Modeling Search and Session Effectiveness

Alfan Farizki Wicaksono (ORCiD: 0000-0003-2561-2778), Alistair Moffat (ORCiD: 0000-0002-6638-0232)

School of Computing and Information Systems, The University of Melbourne, Australia

Abstract

Many information needs cannot be resolved with a single query, and instead lead naturally to a sequence of queries, issued as a search session. In a session test collection, each topic has an associated query sequence, with users assumed to follow that sequence when reformulating their queries. Here we propose a session-based offline evaluation framework as an extension to the existing query-based C/W/L framework, and use that framework to devise an adaptive session-based effectiveness metric, as a way of measuring the overall usefulness of a search session. To realize that goal, data from two commercial search engines is employed to model the two required behaviors: the user conditional continuation probability, and the user conditional reformulation probability. We show that the session-extended C/W/L framework allows the development of new metrics with associated user models that give rise to greater correlation with observed user behavior during search sessions than do previous session metrics, and hence provide a richer context in which to compare retrieval systems at a session level.

Keywords: information retrieval, web search, search effectiveness, user model, user behavior, session effectiveness,

2020 MSC: 00-01, 99-00

1. Introduction

The information retrieval (IR) community has made extensive use of test collection-based evaluation as a complement to (or even replacement for) user-based studies [35, 41]. A test collection and evaluation measure(s), together, can be viewed as a simulation of users, as if they are interacting with search result pages in an operational setting [35]. Many effectiveness metrics have been developed based on such models of user behavior, including rank-biased precision (RBP) [26]; expected reciprocal rank (ERR) [5]; time-biased gain (TBG) [37]; the bejeweled-player model (BPM) [46]; INST [29]; the information foraging model (IFT-model) [1]; and the data-driven model (DDM) [2]. These metrics all result in the calculation of a numeric quality score in regard to a single search engine result page (SERP) returned for a
single query posed in connection with a single information need. Whatever metric is used, the metric scores are then typically aggregated across a set of topics, and, if two or more systems are being compared, a paired statistical test is applied.

More realistically, a user with an information need will typically submit an initial query, and then consider the SERP that is returned. If they fail to find a sufficient number of relevant documents, or feel that their information need has not been adequately resolved for some other reason, they will reformulate their query and examine a second SERP. Such iteration may continue through several – perhaps even many – cycles of refinement/alteration before the search session is concluded. Jansen and Spink [8] studied a sample of Excite.com and AltaVista.com search query logs collected in 2001 and 2002 respectively, and found that 55% of Excite.com users reformulated their initial queries, and that 47% of AltaVista.com queries were submitted in the context of similar queries. Jansen et al. [9] define six types of query reformulation, the most important of which are specialization and generalization. Specialization occurs when the reformulated query contains additional terms, in order to seek more precise information; while generalization arises if the reformulated query contains fewer terms than the previous one. A search session might include both types of reformulation.

There are several underlying reasons for reformulation. Turpin and Hersh [39] and Smith and Kantor [36] demonstrate that users are able to compensate for the reduced effectiveness of search engine systems by adapting their behavior, and one such adaptation is submitting more queries. Järvelin et al. [11] describe a laboratory-based interactive searching study, finding that in some cases initial queries give poor results because users submit query terms that do not accurately cover the topic description, and that reformulation is part of a learning experience. Other experimentation suggests that users tend to pose several short queries in a session, rather than one comprehensive one [10, 17]; and that user engagement is also positively correlated with search success [30].

Given that users carry out their searching activities via sessions, a range of proposals for session-based effectiveness evaluation have emerged [7, 11, 15, 19, 40, 45]. Such mechanisms require session-based test collections, such as the one used in the TREC 2010 Session Track [15, 16]. As with query-oriented test collections, session-based ones consist of three components: a collection of documents; a set of topics; and a set of relevance judgments. In a query-oriented test collection we think of each topic as having one query associated with it (or one query per user per topic; see, for example, Bailey et al. [3]). In a session-oriented test collection each topic is associated with a sequence of queries, with simulated users assumed to follow that fixed ordering when reformulating queries, deciding whether to continue from one query to the next, without control of the ordering [11, 15, 16, 19]. Figure 1 illustrates this idea. Note that the number of queries actually posed by each simulated user is not known, and that they might become satisfied (or disillusioned) at different points in the sequence.

In the framework shown in Figure 1 the score for topic \( \ell \) is a probability-weighted summa-
Figure 1: A session-based test collection in which \( k \) independent topics are each associated with a fixed sequence of \( m \) queries. The simulated user is assumed to always commence their \( \ell \)th session with the first query for that topic, \( Q_{\ell,1} \), and then, having issued query \( Q_{\ell,j} \), to reformulate to \( Q_{\ell,j+1} \) with some probability between zero and one. The variations in color intensity represent the probability that any given user poses the \( j \)th query (and hence examines the \( j \)th SERP) for topic \( \ell \) before they end their session.

It is those overall population characteristics that we study in this paper. Table 1 provides further explanation, comparing the inputs, the parameter estimation (fitting), and the modeling (predicting) components of single query search (center column) and multi-query session-based search (right column). When evaluating a single SERP, the main inputs are the ranking of documents and a set of relevance judgments (qrels), and in most effectiveness metrics the user interaction with that SERP is modeled rather than observationally recorded. That is, the SERP effectiveness score in the center column is determined from the modeled interactions, rather than a set of individual actual interaction. A similar arrangement arises...
Table 1: Observational and model goals for single query evaluation (left) and session evaluation (right). Our focus here is on modeling and evaluation for multi-query sessions (bottom row, right-hand column).

<table>
<thead>
<tr>
<th>Facet</th>
<th>Single query</th>
<th>Session of queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed data</td>
<td>User queries and the resultant SERPs; plus qrels to indicate item relevance.</td>
<td>Sequences of same-intent queries, each a user’s search session, and the sequences of resultant SERPs; plus qrels.</td>
</tr>
<tr>
<td></td>
<td>Perhaps with eye tracking, or item clicks, or dwell times, or user-generated SERP-level satisfaction ratings.</td>
<td>Perhaps with query-level data, as listed in the column at left, and perhaps with user-generated session-level satisfaction ratings.</td>
</tr>
<tr>
<td></td>
<td>May involve one query per topic, or might involve multiple different queries per topic, generated by different users</td>
<td>May involve one session per topic, or might involve multiple sessions per topic, generated by different users</td>
</tr>
<tr>
<td>Observational goal</td>
<td>To establish a fitted relationship between query satisfaction (or a surrogate for it) and factors available in the observed data [21]</td>
<td>To establish a fitted relationship between session satisfaction (or a surrogate for it) and factors available in the observed data [12, 21, 43, 47]</td>
</tr>
<tr>
<td>Model and evaluation goal</td>
<td>To identify a predictive relationship that allows computation of numeric system scores that correlate well with user task performance for future queries when the SERP exit point is not known [1, 2, 5, 29, 37, 46]</td>
<td>To identify a predictive relationship that allows computation of numeric session scores that correlate well with user whole-session task performance for future sessions when the query sequence exit point is not known [11, 15, 16, 19, 45]</td>
</tr>
</tbody>
</table>

in the right column in Table [1]. Not only is the per-SERP behavior of each individual user modeled, rather than observed, for the purpose of the evaluation; but also the number of queries issued by that individual is modeled, rather than observed.

To connect between observation and predictive model, the middle row of Table [1] notes that the model parameters can be fitted to, and model structure informed by, detailed observations of users. That intermediate “observational fitting” goal of session evaluation has been
considered by a range of authors [12, 20, 21, 43, 47]. Given the sequence of queries submitted by each individual user in a session, the challenge is to aggregate the individual query scores via a weighting scheme tuned to optimize a certain quantity, such as user-reported satisfaction [12, 20, 21, 43, 47]. When addressing this observational goal, there is no probability-weighted sum and no reformulation probabilities, since what queries the user posed and how they reformulated them are both known. For example, Zhang et al. [47] consider the notion of forgetfulness, with users tending to increasingly discount the utility gathered from earlier queries; that mechanism relies on knowledge of which query in the sequence was the user’s last one. Similarly in a single-SERP observational evaluation (the center cell in the center column in Table 1) it is possible to connect user-reported satisfaction with knowledge of the particular documents viewed (and hence their utilities), including the order in which they were encountered, and which one was viewed last.

Our goal in this paper is to develop an adaptive user model for session evaluation, capturing overall population behaviors rather than individual user actions. In particular, we extend the work of Moffat et al. [27, 29], who connect metrics, user models, and user behaviors. We start by collecting and reporting empirical evidence that adds further support to the observations of Moffat et al. [27, 29] regarding query-level user behaviors, using three large search interaction logs. We then develop a model for session-level behaviors (that is, query reformulation behaviors), and incorporate that extension into session-based search effectiveness measures, capitalizing on the framework described in Figure 2. In that framework, each user commences their search session by posing a first query, \( j = 1 \); and then sequentially examining the ranked list of items from its top position, \( i = 1 \). At each rank position \( i \) of the \( j \) th SERP a decision is made: to continue to rank \( i + 1 \); or to exit. In the latter case, a further choice is made: to issue a reformulated \( j + 1 \) th query; or to end the session entirely.

As a result there are two important functions to be estimated:

1. The continuation function, \( C(j, i) \), describing the conditional probability of the user inspecting the item at rank \( i + 1 \) in the \( j \) th SERP, given that they have just examined the item at rank \( i \); and
2. The reformulation function, \( F(j) \), describing the conditional probability of the user issuing a \( j + 1 \) th query, given that they have just ended their inspection of the \( j \) th SERP.

The first of these two functions describes the user’s query-level behavior; the second their session-level behavior. A wide range of session-based metrics can then be characterized via these two functions. The challenge is thus to find models for \( C(j, i) \) and \( F(j) \) that accurately predict observed behavior – the goal expressed in the bottom right element of Table 1.

Research Questions. In the context of that background, we consider the following research questions:
• Can the single-query C/W/L framework be extended to query sessions (Section 4)?

• Can observations in regard to user behavior such as impressions and click-throughs be used to infer values for the corresponding parameters of C/W/L models for single- and multi-query sessions and if so, how well do the fitted models anticipate those observations compared to alternative models (Section 5)?

• Can a C/W/L-based session evaluation metric be derived that correlates with measured user session satisfaction data, and if so, what practical implications does that definition hold for system evaluation (Section 6)?

2. Related Work

2.1. Query-Based C/W/L Framework

One possible goal of effectiveness measurement is for the scores generated by the metric to correspond to the expected benefit (or utility) derived from the SERP by a user. In the C/W/L framework [26, 27, 29] a user model is characterized by any one of three interconnected functions: (1) the conditional continuation probability that the user will continue from the document at rank \(i\) to the one at rank \(i+1\), denoted by \(C(i)\); (2) the weight function that describes the proportion of user attention associated with the document at rank \(i\), denoted by \(W(i)\); or (3) the last probability, the probability that the document at rank \(i\) is the final one in that SERP that is inspected by the user, denoted by \(L(i)\).

Two query-level metrics can be derived from any such choice of \(C(\cdot)\), \(W(\cdot)\), or \(L(\cdot)\): an expected rate of gain metric \((M_{ERG})\), and an expected total gain metric \((M_{ETG})\):

\[
M_{ERG}(\vec{r}) = \sum_{i=1}^{\infty} W(i) \cdot r_i, \quad \text{and} \quad M_{ETG}(\vec{r}) = \sum_{i=1}^{\infty} \left( L(i) \cdot \sum_{j=1}^{i} r_j \right),
\]

where \(0 \leq r_i \leq 1\) is the gain accrued when the user inspects the document at rank \(i\) in the SERP. A static C/W/L metric in one in which \(C(i)\) is independent of \(r_j\) for \(1 \leq j \leq i\); whereas an adaptive metric allows previous gain values to influence \(C(i)\) [29].
One important underlying assumption that needs to be noted is that $r_i$ is the gain accrued by this particular user in the context that they are in at this moment in their perusal of this particular SERP, and hence has many contextual dependencies: it will change if a similar or identical document has already been viewed; it will vary depending on whether this document corroborates previous information, or contradicts it, or complements it; and so on. That is, in reality $r_i$ is highly subjective, and not pre-determined in any way. This reality is somewhat at odds with the notion of a static test collection and invariant relevance judgments that assign fixed and absolute values for each $r_i$, and is a tension that applies to the great majority of off-line evaluation protocols employed in IR.

Putting aside the question as to the extent to which $r_i$ can be accurately captured by fixed relevance judgments, Moffat et al. [27, 29] further argue that, all other things being equal, $C(\cdot)$:

1. Increases with rank position $i$;
2. Increases with $T$, the volume of relevance anticipated at the time they posed their query; and
3. Decreases as relevant items are observed in the ranking.

These postulates were validated in a user study with 34 participants [27], and incorporated into an effectiveness metric named INST [29], thereby demonstrating that they were not mutually incompatible. A first contribution of the current work is to show that those three relationships also hold for users of two types of commercial search engine, based on their behavior as captured by interaction logs. That then provides a template in which we add the “session” dimension via the conditional reformulation probability $F(j)$ that was already outlined in connection with Figure 2.

### 2.2. Session-Based Effectiveness Metrics

Kanoulas et al. [16] describe a session test collection as having a set of topics, with each topic associated with two components: (1) a set of relevance judgments; and (2) a sequence of static queries. A number of effectiveness measures have been proposed for this kind of test collection, mostly based on ad-hoc assumptions about how users behave in a search session [11, 16, 19].

**Session-Based Discounted Cumulative Gain.** Järvelin et al. [11] propose *session-based discounted cumulative gain* (sDCG), computed as:

$$sDCG(\vec{r}) = \sum_{j=1}^{m} \frac{DCG(\vec{r}_j)}{(1 + \log bq_j)} = \sum_{j=1}^{m} \sum_{i=1}^{n} \frac{1}{(1 + \log bq_j) \cdot (1 + \log_i)} \cdot r_{j,i},$$

where $\vec{r} = \langle r_1, r_2, \ldots, r_m \rangle$ is the sequence of $m$ gain vectors associated with the SERPs in a session of $m$ queries; $n$ is the evaluation depth within each SERP; $1 < bq < 1000$ is the
parameter governing the extent to which users reformulate queries; and $b$ is the log base of DCG, with $b = 2$ the usual value. Järvelin et al. [11] further suggest using $bq = 4$.

In this formulation $1/(1 + \log_{bq} j)$ is a session-level discount function. In support of their proposal, Järvelin et al. [11] argue that query reformulation involves additional effort, and thus the SERPs returned by reformulated queries are less valuable that the one returned by the initial query. Note that truncation at session length $m$ and ranking depth $n$ is a necessary part of this metric, since the infinite sum of the two discount functions is divergent. That is, the maximum score of sDCG increases as either ranking depth or session length increase, requiring application of a scaling mechanism if the metric score is to be brought into the range from zero to one. Note also that the use of $m$ and $n$ is does not imply that the $n$th result of each of the $m$ SERPs is inspected by every user; but does imply that no users proceed beyond those limits.

As will be shown in Section 4, the SERP-level continuation probability of sDCG increases with rank position $i$ (the sunk-cost property); and at the session-level, its reformulation probability also increases. However, neither of the two probabilities are adaptive.

Session-Based Rank-Biased Precision. Lipani et al. [19] have similarly suggested a session-based version of rank-biased precision (RBP) [26]:

$$\text{LCY-sRBP}\left(\overrightarrow{r}\right) = (1 - p) \cdot \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \left( \frac{p - q \cdot p}{1 - q \cdot p} \right)^{j-1} \cdot (q \cdot p)^{i-1} \cdot r_{j,i},$$

where $0 \leq p \leq 1$ is the user persistence parameter; where $0 \leq q \leq 1$ controls the balance between reformulating queries or continuing to inspect items in the current SERP; and where $0^0$ is taken to be 1.

When $q = 1$, the simulated user never reformulates, and LCY-sRBP calculates an RBP score for the first SERP, RBP($\overrightarrow{r}_1 \mid p$). At the other extreme, when $q = 0$, the simulated user always reformulates immediately after inspecting the first item in each SERP. The reformulation and continuation probabilities are presumed to be constant during the course of each session.

In contrast to sDCG, the discount function of LCY-sRBP gives a convergent sum, making the session length $m$ and ranking depth $n$ unnecessary. That is, LCY-sRBP allows sessions and rankings of arbitrary length; but even so, is still not adaptive.

Other Session-Based Metrics. Kanoulas et al. [16] also propose a session-based smoothing mechanism, esM, computed as an expected score over all possible browsing paths in a session, and applying a standard query-level metric to each possible path. If $\Psi$ is the set of possible browsing paths, and if each path $\psi \in \Psi$ has a probability of $Pr(\psi)$ of being followed by a user, then

$$esM(\Psi) = \sum_{\psi \in \Psi} Pr(\psi) \cdot M(\psi),$$
where \( M(\psi) \) is the score of metric \( M() \) computed for the browsing path \( \psi \), and \( M \) might, for example, be normalized discounted cumulative gain (NDCG) or average precision (AP). Two RBP-like geometric distributions define \( Pr(\psi) \): the first, \( p_{re}^{-1}(1 - p_{re}) \), represents the probability that the \( j \)th query is the last one before the user ends their session; and the second, \( p_{down}^{-1}(1 - p_{down}) \) is the probability that rank position \( i \) is the last one inspected by the user in each SERP. In their proposal, Kanoulas et al. [16] suggest \( p_{re} = 0.5 \) and \( p_{down} = 0.8 \) as suitable values. Yang and Lad [45] describe a similar approach, but without allowing the simulated users to exit early queries. As is demonstrated by our experiments, below, user models with fixed reformulation and continuation probabilities are inflexible, and can be improved upon.

Sakai and Dou [33] demonstrate that U-measure can be used to evaluate multi-query sessions when document lengths (measured in characters) are available. In a similar approach, if document reading times are known, the Cube Test metric can be employed to measure the effectiveness of a session [23]. The Cube Test and U-measure metrics do not have a parameter that accommodates the user’s reason for searching (such as a navigational or informational goals, characterized by a relevance target volume \( T \)); and neither is adaptive.

In recent work, van Dijk et al. [40] (see also Ferrante and Ferro [7]) propose a model-based session effectiveness metric using Markov chains, involving four transition probabilities: the probability of moving forward to the next item in a SERP; the probability of moving backward in a SERP; the reformulation probability; and the probability of stopping a search session. However, these probabilities are fixed, and do not change as the simulated user accumulates answers toward their goal; thus, their methods are not adaptive. This approach has the advantage of reflecting the observations of Thomas et al. [38] and Sakai and Dou [33] that user actions are far from monotone within the SERP, and that typical behavior is more like one step forward, one step backward, two steps forward.

### 2.3. Query-to-Session Aggregation Functions

When the sequence of queries submitted by each individual user are observed, and corresponding session satisfaction ratings captured in some way, a fitted relationship between them can be sought (the middle row in Table 1). One approach is to compute the coefficients for a combining function over the session’s query scores that best correlates with the corresponding session-level satisfaction ratings [12, 20, 21, 43, 47].

Suppose a search session observed from a user consists of \( m \) queries, and hence \( m \) SERPs and \( m \) relevance vectors, \( \overrightarrow{r} = (\overrightarrow{r_1}, \overrightarrow{r_2}, \overrightarrow{r_3}, \ldots, \overrightarrow{r_m}) \). A linear combination session score can be defined as:

\[
    sM(\overrightarrow{r}) = \sum_{j=1}^{m} \theta(j) \cdot M(\overrightarrow{r_j}),
\]

where \( 0 \leq \theta(j) \leq 1 \) is a weight corresponding to the \( j \)th query, and \( M(.) \) is a query-based effectiveness measure. Jiang and Allan [12] consider aggregation functions such as summation.
(that is, \(\theta(j) = 1\)) and \textit{mean} (that is, \(\theta(j) = 1/m\)). Liu et al. \[20, 21\] suggest that the \textit{recency} has a strong influence on session satisfaction, and suggest that \(\theta(j)\) should be increasing with \(j\). Zhang et al. \[47\] similarly argue that to a certain extent users \textit{forget} earlier queries, and propose \(\theta(j) = e^{-\delta(m-j)}\), where \(\delta\) is the loss rate. Finally, Wicaksono and Moffat \[43\] suggest that session \textit{middle} queries are relatively unimportant, and explore a weighting function with a \textit{U-shape} in which late queries have slightly more weight than the early ones, and both are preferred to queries in the middle of the session.

As noted already (Table [1]) our study here has a different goal, and does not seek to fit coefficients to known data, the purpose of the methods summarized in this subsection. Rather, we seek to define predictive models that can be applied to sessions in general, where neither the number of queries, nor the end point of each user sequence, are defined in advance. The next section describes the datasets that we will use in support of that objective.

3. Interaction Logs

This section describe a total of six resources that are used: three from commercial search providers; and then three from laboratory studies.

3.1. Industry-Based Datasets

Table [2] describes three commercial search interaction logs. The first two are from the Australasian job search site, Seek.com.au; the third set is a publicly-available Web search log of Russian search engine Yandex.ru.

The Seek.com.au data was sampled through an eight-week period (30 July to 23 September 2018), and covers two types of queries: via a mobile application (app) running under iOS/Android that has infinite scrolling without pagination; and via a traditional browser-based Web service that has fixed pages containing 20 results. The first page in each SERP might also contain up to two further paid items.

The two modalities employ different definitions of search session. On app-based mobile search a new session starts each time the app is opened, whereas with browser-based search a session is associated with a long-lived cookie that could last for months. In terms of search goals, Seek.com.au users might be intending to apply for one or more jobs, or might just collating information about jobs and salaries, with no intention of applying.

In the Seek.com.au logs user activities are recorded in chronologically-ordered action sequences, where each action is one of: (1) \textit{impression}; (2) \textit{click-through}; or (3) \textit{application}. An \textit{impression} is recorded when a job summary is fully visible on screen for 500 milliseconds or more, suggesting that the user has inspected it, and noting that the SERP structures employed in job search typically only have one or two job summaries fully visible at any given time. A collection of impressions is thus a valuable resource for understanding user viewing behavior.
Table 2: Interaction logs from three commercial search engines (app-based and browser-based Seek.com.au job search; and Yandex.ru Web search). Note that these logs are used to investigate search interaction patterns, and not to address system performance in any way.

<table>
<thead>
<tr>
<th></th>
<th>Seek.com.au</th>
<th>Yandex</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mobile</td>
<td>browser</td>
<td>Yandex</td>
</tr>
<tr>
<td>Users</td>
<td>4,962</td>
<td>4,868</td>
<td>unknown</td>
</tr>
<tr>
<td>Sessions</td>
<td>56,737</td>
<td>19,269</td>
<td>1,000,000</td>
</tr>
<tr>
<td>Queries</td>
<td>121,840</td>
<td>75,401</td>
<td>1,362,421</td>
</tr>
<tr>
<td>SERP size</td>
<td>unlimited</td>
<td>unlimited</td>
<td>truncated,10</td>
</tr>
<tr>
<td>Pagination</td>
<td>no</td>
<td>20–22</td>
<td>no</td>
</tr>
<tr>
<td>Domain</td>
<td>jobs</td>
<td>jobs</td>
<td>web</td>
</tr>
<tr>
<td>Clicks</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Impressions</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Rel. judg.</td>
<td>no</td>
<td>no</td>
<td>binary</td>
</tr>
</tbody>
</table>

Figure 3: Fraction of sessions for job search and Web search, as a function of session length.

In these Seek.com.au logs a *click-through* action arises when the user selects a job summary in the SERP and accesses its corresponding job details page; similarly, an *application* action arises when the user clicks the “apply” button in that job details page. *Click-throughs* and *applications* are useful signals when measuring the quality of a search ranking, especially in the context of job search [18, 32, 34].

The third set of interaction data was obtained from a publicly-available Web search log of a Russian search engine, Yandex.ru. This dataset contains a collection of ordered click-through sequences (rather than impressions or other actions); plus relevance judgments for a subset of the queries, made a year after the logs had been collected. In general, Yandex.ru Web search users are less inclined to reformulate than Seek.com.au job search users. Figure 3
shows the fraction of sessions, categorized by the session length, in these three datasets.

It is useful to be able to study the query-level behavior of users when they are engaged with organic results and see no paid items – for example, to get a clear understanding of user activities at page boundaries. The two organic-only subsets of the Seek.com.au logs contain, respectively, 58,645 SERPs initiated from 3,970 app-based users, and 40,882 SERPs from 4,003 browser-based users.

### 3.2. Laboratory-Based Datasets

Three laboratory-based session search logs are also provided in support of our analysis and experiments, and are summarized in Table 3. The first lab-based set of interaction data was created by Jiang et al. [13]. It contains a set of eye fixation sequences collected using a Tobii 1750 eye-tracker for a minimum fixation duration of 100 milliseconds [13]. We employ observations from this dataset to infer user viewing behaviors for the Yandex.ru logs, relying on the fact that both arise from typical web search scenarios.

The second lab-based resource was developed by Mao et al. [24]. Their THUIR2 dataset contains 223 multi-query sessions and corresponding SERPs, for which the top-five documents in each SERP have been assessed using four-level relevance judgments.

The third lab-based dataset, THUIR3, was developed by Liu et al. [21] to investigate query- and session-level user satisfaction. This data shares some characteristics with THUIR2. However, THUIR3 has more sessions compared to THUIR2, and includes relevance judgments for the top-10 results in each SERP, rather than the top-5.

### 3.3. Action Sequences and Sessions

Any interaction log, whether from a search service or a laboratory study, can be viewed as a collection of action sequences, where an action sequence

\[ A = \langle (a_1, p_1), (a_2, p_2), (a_3, p_3), \ldots \rangle \]
traces a user’s behavior when presented with a SERP. Each element \((a_t, p_t)\) has a timestamp \(t\) and two further attributes: (1) an action type, \(a_t\); and (2) a rank position at which the action took place, \(p_t \in \mathbb{Z}_{>0}\). As already noted, the Seek.com.au logs include three types of action: impressions \((a_t = \text{“I”})\), click-throughs \((a_t = \text{“C”})\), and applications \((a_t = \text{“A”})\). In the Yandex.ru log the action sequences contain only “C” actions.

A session \(S\) is a chronologically-ordered sequence of queries, \(Q_1 \rightarrow Q_2 \rightarrow \cdots \rightarrow Q_m\), where the \(i\)th query \(Q_i\) corresponds action sequence \(A_i\). For example, consider a three-query session \(S\):

\[
A_1 = (\text{“I”}, 1), (\text{“I”}, 2), (\text{“C”}, 4), (\text{“I”}, 2), (\text{“I”}, 3)
\]

\[
A_2 = (\text{“I”}, 1), (\text{“I”}, 2), (\text{“C”}, 2), (\text{“A”}, 2), (\text{“I”}, 3), (\text{“I”}, 5), (\text{“I”}, 6)
\]

\[
A_3 = (\text{“I”}, 1), (\text{“I”}, 3), (\text{“C”}, 3), (\text{“A”}, 3), (\text{“I”}, 4), (\text{“I”}, 7), (\text{“I”}, 5)
\]

In the first query the user examined the items at ranks 1, 2, and 4, then clicked at rank 4, then viewed the items at rank 2 (again) and rank 3. In the third query, after their second reformulation, the user viewed two items, clicked at rank 3 and started an application, and then examined three further items before completing their session.

4. A Session-Based C/W/L Framework

To obtain a session-based framework we add a second dimension to the query-level C/W/L definitions, as was anticipated by Figure 2. Figure 3 explains this idea, unrolling the loops in Figure 2 and showing the browsing paths a user might follow through a search session. It is assumed as a first base case that the user starts the session by submitting a first query \(Q_1\), and then inspecting the first item in the corresponding SERP, to accumulate the gain \(r_{1,1}\). It is also assumed, as a second base case, that if the user reformulates \(j - 1\) times and issues query \(Q_j\), then they always look at the first document in the corresponding \(j\)th SERP, to obtain the gain \(r_{j,1}\).

More generally, suppose that the user has just examined the document at rank \(i\) in the SERP for query \(Q_j\). They must next decide whether to continue in the same SERP and examine the item at rank \(i + 1\), or whether to exit this SERP, and do something else. This choice is modeled by \(C(j, i)\), the conditional continuation probability of the user proceeding from the \(i\)th to the \(i + 1\)th item in the SERP for the \(j\)th query.

If the user decides to abandon query \(Q_j\), they might issue a reformulated \(j + 1\)th query \(Q_{j+1}\), assumed to happen with conditional probability \(F(j)\); or, with conditional probability \(1 - F(j)\), they might end their search session. That is, \(C(j, i)\) and \(F(j)\) provide conditional continuation probabilities that together determine the fraction of user attention spent looking at the \(i\)th document in the \(j\)th SERP.

As with the one-dimensional C/W/L arrangement, it is also possible to compute the probability that the \(i\)th item in the \(j\)th SERP is the last one inspected by the user, denoted
Figure 4: Unrolling Figure 2 to obtain possible browsing paths. The resulting automaton has one start state (at top left) and an arbitrary number of final states.

by \( L(j, i) \); and likewise compute the weight function \( W(j, i) \) that describes the fraction of user attention that is given to the \( i \) th item in the \( j \) th SERP. Both \( W(j, i) \) and \( L(j, i) \) sum to one over their two (possibly infinite) dimensions: \( \sum_j \sum_i W(j, i) = \sum_j \sum_i L(j, i) = 1 \). It is also possible for \( C(j, i) \) and \( F(j) \) to be either static or adaptive.

**ERG Session Metrics.** Given these definitions, the session-based expected rate of gain (ERG) metric is defined via:

\[
s_{M_{ERG}}(\vec{r}) = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} W(j, i) \cdot r_{j,i},
\]

where \( \vec{r} = \langle r_1, r_2, \ldots \rangle \) are the gain vectors corresponding to the sequence of SERPs generated in the course of the session, and where \( W(j, i) \) is the proportion of time the user spends examining the \( i \) th document in the \( j \) th SERP. As with the query-based one-dimensional C/W/L framework, \( W(j, i) \), \( L(j, i) \), and \( C(j, i) \) for any single SERP can be computed from each other. For example, \( C(j, i) \) can be computed from \( W(j, i) \) using \( C(j, i) = W(j, i + 1) / W(j, i) \).

To determine the \( W(j, i) \) values, let \( V(j, i) \) be the proportion of users that inspect the \( i \) th item in the ranking associated with the \( j \) th SERP, with, by definition, \( V(1, 1) = 1 \). If function \( F(j) \) is static, and unaffected by \( r_{j,i} \), then conservation of flow (see Figure 4) means that \( V(j, i) \) can be computed as:

\[
V(j, i) = \begin{cases} 
1 & j = 1, \ i = 1 \\
F(j - 1) \cdot V(j - 1, 1) & j > 1, \ i = 1 \\
C(j, i - 1) \cdot V(j - 1, i) & i > 1.
\end{cases}
\]
The weight function \( W(j, i) \) is then a normalization of \( V(j, i) \):

\[
W(j, i) = \frac{V(j, i)}{\sum_k \sum_l V(k, l)},
\]

and computation of the metric value requires calculating sufficiently many values of \( V(j, i) \) that the sum in the denominator is accurately known. On the other hand, if \( F(j) \) of \( C(j, i) \) are influenced by the gain values in the rankings, the metric is adaptive, and Equation 3 cannot be applied. In this case SERP-specific \( V(j, i) \) values should be computed by counting what happens through the course of a large number of randomized trials, with each Monte-Carlo trial tracing a path through the automaton shown in Figure 4.

**ETG Session Metrics.** The expected total gain (ETG) value can then be computed as \( sM_{ETG}(\vec{r}) = sM_{ERG}(\vec{r})/W(1,1) \), with \( 1/W(1,1) \) the expected number of items viewed, matching the corresponding computation in the single-query C/W/L framework, where the expected number of items viewed is \( 1/W(1) \).

**Existing Session Metrics.** It is now possible to explain two existing session-based metrics: session-based DCG (sDCG) \[11\], and LCY-sRBP \[19\], with the latter an ERG metric defined as:

\[
C_{LCY-sRBP}(j, i) = qp \quad \text{and} \quad F_{LCY-sRBP}(j) = \frac{p - qp}{1 - qp}.
\]

For sDCG, truncation at a defined evaluation depth is required, since the original discount function does not have a limiting sum. We use \( sDCG@(m,n) \) to denote sDCG computed over a sequence of \( m \) SERPs, and to depth of \( n \) in each SERP. Then \( sDCG@(m,n) \) is an ETG metric specified by four parameters – the session- and query-level persistence coefficients, \( bq \) and \( b \); the session depth \( m \); and the ranking depth \( n \):

\[
C_{sDCG}(j, i) = \begin{cases} 
\frac{1 + \log_b(i)}{1 + \log_b(i+1)} & i < n \\
0 & i \geq n
\end{cases} \quad \text{and} \quad F_{sDCG}(j) = \begin{cases} 
\frac{1 + \log_{bq}(j)}{1 + \log_{bq}(j+1)} & j < m \\
0 & j \geq m
\end{cases}
\]

**5. Search Behaviors**

The previous section described the session-based C/W/L structure and its two key factors: the conditional continuation probability (determining query-level behavior), and the conditional reformulation probability (governing session-level behavior). It also showed that two existing session evaluation metrics could be described within that framework, hence establishing the generality of the proposed approach.

In this section we employ interaction logs from two commercial search engines, Seek.com.au and Yandex.ru, to investigate factors contributing to the two key behaviors – the user’s conditional continuation probability within each SERP, and their conditional reformulation
probability as they transition between SERPs. To achieve that goal, Section 5.1 considers the extent to which observable user actions on a per-SERP basis, such as impression sequences and click-through data, allow conditional continuation probabilities at the per-query level to be inferred. In particular, because the datasets that are available to us are absent of explicit continuation signals, we first summarize our previous work in the area, and confirm that those estimators are a good fit. Section 5.1 also explores the extent to which item relevance can be inferred from user actions, since relevance is another signal required for our proposed model that is not directly available in all of the test datasets. Our model also requires as input the value $T$, the user’s anticipated volume of required relevance over the session; that value is the final one examined in Section 5.1.

Section 5.2 then considers similar session-level attributes, primarily the reformulation probability, and shows that it can be similarly estimated based on user observations. With those estimation procedures in place, Section 5.3 is then able to carry out an analysis of the extent to which those inferred behaviors match the corresponding predictions from the C/W/L query-level framework, seeking trend correlations, and validating previous work that used different data sources. Section 5.4 then steps to a session-level analysis. It considers the extent to which the available data provides evidence of factors that affect session level behavior, again seeking confirmation of general principles in regard to which factors exert an influence over visible user actions. That is, each of the next four subsections has as its purpose the fitting of various types of available data to the C/W/L model of user behavior queries and sessions, and the development of “big picture” relationships that are then refined into a specific session-metric proposal in Section 6.

5.1. Inferring Non-Measured User Query Behavior

The analysis we seek to carry out must, of necessity, be based on available information. This section discusses way in which other required values can be inferred from those variables that have been captured in the various logs.

*Inferring $C(\cdot)$ From Impression Data.* To measure the influence of those three factors we compute empirical estimates from the action sequences. Wicaksono and Moffat [42] assign a binary indicator variable $c_i$ to each observation in each action sequence, with the indicator “1” if a continuation occurs at that point, and “0” otherwise. Wicaksono and Moffat further describe “rule G”, which assigns a continuation of “1” to each action that is followed (at any subsequent time-step in that action sequence) by any action at a deeper rank position. The indicator values at rank $i$ can then be aggregated across all of the available action sequences to compute an estimate $\hat{C}(i)$ of the underlying value $C(i)$.

Consider the example action sequences $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{A}_3$ listed in Section 3.3. Using rule G, the sequence $\mathcal{A}_1$ contains non-continuations at rank positions 4 and then 3, since neither are followed by an action at a higher rank. Complete continuation vectors for the three action
Table 4: Calculation of $T_0$, $T_j$, $T_{j,i}$, and a continuation indicator $c_i$, for a session of three action sequences, and assuming $T_α = 0.5$. Note that $c_i$ is computed only for impression actions, since a click and/or a job application imply an impression.

<table>
<thead>
<tr>
<th>$A_1$, $j = 1$</th>
<th>$T_0$</th>
<th>$T_j$</th>
<th>$T_{j,i}$</th>
<th>$i$</th>
<th>$a_i$</th>
<th>$c_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
<td>1</td>
<td>&quot;T&quot;</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
<td>2</td>
<td>&quot;T&quot;</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
<td>4</td>
<td>&quot;T&quot;</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
<td>4</td>
<td>&quot;C&quot;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
<td>2</td>
<td>&quot;T&quot;</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
<td>3</td>
<td>&quot;T&quot;</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$A_2$, $j = 2$</th>
<th>$T_0$</th>
<th>$T_j$</th>
<th>$T_{j,i}$</th>
<th>$i$</th>
<th>$a_i$</th>
<th>$c_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
<td>1</td>
<td>&quot;T&quot;</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
<td>2</td>
<td>&quot;T&quot;</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
<td>2</td>
<td>&quot;C&quot;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>2.5</td>
<td>1.5</td>
<td>2</td>
<td>&quot;A&quot;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>2.5</td>
<td>1.5</td>
<td>3</td>
<td>&quot;T&quot;</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>2.5</td>
<td>1.5</td>
<td>5</td>
<td>&quot;T&quot;</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>2.5</td>
<td>1.5</td>
<td>6</td>
<td>&quot;T&quot;</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$A_3$, $j = 3$</th>
<th>$T_0$</th>
<th>$T_j$</th>
<th>$T_{j,i}$</th>
<th>$i$</th>
<th>$a_i$</th>
<th>$c_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5</td>
<td>1.5</td>
<td>1.5</td>
<td>1</td>
<td>&quot;T&quot;</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>1.5</td>
<td>1.5</td>
<td>3</td>
<td>&quot;T&quot;</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>1.5</td>
<td>1.5</td>
<td>3</td>
<td>&quot;C&quot;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>1.5</td>
<td>0.5</td>
<td>3</td>
<td>&quot;A&quot;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>1.5</td>
<td>0.5</td>
<td>4</td>
<td>&quot;T&quot;</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>1.5</td>
<td>0.5</td>
<td>7</td>
<td>&quot;T&quot;</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>1.5</td>
<td>0.5</td>
<td>5</td>
<td>&quot;T&quot;</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

sequences are shown in the last column of Table 4 (the other columns are discussed shortly). The empirical $\hat{C}(\cdot)$ is obtained by dividing the number of continuations at rank $i$ by the number of actions at rank $i$. In the example, $\hat{C}(1) = 3/3$, $\hat{C}(3) = 2/3$, and $\hat{C}(5) = 1/2$.

**Inferring $C(\cdot)$ From Click-Through Data.** The Yandex.ru data only contains click-throughs and no other signals, and hence potentially underestimates the notion of “examine rank $i$”. While it is clear that, on the whole, users scan the ranking sequentially [6, 14], and hence that it is possible to assume that exactly the items through to the deepest click position have been inspected by the user [19], other models for estimating impressions are also possible. Figure 5, based on information contained in the J&A dataset, shows that users tend to examine the majority of the items prior to the deepest click rank position, plus a smaller number of results
Figure 5: Percentage of action sequences in which each rank position is viewed, stratified by the deepest click rank (J&A dataset).

Table 5: Estimated probability of relevance when the user clicks on a particular item ($\Theta = 1$), and when the user views an item but does not click on it ($\Theta = 2$), for the J&A dataset with three-level graded relevance judgments ($r \in \{0, 1, 2\}$). The last column shows the expected relevance grade for both conditions. The difference between the two conditions is significant ($\chi^2 = 125.7, p < 0.01$).

<table>
<thead>
<tr>
<th>Condition</th>
<th>$\Theta$</th>
<th>$\hat{P}(r = 0 \mid \Theta)$</th>
<th>$\hat{P}(r = 1 \mid \Theta)$</th>
<th>$\hat{P}(r = 2 \mid \Theta)$</th>
<th>$E[r]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viewed and clicked</td>
<td>1</td>
<td>0.263</td>
<td>0.301</td>
<td>0.436</td>
<td>1.173</td>
</tr>
<tr>
<td>Viewed only</td>
<td>2</td>
<td>0.518</td>
<td>0.211</td>
<td>0.271</td>
<td>0.753</td>
</tr>
</tbody>
</table>

Beyond that point, confirming similar analysis by Zhang et al. [48].

To infer gaze distributions for the Yandex.ru click-through sequences, including beyond the deepest click, we employ the first model (of two) proposed by Wicaksono et al. [44], using the J&A dataset as a basis for parameter estimation. This approach has one parameter $K$ that needs to be trained. Suppose that $\hat{P}(\text{view} = i \mid u, q)$ is the probability that user $u$ inspects the item at rank $i$ for query $q$, and $DC(u, q)$ is the known deepest click rank for that same user. Then $\hat{P}(\text{view} = i \mid u, q)$ is estimated as follows:

$$\hat{P}(\text{view} = i \mid u, q) = \begin{cases} 1 & \text{if } i \leq DC(u, q) \\ e^{\frac{DC(u, q) - i}{K}} & \text{otherwise} \end{cases}$$

Based on a best-fit search using the J&A dataset, the value $K = 1.4$ is chosen for use with the Yandex.ru click sequences. Using this model, the continuation variable, $c_i$, is given by:

$$c_i = \frac{\hat{P}(\text{view} = i + 1 \mid u, q)}{\hat{P}(\text{view} = i \mid u, q)} = \begin{cases} 1 & \text{if } i \leq DC(u, q) \\ e^{-1/K} & \text{if } i > DC(u, q) \end{cases}$$

We further assume that item that are viewed but not clicked are non-relevant [4, 14, 31]. Table 5, also derived from the J&A data, provides support for this statement. The assumption...
Table 6: Multiplicative effect sizes for did_click and did_apply, optimized to fit the editorial relevance values using a logistic regression.

<table>
<thead>
<tr>
<th>Factor</th>
<th>coef.</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>2.721</td>
<td>0.00</td>
</tr>
<tr>
<td>did_click</td>
<td>0.527</td>
<td>0.00</td>
</tr>
<tr>
<td>did_apply</td>
<td>0.173</td>
<td>0.00</td>
</tr>
</tbody>
</table>

that clicks correspond to relevance is useful in the context of the Yandex.ru data, because it only contains judgments for clicked items (Table 2), and it is helpful to assume that unclicked documents are non-relevant.

Inferring Relevance. The Seek.com.au dataset does not include any explicit editorial relevance judgments, and we are even more reliant on implicit signals. Fortunately, in the context of job search, the “application” action, in which a user sends their details to an employer, provides a second signal, one that is more tightly coupled with relevance than click-throughs alone.

To support that claim, a logistic regression analysis was carried out with relevance value \( r_i \) as the response variable, and two binary indicators, did_click and did_apply associated with click-through and job application actions, as the explanatory variables. This regression analysis employs a set of relevance judgments (qrels) consisting of 5,145 \( \langle \text{query}, \text{document}, r_i \rangle \) triples on a four-point relevance scale. These qrels were gathered for 205 common job search queries, via a crowd-sourcing platform between June and August 2017\(^1\) prior to the collection of the data described in Table 2. Using the Seek.com.au qrels, we then collected a sample of search logs (corresponding to the qrels) containing around 7,000,000 tuples \( \langle \text{query}, \text{document}, r_i, \text{did_click}, \text{did_apply} \rangle \) observed between February and November 2017. Let this collection of tuples be denoted by \( H \). To estimate the coefficients of the regression model, we only used a subset of the logs, \( H' \subset H \), that contains 5,970,120 tuples, for which the query-document pairs are either fully-relevant \( (r_i = 1) \) or non-relevant \( (r_i = 0) \).

Table 6 shows the effect sizes for did_click and did_apply, when optimized to fit the available editorial relevance values. Note that a job application is always preceded by a click-through, and hence:

\[ \text{did_apply} = 1 \implies \text{did_click} = 1. \]

The relatively large positive intercept value might be caused by the fact that the majority of the query-document pairs in the sample are fully-relevant. The corresponding odds ratios

\(^1\)These qrels were collected by Damiano Spina (RMIT University) and Bahar Salehi (The University of Melbourne) as part of the overall collaboration between the two institutions and Seek.com.au.
Figure 6: Click-through rate (top row) and distribution of $P(\text{App} \mid \text{Click})$ (bottom row), each as a function of relevance grade ($r_i$), computed from the collection of tuples $H$. The small green triangle in each box-whisker element is the mean. Mobile-based queries are shown in the left-hand column, and browser-based queries in the right-hand column. The vertical scale is linear; but for commercial-in-confidence reasons is not labeled.

are exp(0.527) = 1.694 (a 69% increase in the odds of being relevant when clicked) and exp(0.527 + 0.173) = 2.014 (in combination, double odds of being relevant when then applied for).

To reinforce the results presented in Table 6, we also compute click-through rate (CTR) and application probability conditioned on click action, $P(\text{App} \mid \text{Click})$, using the original collection of tuples $H$. The results are then stratified by the relevance grade ($r_i$). Click-through rate and $P(\text{App} \mid \text{Click})$ for a query-document pair are estimated as follows:

$$\text{CTR}((\text{query}, \text{document})) = \frac{\#\text{Click}((\text{query}, \text{document}))}{\#\text{Count}((\text{query}, \text{document}))},$$

$$P(\text{App} \mid \text{Click}, (\text{query}, \text{document})) = \frac{\#\text{App}((\text{query}, \text{document}))}{\#\text{Click}((\text{query}, \text{document}))},$$

where $\#\text{Count}(x)$ is the frequency of the query-document pair $x$ in the set of tuples $H$, and $\#\text{Click}(x)$ and $\#\text{App}(x)$ are the numbers of occurrences of the query-document pair $x$ with, respectively, $\text{did}\_\text{click} = 1$ and $\text{did}\_\text{apply} = 1$. Figure 6 shows the resultant distributions. In general the click-through rate CTR() and $P(\text{App} \mid \text{Click})$ tend to increase as a function of document relevance $r_i$. Hence:

**Assumption 1.** For job search, the item at rank position $i$ in sequence $A$ is relevant if and
only if a job application action was observed at $i$:

$$\{A, i\} \in A \iff r_i = 1.$$  

Inferring Relevance Target $T$. To estimate the user’s anticipated relevance target $T$ from interaction logs, we suppose that:

Assumption 2. Users complete their search session at some point after they have met their expected volume of relevance, that is, at some point after $T_i$ reaches 0.

Moffat et al. [27] conducted a user study that, to some extent, supports Assumption 2. In connection with a laboratory-based search experiment, participants were asked to indicate the number of useful web pages they would expect to need to see to respond to the search task. While not arguing that the users only exit their search once $T_i = 0$, Moffat et al. [27] show that the odds of stopping do increase as $T_i$ decreases.

With Assumptions 1 and 2, we might infer that $T$ at the beginning of a job search session is equal to the number of job applications observed in the session. However, this creates a problem for sessions with no applications, since a user with $T = 0$ has no need to search, clearly a contradiction. To smooth this discontinuity, $T_\alpha > 0$ is added to the number of job applications triggered in the session, as a background expectation. In the absence of any specific knowledge, $T_\alpha = 0.5$ can be used, and that is what is done in all of the experiments reported in this paper.

5.2. Inferring Non-Measured User Session Behavior

It is also necessary to infer certain values as they evolve through a search session, most notably $T$, the anticipated volume of relevance. By definition, $T$ is constant for any given query. But it is also desirable to have it vary across the sessions’ queries. To allow that, we distinguish between the session-level value, denoted by $T_0$; and a per-query value $T_j$, representing the remaining anticipated relevance at the commencement of the $j$th query in the session. We further define $T_{j,i}$ to be the remaining anticipated relevance after the $i$th item in the $j$th SERP has been viewed (with $T_{j,0} \equiv T_j$ useful as a notational convenience); and $T_{j,*}$ to be the remaining anticipated relevance when the $j$th SERP is exited (with $T_{0,*} \equiv T_0$, similarly). Further, while $T_{j,i}$ might in general be negative, it is not appropriate for $T_j$ to be zero or less if the $j$th query is being assumed to have been issued. These considerations lead to:

$$T_{j,i} = \begin{cases} 
\max(T_{j-1,*}, T_\alpha) & i = 0 \\
T_{j,i-1} - r_{j,i} & i > 0.
\end{cases} \tag{4}$$

Finally, if $napp(A)$ is the number of jobs applied for in one of the Seek.com.au action sequences, then $T_0$ is estimated via $\hat{T}_0 = T_\alpha + \sum_{k=1}^{\infty} napp(A_k)$. Table 4 (Section 3) shows the computation of these various values for the three-query example session described in Section 3.3.
Table 7: Effect sizes for factors in a fitted model of a binary continuation indicator, $c_i$, across all of the queries in the sessions, with each factor computed independently of all other factors.

<table>
<thead>
<tr>
<th>Factor</th>
<th>iOS/Android</th>
<th>browser</th>
<th>yandex</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>coef.</td>
<td>$p$</td>
<td>coef.</td>
</tr>
<tr>
<td>$i$</td>
<td>0.02</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td>$T_j$</td>
<td>0.11</td>
<td>0.00</td>
<td>0.04</td>
</tr>
<tr>
<td>$T_0$</td>
<td>0.06</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td>$i % 20 = 0$</td>
<td>–</td>
<td>–</td>
<td>$-1.02$</td>
</tr>
<tr>
<td>$T_{j,i}$</td>
<td>0.11</td>
<td>0.00</td>
<td>0.03</td>
</tr>
</tbody>
</table>

For the Yandex.ru data, $napp(A)$ is replaced by $nre(A)$, the number of distinct relevant items clicked in the action sequence.

5.3. Query-Level Evaluation

With that machinery in place, the goal now is to use the interaction logs that are available to explore user behaviors through the lens of C/W/L, first at the per-query level, to confirm the previous observations that have been made; and then at the pan-query level, across search sessions, to discover new behaviors. The clear focus is how $C(\cdot)$ and its session-level counterpart $C(\cdot, \cdot)$ varies with respect to each of: (1) the current rank position, $i$; (2) the initial anticipated volume of relevance, $T$; and (3) the adjusted amount of relevance still unfulfilled after $i$ items in the ranking have been considered, $T_i = T - \sum_{k=1}^{i} r_k$.

A sequence of logistic regressions was used to obtain an overview of how these various factors contribute to $C(\cdot, \cdot)$, using the binary continuation indicator as a response variable in independent one-variable analyses. An additional input indicator “$i \% 20 = 0$” was also included, to allow for pagination, a likely issue with the browser-based queries in the Seek.com.au logs. The organic-results-only subset of the Seek.com.au results was used, with the two alternatives of inferring $T$, namely $T_0$ and $T_j$, tested separately. This separation was necessary because $T_0$ and $T_j$ are highly correlated (Pearson-$r > 0.8$), and putting them together in a generalized linear model would generate unreliable estimates of their coefficients, since they largely explain the same variance.

Table 7 gives multiplicative effect sizes for $i$, $T_0$ or $T_j$, and $T_{j,i}$, aggregated over queries and all sessions. Intercepts are not shown since they are redundant in this case ($i$ is always greater than zero). The indicator “$i \% 20 = 0$” was not included when modeling the continuous-scroll data. In these results the signs are of more interest than the magnitudes, and show that all of the rank position $i$, the user goal $T$ (when expressed either as $T_0$ or as $T_j$), and $T_{j,i}$ are all positively correlated with $c_i$. Pagination has a strong negative relationship with continuation.
Table 8: Effect sizes in a fitted model of the continuation indicator, \( c_i \), tabulated separately for first three queries in each sessions, and again computed as a sequence of independent regressions.

<table>
<thead>
<tr>
<th>Query</th>
<th>Factor</th>
<th>iOS/Android</th>
<th>browser</th>
<th>yandex</th>
</tr>
</thead>
<tbody>
<tr>
<td>( j = 1 )</td>
<td>( i )</td>
<td>0.02</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>( T_j )</td>
<td>0.12</td>
<td>0.08</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>( T_0 )</td>
<td>0.12</td>
<td>0.08</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>( i%20 )</td>
<td>–</td>
<td>–1.05</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>( T_{j,i} )</td>
<td>0.14</td>
<td>0.07</td>
<td>1.06</td>
</tr>
<tr>
<td>( j = 2 )</td>
<td>( i )</td>
<td>0.02</td>
<td>0.02</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>( T_j )</td>
<td>0.09</td>
<td>0.09</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>( T_0 )</td>
<td>0.07</td>
<td>0.07</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>( i%20 )</td>
<td>–</td>
<td>–1.03</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>( T_{j,i} )</td>
<td>0.08</td>
<td>0.08</td>
<td>0.76</td>
</tr>
<tr>
<td>( j = 3 )</td>
<td>( i )</td>
<td>0.02</td>
<td>0.02</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>( T_j )</td>
<td>0.12</td>
<td>0.07</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>( T_0 )</td>
<td>0.08</td>
<td>0.06</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>( i%20 )</td>
<td>–</td>
<td>–0.94</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>( T_{j,i} )</td>
<td>0.11</td>
<td>0.07</td>
<td>0.51</td>
</tr>
</tbody>
</table>

behavior, and in the browser-based Seek.com.au queries, users are much more likely to end their inspection of the SERP on page boundaries than at other ranks.

Table 8 separates the effects according to each SERP’s position in the session, covering the first three queries in each session. Similar trends arise in the first two queries, with some weakening of those patterns in the third queries. In terms of overall behavior, each of the SERPs is treated broadly the same by the users viewing them. Figure 7 illustrates the effect of the factors \( i \) and “\( i%20 = 0 \)” that are described in Table 8. It is clear that \( C(\cdot, \cdot) \) increases with rank position \( i \), and exhibits a significant drop at page boundaries in browser-based queries. It is also clear that user behavior is broadly the same in first queries as it is in second and third queries.

Overall, these results confirm that, as hypothesized, the conditional continuation probability at the query level is positively correlated with all of the rank position \( i \), the user’s initial relevance target \( T \), and the user’s progress towards resolving their information need, expressed as the unfulfilled relevance goal \( T_i \) or \( T_{j,i} \).
Figure 7: Empirical conditional continuation probabilities for ranks \( i \leq 30 \), for the first three queries in each session \( (j \leq 3) \), for Seek.com.au app-based (top) and browser-based (bottom) queries.

5.4. Session-Level Evaluation

Users can exit from each SERP at any rank position. Once that happens, they either also end their session, or reformulate and issue a new query. The factors that drive that decision are also of interest, and in this section we model the user’s propensity to reformulate to \( Q_{j+1} \), given they have exited from the SERP for query \( Q_j \). This is the conditional reformulation probability \( F(j) \) that was introduced in connection with Figure 2.

Possible explanatory factors include the query’s position in the session, \( j \); the initial relevance target, \( T_0 \); and the remaining volume of anticipated relevance upon exit from the \( j \)th SERP, \( T_{j,*} \). Our hypotheses in regard to these factors are as follows:

1. \( F(j) \) increases with query count \( j \), so that the more the user reformulates their queries, the more likely it is that they reformulate again in the future.

2. \( F(j) \) increases with \( T_0 \), so that the likelihood of reformulating increases as the total anticipated relevance increases. This hypothesis suggests that a user model should provide parameters to allow different numbers of reformulations, and both sDCG and LCY-sRBP support this to some extent via their parameters.

3. \( F(j) \) decreases as \( T_{j,*} \) decreases, so that as the user accumulates answers toward their goal, they are more likely to end their search session. Neither sDCG nor LCY-sRBP are compliant with this hypothesis.
Table 9: Independent-regression effect sizes and p values for factors in a fitted model for the binary reformulation indicator $f_j$.

<table>
<thead>
<tr>
<th>Factor</th>
<th>iOS/Android</th>
<th>browser</th>
<th>yandex</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>coef.</td>
<td>p</td>
<td>coef.</td>
</tr>
<tr>
<td>$j$</td>
<td>0.14</td>
<td>0.00</td>
<td>0.08</td>
</tr>
<tr>
<td>$T_0$</td>
<td>0.16</td>
<td>0.00</td>
<td>0.15</td>
</tr>
<tr>
<td>$T_{j,*}$</td>
<td>6.83</td>
<td>0.00</td>
<td>4.00</td>
</tr>
</tbody>
</table>

Logistic regression is again used, now to model the query reformulation decision at the end of query $j$, denoted by $f_j$, a dichotomous variable, with $f_j = 1$ and $Q_j$ followed by $Q_{j+1}$), or $f_j = 0$ and $Q_j$ the last query for the session. The explanatory variables are again fitted independently via a sequence of regressions, with the sign of the resultant coefficient of primary interest.

Table 9 lists the resultant effect coefficients. In the case of the Yandex.ru data, L1 regularization is used, since a quasi-complete separation with respect to $T_{j,*}$ led to a very large coefficient value. In general, the three hypotheses regarding $F(j)$ are validated, with positive coefficients emerging for all three explanatory variables, albeit partially as a result of the way in which the explanatory variable $T_0$ was inferred.

Figure 8 shows the positional distribution of last application (Seek.com.au, left) and last relevant click (Yandex.ru, right) as a function of query number in the session (horizontal axis), stratified by the number of queries in the session (vertical axis). The values across each row sum to one. The Seek.com.au browser-based users have a similar trend.

Finally in this section, Figure 9 plots empirical conditional reformulation probabilities,
Recall that $\hat{F}(j)$ is the empirical probability of users reformulating the $j$th query. The vertical scale is linear; but for commercial-in-confidence reasons associated with the Seek.com.au data, is not labeled.

$\hat{F}(j)$, computed as:

$$\hat{F}(j) = \frac{\sum_{S} \mathbb{I}(Q_j \in S \land f_j = 1)}{\sum_{S} \mathbb{I}(Q_j \in S)},$$

where the indicator function $\mathbb{I}(P)$ yields 1 if $P$ holds, and 0 if not. The upward trend of $\hat{F}(j)$ is clear, and matches the within-query continuation behavior illustrated in Figure 7. Note also how Yandex.ru web search users are far less inclined to reformulate queries than are Seek.com.au job search users.

6. A Model-Based Session Metric

Section 5 established a number of behaviors that can be associated with $C(i)$ and $F(j)$ in the extended C/W/L framework shown in Figure 2 (see Table 9 and Figure 9). We now crystallize those observed relationships trends into a specific model for session evaluation, and measure its fit against observation compared to two previous session-based evaluation proposals. While the model might still employ parameters such as the initial session target $T_0$, it should not depend on additional information other than relevance judgments, since the main goal is to use the model to develop collection-based whole-of-session effectiveness metric. That is, just as the query-level C/W/L metrics already developed do not rely on exact knowledge of individual users and individual SERPs, we wish to develop a session-level C/W/L metric that is similarly not dependent on knowledge of how many queries any particular user issued, nor how deep in each of their SERPs that user browsed.
6.1. Session-Based INST

We again refer to the C/W/L framework, and develop a “two-dimensional” version of the metric INST. The first step is to demonstrate by example that it is possible to have a reformulation function that displays the required properties.

Modeling $F(j)$. Taking our lead from the one-dimensional (that is, single query) proposal of Moffat et al. [27, 29], we suggest the following idealized model for $F(j)$:

$$F(j) = \left( \frac{j + T_0 + T_{j,*}}{j + T_0 + T_{j,*} + \kappa} \right)^2,$$

where $\kappa > 0$ is an constant that controls the rate at which the reformulation probability increases with $j$. As required, all of $j$ (length of session so far), $T_0$ (initial anticipated relevance volume), and $T_{j,*}$ (remaining volume after the $j-1$ th SERP is exited) are then positively correlated with $F(j)$.

A Whole of Session Metric. To then formulate a session-based effectiveness metric we take Equation 5 and add a $C(j, i)$ function, to provide within-query conditional continuation probabilities. Again drawing on the work of Moffat et al. [27, 29], we define “session INST” (sINST) via:

$$C_{\text{sINST}}(j, i) = \left( \frac{i + T_j + T_{j,i} - 1}{i + T_j + T_{j,i}} \right)^2,$$

in effect treating the $j$ th query as an independent entity within the session for the purpose of predicting user behavior, with the user assumed to undertake their inspection of the corresponding SERP in the anticipation of gaining $T_j$ units of relevance. The $W(j, i)$ computation that defines the value of the ERG metric is then computed by aggregating over the entire sequence of queries (Equations 2 and 3 in Section 4), so that the result is an over-arching ERG score that indicates the effectiveness of the complete session when the sequence of SERPs is interpreted through the eyes of a probabilistic user.

Figure 10 shows the distribution of $V(j, i)$ for sINST and one choice of parameters ($T = 8$ and $\kappa = 3$), and four extreme scenarios. As can be seen, $V(j, i)$, which is directly proportional to $W(j, i)$, varies depending on the relevance accrued from the SERP entries that have been inspected by the user. When first SERP in the session is rich in relevance (the two left-hand plots), users make good progress towards their anticipated target $T_0$ in that SERP, and are modeled as having reduced motivation to reformulate and look at further SERPs. However, when the opposite happens (the two right-hand plots), users examine the first SERP and are then modeled as reformulating one or more followup queries, because $T_{1,*}$ is still close to $T_0$ and hence $T_2$ is still close to $T_0$.

That is, sINST is an adaptive metric, and the modeled user behavior alters according to the gain values $r_{j,i}$ of the documents each user has encountered, with both $T_j$ and $T_{j,i}$ evolving. Equation 4 completely defines $T_{j,i}$ for each individual user. However, the probabilistic nature
Figure 10: Distribution of $V(j, i)$ for sINST ($T = 8$ and $\kappa = 3$), for four extreme scenarios, where $V(j, i)$ is the proportion of users that view the $i$th document in the $j$th SERP. The four scenarios are $r_{j,i} = 1$ throughout the rankings in the session (top left); $r_{j,i} = 0$ throughout (top right); $r_{1,i} = 1$ and $r_{j,i} = 0$ for $j > 1$ (bottom left); and $r_{1,i} = 0$ and $r_{j,i} = 1$ for $j > 1$ (bottom right).

of the user model (Figure 2), and the fact that the metric represents a population of users, means that the last rank position inspected in each SERP is a random variable that itself has a distribution. Hence, $T_{j, \ast}$ has a distribution that should be allowed for when $T_{j+1,0}$ is being computed (Equation 4), with the users making up the population having different $T_j$ values at the commencement of the same $j$th query. That, in turn, means that a Monte Carlo simulation (tracing the actions of a large set of random users) is required to obtain an approximation of the $V(j, i)$ values, and hence the session metric score.

To avoid the expense of such computations, we suggest that in practice the evolution of $T_j$ be computed using expectations:

$$T_j = \max(T_{j-1} - M_{ETG}(\text{SERP}_{j-1}), T_\alpha),$$  \hspace{1cm} (7)

where $M_{ETG}(\text{SERP}_{j-1})$ is the expected total gain derived from the $j - 1$th SERP when a probabilistic user modeled by INST examines it with the anticipation of accumulating $T_{j-1}$ units of relevance. That is, as each SERP in the session is considered, the suggestion is that the distribution over $T_{j, \ast}$ be condensed to a single representative value for the purposes of
computing both $T_j$ (Equation 4) and $F(j)$ (Equation 5), so that consideration of the next SERP can be undertaken from a point starting position, rather than from a distribution.

### 6.2. Meta-Evaluation Via Held-Out Data

Wicaksono and Moffat [43] measure the accuracy of continuation probabilities through the lens of the C/W/L framework, making use of the weighted mean squared error to measure the fit between the function $C(\cdot)$ associated with a metric and its empirical value, $\hat{C}(\cdot)$, observed via system interactions. The weights on the mean squared error, which correspond to the relative frequency of the $i$th item being examined, are required since $C(\cdot)$ is not a probability distribution that sums to one. We extend this approach to meta-evaluate the accuracy of $C(j,i)$ and $F(j)$ in the context of session-based user models. Because $C(j,i)$ and $F(j)$ might have shared parameters, their optimal values are identified with respect to a single error function:

$$WMSE(\omega) = \sum_j \sum_i w_c(j,i) \cdot (C(j,i;\omega) - \hat{C}(j,i))^2 + \sum_j w_f(j) \cdot (F(j;\omega) - \hat{F}(j))^2,$$

where $\omega$ is a set of parameters of a particular user model; $w_c(j,i)$ is the relative frequency of the item at $i$th item in the $j$th SERP being viewed; and $w_f(j)$ is the weight associated with the fraction of sessions that contains at least $j$ queries.

Held-out datasets containing sessions initiated from 1,000 users for both browser- and mobile-based Seek.com.au data, and 100,000 search sessions for Yandex.ru data, were used to compute the $WMSE(\omega)$ function. Table 10 shows the parameter combinations that provide the best fit for three session-based metrics (that is, minimizing the $WMSE(\omega)$ function), across the first five queries in each session, and the first 50 results in each SERP comparing $sINST$ with the two prior session metrics described in Section 2.2, $sDCG$ [11] and LCY-sRBP [19].

As can be seen, among the three session-based user models, $sINST$ provides notably better agreement with the empirical observations, with a value of $T$ (the initial anticipated relevance) that is smaller for web search than it is for job search.

**Computation of $sINST$.** Figure 11 shows the relationship between $sINST$ when computed by the expectation method (Equation 7, assumed to have converged when computed to a depth of 2,000 over an assumed 50 SERPs, see Moffat et al. [28] for a description of how to compute $INST$-like adaptive C/W/L metrics) and the $sINST$ scores computed via a Monte Carlo simulation over 50,000 random users. Both are approximations to the true value, with the expectation method generating scores that are up to 10% higher than the more accurate measurements generated by the Monte-Carlo method. Note, however, that the estimation process provides the same relative outcomes as the Monte-Carlo approach, while executing around
Table 10: Best-fit parameters, found by minimizing $\text{WMSE}(\omega) \left( \times 10^{-2} \right)$ for three session-based user models, across the first 5 queries in each session, and the first 50 results in each query’s SERP, comparing $\text{sINST}$ with the two prior session metrics described in Section 2.2 $\text{sDCG}$ \cite{11} and LCY-sRBP \cite{19}.

<table>
<thead>
<tr>
<th>User Model</th>
<th>Best-fit $\omega$</th>
<th>WMSE($\omega$)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Seek.com.au</strong>, mobile app-based</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{sDCG}$</td>
<td>$b, q, b, M, N = 2.0, 10.0, 6, 51$</td>
<td>4.63</td>
</tr>
<tr>
<td>$\text{LCY-sRBP}$</td>
<td>$q, p = 0.93, 0.94$</td>
<td>0.40</td>
</tr>
<tr>
<td>$\text{sINST}$</td>
<td>$T, \kappa = 4.5, 4.0$</td>
<td>0.19</td>
</tr>
<tr>
<td><strong>Seek.com.au</strong>, browser-based</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{sDCG}$</td>
<td>$b, q, b, M, N = 3.0, 4.5, 6, 51$</td>
<td>2.76</td>
</tr>
<tr>
<td>$\text{LCY-sRBP}$</td>
<td>$q, p = 0.88, 0.95$</td>
<td>0.59</td>
</tr>
<tr>
<td>$\text{sINST}$</td>
<td>$T, \kappa = 4.0, 2.0$</td>
<td>0.28</td>
</tr>
<tr>
<td><strong>Yandex.ru</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{sDCG}$</td>
<td>$b, q, b, M, N = 1.5, 2.5, 2, 10$</td>
<td>7.56</td>
</tr>
<tr>
<td>$\text{LCY-sRBP}$</td>
<td>$q, p = 0.88, 0.74$</td>
<td>0.78</td>
</tr>
<tr>
<td>$\text{sINST}$</td>
<td>$T, \kappa = 2.0, 4.5$</td>
<td>0.43</td>
</tr>
</tbody>
</table>

forty times faster. That is, the high correlations shown in Figure 11 provide support for the computationally less-expensive methodology.

An additional benefit of weighted precision metrics is the ability to measure the coverage of the relevance judgments via a residual \cite{22, 26}, calculated as the difference in score that arises when all unjudged documents are deemed to be fully relevant, compared to the score that arises when all unjudged documents are deemed to be non-relevant. That is, the residual quantifies the maximum extent to which any unjudged documents in the SERP can alter the computed metric score. Figure 12 shows that the $\text{sINST}$ residuals are relatively modest across almost all of sessions in the J&A, THUIR2, and THUIR3 datasets used in our evaluations for small anticipated relevance targets $T \leq 3$; and that as $T$ increases, either the residuals and hence measurement uncertainty must of necessity increase, or more judgments must be performed. This type of analysis is a useful tool that can be applied at the time researchers are designing and costing experimental evaluations, as well as in post-evaluation checking as was done here.

7. Context and Conclusions

We have extended the C/W/L framework to session-based effectiveness evaluation, and demonstrated that existing session-based user models can be explained by this generalized
Figure 11: Monte Carlo simulation method (y-axis) versus expectation method (Equation 7, x-axis) for computing sINST score (ERG version, $T = 8$ and $\kappa = 3$), and Pearson’s correlation coefficients. The three plots show 80 sessions from the J&A dataset (top-left), 223 sessions from the THUIR2 dataset [24] (top-right), and 450 sessions from the THUIR3 dataset [21] (bottom). The plots also show lines of best fit: $y = 0.90x + 0.01$ (top-left), $y = 0.97x + 0.00$ (top-right), and $y = 0.90x + 0.01$ (bottom).

evaluation framework. In the session-based C/W/L approach a user model (describing a universe of users) is characterized by two behaviors: their conditional continuation probability at rank $i$ when examining the $j$th SERP, $C(j,i)$; and their conditional reformulation probability, $F(j)$.

Three commercial search interaction logs were used to identify factors (including $j$ and $i$ themselves) contributing to $C(j,i)$ and to $F(j)$. Results from those experiments were used first to confirm previous finding in regard to the one-dimensional conditional continuation probability that applies within each query, $C(i)$ [27, 29], namely that it is correlated with at least three different factors: $i$ itself; the initial relevance target $T$; and the amount of that anticipated target that remains to be accumulated after the document at rank $i$ has been viewed.

Treating the interaction logs as sessions of queries then confirmed that (at least) three factors similarly affect the reformulation probability $F(j)$: the query position $j$ in the session; the user’s initial anticipated relevance target $T$ at the commencement of the session; and the gap $T_{j,*}$ between the target $T$ and the amount of relevance accrued at the time the $j + 1$st query is issued. Combining the session-based C/W/L framework and the relationships that influence $F(j)$, we proposed session-based INST (sINST), and showed that it allows plausible
parameter combinations to be identified that provide a better fit to observed user behavior than do session DCG ($sDCG$, [11]) and the recently proposed session-based RBP (LCY-sRBP, [19]). That is, the new models we propose here more closely match observed user data than do the models associated with previous session metrics, and hence provide a better description of user actions and expectations. At a meta-evaluation level, we can thus conclude that the new models allow improved understanding of user behavior, and hence that the corresponding metrics will lead to system scores that more closely reflect the perceptions of system users as captured by satisfaction ratings. Ultimately, these new models will thus allow greater fidelity in system measurement, and more precise differentiation between alternative systems.

A clear future direction now is to incorporate ideas in regard to rate based user behavior [1]. For example allowing the rate at which gain has been accrued in query $j$ (rather than through the whole session so far) is an additional signal that might allow more precise estimation of $F(j)$, and at the same time could also draw on the notion of egregious non-relevance described by Moffat and Wicaksono [25]. A second direction is to carry out a more fine-grained analysis of the relationship between individual queries and user-reported satisfaction ratings using the notion of forgetfulness [20, 21, 47], with one potential argument being that forgetfulness is affected not only by query position but also by query quality.

Finally, it is appropriate to comment on what might loosely be called “usefulness”. We have justified our work here – and $sINST$ in particular – by showing that it better models, and
more closely fits, observed user behaviors over query sequences than do previous approaches. But it is also important to establish the extent to which measurements made using our proposed framework lead to different experimental outcomes compared to previous query session measurement regimes. For example, if $s_{INST}$ is 100% correlated with some other session metric for all cases of interest, then the argument for its existence is weakened. On the other hand, if $s_{INST}$ is not 100% correlated with previous approaches – by far the more likely outcome of such an experiment – then understanding the reasons for the differences that arise will help us refine our understanding of how to measure users and the way that they interact with IR systems. We plan to undertake such a detailed evaluation as future work, drawing on further data resources beyond those already used here.

Acknowledgements

This work was in part supported by the Australian Research Council (grant LP150100252, held in collaboration with Seek.com; and grant DP190101113). We gratefully acknowledge the assistance of Bahar Salehi and Justin Zobel (The University of Melbourne); Damiano Spina (RMIT University); and Sargol Sadeghi and Vincent Li (Seek). We also thank the anonymous referees, who provided helpful comments that improved the paper.

References


session-based IR systems. In Proc. European Conf. on Information Retrieval (ECIR),

[41] E. M. Voorhees. The philosophy of information retrieval evaluation. In Working Notes

Proc. Int. Conf. on Information and Knowledge Management (CIKM), pages 1571–1574,
2018.

Conf. on Web Search and Data Mining (WSDM), pages 654–662, 2020.


a bejeweled player model. In Proc. Ann. Int. Conf. on Research and Development in

[47] F. Zhang, J. Mao, Y. Liu, W. Ma, M. Zhang, and S. Ma. Cascade or recency: Construct-
ing better evaluation metrics for session search. In Proc. Ann. Int. Conf. on Research