Meta-evaluation of Online and Offline Web Search Evaluation Metrics

Ye Chen[†], Ke Zhou[‡], Yiqun Liu[†*], Min Zhang[†], Shaoping Ma[†]
[†]Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science & Technology, Tsinghua University, Beijing, China
[‡]University of Nottingham, Nottingham, U.K.
yiqunliu@tsinghua.edu.cn

ABSTRACT

As in most information retrieval (IR) studies, evaluation plays an essential part in Web search research. Both offline and online evaluation metrics are adopted in measuring the performance of search engines. Offline metrics are usually based on relevance judgments of query-document pairs from assessors while online metrics exploit the user behavior data, such as clicks, collected from search engines to compare search algorithms. Although both types of IR evaluation metrics have achieved success, to what extent can they predict user satisfaction still remains under-investigated. To shed light on this research question, we meta-evaluate a series of existing online and offline metrics to study how well they infer actual search user satisfaction in different search scenarios. We find that both types of evaluation metrics significantly correlate with user satisfaction while they reflect satisfaction from different perspectives for different search tasks. Offline metrics better align with user satisfaction in homogeneous search (i.e. ten blue links) whereas online metrics outperform when vertical results are federated. Finally, we also propose to incorporate mouse hover information into existing online evaluation metrics, and empirically show that they better align with search user satisfaction than click-based online metrics.

KEYWORDS

Search satisfaction, evaluation metrics, online evaluation

1 INTRODUCTION

Search engine evaluation is important in both academic and industrial IR research. The goal of IR researchers is to bulid search engine systems which can satisfy users' information needs. Both offline and online metrics have been adopted to measure how well the system serves real users. Offline metrics mainly originated from Cranfield approach [12] and are based on editorial judgments of the relevance of search results. Typical offline metrics include Average Precision (AP), Normalized Discounted Cumulative Gain (NDCG) and Rank-Biased Precision (RBP) [36]. These metrics are widely used to measure the quality of ranking algorithms [46] and are of great value in guiding search algorithm designing. However, although they may provide easily interpretable outcomes, offline search evaluation has encountered two major problems. The first one lies in that editorial judgments are often less credible when measuring actual user experience. Recent studies show that assessors' judgments may significantly differ from users' assessments [32]. The second problem is that the evaluation results based on offline metrics can be biased because they are usually generated with a small and incomplete dataset [13].

Rather than relying on offline metrics with relevance judgments, a popular contrasting approach is to use online metrics based on the simple fact that the interactions between users and search engines reflect the actual users' experiences in a natural usage environment. Such metrics are calculated based on practical users' behavior logs and can give us straightforward descriptions on how users interact with search engines. In addition, it is often cheap and fast to collect such data in modern search engines, making it particularly easy to scale up online evaluation. Typical online metrics include click-based metrics such as CTR (click through rate), UCTR [11] (binary value representing click) and PLC [9] (number of clicks divided by the position of the lowest click) as well as dwell time-based metrics such as query dwell time, average of click dwell time [22] and so on. Although easily scalable and arguably more truthful indication of operational IR effectiveness, online metrics can suffer from various biases present in typical search logs. Online behavior of users can be affected by many factors, with position bias being the most widely recognized effect, which requires de-biasing when inferring search success. In addition, online metrics may not be as reusable as offline metrics [44].

Both online and offline metrics have been widely used to measure search performance in the past years. Nonetheless, they are usually poorly correlated [11] because they measure IR systems from different perspectives. Which measures better reflect the ultimate actual user satisfaction remains an open research question. Therefore, in this work we investigate the relationship between offline/online metrics and actual user satisfaction, aiming to establish a thorough understanding of the effectiveness of those metrics in various search scenarios. We meta-evaluate a series of existing evaluation metrics based on two public datasets with more than two thousand search sessions. Query-document relevance assessments, users' interaction behaviors and their explicit satisfaction feedback are all included in the datasets, which makes us able to compute most of the widely-used online and offline metrics. With more

[*] Corresponding author

SIGIR '17, Shinjuku, Tokyo, Japan
© 2017 ACM. 978-1-4503-5022-8/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3077136.3080804

than thirty metrics, we calculate Pearson correlation and conduct concordance test [42] to study how well each metric infers actual search user satisfaction. We found that while online and offline metrics measure users' search experience from different perspectives, they generally both significantly correlate with actual user satisfaction.

To enable thorough meta-evaluation with different information needs, we categorize the search tasks according to two existing taxonomies, with respect to search goal types [5] and cognitive level [3]. We find that top-weighted offline metrics correlate extremely well with user satisfaction in *navigational* search while online metrics perform comparatively better in *informational* and *transactional* search tasks. Inspired by previous studies on federated search [10, 33], we also compare the performance of evaluation metrics in both homogeneous and heterogeneous search environment as users search behaviors as well as satisfaction perception may be affected by vertical results. We find that online metrics perform better than offline metrics in heterogeneous search environment. This is probably because offline metrics mainly rely on relevance assessments while the interaction-based online metrics may be more sensitive to the effect of vertical results. In addition, inspired by research on users' fine-grained interaction behaviors such as satisfied clicks [53] and mouse hovers [16], we also investigate the differences between online metrics calculated based on different interaction behavior signals (clicks, satisfied clicks, hovers). The results show that online metrics can better estimate user satisfaction when mouse hover information is incorporated.

Our contributions in this paper are three-folds:

- We present a thorough meta-evaluation of online/offline metrics from the perspective of their relationship with user satisfaction for various types of information needs. The results provide insights for both evaluation metrics study and user satisfaction understanding.
- We investigate the differences and applicabilities of different evaluation metrics in both homogeneous and heterogeneous search environment. We demonstrate that offline metrics work better in homogeneous search while online metrics outperform in heterogeneous search environment.
- We propose to incorporate mouse hover information into existing online evaluation metrics and empirically demonstrate that they correlate better with user satisfaction.

2 RELATED WORK

Of particular interest to our research is the extensive body of work on (i) meta evaluation of IR metrics, and (ii) search satisfaction.

2.1 Meta Evaluation of IR Metrics

As evaluation serves as an important part in IR-related research, the meta-evaluation of evaluation metrics has also been widely studied in recent years and different criteria have been adopted to compare different evaluation metrics [34].

One widely-used method is to use "discriminative power" to measure evaluation metrics. Early in 2000, Buckley and Voorhees proposed to use error rate, which is the likely error of concluding "System A is better than system B", to compare between different metrics [6]. They also adopted "fuzziness value" to examine "the power of a measure to discriminate among systems". This idea was further formalized to be "discriminative power" by Sakai in 2006 [41]. He pointed out that mildly top-weighted metrics, such as AP, NDCG and RBP(0.95) usually have higher discrimination ratios than those strongly-weighted metrics, such as Prec@5 and RBP(0.5) [43]. In 2010, Yilmaz and Robertson compared AP and NDCG based on their statistical ability to predict outcomes on held-out data [52]. Another criteria to compare metrics is to calculate the correlations between the system orderings generated by different metrics. Correlation coefficients such as Kendall's τ and Spearman's ρ are widely used [44]. However, a weakness is that such coefficients introduce the same penalty for the discords at different ranking positions, whereas the top positions are usually more important in IR systems. To shed light on these considerations, Yilmaz et al. [51] introduced an AP-based correlation coefficient called τ_{ap} to achieve a top-weighted emphasis. Webber et al. [50] also proposed Rank Biased Overlap (RBO) to operate over indefinite and non-conjoint rankings. Other evaluation criteria including evaluation based on judgment cost [7, 8, 36, 51], coverage [40], inversions, interpretation and volatility matrix were also proposed in the past few years [34].

While A/B tests [28] and interleaving [24] as well as their variants [31, 45] are also widely-used approaches for search system evaluation [38], we choose user satisfaction as the ground truth for evaluating different evaluation metrics because satisfaction reflects users' search experience directly. Since the goal of IR systems is to satisfy users' information needs, it is important to investigate to what extent can existing evaluation metrics measure practical users' search experience. In this paper, we make a deep analysis of how different evaluation metrics correlate with user satisfaction. Because satisfaction has been regarded as the gold standard in search performance evaluation, we try to find metrics which can surrogate satisfaction, and introduce metrics which can better estimate user experience.

2.2 Search Satisfaction

Search satisfaction has become one of the major concerns in search evaluation studies. The concept of satisfaction was first proposed by Su et al. [47] and was defined as "the fulfillment of a specified desire or goal" by Kelly [26]. To evaluate a search system, satisfaction can be considered as regarding not only to the whole search experience but also to some specific aspects [48], such as the precision or completeness of search results, response time and so on. Since satisfaction is important for both search engine evaluation and optimization, a number of research studies have tried to quantify user satisfaction in both desktop search [21, 49] and mobile search [27, 29], and in both homogeneous [32] and heterogeneous search environment [10]. In recent years, a number of works (e.g, [22, 23]) have started using the benefit-cost framework to analyze the satisfaction judgement process of users. In this framework, both the benefit factors (document relevance) and cost factors (the effort users spend on examining search engine result pages (SERPs) and landing pages) are used to estimate satisfaction.

Although satisfaction can be regarded as the gold standard in search performance evaluation, as mentioned, it is not easy to be collected. This makes it urgent to find a reliable and reusable metric to estimate user satisfaction. However, as indicated by recent studies, relevance-based evaluation metrics, such as MAP and nDCG, may not be perfectly correlated with users' search experience [2, 21]. Recently, Mao et al. [35] further studied the relationship between relevance, usefulness and satisfaction and also suggested that traditional system-centric evaluation metrics are not well aligned with user satisfaction.

Different from these existing works, we study the relationships between user satisfaction and both offline and online evaluation metrics. Our work is complementary to those research on satisfaction understanding and prediction, but also provides additional insights from the evaluation measure perspective. We investigate how the existing metrics perform in different search scenarios and compare the applicabilities of different metrics in homogeneous and heterogeneous search environment. Meanwhile, we also investigate how mouse hover information can be incorporated into online metrics to obtain better alignment with user satisfaction. The major contribution of our work lies in that we comprehensively meta-evaluate over thirty most popular IR metrics for many search tasks. The obtained insights can be used for guiding search evaluation practitioners.

3 DATASET AND METHODOLOGY

In this section, we describe the datasets as well as the meta-evaluation methods used throughout this paper. We also perform an analysis of user satisfaction distribution according to the characteristics of the datasets.

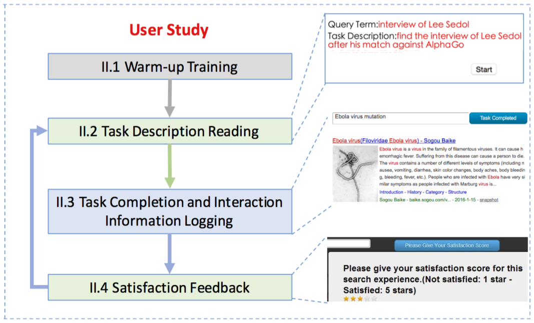
3.1 Overview

Our study aims to meta-evaluate different metrics based on two datasets which we have made publicly available[1]. These two datasets contain more than 2400 search sessions collected under 56 search search tasks in total. The detailed statistics are shown in Table 1.

Table 1: Characteristics of Datasets

	# queries	# different rankings per query	# users	# sessions
Dataset #1	26	3	40	1038
Dataset #2	30	6∼10	58	1397

These two datasets are generated under the same experimental process which is shown in Figure 1. Each participant completed a series of no more than 30 tasks in the datasets and they were required to perform two warm-up search tasks first to get familiar with the search process. Before each task, the participant was shown the search query and task explanations (see the card on the top right corner of Figure 1 as an example) first to avoid ambiguity. After that, the participant would be guided to search result page where the query is not allowed to change. Each participant was asked to examine the 10 fixed search results provided by the system and end the search session either if the search goal was completed or he/she was disappointed with the results. The provided search result lists were pre-crawled from commercial search engines and we made sure that the participant is able to complete the search goal as long as he/she examines all the provided search results. Each time the participant completed a search session, he/she was required to label a satisfaction score to reflect his/her search experience. Then they would be guided to continue to the next search task.


User Study
II.1 Warm-up Training
II.2 Task Description Reading
II.3 Task Completion and Interaction Information Logging
II.4 Satisfaction Feedback
Start
Task Completed
Introduction - History - Category - Structure
Please Give Your Satisfaction Score

Query Term:interview of Lee Sedol
Task Description:find the interview of Lee Sedol after his match against AlphaGo

Ebola virus mutation

Ebola virus(Filoviridae Ebola virus) - Sogou Baike
Ebola virus is a virus in the family of filamentous viruses. It can cause hemorrhagic fever. Suffering from this disease can cause a person to die. The virus contains a number of different levels of symptoms (including nausea, vomiting, diarrhea, skin color changes, body aches, body bleeding, bleeding, fever, etc.). People who are infected with Ebola have very similar symptoms as people infected with Marburg virus is...
Sogou Baike - baike.sogou.com/v... - 2016-1-15 - snapshot

Please give your satisfaction score for this search experience.(Not satisfied: 1 star - Satisfied: 5 stars)
★★★☆☆

Figure 1: Data Collection Procedure

Note that no query reformulation are allowed and the number of search results are fixed to 10 for the consistency of result sets across users. Such data collection settings are similar to previous studies such as [37]. Meanwhile, we adopt SERP-level satisfaction rather than session-level satisfaction in this paper because most offline metrics are designed to measure the quality of only one search result page. While there may be ways to adjust or merge the metrics to measure session-level result quality, the metric adaptation may cause other uncontrollable effects and is out of the scope of this paper.

The search results in Dataset #1 are all organic search results and the ones in Dataset #2 are mainly federated search results. The vertical results included in Dataset #2 contain various types, including image, encyclopedia, news and download. The combination of Dataset #1 and Dataset #2 is consistent with real-life settings because not all SERPs provided by commercial search engines contain vertical results. Furthermore, such composition of our datasets also provides the advantage for us to evaluate how metrics perform differently in homogeneous and heterogeneous search environment (see section 4.4).

In these two datasets, each task is accompanied by several different search result pages provided by several ranking mechanisms with a same pre-specified query, which is to ensure that all users saw the same page under the same ranking mechanism. This makes it possible for us to meta-evaluate the performance of different evaluation metrics. Both datasets contain the following information for each search session: (1) Query and corresponding task descriptions. (2) Information of ranked search results as shown on SERPs. (3) 4-scaled relevance assessments of all search results. (4) 5-scaled user satisfaction annotations. (5) Users' interaction behaviors during the search process, including click-through, mouse hover and dwell time information.

With the rich information provided by the datasets, we can compute most widely-used offline / online metrics and hence meta-evaluate the relationship between these metrics and users' percieved satisfaction scores.

3.2 Search Task Taxonomy

To further evaluate the performance of different evaluation metrics in different search scenarios, we classify the search sessions into different categories based on the query and corresponding task descriptions provided in the dataset. We organize the search sessions into the following two most widely-used task taxonomies in this paper:

[1] The link of the datasets and detailed data descriptions is https://1drv.ms/f/s!AqRbaaorUiT1avvHQdagVv9uOPM

Table 2: Examples of Search Queries and Corresponding Taxonomies

Query	Task Description	Search Goal	Cognitive Level
Meizu official website	find the official website of Meizu	Navigational	Remember
Stramaccioni	find a biographical sketch of Stramaccioni	Informational	Remember
interview of Lee Sedol	find the interview of Lee Sedol after his match against AlphaGo	Informational	Understand
yesterday once more	find the online audition of "yesterday once more" sung by carpenters	Transactional	Remember
dunk video	find online videos about dunking	Transactional	Understand

- Search Goal [5]: This taxonomy classifies search tasks from the perspective of search goals. The search queries are classified into navigational queries, informational queries and transactional queries.
- Cognitive Level [3]: This taxonomy is proposed by Anderson and Krathwohl, which identifies six cognitive process dimensions: remember, understand, apply, analyze, evaluate and create.

Table 2 shows an example set of the search tasks and their corresponding taxonomies while Table 3 presents the numbers of search queries / sessions within different types of search tasks. Note that we only include a subset of all the task types described in the task taxonomies due to the composition of our dataset. With respect to the cognitive level taxonomy, we only classify search sessions as either "remember" or "understand" since it is difficult and unrealistic to classify the fixed 10 result based search sessions into the other four types of search tasks. The data size of the navigational search tasks is comparatively small. Although the meta-evaluation results may be potentially less reliable for this navigational search task, we believe this can still provide useful preliminary insights while the more thorough analysis with more data points are left for future work.

Table 3: Distribution of Queries and Sessions of Different Types of Search Tasks

	Navigational	Informational	Transactional
Queries	10	23	23
Sessions	400	1047	988

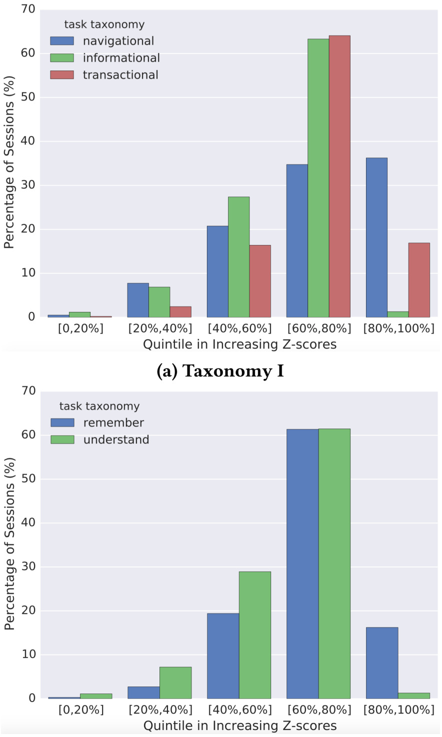
	Remember	Understand
Queries	31	25
Sessions	1325	1110

3.3 Analysis of User Satisfaction

In this section, we try to compare the satisfaction distribution across different search task taxonomies. Inspired by [26], which says satisfaction judgement may be quite subjective and different users may have different opinions, we regularize the satisfaction scores labelled by each user into Z-scores according to equation (1), where sat_i is one particular satisfaction score given by one user and $Avg(Sat)$ is the average of all satisfaction scores he/she labelled. $Var(Sat)$ in equation (1) refers to the variance of the satisfaction scores of this user.

$$Z\text{-}score_i = \frac{sat_i - Avg(Sat)}{Var(Sat)} \tag{1}$$

Figure 2 shows the distribution of quintiles in increasing Z-scores based on different search task taxonomies. Different colors show satisfaction scores from search sessions originated from different task categories. We can see that in all types of search tasks, the general trend is that users tend to give a high satisfaction score in most cases, which indicates that commercial search engines generally provide promising results for these non-long-tailed search tasks.



(a) Taxonomy I

(b) Taxonomy II

Figure 2: Satisfaction Distribution Based on Different Search Task Taxonomies with Different Quintiles

With respect to the first taxonomy based on query intent, we can see that users tend to be the least satisfied if they are searching with informational queries because the percentage of sessions with lower Z-scores (less than 60%) is comparatively higher in the informational case (35.4%) than in navigational (29.0%) and transactional (18.9%) cases. Also, the percentage of sessions of the highest 20 percent of Z-scores is extremely low (1.3%) for informational search tasks compared with the other two types of search tasks. This is reasonable because in most navigational and transactional search tasks, users intended to find a specific website or information resource, which can often be satisfied with only one search result. While in the case of informational search tasks, users often have to read quite a number of search results to get a comprehensive understanding of the information need, which may be more difficult and time consuming.

From the perspective of the second task taxonomy based on task cognitive level, we can see that users give lower satisfaction scores in search tasks which belong to "understand" categories. This is in line with our expectation because search tasks identified as "understand" category are considered to be more difficult than

```
$Total = 0; Correct_1 = 0; Correct_2 = 0;$
foreach pair of search sessions $(s_1, s_2)$ do
    $Total++;$
    $\delta M = M(s_1) - M(s_2);$
    $\delta M^* = M^*(s_1) - M^*(s_2);$
    if $(\delta M \times \delta M^*) > 0$ then // $M$ and $M^*$ positively agree
        $Correct_1++;$
    if $(\delta M \times \delta M^*) < 0$ then // $M$ and $M^*$ negatively agree
        $Correct_2++;$
    if $(\delta M = 0$ and $\delta M^* = 0)$ then // $M$ and $M^*$ agree
        $Correct_1++;$
        $Correct_2++;$
$Correct = Max(Correct_1, Correct_2)$
$Concordance(M, M^*) = Correct/Total;$
```

Algorithm 1: Computing the concordance of metrics M and golden standard metric M^* (user satisfaction feedback) based on preference agreement.

those identified to be "remember" tasks. In general, from the results shown in Figure 2, we can see that users perceive different levels of satisfaction in different search scenarios, which inspires us to study the relationship between different evaluation metrics and satisfaction across different search task taxonomies.

3.4 Meta-Evaluation Methods

With satisfaction widely regarded as the gold standard of user-centric evaluation metrics, we analyze which metrics can better reflect user satisfaction based on the datasets described in this section. We do not consider session-based SAT in our work, rather we assume one SERP page interaction, which consists of most of the search sessions. With respect to the meta-evaluation methodology, we use both pearson correlation coefficient [4] and concordance test to compare different evaluation metrics. The idea of using concordance test is inspired by [42] and the method is described in Algorithm 1. We use s_1 and s_2 to denote different systems of the same task in our dataset and $M(s_i)$ to denote the averaged value of metric M computed on all sessions under s_i. The gold standard metric M^* is user satisfaction. Our algorithm differs from the algorithm used in [42] in that we take the possibility of both positive and negative correlation into consideration. Furthermore, we use strict > and < instead of ≥ and ≤ in the concordance test [42] because loose restrictions cannot work properly for two-valued metrics such as UCTR.

Table 4: Numbers of data points / pairs for meta-evaluation

	# data points for Pearson Correlation	# data pairs for Concordance Test
All	291	744
Navigational	30	30
Informational	130	348
Transactional	131	366
Remember	152	378
Understand	139	366
Homogeneous	78	78
Heterogeneous	213	666

The number of data points for computing pearson correlation and data pairs for concordance test in different search scenarios are shown in Table 4. We should note that the size of data points/pairs of navigational search and homogeneous search (due to the limited ranking mechanisms of Dataset #1) is comparatively small, which may lead to the insignificant results in these two taxonomies. We believe our results (shown in Sec. 4) are still informative and evaluation on a larger dataset can be carried out in the future.

4 EXPERIMENTAL RESULTS

In this section, we meta-evaluate the performance of different evaluation metrics in different search scenarios based on the algorithm described in section 3.4 to obtain thorough insights into how different metrics perform according to different information needs. We first evaluate the performance of offline metrics and online metrics in section 4.1 and 4.2, respectively. In section 4.3, we take a deep insight into how some main offline/online metrics infer user satisfaction in different search scenarios. Finally, we incorporate mouse hover information into some existing online metrics and demonstrate its effectiveness in section 4.4.

4.1 Comparison Across Offline Metrics

Table 5: Comparison of Pearson Correlations / Concordance between Satisfaction and Offline Metrics (* indicates t-test statistical significance at $p < 0.01$ level)

	Pearson Correlation	Concordance
CG	0.354*	45.8%
DCG@3	0.356*	61.6%*
DCG@5	0.411*	65.7%*
DCG@10	0.421*	65.3%*
AP	0.396*	60.2%*
RBP(0.1)	0.389*	66.7%*
RBP(0.5)	0.438*	66.5%*
RBP(0.8)	**0.445***	65.7%*
RBP(0.95)	0.384*	63.4%*
ERR	0.433*	**66.8%***

Based on the search result relevance assessments provided by the datasets, we compute some widely-used traditional offline metrics and investigate how they align with user satisfaction, including cumulative gain (CG), discounted cumulative gain (DCG), average precision (AP) and rank-biased precision (RBP). For DCG, we compare the performance of the metric calculated at different ranking lengths to investigate the effect of evaluation depth. For RBP, we compare the metric performance when the persistence parameter p is set as 0.1, 0.5, 0.8 and 0.95, which is suggested to be appropriate for impatient, neutral, patient and extremely patient users [36], respectively. We also include expected reciprocal rank (ERR), which is based on the "cascade" user model and is suggested to better correlates with click-based metrics compared to DCG and other editorial metrics [9]. The pearson correlation coefficients and concordance test results between these metrics and user satisfaction are shown in Table 5. The best pearson correlation and concordance results are bolded.

We can see that ERR achieves the highest concordance with user satisfaction based on the results in Table 5, which is in line with the findings in [9]. This may be because the "cascade" user model utilized in ERR better models user behavior that capture satisfaction. RBP(0.8) achieves the highest pearson correlation among all the metrics, which may indicate that a patient (but not extremely patient) rank-biased user model can best describe the characteristic of the tested user group. From the perspective of DCG, we can see that DCG@10 correlates user satisfaction slightly better than DCG@3 and DCG@5, which probably indicates that metrics calculated based on a longer ranking length can capture more information and may have a better estimation of user satisfaction. Among all these offline metrics, CG has the lowest pearson correlation and concordance with user satisfaction. This is due to the fact that DCG, AP as well as ERR are all top-weighted metrics while CG is not. Relevant results placed at top positions may be much more important than those placed at bottom. Futhermore, we can note that the poor performance of CG is especially remarkable in the case of navigational search scenario (shown in Table 7) where one top-ranked relevant result is usually sufficient to complete the search goal.

Overall, we can observe that many offline metrics (DCG, RBP and ERR) have significant and moderate correlations (0.4 to 0.6 [14]) with user satisfaction.

4.2 Comparison Across Online Metrics

Table 6: Comparison of Pearson Correlations / Concordance between Satisfaction and Online Metrics (* indicates t-test statistical significance at $p < 0.01$ level)

	Pearson Correlation	Concordance
UCTR	-0.069	24.2%*
QCTR	-0.330*	57.9%*
PCTR@3	0.043	50.3%*
PCTR@5	-0.092	44.5%*
PCTR@10	-0.226*	33.5%*
MaxRR	0.095	50.1%*
MinRR	0.330*	**61.2%***
MeanRR	0.266*	59.5%*
PLC	0.222*	58.1%*
MaxScroll	**-0.519***	60.9%*
SumClickDwell	-0.417*	58.9%*
AvgClickDwell	-0.109	50.9%*
QueryDwellTime	**-0.559***	62.6%*
TimeToFirstClick	-0.432*	65.6%*
TimeToLastClick	-0.504*	**67.3%***
DsatClickCount	-0.170*	56.0%*
DsatClickRatio	-0.130*	52.3%*

While offline metrics are especially valuable when evaluating a system in prior to its deployment [13, 35], online metrics have been widely adopted for modern search engines because such metrics are calculated based on the interactions between practical users and systems. Inspired by previous research on metrics meta-evaluation [9, 11, 15, 20, 39], we compare the evaluation performance of some most widely-used online metrics, including:

- Mouse-based (clicks or scroll) metrics
  - **UCTR** - Binary variable indicating whether there was a click or not in the session (the opposite of abandonment).
  - **QCTR** - Number of clicks in a session.
  - **PCTR** - Page click-through rate as defined in [53]. We use all clicks rather than satisfied clicks to calculate PCTR in Table 6, which is different from [53] where only satisfied clicks are used. Comparisons between metrics based on other user interaction signals (e.g. satisfied clicks) are further discussed in section 4.3.
  - **MaxRR, MinRR, MeanRR** - Respectively maximum, minimum and mean reciprocal ranks of the clicks. Zero if no clicks.
  - **PLC** - Number of clicks divided by the position of the lowest click.
  - **MaxScroll** - Maximum of scroll distance.
- Dwelltime-based metrics
  - **SumClickDwell, AvgClickDwell** - Respectively sum and average of click dwell time in a query.
  - **QueryDwellTime** - Query dwell time.
  - **TimeToFirstClick, TimeToLastClick** - Time delta between the start of search session and the first click and last click in the session, respectively.
  - **DsatClickCount, DsatClickRatio** - Previous studies divide clicks into satisfied clicks and dissatisfied clicks based on various dwell time thresholds [18, 22]. We tested different thresholds and choose to define clicks with a dwell time <15s as dissatisfied clicks because it performs the best on our dataset. Previous work [22] also analyzed the threshold of 15s to differentiate satisfied clicks. We calculate the number and ratio of dissatisfied clicks, respectively.

The online metrics we discuss in this section are mostly based on mouse (click and scroll) behaviors and dwell time information, which can be easily computed based on users' behavior logs. There are also several studies tried to utilize users' behavior information to quantify or predict user satisfaction (e.g. [1, 17, 19]). We do not include these methods in our work because they are more complicated prediction models, rather than the simple and easy-to-interpret evaluation metrics of our interests. Meanwhile, we do not include the session-based metrics discussed in [18, 25] because there are no query reformulations included in the datasets.

The correlations / concordances between the online evaluation metrics and user satisfaction are shown in Table 6, whereas the highest correlation based on each interaction signals are shown in bolded terms. The results in reveal a number of interesting findings:

(1) In contrast to the positive correlation between offline metrics and user satisfaction, online metrics generally correlates with user satisfaction negatively. This is reasonable because the offline metrics measure the quality of search result page based on relevance assessments and users usually tend to feel more satisfied if the search results are of high quality [35]. On the contrary, the interactions signals which online metrics adopted usually reflects search effort and high search effort can reduce user satisfaction [10, 22]. MaxScroll also correlates with satisfaction negatively because a longer scroll distance may also indicate more search effort. MaxRR, MinRR and MeanRR compute the reciprocal ranks of the clicked results and hence correlate with satisfaction positively. It is in line with the findings in previous studies that PLC correlates with satisfaction positively as PLC is regarded as approximately the precision of examined results [9].

(2) The metrics based on click behaviors in general correlates more weakly with user satisfaction, compared with dwelltime-based metrics. This may be because a clicked result does not always necessarily mean a high quality document hence the click-based metrics may fail. Meanwhile, previous studies [38, 45] pointed out that approximately one order of magnitude more online samples are required to match corresponding offline metrics' reliability, which may also explain the reason of the comparatively poor performance of click-based metrics. In contrast, some metrics based on scroll (MaxScroll) and dwelltime information (SumClickDwell, QueryDwellTime and TimeToLastClick) have stronger (moderate) negative correlation with user satisfaction, which means scrolls and dwelltime information are quite important behavior signals to infer user satisfaction.

(3) Among all these online metrics, TimeToLastClick has the best concordance with user satisfaction. The last click in a search session is usually considered as satisfied click [53] and therefore TimeToLastClick measures the time "wasted" before the user find a satisfactory document, which may account for the good performance of TimeToLastClick. We can also note that the concordance between TimeToLastClick and user satisfaction is even better than the offline metrics, which implies that online metrics can be as available as offline metrics even without offline relevance judgments. QueryDwellTime has the strongest (moderate) pearson correlation but relatively low concordance with user satisfaction. This may be because query dwell time has the largest value range among all these metrics, which may take an advantage during the computing process of pearson correlation.

(4) From the perspective of ranking length, we can observe that PCTR@10 correlates user satisfaction better than PCTR@3 and PCTR@5, which is consistent with the findings of DCG and further indicates that metrics calculated based on a longer ranking length can better estimate user satisfaction. We can also note that PCTR has very low concordance with user satisfaction compared with other metrics. This is because there are quite a number of SERPs with the same PCTR metric values in our dataset while there are minor changes within the satisfaction judgements. Therefore, this can result in discordance according to Algorithm 1.

(5) It is in line with our expectation that there is almost no correlation between user satisfaction and simple metrics such as UCTR. UCTR also has a very poor concordance with satisfaction, which is because it is a two-valued metric while we require strict $<$ or $>$ in Algorithm 1.

In general, we can observe that several online metrics (MaxScroll, QueryDwellTime,TimetoFirstClick and TimeToLastClick) maintain significant and moderate correlations (0.4 to 0.6) with satisfaction.

4.3 Online Metrics v.s. Offline Metrics

Based on the findings in section 4.1 and 4.2, we select out some well-behaved metrics and investigate their correlations with user satisfaction in different task taxonomies described in section 3.2 to make a more detailed comparison. The results are shown in Table 7. The highest correlation achieved by offline metrics/online metrics in each task taxonomy are in bolded terms.

The results in Table 7 show that generally online metrics have as good correlation and concordance as offline metrics, which further verifies the value of using online metrics in guiding search engine development because they achieve comparatively good performance without external relevance assessments. In most task scenarios, RBP and ERR perform the best among all offline metrics while TimeToLastClick along with QueryDwellTime perform the best among all online metrics, which is consistent with the findings in section 4.1 and 4.2.

From the search goal-based task taxonomy, we should note that top-weighted offline metrics correlate quite well with user satisfaction in navigational search tasks, especially for RBP(0.1) which models the search behavior of impatient users (strong pearson correlation, 0.6 to 0.8 [14]). While the concordance test results may be discrete and not so reliable due to the limited number of data pairs, the pearson correlation coefficients are extremely significant. This is because in navigational search scenario, where the user is usually required to reach a particular site [5], high quality results placed at the top positions are especially important. In contrast, online metrics perform slightly worse in navigational tasks than in informational and transactional tasks. This is probably because the search effort required in navigational search is usually less than that required in informational and transactional search, in which case the online metrics can hardly capture the differences of user's satisfaction perception.

We do not observe too much difference of the performance of different metrics from the perspective of cognitive level based taxonomy. The dwelltime based online metrics perform slightly better in "understand" search scenario but the difference is not remarkable. Such findings may suggest that users do not behave significantly different in such two types of search tasks. An analysis in tasks with more deep cognitive levels such as "analyze" and "create" [3] can be carried out in the future.

We also observe the poor concordance of PCTR. This is because the test is conducted within the same task across different SERPs generated by different ranking mechanisms. The results on different SERPs are the same in most cases while the rankings are different. Therefore, there might exist many identical metric PCTR values for the two SERPs of many data pairs while users' perceived satisfaction is different. According to Algorithm 1, such case will be regarded as discordance, which is the reason for the poor concordance of PCTR. While most of the results in Table 7 are informative, we must admit the comparatively small-scaled dataset in Navigational search scenarios is an inevitable limitation. A small number of data pairs for concordance test may result in discrete and unavailable results. We may need a larger dataset for more robust results in the future.

4.4 Metric Evaluation in Homogeneous / Heterogeneous Search

With data collected from both homogeneous search environment (Dataset #1) and heterogeneous search environment (Dataset #2), we investigate how different metrics perform in different search environments as shown in Table 8. The highest correlation achieved by different metrics in different search scenarios are bolded.

Different metrics are categorized into three groups and it is obvious from Table 8 that different groups of metrics reveal different characteristics in different search environments. Offline metrics better align with user satisfaction in homogeneous search environment while in heterogeneous search tasks, online metrics, especially

Table 7: Comparison of Pearson Correlations / Concordance between Satisfaction and Offline and Online Metrics in Different Search Scenarios(* indicates t-test statistical significance at $p < 0.01$ level)

	Search Goal			Cognitive Level	
	Navigational	Informational	Transactional	Remember	Understand
CG	0.335 / 53.3%	0.405* / 47.1%	0.354* / 44.0%*	0.414* / 44.7%*	0.389* / 47.0%*
DCG@10	0.543* / 76.7%*	0.454* / 64.1%*	0.403* / 65.6%*	0.475* / 67.5%*	0.437* / 63.1%*
RBP(0.1)	**0.653*** / **80.0%***	0.400* / 66.4%*	0.314* / **65.8%***	0.407* / **68.0%***	0.371* / 65.3%*
RBP(0.8)	0.566* / **80.0%***	**0.475*** / 65.8%*	**0.419*** / 64.5%*	**0.492*** / 66.7%*	**0.454*** / 64.8%*
ERR	0.625* / **80.0%***	0.451* / **66.7%***	0.348* / **65.8%***	0.432* / **68.0%***	0.430* / **65.6%***
QCTR	-0.187 / 53.3%	-0.345* / 58.0%*	-0.290* / 58.2%*	-0.272* / 58.2%*	-0.367* / 57.7%*
PCTR@10	-0.135 / 0.0%*	-0.358* / 33.9%*	0.018 / 35.8%*	-0.005 / 33.6%*	-0.392* / 33.3%*
MinRR	0.454* / **73.3%***	0.323* / 61.5%*	0.281* / 59.8%*	0.318* / 59.8%*	0.344* / 62.6%*
PLC	0.408* / 70.0%*	0.195* / 57.2%*	0.181* / 57.9%*	0.242* / 57.7%*	0.213* / 58.5%*
MaxScroll	**-0.477*** / **73.3%***	-0.577* / 59.2%*	-0.465* / 61.5%*	-0.471* / 63.2%*	-0.547* / 58.5%*
QueryDwellTime	-0.392* / 50.0%	**-0.600*** / 62.1%*	**-0.479*** / 64.2%*	**-0.485*** / 64.6%*	-0.572* / 60.7%*
TimeToLastClick	-0.351 / 50.0%	-0.525* / **68.1%***	-0.428* / **68.0%***	-0.445* / **67.5%***	**-0.551** / **67.2%***
DsatClickCount	-0.020 / 53.3%	-0.187 / 55.7%*	-0.145 / 56.6%*	-0.134 / 56.1%*	-0.223* / 56.0%*

Table 8: Comparison of Pearson Correlations / Concordance between Satisfaction and Offline and Online Metrics in Homogeneous and Heterogeneous Search Environment(* indicates t-test statistical significance at $p < 0.01$ level)

		homogeneous search		heterogeneous search	
		Pearson Correlation	Concordance	Pearson Correlation	Concordance
Offline Metrics	CG	0.483*	55.1%	0.321*	44.7%
	DCG@10	**0.535***	70.5%*	0.392*	64.7%*
	RBP(0.1)	0.433*	**71.8%***	0.383*	66.1%*
	RBP(0.8)	**0.535***	**71.8%***	0.418*	65.0%*
	ERR	0.462*	**71.8%***	**0.429***	**66.2%***
Online Metrics (Mouse-based)	QCTR	-0.137	55.1%	-0.401*	58.3%*
	PCTR@10	-0.062	2.6%*	-0.284*	37.4%*
	MinRR	0.417*	**64.1%***	0.325*	**60.8%***
	PLC	0.366*	61.5%*	0.176*	57.7%*
	MaxScroll	**-0.510***	**64.1%***	**-0.540***	60.5%*
Online Metrics (Dwelltime-based)	QueryDwellTime	-0.224*	**57.7%**	**-0.637***	65.0%*
	TimeToLastClick	**-0.235***	50.0%	-0.570*	**69.4%***
	DsatClickCount	-0.034	53.8%	-0.209*	56.3%*

the dwelltime-based metrics perform much better. This is reasonable because most existing offline metrics do not take the effect of vertical results into consideration during the evaluation process. While both organic and vertical search results can provide relevant information, vertical results are presented in different styles and can help satisfy users' information need from various dimensions [10]. Offline metrics are solely based on relevance assessments and do not consider the effect of vertical results, which may account for their comparatively poor performance in heterogeneous search environment. Online metrics are mainly based on users' interaction behaviors, which may also be sensitive to the existence of vertical results. Therefore, the online metrics may be more effective in the heterogeneous search environment. Dwelltime-based metrics achieves a pearson correlation of around -0.6 (strong) in heterogeneous search while the best offline metrics is only moderately correlated (around 0.4). In addition, note that since high quality vertical results can provide users sufficient information to complete the search goal without a need to click (called "good abandonment" in previous work [30]), the number of clicks in heterogeneous search may not be sufficient and hence may lead to a drop of performance of click-based online metrics. In our datasets, there are on average 2.83 clicks in a homogeneous search session while only 1.81 clicks in a heterogeneous one. This may be the reason that click-based online metrics do not perform so well in heterogeneous search.

4.5 Click and Hover based Online Metrics

Previous research [10, 16] suggested that fine-grained user interaction information can help better model user behaviors and estimate user satisfaction. With the rich interaction informational provided by the datasets, we further investigate the performance of online metrics calculated based on the following four types of user interaction signals:

All click-based: This group of evaluation metrics are calculated based on all clicks in a search session, which is the same as the metrics used in the previous sections.

Satisfied click-based: Previous studies pointed out that sometimes clicks do not imply the high relevance of a result document in web search. This is because although the result snippet may appear to be relevant and attractive, the landing page may be of low quality. For example, a click is defined as a satisfied click if the user spent 30 seconds or more reading the clicked document or if it was the last click in the search session [53]. Satisfied click is widely regarded as the signal of a relevant document. For this group of metrics, we use satisfied clicks to replace the all clicks used in "all click-based" metrics.

Hover-based: Hovers are also regarded to be important behavior signal because there have been various types of results which do not require a click to provide users with necessary information [10, 30]. We use hovers to replace the all clicks used in "all click-based" metrics to achieve the "hover-based" metrics.

Click and Hover-based: We combine the click and hover information in this group of evaluation metrics. If a search results is either clicked or hovered on, then this result will be regarded as "clicked" as in the calculation process of "all click-based" metrics. In this way, we get the "click and hover-based" metrics.

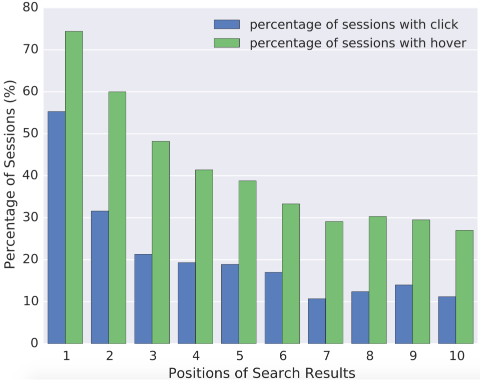


Figure 3: Percentage of all search sessions with clicks/hovers at different result positions

We use suffix "_ac", "_sc", "_h" and "_ch" to represent the "all click-based", "satisfied click-based", "hover-based" and "click and hover-based" metrics, respectively. We choose MinRR and PLC as examples because they correlate with user satisfaction better than other click-based metrics based on the findings in previous sections. The correlations between user satisfaction and metrics computed based on different information signals are shown in Table 9. The best correlation / concordance achieved by each metric in different search scenarios are bolded.

We can see from the results that in almost all search scenarios, both of these two click-based metrics can better estimate user satisfaction when hover information is incorporated. This is probably because in today's search engine, various types of results such as instant answers, verticals and even result snippets contain sufficient information to satisfy the users, which sometimes makes clicks unnecessary. In such case, hovers can help capture more information than clicks. It is not surprising that the performance of MinRR_h and MinRR_ch are the same because in most cases the clicked results are usually a subset of hovered results. Furthermore, we can see that the hover information is especially effective in heterogeneous search according to the performance of MinRR. When hover information is incorporated, the pearson correlation coefficient improves by 0.169(from 0.325 to 0.494) in heterogeneous search and only 0.025(from 0.417 to 0.442) in homogeneous search. The improvement of concordance test result is also larger in heterogeneous search environment. This is reasonable because sometimes users can accomplish their search tasks by interacting with the vertical results on heterogeneous SERPs. An example of the distribution differences between hovers and clicks are shown in Figure 3. For each rank position of search results, we compute the percentage of sessions with clicks and hovers respectively. It is apparent that there are more hovers than clicks in all positions, which may further help confirm that clicks solely may not be sufficient to capture users' interaction information. There appears to be at least 27% sessions with hovers in all positions while only the first two positions have a probability of more than 27% to be clicked. In this way, hovers may contain much more valuable information than clicks. From the perspective of MinRR and PLC, a metric which combines click and hover information may be the most reliable because neither click nor hover information alone (PLC_h) can achieve the best correlation with satisfaction.

5 CONCLUSIONS

Search engine evaluation is essential in both academic and industrial IR research. Both offline and online evaluation metrics are adopted to measure the performance of search engines. While search satisfaction is widely regarded as the gold standard in search performance evaluation, the relationship between different evaluation metrics and satisfaction remains under-investigated.

In this work, we meta-evaluate the performance of different offline/online evaluation metrics based on two datasets. We investigate how different metrics align with user satisfaction in different search scenarios using both pearson correlation and concordance test. We find that different types of evaluation metrics estimate user satisfaction from different perspectives. Offline metrics work better in homogeneous search environment while online metrics are more consistent with user satisfaction in heterogeneous search environment. We further compare the effectiveness of metrics calculated based on different user interaction signals. We propose to incorporate hover information into traditional click-based online metrics because they can help better estimate user satisfaction.

There are still some limitations of our work which we would like to list as our future work directions. Due to the nature of the utilized laboratory-based datasets, compared to the commercial search engine settings, online metrics are calculated based on relatively small-scale search sessions (from which our conclusions are drawn). Meanwhile, our datasets are based on a fixed search result page and no ability to reformulate the query, which may also affect users' interaction behaviors. While metrics calculated based on multi-query sessions can be developed, it will be another challenging research question. Our current aim is to meta-evaluate all the metrics on a single search page level, and we leave the session-level or even task-level evaluation as future work. Finally, all the participants within the datasets are undergraduate students. We think this may help reduce potential distractions and make the collected data more consistent. However, such specific age distribution may also cause potential bias. The study of large-scale online/offline metrics comparison, online metric sensitivity and more evaluation measures (such as interleaving) are left for future work.

REFERENCES

[1] Mikhail Ageev, Qi Guo, Dmitry Lagun, and Eugene Agichtein. 2011. Find it if you can: a game for modeling different types of web search success using interaction data. In *SIGIR'11*. ACM, 345–354.

[2] Azzah Al-Maskari, Mark Sanderson, and Paul Clough. 2007. The relationship between IR effectiveness measures and user satisfaction. In *SIGIR'07*. ACM, 773–774.

[3] Lorin W Anderson, David R Krathwohl, and Benjamin Samuel Bloom. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Allyn & Bacon.

[4] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*. Springer, 1–4.

Table 9: Comparison of Pearson Correlations / Concordance between Satisfaction and Online Metrics Based on Different User Interaction Signals (* indicates t-test statistical significance at $p < 0.01$ level)

	Search Goal			Cognitive Level		Search Environment	
	Navigational	Informational	Transactional	Remember	Understand	Homogeneous	Heterogeneous
MinRR_ac	0.454* / 73.3%*	0.323* / 61.5%*	0.281* / 59.8%*	0.318* / 59.8%*	0.344* / 62.6%*	0.417*/64.1%*	0.325*/60.8%*
MinRR_sc	0.442* / **76.7%***	0.322* / **62.6%***	0.209* / 58.2%*	0.254* / 58.7%*	0.339* / **63.4%***	0.385*/**67.9%***	0.278*/60.2%*
MinRR_h	**0.464*** / 70.0%	**0.474*** / 61.8%*	**0.440*** / **66.7%***	**0.462*** / **67.5%***	**0.486*** / 61.5%*	**0.442***/65.4%*	**0.494***/**64.4%***
MinRR_ch	**0.464*** / 70.0%	**0.474*** / 61.8%*	**0.440*** / **66.7%***	**0.462*** / **67.5%***	**0.486*** / 61.5%*	**0.442***/65.4%*	**0.494***/**64.4%***
PLC_ac	0.408* / 70.0%*	0.195* / 57.2%*	0.181* / 57.9%*	0.242* / 57.7%*	0.213* / 58.5%*	0.366*/61.5%	0.176*/57.7%*
PLC_sc	0.416* / 73.3%*	0.198* / 56.9%*	0.119* / 57.1%*	0.190* / 57.7%*	0.215* / 58.5%*	0.360*/67.9%*	0.139/57.7%*
PLC_h	0.302 / 56.7%	0.113 / 48.0%	0.176 / 46.4%	0.198 / 47.6%	0.108 / 47.5%	0.179/51.3%	0.142/47/1%
PLC_ch	**0.444** / **80.0%***	**0.262*** / **59.2%***	**0.302*** / **61.7%***	**0.347*** / **63.8%***	**0.271*** / **58.7%***	**0.429***/**69.2%***	**0.265***/**60.4%***

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

[6] Chris Buckley and Ellen M Voorhees. 2000. Evaluating evaluation measure stability. In *SIGIR'00*. ACM, 33–40.

[7] Stefan Büttcher, Charles LA Clarke, Peter CK Yeung, and Ian Soboroff. 2007. Reliable information retrieval evaluation with incomplete and biased judgements. In *SIGIR'07*. ACM, 63–70.

[8] Ben Carterette, Evangelos Kanoulas, and Emine Yilmaz. 2010. Low cost evaluation in information retrieval. In *SIGIR'10*. ACM, 903–903.

[9] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *CIKM'09*. ACM, 621–630.

[10] Ye Chen, Yiqun Liu, Ke Zhou, Meng Wang, Min Zhang, and Shaoping Ma. 2015. Does Vertical Bring more Satisfaction?: Predicting Search Satisfaction in a Heterogeneous Environment. In *CIKM'15*. ACM, 1581–1590.

[11] Aleksandr Chuklin, Pavel Serdyukov, and Maarten De Rijke. 2013. Click model-based information retrieval metrics. In *SIGIR'13*. ACM, 493–502.

[12] Cyril Cleverdon, Jack Mills, and Michael Keen. 1966. FACTORS DETERMINING THE PERFORMANCE OF INDEXING SYSTEMS VOLUME 1. DESIGN. (1966).

[13] Alex Deng and Xiaolin Shi. 2016. Data-Driven Metric Development for Online Controlled Experiments: Seven Lessons Learned. In *SIKDD'16*. ACM.

[14] Geoffrey Evans, Anthony Heath, and Mansur Lalljee. 1996. Measuring left-right and libertarian-authoritarian values in the British electorate. *British Journal of Sociology* (1996), 93–112.

[15] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating implicit measures to improve web search. *TOIS'05* 23, 2 (2005), 147–168.

[16] Qi Guo, Dmitry Lagun, and Eugene Agichtein. 2012. Predicting web search success with fine-grained interaction data. In *CIKM'12*. ACM, 2050–2054.

[17] Ahmed Hassan, Rosie Jones, and Kristina Lisa Klinkner. 2010. Beyond DCG: user behavior as a predictor of a successful search. In *WSDM'10*. ACM, 221–230.

[18] Ahmed Hassan, Xiaolin Shi, Nick Craswell, and Bill Ramsey. 2013. Beyond clicks: query reformulation as a predictor of search satisfaction. In *CIKM'13*. ACM, 2019–2028.

[19] Ahmed Hassan, Yang Song, and Li-wei He. 2011. A task level metric for measuring web search satisfaction and its application on improving relevance estimation. In *CIKM'11*. ACM, 125–134.

[20] Katja Hofmann, Lihong Li, Filip Radlinski, and others. 2016. Online evaluation for information retrieval. *Foundations and Trends® in Information Retrieval* 10, 1 (2016), 1–117.

[21] Scott B Huffman and Michael Hochster. 2007. How well does result relevance predict session satisfaction?. In *SIGIR'07*. ACM, 567–574.

[22] Jiepu Jiang, Ahmed Hassan Awadallah, Xiaolin Shi, and Ryen W White. 2015. Understanding and predicting graded search satisfaction. In *WSDM'15*. ACM, 57–66.

[23] Jiepu Jiang, Daqing He, and James Allan. 2014. Searching, browsing, and clicking in a search session: changes in user behavior by task and over time. In *SIGIR'14*. ACM, 607–616.

[24] Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *SIGKDD'02*. ACM, 133–142.

[25] Evangelos Kanoulas, Ben Carterette, Paul D Clough, and Mark Sanderson. 2011. Evaluating multi-query sessions. In *SIGIR'11*. ACM, 1053–1062.

[26] Diane Kelly. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval* 3, 1fi!?2 (2009), 1–224.

[27] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Imed Zitouni, Aidan C Crook, and Tasos Anastasakos. 2016. Predicting User Satisfaction with Intelligent Assistants. In *SIGIR'16*. 495–505.

[28] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M Henne. 2009. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery* 18, 1 (2009), 140–181.

[29] Dmitry Lagun, Chih-Hung Hsieh, Dale Webster, and Vidhya Navalpakkam. 2014. Towards better measurement of attention and satisfaction in mobile search. In *SIGIR'14*. ACM, 113–122.

[30] Jane Li, Scott Huffman, and Akihito Tokuda. 2009. Good abandonment in mobile and PC internet search. In *SIGIR'09*. ACM, 43–50.

[31] Lihong Li, Jin Young Kim, and Imed Zitouni. 2015. Toward predicting the outcome of an A/B experiment for search relevance. In *WSDM'15*. ACM, 37–46.

[32] Yiqun Liu, Ye Chen, Jinhui Tang, Jiashen Sun, Min Zhang, Shaoping Ma, and Xuan Zhu. 2015. Different users, different opinions: Predicting search satisfaction with mouse movement information. In *SIGIR'15*. ACM, 493–502.

[33] Zeyang Liu, Yiqun Liu, Ke Zhou, Min Zhang, and Shaoping Ma. 2015. Influence of vertical result in web search examination. In *SIGIR'15*. ACM, 193–202.

[34] Xiaolu Lu, Alistair Moffat, and J Shane Culpepper. 2016. The effect of pooling and evaluation depth on IR metrics. *Information Retrieval Journal* 19, 4 (2016), 416–445.

[35] Jiaxin Mao, Yiqun Liu, Ke Zhou, Jian-Yun Nie, Jingtao Song, Min Zhang, Shaoping Ma, Jiashen Sun, and Hengliang Luo. 2016. When does Relevance Mean Usefulness and User Satisfaction in Web Search?. In *SIGIR'16*. ACM.

[36] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *TOIS'08* 27, 1 (2008), 2.

[37] Vidhya Navalpakkam, LaDawn Jentzsch, Rory Sayres, Sujith Ravi, Amr Ahmed, and Alex Smola. 2013. Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In *WWW'13*. ACM, 953–964.

[38] Filip Radlinski and Nick Craswell. 2010. Comparing the sensitivity of information retrieval metrics. In *SIGIR'10*. ACM, 667–674.

[39] Filip Radlinski, Madhu Kurup, and Thorsten Joachims. 2008. How does clickthrough data reflect retrieval quality?. In *CIKM'08*. ACM, 43–52.

[40] Sri Devi Ravana and Alistair Moffat. 2010. Score estimation, incomplete judgments, and significance testing in IR evaluation. In *AIRS'10*. Springer, 97–109.

[41] Tetsuya Sakai. 2006. Evaluating evaluation metrics based on the bootstrap. In *SIGIR'06*. ACM, 525–532.

[42] Tetsuya Sakai. 2013. How intuitive are diversified search metrics? Concordance test results for the diversity U-measures. In *AIRS'13*. Springer, 13–24.

[43] Tetsuya Sakai and Noriko Kando. 2008. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval* 11, 5 (2008), 447–470.

[44] Mark Sanderson. 2010. *Test collection based evaluation of information retrieval systems*. Now Publishers Inc.

[45] Anne Schuth, Katja Hofmann, and Filip Radlinski. 2015. Predicting search satisfaction metrics with interleaved comparisons. In *SIGIR'15*. ACM, 463–472.

[46] Hinrich Schütze. 2008. Introduction to Information Retrieval. In *Proceedings of the international communication of association for computing machinery conference*.

[47] Louise T Su. 1992. Evaluation measures for interactive information retrieval. *Information Processing & Management* 28, 4 (1992), 503–516.

[48] Louise T Su. 2003. A comprehensive and systematic model of user evaluation of Web search engines: II. An evaluation by undergraduates. *Journal of the American Society for Information Science and Technology* 54, 13 (2003), 1193–1223.

[49] Hongning Wang, Yang Song, Ming-Wei Chang, Xiaodong He, Ahmed Hassan, and Ryen W White. 2014. Modeling action-level satisfaction for search task satisfaction prediction. In *SIGIR'14*. ACM, 123–132.

[50] William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *TOIS'10* 28, 4 (2010), 20.

[51] Emine Yilmaz, Evangelos Kanoulas, and Javed A Aslam. 2008. A simple and efficient sampling method for estimating AP and NDCG. In *SIGIR'08*. ACM, 603–610.

[52] Emine Yilmaz and Stephen Robertson. 2010. On the choice of effectiveness measures for learning to rank. *Information Retrieval* 13, 3 (2010), 271–290.

[53] Masrour Zoghi, Tomáš Tunys, Lihong Li, Damien Jose, Junyan Chen, Chun Ming Chin, and Maarten de Rijke. 2016. Click-based Hot Fixes for Underperforming Torso Queries. SIGIR.