text{fg},0.8}$) or with the maximum r_i seen (column $A_{\text{PE},0.5}$). The combination CWLA($C_{\text{Prec@}k}, A_{\text{fin}}$) is especially poor – it places the entire evaluation focus upon the k th document in the SERP and, unsurprisingly, is slightly *negatively* correlated with user satisfaction. The change to A_{max} makes a big difference in that first row, suggesting that Succ@k – which, like Prec@k, only requires shallow relevance judgments – might be a useful member of evaluation toolkits. The $C_{\text{RBP@}\phi}$ version noted in connection with Table 3 also falls into that "possibly of interest" category.

The four $C(\cdot)$ functions in the lower section of Table 4 are all sensitive to r_i, two in an adaptive manner, and two in a clairvoyant manner. Across this group A_{fin} is stronger, in part because in each of those four $C(i)$ functions SERP exit at rank i is more likely if r_i is large, making it in turn more likely that the final item viewed is also the best item seen. The two blended aggregation functions ($A_{\text{fg},0.8}$ and $A_{\text{PE},0.5}$) also lift, because of the same effect. Overall, A_{max} continues to provide strong performance in this group of rows, but with the two AP-based $C(\cdot)$ functions with A_{fin} generating the largest numbers in the table.

Figure 1 helps understand the correlation coefficients listed in Table 4. In the left pane a C/W/L/A metric combination with a moderate correlation with SERP satisfaction is shown, with the five levels of SERP satisfaction showing somewhat irregular patterns across the spectrum of computed metric scores. The right pane shows a C/W/L/A combination with a higher τ_b correlation. In that second plot each SERP grade level has a more distinct pattern of metric scores, and the separations are more pronounced.

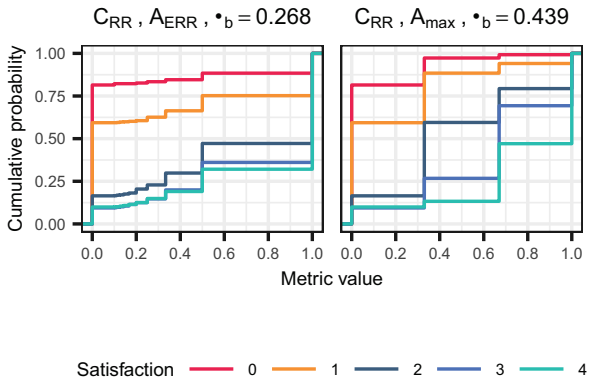

Figure 1: Fractions of SERP satisfaction labels for the Tsinghua dataset, with each of the ordinal SERP satisfaction labels plotted cumulatively as a function of metric score. Two different combinations of $C(\cdot)$ and $A(\cdot)$ are shown, representing expected reciprocal rank and the "novel" metric CWLA($C_{\text{RR}}, A_{\text{max}}$). The version with A_{max} better separates the higher satisfaction levels.

Experiment 2: Commercial Search Data. Table 4 demonstrates that varying $A(\cdot)$ can result in metrics which result in improved prediction of searchers' self-reported satisfaction. In practice we might care instead, or as well, about choosing between two search engines or components thereof. That is, we may want to find which of two systems is better (a relative measure), rather than how good each is (an absolute measure). We now consider whether adding $A(\cdot)$ lets us improve metrics for this task.

We use a set of approximately 26,000 queries collected by Bing. Crowd workers, subject to training and quality control, were told that each query was run by two different engines, producing a pair of SERPs for each query (see Figure 2). They then labeled each pair of SERPs using a preference slider, resulting in a judgment between −1 (strong preference for SERP S_1) to +1 (strong preference for SERP S_2). Each pair was labeled multiple times, with allocation to left and right sides controlled to account for left-right bias [16]. We then reduced the preferences to ternary variables: overall preference for SERP_1, SERP_2, or neither, according to the sign of the mean preference, and used those ternary categories in the analyses reported shortly.

Separately, and via a different set of trained and audited crowd workers, the top three results from each SERP were labeled for relevance on a five-point ordinal scale from "*bad*" to "*perfect*". We mapped the five points of this scale to gains $r_i \in \{0, 1/4, 1/2, 3/4, 1\}$. Computing metric scores across rankings containing just three judged documents is at best a coarse approximation, of course. Compounding that approximation risk is lack of knowledge of $R = \sum_{i=1}^{\infty} r_i$, the total relevance, which is required in order to calculate AP (and hence C_{AP1} and C_{AP2}). We approximated the "collection R" by the "within-run" total relevance (an approach referred to as

Table 5: Correlation between differences in computed metric scores and side-by-side paired SERP ternary preferences, computed as Kendall's τ_b coefficients derived from the Bing dataset. Entries shown as "X" or "XX" are constant functions for which it makes no sense to report a correlation, see Table 3. The parameters used in the various $C(\cdot)$ functions were $k = 3$, $\phi = 0.5$, and $T = 1$. The canonical metric for each $C(\cdot)$ function is marked with a superscript "a", and the largest values in each row are highlighted in blue. All correlations are statistically significant at $p \ll 10^{-10}$, except CWLA(C_{AP1}, A_{ERR}) where correlation is significant at $p < 0.02$; in this setting no combinations significantly improve on the canonical metric in each row.

	A_{ETG}	A_{ERG}	A_{ERR}	A_{avg}	A_{max}	A_{fin}	$A_{fg,0.8}$	$A_{PE,0.5}$
$C_{Prec@k}$	0.081	0.081^a	X	0.081	0.070	0.044	0.073	0.060
$C_{RBP@\phi}$	0.083	0.083^a	X	0.080	0.076	0.083	0.082	0.079
$C_{DCG@k}$	0.083	0.083^a	X	0.082	0.076	0.073	0.082	0.079
C_{AP1}	0.077	0.072^a	−0.012	0.072	0.066	0.060	0.075	0.060
C_{RR}	0.071	0.081	0.071^a	0.081	0.071	0.071	0.071	0.071
$C_{INST@T}$	0.081	0.081^a	0.077	0.078	0.075	0.079	0.080	0.078
C_{AP2}	XX	0.071	−0.017	0.072^a	0.066	XX	0.076	0.061

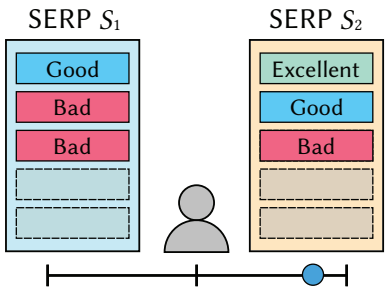

Figure 2: Data collection process for the Bing dataset, showing two SERPs, S_1 and S_2. Assessors view each pair of standard SERPs side by side and indicate their preference between them using a slider. The documents shown at the top of each SERP are judged for relevance on a five-point scale, as a separate activity.

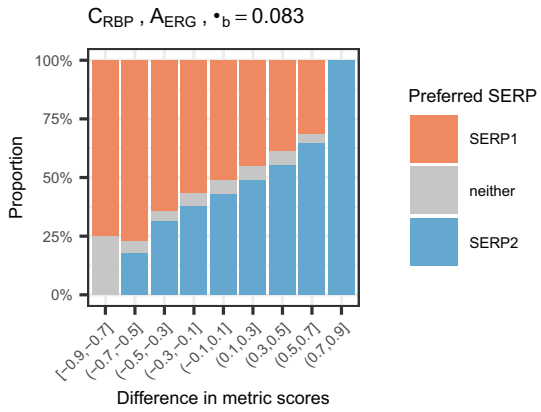

Figure 3: Distribution of preferences between SERPs, as a function of difference between scores computed for each SERP. Score difference is only weakly predictive of preference over the middle part of the score difference range.

"self-normalizing AP" by Moffat [24], and also used for Table 4), acknowledging that this may result in AP scores for SERP preference pairs derived from different assumed recall bases R.

We then processed the side-by-side preference pairs. For each metric and pair of SERPs we calculated two scores, and took their difference, obtaining values from −1 to +1. Finally, for each metric we computed the correlation between the differences in scores and the ternary side-by-side preference category. A metric which is good for this task would have high correlation, that is, the metrics from combining document-level labels would predict the SERP-level preference. Results are summarized in Table 5.

As can be seen, the τ_b scores are lower across the board: in the experimental structure used, score differences correlate with side-by-side preference much less than individual scores correlate with satisfaction, for all metrics. Figure 3 illustrates this with CWLA($C_{RBP@\phi}, A_{ERG}$), the best-performing metric. Although there is a clear SERP preference at large metric differences, across much of the range the preference is more balanced. This lack of correlation is not surprising: any noise in the metrics is increased by measuring twice. Different workers were also given the two tasks. Finally, workers were asked about different objects – single documents in one case, pairs of SERPs in the other – and a SERP may be preferred for reasons such as diversity or captioning, which C/W/L/A – at least in current form – is oblivious to. We note that Sakai and Zeng [32, 33] have carried out similar experiments in which they compare the SERP preference and the *sign* of the metric's score difference. Sanderson et al. [35] have also considered this issue.

Accordingly, in Table 5, none of the alternative $A(\cdot)$ functions improve on the "canonical" metric variants. We can however note one trend: A_{ETG} and A_{ERG} perform better here than in the earlier experiment. This suggests the side-by-side judges were considering the total "weight" of each SERP, perhaps imagining that a SERP with more relevant documents would serve more people or is somehow more impressive. The Tsinghua participants, who labeled their own searches, seem more swayed by single results, and the relative performance across $A(\cdot)$ functions highlights this difference.

5 CONCLUSIONS

We have added an explicit aggregation function to the C/W/L framework, renamed as C/W/L/A, and demonstrated that the separation of user browsing actions from each user's summative assessment of SERP quality leads to a powerful taxonomy that captures almost all current effectiveness metrics, and has the flexibility to suggest a wide range of other combinations not previously considered.

Our primary purpose has been to provide a cohesive structure in which metrics can be formulated and argued about, and via that structure, to illustrate possibilities and opportunities. In particular, we have not sought to try and identify a "best" metric, because metric choice must – of necessity – vary according to the nature of the users, the nature of the task they are performing at the time, the nature of the collection they are searching, the manner in which the search results are presented, and so on. Nevertheless, our experiments with two user-based datasets suggest that the flexible offline evaluation possibilities created by the C/W/L/A framework will be of interest to researchers and practitioners alike.

Acknowledgment. We thank the Tsinghua group for making their data publicly available. This work was supported under the Australian Research Council's Discovery Projects funding scheme (project numbers DP190101113 and DP200103136).

REFERENCES

[1] L. Azzopardi, P. Thomas, and N. Craswell. Measuring the utility of search engine result pages: An information foraging measure. In *Proc. SIGIR*, pages 605–614, 2018.

[2] L. Azzopardi, P. Thomas, and A. Moffat. cwl_eval: An evaluation tool for information retrieval. In *Proc. SIGIR*, pages 1321–1324, 2019.

[3] L. Azzopardi, R. W. White, P. Thomas, and N. Craswell. Data-driven evaluation metrics for heterogeneous search engine result pages. In *Proc. CHIIR*, pages 213–222, 2020.

[4] L. Azzopardi, J. Mackenzie, and A. Moffat. ERR is not C/W/L: Exploring the relationship between expected reciprocal rank and other metrics. In *Proc. ICTIR*, pages 231–237, 2021.

[5] C. Buckley and E. M. Voorhees. Retrieval system evaluation. In E. M. Voorhees and D. K. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, chapter 3, pages 53–78. MIT Press, 2005.

[6] B. Carterette. System effectiveness, user models, and user utility: A conceptual framework for investigation. In *Proc. SIGIR*, pages 903–912, 2011.

[7] B. Carterette, E. Kanoulas, and E. Yilmaz. Incorporating variability in user behavior into systems based evaluation. In *Proc. CIKM*, pages 135–144, 2012.

[8] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proc. CIKM*, pages 621–630, 2009.

[9] J. Chen, J. Mao, Y. Liu, F. Zhang, M. Zhang, and S. Ma. Towards a better understanding of query reformulation behavior in web search. In *Proc. WWW*, pages 743–755, 2021.

[10] Y. Chen, K. Zhou, Y. Liu, M. Zhang, and S. Ma. Meta-evaluation of online and offline web search evaluation metrics. In *Proc. SIGIR*, pages 15–24, 2017.

[11] C. W. Cleverdon. The ASLIB Cranfield research project on the comparative efficiency of indexing systems. In *ASLIB proceedings*. MCB UP Ltd, 1960.

[12] N. Craswell, O. Zoeter, M. J. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proc. WSDM*, pages 87–94, 2008.

[13] G. Dupret and B. Piwowarski. A user behavior model for average precision and its generalization to graded judgments. In *Proc. SIGIR*, pages 531–538, 2010.

[14] M. Ferrante, N. Ferro, and E. Losiouk. How do interval scales help us with better understanding IR evaluation measures? *Inf. Retr.*, 23(3):289–317, 2020.

[15] B. L. Fredrickson and D. Kahneman. Duration neglect in retrospective evaluations of affective episodes. *J. Personality and Social Psych.*, 65(1):45–55, 1993.

[16] H. H. Friedman and T. Amoo. Rating the rating scales. *J. Marketing Manag.*, 9(3): 114–123, 1999.

[17] N. Fuhr. Some common mistakes in IR evaluation, and how they can be avoided. *SIGIR Forum*, 51(3):32–41, 2017.

[18] D. K. Harman. The TREC test collections. In E. M. Voorhees and D. K. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, chapter 2, pages 21–52. MIT Press, 2005.

[19] K. Hofmann, L. Li, and F. Radlinski. Online evaluation for information retrieval. *Found. & Trends in IR*, 10(1):1–117, 2016.

[20] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Sys.*, 20(4):422–446, 2002.

[21] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Found. & Trends in IR*, 3(1-2):1–224, 2009.

[22] M. Liu, Y. Liu, J. Mao, C. Luo, M. Zhang, and S. Ma. "Satisfaction with failure" or "unsatisfied success": Investigating the relationship between search success and user satisfaction. In *Proc. WWW*, pages 1533–1542, 2018.

[23] J. Mao, Y. Liu, K. Zhou, J. Nie, J. Song, M. Zhang, S. Ma, J. Sun, and H. Luo. When does relevance mean usefulness and user satisfaction in web search? In *Proc. SIGIR*, pages 463–472, 2016.

[24] A. Moffat. Seven numeric properties of effectiveness metrics. In *Proc. Asia Info. Retri. Soc. Conf.*, pages 1–12, 2013.

[25] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Sys.*, 27(1):2.1–2.27, 2008.

[26] A. Moffat, P. Thomas, and F. Scholer. Users versus models: What observation tells us about effectiveness metrics. In *Proc. CIKM*, pages 659–668, 2013.

[27] A. Moffat, P. Bailey, F. Scholer, and P. Thomas. Incorporating user expectations and behavior into the measurement of search effectiveness. *ACM Trans. Inf. Sys.*, 35(3):24:1–24:38, 2017.

[28] S. E. Robertson. A new interpretation of average precision. In *Proc. SIGIR*, pages 689–690, 2008.

[29] S. E. Robertson, E. Kanoulas, and E. Yilmaz. Extending average precision to graded relevance judgments. In *Proc. SIGIR*, pages 603–610, 2010.

[30] T. Sakai. On Fuhr's guideline for IR evaluation. *SIGIR Forum*, 54(1):12:1–12:8, 2020.

[31] T. Sakai and Z. Dou. Summaries, ranked retrieval and sessions: A unified framework for information access evaluation. In *Proc. SIGIR*, pages 473–482, 2013.

[32] T. Sakai and Z. Zeng. Which diversity evaluation measures are "good"? In *Proc. SIGIR*, pages 595–604, 2019.

[33] T. Sakai and Z. Zeng. Retrieval evaluation measures that agree with users' SERP preferences: Traditional, preference-based, and diversity measures. *ACM Trans. Inf. Sys.*, 39(2):14:1–14:35, 2021.

[34] M. Sanderson. Test collection based evaluation of information retrieval systems. *Found. & Trends in IR*, 4(4):247–375, 2010.

[35] M. Sanderson, M. L. Paramita, P. D. Clough, and E. Kanoulas. Do user preferences and evaluation measures line up? In *Proc. SIGIR*, pages 555–562, 2010.

[36] M. D. Smucker and C. L. A. Clarke. Stochastic simulation of time-biased gain. In *Proc. CIKM*, pages 2040–2044, 2012.

[37] M. D. Smucker and C. L. A. Clarke. Time-based calibration of effectiveness measures. In *Proc. SIGIR*, pages 95–104, 2012.

[38] P. Thomas, F. Scholer, and A. Moffat. What users do: The eyes have it. In *Proc. Asia Info. Retri. Soc. Conf.*, pages 416–427, 2013.

[39] L. Wang, J. Lin, and D. Metzler. A cascade ranking model for efficient ranked retrieval. In *Proc. SIGIR*, pages 105–114, 2011.

[40] A. F. Wicaksono and A. Moffat. Empirical evidence for search effectiveness models. In *Proc. CIKM*, pages 1571–1574, 2018.

[41] A. F. Wicaksono and A. Moffat. Modeling search and session effectiveness. *Inf. Proc. & Man.*, 58(4):102601, 2021.

[42] E. Yilmaz, M. Shokouhi, N. Craswell, and S. E. Robertson. Expected browsing utility for web search evaluation. In *Proc. CIKM*, pages 1561–1564, 2010.

[43] F. Zhang, Y. Liu, X. Li, M. Zhang, Y. Xu, and S. Ma. Evaluating web search with a bejeweled player model. In *Proc. SIGIR*, pages 425–434, 2017.

[44] F. Zhang, Y. Liu, J. Mao, M. Zhang, and S. Ma. User behavior modeling for web search evaluation. *AI Open*, 1:40–56, 2020.

[45] F. Zhang, J. Mao, Y. Liu, X. Xie, W. Ma, M. Zhang, and S. Ma. Models versus satisfaction: Towards a better understanding of evaluation metrics. In *Proc. SIGIR*, page 379–388, 2020.