Metrics, User Models, and Satisfaction

Alfan Farizki Wicaksono
The University of Melbourne
Melbourne, Australia

Alistair Moffat
The University of Melbourne
Melbourne, Australia

ABSTRACT
User satisfaction is an important factor when evaluating search systems, and hence a good metric should give rise to scores that have a strong positive correlation with user satisfaction ratings. A metric should also correspond to a plausible user model, and hence provide a tangible manifestation of how users interact with search rankings. Recent work has focused on metrics whose user models accurately portray the behavior of search engine users. Here we investigate whether those same metrics then also correlate with user satisfaction. We carry out experiments using various classes of metrics, and confirm through the lens of the C/W/L framework that the metrics with user models that reflect typical behavior also tend to be the metrics that correlate well with user satisfaction ratings.

KEYWORDS
Web search; user model; evaluation; effectiveness metric; session

1 INTRODUCTION
A good IR effectiveness metric should not only capture and quantify the systems it is applied to, but should also reflect the user’s experience as they interact with the search ranking in question. In particular, a metric should correlate well with observable user behaviors, in the sense that it should help explain the visible activities of the user as they search. While some authors have argued that user satisfaction alone is not reliable [12, 26, 35], the concept of user satisfaction has been widely used [10]; and recent work has shown that system effectiveness has a significant relationship with user satisfaction either via the general notion of relevance [2, 6, 14, 19] or the notion of usefulness [25, 27]. That is, a meta-evaluation that involves user satisfaction can be formulated as asking whether the metric under consideration has a strong positive correlation with user satisfaction ratings, assuming that satisfaction ratings, which are usually generated after the users have completed a query or a session, serve as a prima facie feedback of how they feel [24].

Attention has been given to search effectiveness metrics that are parameterized in terms of user behavior, such as rank-biased precision, RBP [28], which describes the behavior of users via their persistence in regard to stepping through each search ranking. Other metrics that similarly embody an explicit user model in this way include expected reciprocal rank, ERR [5]; time-biased gain, TGB [33]; and INSQ [29]. Common across these metrics is the notion of a population of users independently interacting with the ranking, exhibiting certain gross behaviors that can be accurately predicted. A second shared assumption is that each user inspects the ranking in a sequential manner, starting at the beginning.

The notion of adaptivity [30] suggests that the user decision to continue in the ranking is influenced not only by their initial goal, but also the volume of relevance accumulated so far. Adaptive model-based metrics include the bejeweled player model, BPM [42]; the information foraging model, IFT-model [3], and INST [29, 30]. However, we lack evidence connecting user models, especially adaptive ones, with user satisfaction ratings. From the perspective of time-biased gain, Smucker and Jethani [34] contend that Cranfield-style evaluation metrics that embody an inaccurate user model tend to have a weak relationship with human performance. Here we address a similar issue, but via the argument that a good user model is one that reflects three hypothetical probabilities: those associated with viewing, continuing, and stopping.

Contributions. We investigate the relationship between model accuracy and user satisfaction ratings, exploring effectiveness metrics for both query- and session-level evaluation. We also evaluate the accuracy of metric-based user models from the perspective of the C/W/L framework [29, 30]. Finally, we ask whether the accuracy of a metric is connected with its correlation with user satisfaction.

We also examine session-level metrics [21]. Past proposals describe ways of combining the query-level scores into a single session score, such as using a decreasing weight function [16, 23] or simple aggregation function [17, 18]. We explore those and other methods for combining query-level scores [16, 17, 23], seeking session-level scores that also correlate with overall user satisfaction.

We use four pre-existing resources to consider these questions, and in doing so both develop an important new framework for metric meta-evaluation, and at the same time demonstrate that the metrics and parameter settings that correlate well with user satisfaction closely match the user models and parameter settings that best fit observed user behaviors.

2 BACKGROUND
Meta-Evaluation with User Ratings. User satisfaction can be thought of as being the successful fulfillment of an information need [24], and is a key goal of any information retrieval activity [7]. Thus, the effectiveness of an IR system is closely coupled with the concept of user satisfaction [36]. Su [37] argues that quantifying user satisfaction has several advantages, including that it takes into account users and their subjective assessments.
Others have argued that user satisfaction alone does not provide reliable measurement [12, 26, 35]. Al-Maskari and Sanderson [1] suggest that measurement of user satisfaction is affected by two issues: that any operational definition of “satisfaction” is inherently ambiguous; and that it is difficult to devise instruments for measuring it. Similarly, Kelly [24] observes that satisfaction is not observable, and lies in the user’s head; and Soergel [35] notes the “user-distraction” phenomenon, with users perhaps still satisfied by non-relevant items. Soergel also argues that the goal is not subjective user satisfaction (making the user happy), but the objective goal of improved task performance (making the user successful).

A range of recent studies have demonstrated that the effectiveness of an IR system can have a substantial correlation with user satisfaction [2, 14, 19]; and many researchers continue to rely on this relationship [6, 18, 25, 42]. There are two ways of evaluating how well a metric predicts the satisfaction or performance of users:

- Direct computation of correlation coefficients between the scores generated by metrics and user satisfaction ratings; and
- Evaluation of whether or not the model embedded in the effectiveness metric reflects any patterns of observable user behavior that have a relationship with user satisfaction or performance.

Multiple instances of the first of these approaches have been provided through the last decade [2, 6, 14, 17, 18, 25, 27, 42]. In the absence of user ratings, the second approach provides an alternative [3, 5]. For example, user satisfaction can be estimated using online click metrics [6, 31]; and sequences of user actions can be used to predict success [11]. There is also a relationship between implicit measures (such as click-through patterns, or the total time spent on a SERP) and users’ explicit satisfaction ratings [9]. Chapelle et al. [5] use click-throughs to evaluate expected reciprocal rank, ERR, arguing that when the score of a metric is correlated with that of a click-through measure, it indicates that the metric captures satisfaction; and Azzopardi et al. [3] adopt ideas from economics and argue that a good measure should reflect user behavior. Based on the observation that user behavior can be employed as a surrogate for search success [11], Azzopardi et al. propose three meta-evaluation approaches, one of which is to compare model-derived predicted stopping probabilities with empirical distributions of stopping ranks estimated via click-through logs.

User Models and the C/W/L Framework. A test collection-based evaluation can be viewed as the simulation of an artificial user who inspects the systems’ SERPs and rates their performance under constant operational settings [32]. Hence, an effectiveness metric should explain how the artificial user interacts with each SERP, forming the basis of the user model associated with a metric [28]. For example, the model that corresponds to Moffat and Zobel’s rank-biased precision metric (RBP) with parameter \(0 \leq \phi \leq 1\) is that each user inspects the documents in the ranking one-by-one from top down, continuing from position \(i\) to \(i+1\) with probability \(\phi\).

A range of other user models also operate on the assumption that SERP inspection is sequential [3, 5, 28, 30, 33, 42], but differ in their details, including INSQ and INST, based on a user “relevance goal” parameter \(T\) [29, 30]; and information foraging-based metrics that add a rationality parameter \(R\) [3]).

To categorize user models Moffat et al. [29, 30] describe the “C/W/L” framework, in which user models are characterized via any of three interrelated behaviors:

- **Continuation** probability, \(C(i)\): the conditional probability that a user inspects the document at rank \(i+1\), given that they have just examined the one at rank \(i\) (also referred to as the persistence).
- **Weight** function, \(W(i)\): the fraction of user attention associated with position \(i\) in the ranking.
- **Last** probability, \(L(i)\): the probability that a user stops interacting with the SERP after examining the document at rank \(i\).

Note that the formulation of \(C(i)\) assumes that the user inspects the ranked-list of documents sequentially from the top, a general tendency that has been validated in a range of experiments [8, 20, 38, 40]. Effectiveness metrics directly developed in the C/W/L framework include RBP [28], INSQ [29], INST [30], and information foraging-based metrics [3]; other metrics including Prec@K, reciprocal rank (RR), and average precision (AP) [29] also fit the framework.

Note also that \(C(i), L(i), W(i)\) can be computed from each other [29, 30], and hence provide three different ways to link observable behaviors to metrics. For example, after comparing continuation functions of several metrics against the empirical \(C(i)\) observed from commercial job search interaction logs, Wicaksono and Moffat [40] show that with a suitable parameter choice, the \(C(i)\) function associated with INSQ is a better fit than the \(C(i)\) functions corresponding to several other weighted-precision metrics.

**Expected Rate of Gain Versus Expected Total Gain.** Another categorization of metrics arises from the scale that is used. Expected rate of gain (ERG) metrics measure the average gain (or utility) derived per document inspected:

\[
M_{ERG}(\hat{T}) = \sum_{i=1}^{\infty} W(i) \cdot r_i,
\]

where \(W(i)\) is the weight associated with the result at rank \(i\), and \(\hat{T}\) is the relevance vector for the SERP, with \(0 \leq r_i \leq 1\). If relevance is measured in units of (say) “rels”, then ERG metrics give scores in units of “rels per document”. The second family measure expected total gain (ETG), and scores are computed as:

\[
M_{ETG}(\hat{T}) = \sum_{i=1}^{\infty} \left( L(i) \cdot \sum_{j=1}^{i} r_j \right) = \frac{M_{ERG}(\hat{T})}{W(1)},
\]

to give units of “rels”. The experiments in Section 3 compare these two classes of metric; Carterette [4] provides further discussion.

**Static Versus Adaptive.** Another categorization arises in connection with the factors that influence the continuation probability \(C(i)\).

*Static* metrics (also called positional metrics) compute \(C(i)\) solely as a function of the rank position \(i\), paying no heed to the corresponding gain value \(r_i\), nor to the previous gains \(r_j\) for \(1 \leq j < i\).

*Adaptive* metrics allow the gain values \(r_j\) for \(0 \leq j \leq i\) to also influence \(C(i)\). One way in which this can be done is to include the volume of unmet desire at rank \(i\), computed as \(T_i = T - \sum_{j=1}^{i} r_j\), where \(T\) is the user’s initial relevance goal [29, 30]. Examples include INST and information foraging-based metrics [3], with the continuation function of the former defined as \(C_{INST}(i) = ((i + T + T_i - 1)/(i + T + T_i))^2\) [30].
Scoring Sessions. A search session consists of a sequence of queries with a single information goal, each query resulting in a SERP and a corresponding relevance vector, with the user regarding all of the SERPs as contributing to the overall result. Several metrics have been proposed for measuring multi-query sequences. For example, session-based DCG [16], sDCG, increasingly discounts the DCG scores [15] of the reformulated queries, arguing that their SERPs are less valuable than the one returned by the first query because of the additional effort involved:

\[
sDCG(s) = \sum_{j=1}^{|s|} \frac{1}{1 + \log_{10} j} \cdot DCG(q_j),
\]

where \(1 < bq < 1000\) is a parameter representing the extent to which the modeled user reformulates queries, and where \(q_j\) is the \(j\)th query in session \(s\). Small values of \(bq\) reflect users reluctant to reformulate queries; and larger values reflect patient users who regularly reformulate queries; Jarvelin et al. suggest \(bq = 4\). Kanoulas et al. [23] proposed a second version of sDCG, with a modified discount function, \(1/(\log_{10} j + bq - 1)\). Kanoulas et al. also describe the normalized sDCG used in the TREC 2010 session track [22].

In general, scoring a session can be thought of as aggregating the scores of the individual queries comprising the session. Jiang and Allan [17] measure the performance of several simple aggregation functions, combining the NDCG scores of the queries in the session. They demonstrate that simple aggregation functions, such as “mean”, “min”, and “last”, produce session scores that have moderate correlation with session-level satisfaction ratings, indicating that the worst query, and the last one in the session, are predictors of satisfaction. Liu et al. [25] have also concluded that the later queries in a session have a strong effect on session-level satisfaction.

3 METRIC SCORES AND SATISFACTION

It is axiomatic that a good metric is one that has a high correlation with user satisfaction ratings. With that in mind, we now explore the relationship between metric scores and user satisfaction ratings, working with both individual queries and also whole sessions.

Previous Work and Datasets. Many authors have computed correlation coefficients between metric scores and satisfaction ratings [6, 17, 27, 39]. Jiang and Allan [17] compute correlation coefficients between session-level satisfaction ratings and various DCG-based session metrics, including the one proposed by Jarvelin et al. [16]. They found that one of the best session-level metrics is “sDCG divided by the number of queries in the session”; overall they conclude that the number of queries, and the worst query and last queries in the session, are important contributors to session-level satisfaction.

Mao et al. [27] examine the difference between relevance and usefulness. They compute correlation coefficients between click sequence-based metrics and both query and session-level satisfaction ratings, concluding that metric scores, especially click-sequence based ones, better reflect user satisfaction when the gain function is defined based on the document usefulness than when it is defined as a function of the document relevance.

Chen et al. [6] compare four offline metrics (cumulative gain, CG [15], DCG, RBP, and ERR) and various online metrics (based on clicks, mouse-scrolls, and dwell-time), computing correlation coefficients with query-level satisfaction ratings. In their experiments RBP with \(\phi = 0.8\) provided the highest correlation coefficient (Pearson’s \(r = 0.445\)); and some online metrics (query dwell time and maximum scroll distance) were useful surrogates for user satisfaction.

Thomas et al. [39] use data from Bing.com to compute correlations between effectiveness metrics and query non-reformulation rates, finding that shallow metrics such as RBP with \(\phi = 0.1\), Prec@1, and RR provide reasonable predictions of success (non-refor- mulation) rate for a set of (mostly navigational) web queries.

Table 1 summarizes the datasets generated by these three groups. In the experiments described below we compute similar correlations between metric scores and satisfaction ratings, but via the lens of the C/W/L framework. That is, we regard a metric as not only generating scores, but also as describing user behavior. The latter can then be compared with observed behavior, to yield a second and equally-important set of correlation scores.

Notation and Metrics. We work with sets of sessions:

\[\{(s_1, u_1, p_1), (s_2, u_2, p_2), \ldots\}\]

where each triple records a search session \(s_i\), the user \(u_i\) who completed that session, and a corresponding satisfaction rating \(p_i\) generated by the user. Each session \(s_i\) consists of a sequence of queries:

\[q_i = \langle q_{i,1}, q_{i,2}, q_{i,3}, \ldots q_{i,n_i}\rangle\]

where \(q_{i,j}\) might be associated with information such as relevance vector \(r_{i,j}\), a click-through sequence \(c_{i,j}\), or a gaze sequence \(g_{i,j}\). We take \(0 \leq r_{i,j,k} \leq 1\) to be the gain associated with the \(k\)th document in the ranking of the \(j\)th query in the \(i\)th session, with gains computed from relevance grades using \(gain(r) = (2^r - 1)/(2^{max} - 1)\) for all metrics except AP, ERR, and RR. In the case of ERR, the denominator is replaced by \(2^{max}\). For AP and RR, \(gain(r) = 1\) if the corresponding document is fully relevant, otherwise \(gain(r) = 0\).

We did not compute an expected rate of gain (ERG) version for DCG and BFM (bejeweled player model), since they are defined as expected total gain (ETG) metrics. Similarly, no ETG version of AP was considered. Both BFM and IFT-model involve multiple parameters; unless explicitly noted as variations, the default values recommended by their developers were employed [3, 42]; TBG is computed using the implementation of Jiang and Allan [18].

Residuals. Moffat and Zobel [28] introduce the notion of residuals, the uncertainty in a metric score as a result of unjudged items. With a weighted-precision metric, the residual is just the difference between a lower bound and an upper bound score. The former is computed by assuming that all unjudged documents are non-relevant, and the latter by assuming that they are all fully relevant.
were constructed only for the top-5 documents in the ranking. D-BPM is addressed in the initial explorations [6, 17, 27]. This experiment implemented that created the datasets, including several metrics not coefficients (Pearson’s $\rho$ = 0.8, the corresponding residuals raised by unjudged documents are 0.13, 0.11, 0.33, and 0.07 for the ERG versions. The residuals associated with $\rho = 0.1$ are much lower. Figure 1 shows the residuals for two adaptive metrics: INST and IFT with $T = 2, 3$, computed using the 291 queries in the THUIR1 dataset. The overall moderate residuals indicate that the SERP scores can be used with a degree of confidence.

**Query-Level Correlations.** First, we compute correlation coefficients (Pearson’s $r$) between the scores of a range of metrics and the query-level satisfaction ratings generated as part of the experimentation that created the datasets, including several metrics not addressed in the initial explorations [6, 17, 27]. This experiment uses the THUIR1 and THUIR2 datasets (see Table 1); for THUIR2, all rankings were truncated at depth five, since relevance judgments were constructed only for the top-5 documents in the ranking.

Table 2 shows the results, with S-BPM a static version of BPM, and D-BPM an adaptive one [42]. Similarly, Azzopardi et al. [3] describe three types of IFT-model: IFT-C1, which is solely goal sensitive; IFT-C2, which is only governed by the rate of gain so far; and IFT, which a combination of both versions. For static metrics, the ERG formulations yield the same correlations as the corresponding ETG ones, since $W(1)$ is constant across queries and SERPs.

The judgment pooling depths for the J&A, THUIR1, THUIR2, and MS datasets were 9, 10, 5, and 12, respectively. With RBP $\phi = 0.8$, the corresponding residuals raised by unjudged documents are 0.13, 0.11, 0.33, and 0.07 for the ERG versions. The residuals associated with $\phi = 0.1$ are much lower. Figure 1 shows the residuals for two adaptive metrics: INST and IFT with $T = 2, 3$, computed using the 291 queries in the THUIR1 dataset. The overall moderate residuals indicate that the SERP scores can be used with a degree of confidence.

**Query-Level Correlations.** First, we compute correlation coefficients (Pearson’s $r$) between the scores of a range of metrics and the query-level satisfaction ratings generated as part of the experimentation that created the datasets, including several metrics not addressed in the initial explorations [6, 17, 27]. This experiment uses the THUIR1 and THUIR2 datasets (see Table 1); for THUIR2, all rankings were truncated at depth five, since relevance judgments were constructed only for the top-5 documents in the ranking.

Table 2 shows the results, with S-BPM a static version of BPM, and D-BPM an adaptive one [42]. Similarly, Azzopardi et al. [3] describe three types of IFT-model: IFT-C1, which is solely goal sensitive; IFT-C2, which is only governed by the rate of gain so far; and IFT, which a combination of both versions. For static metrics, the ERG formulations yield the same correlations as the corresponding ETG ones, since $W(1)$ is constant across queries and SERPs.

The case of adaptive metrics, the difference between ERG and ETG depends on the search depth of the corresponding user

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>Laboratory experiment</td>
<td>Laboratory experiment</td>
<td>Laboratory experiment</td>
<td>Commercial search engine</td>
</tr>
<tr>
<td># sessions</td>
<td>80 sessions, in which a session can contains multiple queries</td>
<td>2,435 single-SERP sessions, each a combination between SERP ID and User ID</td>
<td>225 sessions, in which a session can contains multiple queries</td>
<td>--</td>
</tr>
<tr>
<td># queries and SERPs</td>
<td>388</td>
<td>291</td>
<td>933</td>
<td>994, mostly popular queries (or navigational)</td>
</tr>
<tr>
<td># results per SERPs</td>
<td>9</td>
<td>10</td>
<td>10, however, only top-5 results plus those clicked were judged</td>
<td>10 – 12</td>
</tr>
<tr>
<td>Relevance judgments</td>
<td>3-level graded judgments</td>
<td>4-level graded judgments</td>
<td>4-level graded judgments</td>
<td>4-level graded judgments</td>
</tr>
<tr>
<td>Click-throughs</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Gaze sequences</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Mouse movements</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Query-level ratings</td>
<td>No</td>
<td>Yes, 5-level ratings</td>
<td>Yes, 5-level ratings</td>
<td>Query-level satisfaction estimated by the query non-ref ormulation rate</td>
</tr>
<tr>
<td>Session-level ratings</td>
<td>Yes, 5-level ratings</td>
<td>No</td>
<td>Yes, 5-level ratings</td>
<td>No</td>
</tr>
</tbody>
</table>

**Table 1**: Datasets used in experiments, and their origins.

**Figure 2**: Correlation coefficients (y-axis) between query-level satisfaction ratings and both ERG and ETG versions of IFT-C1 for $T = 1, 1.5, \ldots, 5$ (x-axis), and for THUIR1 (left) and THUIR2 (right).
We use the THUIR2 dataset to evaluate the use of observable query
representations. The values listed are Pearson correlation coefficients between
the signals and query-level satisfaction ratings.

<table>
<thead>
<tr>
<th>coeff.</th>
<th>0.25</th>
<th>0.38</th>
<th>0.39</th>
<th>0.26</th>
<th>0.40</th>
<th>0.36</th>
</tr>
</thead>
<tbody>
<tr>
<td>nreform numclick</td>
<td>PLC</td>
<td>minRC</td>
<td>maxRC</td>
<td>meanRC</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: User action-based signals and query-level satisfaction ratings. The values listed are Pearson correlation coefficients between the signals and query-level satisfaction ratings.

Comparing Correlation Coefficients. Hotelling’s $t$ test [13] can be used to compare correlation coefficients when there are overlapping variables. In the ERG column in Table 2 the highest correlation for the THUIR dataset is INSO with $T = 1$, which is significantly better than Prec@1 ($p = 0.025$) and RR ($p = 0.040$), but not significantly better to AP; and the highest correlation for the THUIR2 dataset is D-RPM, which significantly outperforms all of Prec@1, RR, and AP ($p < 0.01$ in all three cases).

Surrogates for Satisfaction. The fact that negative coefficients dominate the MS columns in Table 2 suggests that query non-reformulation rates are a poor surrogate for query-level satisfaction. We use the THUIR dataset to evaluate the use of observable query-level user actions in terms of approximating satisfaction ratings. The query non-reformulation rate can be represented as an indicator metric “nreform” which is 1 if the corresponding query is the last one in the session and 0 if not. Other signals tested were the number of distinct items clicked (numclick); the precision at lowest click (PLC), computed as the ratio between the number of clicks and the rank position of the lowest click; and the maximum, minimum, and mean reciprocal ranks of clicked items (minRC, maxRC, and meanRC, all zero if no clicks observed), as originally proposed by Radlinski et al. [31] and then employed by Chapelle et al. [5] to meta-evaluate ERR. Table 3 shows that the two click-based metrics PLC and maxRC are better than the binary indicator nreform, with both differences significant (Hotelling’s $t$ test, $p < 0.01$).

Consistency Across Users. It is also possible to treat each user differently and compute distributions of correlation coefficients. In the THUIR1 dataset, a group of 25 users (user IDs 71–80, 82–95, and 97) evaluated the same set of 21 SERPs. The graphs in Figure 3 represent distributions of correlation coefficients (computed using kernel density estimation) between metric scores (ERG) and satisfaction ratings across those 25 users, in each case computing the correlation coefficient for a user from their 21 data pairs. The distribution for IFT-C1 with $T = 2$ is more skewed toward the high end than the other two, indicating that the users more agree with IFT-C1 with $T = 2$, which is an adaptive metric, than with the traditional ones, such as RR and Prec@1. Besides IFT-C1 with $T = 2$, the other metrics whose distributions are more skewed to the right are RBP with 0.8, INSO with $T = 2$, and INST with $T = 2$.

Metrics For Sessions. Session-based metric scores aggregate individual query scores. One option is via a linear combination:

$$sM(s) = \sum_{j=1}^{n} a_j \cdot M(q_j),$$

where $a_j$ is a weight associated with the $j$th query in session $s$ containing $|s| = n$ queries. Many approaches can be used to define $a_j$, including simple summations, taking $a_j = 1$; and averaging across queries, taking $a_j = 1/n$. More complex approaches can also be employed, such as the discount function proposed by Järvelin et al. [16] (Equation 1); and a RBP-like session aggregation approach, denoted $\text{geom}$, defined as $a_j = (1-\mu)^j \mu^{-1}$, where $\mu$ is a hypothetical user behavior parameter representing “how likely that user is to reformulate their queries”. A revg (reverse geometric) weighting scheme is also an option, assigning the greatest weight to the final query. The first, last, maximum, and minimum of the set of query scores might also be taken to be the session score.

Table 4 explores nine of these many possibilities, listing correlation coefficients between session-level satisfaction ratings and aggregated metric scores using the 80 sessions in the J&A dataset. The differences between pairs of coefficients in Table 4 are not significant under Hotelling’s $t$ test because only 80 data points are involved. However, there are some general patterns to be observed across metrics and across the aggregation techniques. To draw them out, each row and column in Table 4 has an adjusted geometric mean ($\text{gmean}$) associated with it to provide an overall perspective on that row or column, computed as $\text{gmean}(x) = (\prod_{i=1}^{n}(x_i + 0.2))^{1/n} - 0.2$, with the 0.2 adjustment to accommodate negative values $x_i$.

Looking down the final column in Table 4 indicates that AP, RBP with $\beta = 0.8$, and INSO with $T = 2$ provide the better correlations when combined across the suite of aggregation columns, and perform notably better than metrics such as Precision with $k = 1$, ERR, and TGB. The final row in Table 4 provides a number of further observations: that the average across the query scores in a session appears to be better than taking their sum; that “min” appears to be better than “max”; that the last query in each session seems to be more influential than the first one; and that “revg” outperforms “geom”. These agree with the patterns observed by Liu et al. [25] in their recent study. Note also that the “sum” and Järvelin et al. [16] approaches perform poorly, perhaps because they give rise to unbounded values in which the number of queries carried out is itself the strongest determinant of overall session score.

Fitting Query Weights. We next used the query- and session-level ratings of the THUIR2 dataset to compute $a_j$ values. Best-fit weights were identified to linearly combine each sessions’ query-level satisfaction ratings into a single score, with one vector for each different session length $n$. The query-level ratings serve as ground-truth scores that represents users’ experience with the session’s
Combining query-level satisfaction ratings into a single session-level rating for the THUIR2 dataset. The values listed are using the THUIR2 dataset, stratified by each of which considered the same set of SERPs. Two constraints were enforced, that $0 \leq a_j \leq 1$, and that $\sum_{j=1}^{n} a_j = 1$ for each session length $n$. Figure 4 shows the computed best-fit weights $a_j$, stratified by $n$. The first and last query are clearly important when computing session-level scores. A range of aggregation functions using the query-level ratings of the THUIR2 dataset were also measured, shown in Table 5. Functions “mean”, “max”, “min”, “first”, and “last” again provide the best correlations. Based on these findings two further aggregation functions are proposed. The first sets the $a_j$ via an asymmetric U-shape function that favors early queries, and favors late queries somewhat more:

$$a_j = \frac{f(j)}{\sum_{k=1}^{n} f(k)},$$

where $f(x) = (x - n/2)^2 + 1$.

The strength of “max” and “min” aggregations indicates that the best and worst queries in a session have an important effect on overall satisfaction, so a linear combination spanning “first”, “last”, “max”, and “min” is also considered:

$$sM(s) = \theta_0 M(q_1) + \theta_1 M(q_n) + \theta_2 \max\{M(q_j) | 1 \leq j \leq n\} + \theta_3 \min\{M(q_j) | 1 \leq j \leq n\},$$

where the coefficients satisfy $0 \leq \theta_0, \theta_1, \theta_2, \theta_3 \leq 1$, and $\sum_{j=0}^{n} \theta_j = 1$.

The THUIR2 dataset was used to determine best-fit values of $(\theta_0, \theta_1, \theta_2, \theta_3) = (0.140, 0.267, 0.523, 0.070)$. The J&A dataset was then used to evaluate the two proposed aggregation functions. Table 6 shows the results of the same experiment as was carried out for Table 4, using a subset of metrics, a subset of the previous nine aggregation approaches, and the two further approaches. The U-shape.

Table 4: Correlation coefficients (Pearson’s $r$) between session-level satisfaction ratings and metric scores (all reporting expected rate of gain except for DCG) for the 80 sessions in the J&A dataset. The columns compare a range of different ways of aggregating query scores in a session to obtain a session score. The “geom” and “ravg” columns use $\mu = 0.5$.

<table>
<thead>
<tr>
<th>sum</th>
<th>mean</th>
<th>max</th>
<th>min</th>
<th>first</th>
<th>last</th>
<th>jarv</th>
<th>geom</th>
<th>revg</th>
<th>gmean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prec, $k = 1$</td>
<td>-0.01</td>
<td>0.26</td>
<td>0.08</td>
<td>0.27</td>
<td>0.11</td>
<td>0.22</td>
<td>0.03</td>
<td>0.16</td>
<td>0.12</td>
</tr>
<tr>
<td>Prec, $k = 5$</td>
<td>0.02</td>
<td>0.43</td>
<td>0.31</td>
<td>0.39</td>
<td>0.31</td>
<td>0.44</td>
<td>0.08</td>
<td>0.27</td>
<td>0.34</td>
</tr>
<tr>
<td>ERR</td>
<td>-0.05</td>
<td>0.39</td>
<td>0.19</td>
<td>0.36</td>
<td>0.20</td>
<td>0.34</td>
<td>-0.01</td>
<td>0.18</td>
<td>0.15</td>
</tr>
<tr>
<td>DCG</td>
<td>-0.02</td>
<td>0.40</td>
<td>0.30</td>
<td>0.39</td>
<td>0.29</td>
<td>0.41</td>
<td>0.04</td>
<td>0.23</td>
<td>0.27</td>
</tr>
<tr>
<td>NDCG</td>
<td>-0.02</td>
<td>0.35</td>
<td>0.27</td>
<td>0.35</td>
<td>0.26</td>
<td>0.37</td>
<td>0.04</td>
<td>0.22</td>
<td>0.25</td>
</tr>
<tr>
<td>RBP, $\phi = 0.8$</td>
<td>-0.01</td>
<td>0.41</td>
<td>0.31</td>
<td>0.39</td>
<td>0.30</td>
<td>0.43</td>
<td>0.05</td>
<td>0.24</td>
<td>0.29</td>
</tr>
<tr>
<td>INSQ, $T = 3$</td>
<td>-0.01</td>
<td>0.41</td>
<td>0.31</td>
<td>0.39</td>
<td>0.29</td>
<td>0.42</td>
<td>0.05</td>
<td>0.24</td>
<td>0.29</td>
</tr>
<tr>
<td>TFG</td>
<td>-0.15</td>
<td>0.36</td>
<td>0.29</td>
<td>0.37</td>
<td>0.27</td>
<td>0.35</td>
<td>-0.12</td>
<td>0.08</td>
<td>0.09</td>
</tr>
<tr>
<td>S-BPM</td>
<td>-0.07</td>
<td>0.32</td>
<td>0.22</td>
<td>0.35</td>
<td>0.27</td>
<td>0.32</td>
<td>-0.03</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>AP</td>
<td>0.04</td>
<td>0.49</td>
<td>0.27</td>
<td>0.39</td>
<td>0.30</td>
<td>0.43</td>
<td>0.10</td>
<td>0.31</td>
<td>0.30</td>
</tr>
<tr>
<td>RR</td>
<td>0.02</td>
<td>0.39</td>
<td>0.17</td>
<td>0.35</td>
<td>0.24</td>
<td>0.33</td>
<td>0.08</td>
<td>0.25</td>
<td>0.21</td>
</tr>
<tr>
<td>D-BPM</td>
<td>-0.09</td>
<td>0.35</td>
<td>0.27</td>
<td>0.35</td>
<td>0.24</td>
<td>0.32</td>
<td>-0.05</td>
<td>0.12</td>
<td>0.11</td>
</tr>
<tr>
<td>IFT-C1, $T = 3$</td>
<td>0.04</td>
<td>0.33</td>
<td>0.27</td>
<td>0.33</td>
<td>0.26</td>
<td>0.33</td>
<td>0.09</td>
<td>0.20</td>
<td>0.25</td>
</tr>
<tr>
<td>IFT-C2</td>
<td>-0.10</td>
<td>0.40</td>
<td>0.28</td>
<td>0.38</td>
<td>0.23</td>
<td>0.36</td>
<td>-0.06</td>
<td>0.14</td>
<td>0.12</td>
</tr>
<tr>
<td>IFT, $T = 3$</td>
<td>0.01</td>
<td>0.33</td>
<td>0.24</td>
<td>0.34</td>
<td>0.25</td>
<td>0.34</td>
<td>0.06</td>
<td>0.20</td>
<td>0.25</td>
</tr>
<tr>
<td>INST, $T = 3$</td>
<td>0.01</td>
<td>0.40</td>
<td>0.31</td>
<td>0.37</td>
<td>0.29</td>
<td>0.41</td>
<td>0.07</td>
<td>0.25</td>
<td>0.30</td>
</tr>
</tbody>
</table>

Table 5: Combining query-level satisfaction ratings into a single session-level rating for the THUIR2 dataset. The values listed are Pearson correlation coefficients between the computed session-level ratings and the actual session-level satisfaction ratings.

<table>
<thead>
<tr>
<th>mean</th>
<th>max</th>
<th>min</th>
<th>first</th>
<th>last</th>
<th>jarv</th>
<th>revg</th>
<th>coeff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.71</td>
<td>0.64</td>
<td>0.48</td>
<td>0.55</td>
<td>0.58</td>
<td>0.18</td>
<td>0.45</td>
<td></td>
</tr>
</tbody>
</table>
function is no better than the simpler “last” and “mean” techniques, and the linear combination of positional and non-positional factors fares no better either.

4 USER MODELS AND USER BEHAVIOR

Section 3 explored correlation coefficients between metric scores and satisfaction ratings at query and session levels. This section explores the dual of that relationship – the relationship between user models (corresponding to metrics in the C/W/L framework), and observed user behavior (corresponding, perhaps, to satisfaction).

We model user behavior by calculating the extent to which the model-predicted behavior matches observed user behavior, computing the distances between \( W(\cdot) \), \( L(\cdot) \), and \( C(\cdot) \) for each given metric, and their corresponding observed distributions, denoted by \( \hat{W}(\cdot) \), \( \hat{L}(\cdot) \), and \( \hat{C}(\cdot) \). In the model, \( W(\cdot) \), \( L(\cdot) \), and \( C(\cdot) \) can be computed from each user under the assumption that the users inspect documents sequentially in each SERP. In reality that relationship cannot be assumed for \( W(\cdot) \), \( \hat{L}(\cdot) \), and \( \hat{C}(\cdot) \), and we estimate each of them independently from interaction logs.

Suppose a dataset covers a set of users \( U = \{u_1, u_2, \ldots, u_{|U|}\} \), and each user \( u_i \) is associated with a set of view vectors \( \mathcal{V}(u_i) = \{\overrightarrow{v}_1, \ldots, \overrightarrow{v}_{|\mathcal{V}(u_i)|}\} \), with each view vector a sequence \( (v_1, \ldots, v_N) \), where \( N \) is the number of elements in the SERP, and \( v_1 = 1 \) if the user inspected the result at rank position \( i \) in that SERP, and \( v_1 = 0 \) otherwise. Ideally, the view vectors are constructed using gaze sequences obtained from an eye-tracking experiment.

**Evaluating W.** The data likelihood is maximized when the observed \( \hat{W}(i) \) is estimated as:

\[
\hat{W}(i) = \frac{\sum_{u \in U} \sum_{\overrightarrow{v} \in \mathcal{V}(u)} v_i}{\sum_{u \in U} \sum_{\overrightarrow{v} \in \mathcal{V}(u)} \sum_{j=1}^{N} v_j}.
\]

To measure the closeness between a model-generated \( W(\cdot) \) for a metric and the observed \( \hat{W}(\cdot) \) estimated from a dataset, we employ a mean squared error (MSE) function. The closer the MSE value is to zero, the stronger the empirical evidence for that \( W(\cdot) \) formulation.

**Evaluating L.** Assuming that the users sequentially scan down the ranking, the position of the last item inspected is also the deepest rank position examined. Suppose \( I(i, \overrightarrow{v}) \) is an indicator function that returns 1 if the rank position \( i \) is the deepest one observed in the view vector \( \overrightarrow{v} \), and 0 if otherwise. A maximum likelihood estimator for \( \hat{L}(i) \) is then:

\[
\hat{L}(i) = \frac{\sum_{u \in U} \sum_{\overrightarrow{v} \in \mathcal{V}(u)} I(i, \overrightarrow{v})}{\sum_{u \in U} |\mathcal{V}(u)|}.
\]

We again employ MSE to compute the distance between the model \( L(\cdot) \) and the observed distribution \( \hat{L}(\cdot) \).

**Evaluating C.** In contrast to \( W(\cdot) \) and \( L(\cdot) \), which are both probability distributions (that is, \( \sum_{i=1}^{\infty} W(i) = \sum_{i=1}^{\infty} L(i) = 1 \)), the continuation function \( C(\cdot) \) is a set of independent values between zero and one. To compute \( \hat{C}(\cdot) \) we consider view sequences across all users and queries. For example, in the view sequence \( (2, 1, 3, 2, 3, 5, 3, 6, 7) \) the user first inspects the document at rank 2, then the one at rank 1, and so on, until the one at rank 7. A continuation is deemed to occur at rank \( i \) if an inspection at rank \( i \) is followed by another at a higher-numbered rank \( |U| \), and an empirical \( \hat{C}(\cdot) \) function is computed from the continuation indicators. The distance between \( C(\cdot) \) and \( \hat{C}(\cdot) \) is then measured via a weighted mean squared error function, weighted at rank position \( i \) by the relative frequency with which documents at rank \( i \) were viewed, and with the weighting required because \( C(\cdot) \) is itself not a probability distribution.

**Inferring Gaze Sequences.** The J&A dataset (Table 1) contains gaze sequences generated by eye-tracking, but the other three do not. In the absence of gaze sequences, click-through data can be employed. Using click-through data as a direct surrogate for gaze sequences is not effective, but inferred gaze distributions are [41, 43]. Suppose that \( P(\text{view} = i \mid u, q) \) is the probability that user \( u \) views the item listed at rank \( i \) for query \( q \). Using click-through data, Wicaksono et al. [41] estimate \( P(\text{view} = i \mid u, q) \) as follows:

\[
P(\text{view} = i \mid u, q) = \begin{cases} 
1 & \text{if } i \leq DC(u, q) \\
\exp(g(K(u, q))) & \text{otherwise},
\end{cases}
\]

where \( K(u, q) = w_0 + w_1 \cdot DC(u, q) + w_2 \cdot NC(u, q) \) is the deepest rank position clicked; \( NC(u, q) \) is the number of distinct items clicked; \( g(x) = \ln(1 + e^x) \) is a softplus function; and \( (w_0, w_1, w_2) \) are parameters estimated from the data. This model assumes that the user scans down the ranking one-by-one from top to bottom, and that if they click the one at rank \( i \), they have viewed ranks \( 1 \) to \( i \) previously.

Suppose that each user \( u_i \) can also be thought of as having an association with a set of queries \( Q(u_i) = \{q_1, q_2, \ldots, q_{|Q(u_i)|}\} \). Wicaksono et al. [41] then estimate \( \hat{C}(\cdot) \) using \( \hat{P}(\text{view} = i \mid u, q) \). Similarly, by considering \( \hat{P}(\text{view} = i \mid u, q) \) as an expected count, we can estimate \( \hat{W}(i) \) as follows:

\[
\hat{W}(i) = \frac{\sum_{u \in U} \sum_{q \in Q(u)} \hat{P}(\text{view} = i \mid u, q)}{\sum_{u \in U} \sum_{q \in Q(u)} \sum_{j=1}^{N} \hat{P}(\text{view} = j \mid u, q)}.
\]

Finally, to estimate \( \hat{L}(i) \), we make use of the assumption that the users sequentially inspect the ranking:

\[
\hat{L}(i) = P(\text{view} = i) = P(\text{view} = i + 1),
\]

where \( P(\text{view} = i) = \prod_{j=1}^{i+1} C(j) \) is the examination probability at rank \( i \). By defining

\[
\hat{P}(\text{last} = i \mid u, q) = \hat{P}(\text{view} = i \mid u, q) - \hat{P}(\text{view} = i + 1 \mid u, q),
\]

\( \hat{L}(i) \) can be estimated via:

\[
\hat{L}(i) = \frac{\sum_{u \in U} \sum_{q \in Q(u)} \hat{P}(\text{last} = i \mid u, q)}{\sum_{u \in U} \sum_{q \in Q(u)} \sum_{j=1}^{N} \hat{P}(\text{last} = j \mid u, q)}.
\]
These alternatives can be used if gaze information is not available but click information is (the THUIR1 and THUIR2 datasets).

**User Model and Satisfaction.** The accuracy of a range of C/W/L user models was measured using the J&A and THUIR datasets, with the THUIR dataset not suitable because of its shallower relevance judgments. The THUIR dataset does not include gaze sequences, and hence the alternative $\hat{p}(\text{view} = i | u, q)$-based formulations for $\hat{C}(\cdot), \hat{L}(\cdot),$ and $\hat{W}(\cdot)$ were employed, based on parameters $(w_0, w_1, w_2)$ fitted using gaze sequences and click-through information associated with the J&A dataset:

$$K(u, g) = 3.48 - 0.46 \cdot DC(u, g) + 0.20 \cdot NC(u, g).$$

Tables 7 and 8 show the results. Recall that in adaptive models $C(\cdot)$ depends on the relevance vector. Therefore, $C(i), W(i)$ and $L(i)$ for an adaptive user model are determined by averaging their values across all relevance vectors, $R,$ in the dataset. For example, $C_{\text{INST}}(3) = 1/|R| \sum_{R} C_{\text{INST}}(3, \vec{r})$, where $C_{\text{INST}}(i, \vec{r})$ is the INST continuation probability at rank $i$ with respect to the relevance vector $\vec{r}$. In general, RBP with $\phi = 0.8$, INSQ with $T = 3$, and the three adaptive metrics INST with $T = 3$, IFT-C1 with $T = 3$, and IFT with $T = 3$ are more accurate than Precision with $k = 1$ or $k = 5$, and RR, for all of $W(i), L(i),$ and $C(i).$ Note also that being accurate in one characteristic tends to correspond to accuracy in the other two characteristics. That is, the relationship between $W(i), L(i),$ and $C(i)$ is reflected in their observed versions $\hat{W}(i), \hat{L}(i),$ and $\hat{C}(i)$.

Table 7: User model accuracy using the J&A dataset. Small numbers are better; INSQ with $T = 1$ and with $T = 2$ were inferior to the $T = 3$ version listed; similarly all of INST, IFT-C1, and IFT with $T = 1$ and $T = 2$ were inferior to their corresponding $T = 3$ versions.

<table>
<thead>
<tr>
<th>$\text{Parameter}$</th>
<th>$\text{WMSE}(\hat{C}, C)$</th>
<th>$\text{MSE}(\hat{L}, L)$</th>
<th>$\text{MSE}(\hat{W}, W)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{(x10^{-3})}$</td>
<td>$\text{(x10^{-3})}$</td>
<td>$\text{(x10^{-3})}$</td>
<td></td>
</tr>
<tr>
<td>$\text{Prec, } k = 1$</td>
<td>674.2</td>
<td>1089.3</td>
<td>790.1</td>
</tr>
<tr>
<td>$\text{Prec, } k = 5$</td>
<td>243.1</td>
<td>1113.7</td>
<td>59.5</td>
</tr>
<tr>
<td>$\text{RBP, } \phi = 0.8$</td>
<td>3.6</td>
<td>118.7</td>
<td>8.9</td>
</tr>
<tr>
<td>$\text{INSQ, } T = 3$</td>
<td>7.4</td>
<td>133.2</td>
<td>17.5</td>
</tr>
<tr>
<td>$\text{RR}$</td>
<td>244.7</td>
<td>218.2</td>
<td>210.2</td>
</tr>
<tr>
<td>$\text{INST, } T = 3$</td>
<td>11.1</td>
<td>180.3</td>
<td>16.3</td>
</tr>
<tr>
<td>$\text{IFT-C1, } T = 3$</td>
<td>27.8</td>
<td>21.8</td>
<td>21.8</td>
</tr>
<tr>
<td>$\text{IFT, } T = 3$</td>
<td>40.8</td>
<td>112.5</td>
<td>19.3</td>
</tr>
</tbody>
</table>

Table 8: User model accuracy using the THUIR dataset. All other details are as for Table 7, including the relationships between the $T = 1$ and $T = 2$ versions of INSQ, INST, IFT-C1, and IFT, and their $T = 3$ versions. Here gaze sequences are inferred from click data.

$\text{WMSE}(\hat{C}, C)$ | $\text{MSE}(\hat{L}, L)$ | $\text{MSE}(\hat{W}, W)$ |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{(x10^{-3})}$</td>
<td>$\text{(x10^{-3})}$</td>
<td>$\text{(x10^{-3})}$</td>
</tr>
<tr>
<td>$\text{Prec, } k = 1$</td>
<td>725.9</td>
<td>957.0</td>
</tr>
<tr>
<td>$\text{Prec, } k = 5$</td>
<td>264.7</td>
<td>926.6</td>
</tr>
<tr>
<td>$\text{RBP, } \phi = 0.8$</td>
<td>7.8</td>
<td>34.8</td>
</tr>
<tr>
<td>$\text{INSQ, } T = 3$</td>
<td>13.7</td>
<td>59.3</td>
</tr>
<tr>
<td>$\text{RR}$</td>
<td>232.6</td>
<td>128.5</td>
</tr>
<tr>
<td>$\text{INST, } T = 3$</td>
<td>19.1</td>
<td>82.0</td>
</tr>
<tr>
<td>$\text{IFT-C1, } T = 3$</td>
<td>57.0</td>
<td>31.3</td>
</tr>
<tr>
<td>$\text{IFT, } T = 3$</td>
<td>62.2</td>
<td>25.8</td>
</tr>
</tbody>
</table>

Figure 5 shows the relationship between correlation with satisfaction ratings and user model accuracy for several parameter values. The plots are for Precision (parameter $k$), RBP (parameter $\phi$), and INST (parameter $T$), respectively, using the J&A dataset and “U-shape” aggregation.

Figure 5: Joint plots between correlation with session satisfaction ratings and user model accuracy for several parameter values. The plots are for Precision (parameter $k$), RBP (parameter $\phi$), and INST (parameter $T$), respectively, using the J&A dataset and “U-shape” aggregation.
observed user actions. This is an important new way of thinking about meta-evaluation of metrics, and one that we plan to continue exploring. Note also that the relationship works in both directions — metrics that have accurate user models (in terms of \( C(i), W(i), \) and \( L(i) \)) being good fits to observed behavior) can then be argued as being the ones that should be used as the most appropriate surrogates for user satisfaction.

As part of the consideration of user satisfaction, and the indicators associated with it, we have also considered ways in which individual query scores can be combined to derive session scores. Our results show that the previous Järvelin et al. aggregation approach may perform relatively poorly, and that the last query in each session is an important factor when computing session scores. We plan to extend those observations to consider the implications in terms of adapting user models to the important behavioral issue of query reformulation.

Acknowledgments. We gratefully acknowledge the generosity of the creators of the datasets we have used, for making them available; and Paul Thomas (Microsoft) in particular for his assistance. This work was supported by the Australian Research Council’s Linkage Projects scheme (project number LP150100252), by the Australian Research Council’s Discovery Project Scheme (project number DP190101113), and by Seek.com.

REFERENCES