Cascade or Recency: Constructing Better Evaluation Metrics for Session Search*

Fan Zhang, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Min Zhang, Shaoping Ma†
Department of Computer Science and Technology, Institute for Artificial Intelligence,
Beijing National Research Center for Information Science and Technology,
Tsinghua University
Beijing, China
yiqunliu@tsinghua.edu.cn

ABSTRACT

Recently session search evaluation has been paid more attention as a realistic search scenario usually involves multiple queries and interactions between users and systems. Evolved from model-based evaluation metrics for a single query, existing session-based metrics also follow a generic framework based on the cascade hypothesis. The cascade hypothesis assumes that lower-ranked search results and later-issued queries receive less attention from users and should therefore be assigned smaller weights when calculating evaluation metrics. This hypothesis gains much success in modeling search users' behavior and designing evaluation metrics, by explaining why users' attention decays on search engine result pages. However, recent studies have found that the recency effect also plays an important role in determining user satisfaction in search sessions. Especially, whether a user feels satisfied in the later-issued queries heavily influences his/her search satisfaction in the whole session. To take both the cascade hypothesis and the recency effect into the design of session search evaluation metrics, we propose Recency-aware Session-based Metrics (RSMs) to simultaneously characterize users' examination process with a browsing model and cognitive process with a utility accumulation model. With both self-constructed and public available user search behavior datasets, we show the effectiveness of proposed RSMs by comparing them with existing session-based metrics in the light of correlation with user satisfaction. We also find that the influence of the cascade and the recency effects varies dramatically among tasks with different difficulties and complexities, which suggests that we should use different model parameters for different types of search tasks. Our findings highlight the importance of investigating and utilizing cognitive effects besides examination hypotheses in search evaluation.

*This work is supported by the National Key Research and Development Program of China (2018YFC0831700), Natural Science Foundation of China (Grant No. 61732008, 61532011, 61902209) and Beijing Academy of Artificial Intelligence (BAAI). Dr Weizhi Ma has been supported by Shuimu Tsinghua Scholar Program.
†Corresponding Author: Yiqun Liu

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401163

CCS CONCEPTS

• **Information systems** → **Evaluation of retrieval results**.

KEYWORDS

recency effect, user behavior, session search, evaluation metrics

ACM Reference Format:
Fan Zhang, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Min Zhang, Shaoping Ma. 2020. Cascade or Recency: Constructing Better Evaluation Metrics for Session Search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3397271.3401163

1 INTRODUCTION

To help improve the performance of search engines, researchers in Information Retrieval (IR) community have been focused on developing evaluation methods for many years. Traditionally, the Cranfield evaluation paradigm [8], which plays a critical role in search evaluation, is mainly designed for evaluating the performance of search engines given single queries. However, as users' information needs become more and more complex in realistic search scenarios, users usually reformulate their queries as search sessions proceed until they are satisfied or frustrated. How to evaluate the quality of search sessions, rather than single queries, has become a great challenge and received much attention in recent years.

TREC Session Track [6] and Dynamic Domain Track [32] are two important attempts to evaluate system performance over an entire search session. The session-based metrics they adopt (e.g. Session-based DCG [12] and Cube Test [22]) are based on the cascade hypothesis [9], which is also adopted in the designing of most query-level metrics. Given this hypothesis in a search session, lower-ranked search results and later-issued queries receive less attention and smaller weight. However, previous studies [13, 19] suggest that the last query within a session may have a stronger correlation with users' search satisfaction compared with other queries. Liu et al. [20] further investigate the influence of cognitive effects on users' satisfaction in terms of whole search sessions. They find that the recency effect has a stronger influence on users' satisfaction in a session. It indicates that users' impressions of later queries receive greater weights in forming their satisfaction perceptions.

An evaluation metric can be viewed as a measurement of a simulated user's search experience based on a user model. Moffat et al. [24] explore the relationship between user models and metrics.

They describe the conditional probability that users proceed to the next result once they have reached the current result in the ranking. Further, Carterette [4] summarizes the three underlying models composing model-based measures:

- a *browsing model* that describes how a user interacts with results;
- a *document utility model* that describes how a user derives utility from individual relevant documents;
- a *utility accumulation model* that describes how a user accumulates utility in the course of browsing.

While most related studies focus on the browsing model and the document utility model for search evaluation, few studies have investigated the utility accumulation model within this framework. Carterette [4] describes four utility accumulation models in existing measures. However, these four models have similar forms, which are the expected utility/total utility/effort/average utility at the stopping rank, given a probability distribution of users' stopping behavior. Inspired by Liu et al. [20], we assume that the utility accumulation model should be affected by cognitive effects. Specifically, in a search session, the accumulation of utility perceived by users should be decided not only by user's examination process (which determines the probability that a result is examined), but also by his/her cognitive process (which is related to user's impression of a result). Therefore, in this paper, we combine user's examination process and cognitive process to enhance the performance of evaluation metrics for session search. Incorporating the recency effect into the framework of session-based metrics, we propose Recency-aware Session-based Metrics (RSMs).

To verify the effectiveness of proposed metrics in a more realistic search scenario, we conduct a field study involving 30 participants for one month. During this field study, participants' daily search logs are collected and they are required to provide explicit feedbacks for their search experiences. Besides this self-constructed dataset, we also compare the performance of different metrics based on a public available search dataset [1] from an existing user study [14]. Our results on the above two datasets show that the proposed RSMs have stronger correlations with user satisfaction compared with existing session-based metrics. We further investigate the influence of two cognitive factors, task difficulty and task complexity, on users' examination process and cognitive process. It is found that users' examination and cognitive behaviors vary among tasks with different difficulties and complexities, which means we should use different model parameters for different types of search tasks. To summarize, the main contributions of our work are three folds:

- We propose Recency-aware Session-based Metrics (RSMs), which extending existing session-based metrics by considering cognitive effects. Based on both self-constructed and public-available search user behavior datasets, we show stronger correlations between user satisfaction and RSMs, compared with existing session-based metrics.
- We investigate the differences in users' examination process and cognitive process between tasks of different difficulties and complexities. The results suggest that it is important to consider task features in the design of session-based metrics.
- We construct a field study based session search dataset, which contains more abundant behaviors and feedbacks from users in their daily search sessions than previous studies. This dataset will be publicly available upon publication of the paper. We hope it can provide more convenience for researchers working on session search related studies in a more realistic view.

[1] https://github.com/jiepujiang/ir_metrics

2 RELATED WORK

2.1 Model-based Evaluation Metrics

Evaluation has been sitting at the center of IR researches for many years. To compare different search systems repeatedly and automatically, numerous evaluation metrics have been proposed based on the widely used Cranfield evaluation paradigm [8]. Under this paradigm, the test collection-based evaluation with chosen metrics provide a simulation of users of a searching system in an operational setting [27]. Given the simulation, an evaluation metric can be viewed as a measurement of users' search experience based on a user model describing how a simulated user interacts with the system. For example, Moffat and Zobel [25] formally use the idea of a user model in Rank-Biased Precision (RBP) and assume that users will examine the results one-by-one from top to bottom and end the current browsing with a certain probability. Considering different constraints for the continuation probability, user models behind some other metrics also depend on the cascade hypothesis [9] that the examination pattern of users is sequential and unidirectional. These metrics include Expected Reciprocal Rank (ERR) [7], Expected Browsing Utility (EBU) [34], Time-Biased Gain (TBG) [28], U-measure [26], INST [3], Height-Biased Gain (HBG) [21], Bejeweled Player Model (BPM) [35] and Information Foraging Based Measure (IFT) [1]. Further, Moffat et al. [24] introduce the C/W/L Framework to formalize user models with three different but interrelated ways: Continuation (C) probability, Weight (W) function and Last (L) probability. Choosing different user models, we will get different continuation probability and weight distributions, thus different metrics.

These studies provide a perspective of simulated users to evaluate users' search experience. However, they mainly focus on search evaluation for single queries. Whether the framework in this line of research is valid for session search attracts the attention of researchers, which also motivates our work in this paper.

2.2 Session Search Evaluation

Compared with evaluation for single queries, session search evaluation focuses on a more realistic search scenario since it considers multiple queries and interactions between users and systems. It also presents a challenge to evaluate the performance of the whole search session involving multiple queries.

As an intuitive extension of the Discounted Cumulative Gain (DCG) [11], the Session-based DCG (sDCG) [12] metric is proposed for multiple interactive queries. It incorporates query sequence as another dimension and assumes that the results in later queries are less likely to be examined by users. Inspired by sDCG, Lipani et al. [17] develop a generalization of RBP [25], which they call session RBP (sRBP). Besides the persistence parameter similar to RBP, sRBP introduces a new parameter, balancer, to quantify the

balance between reformulating queries and examining more results in the current SERP. They show that sRBP better characterizes the observed user behavior compared to sDCG.

Taking novelty into account, Yang and Lad [33] propose to compute Expected Utility (EU) by simulating all the possible browsing patterns. It considers the information nuggets and discount the utility of a result the nugget of which has been encountered before. Similarly, Cube Test [22] use subtopics to refer to the similar idea of nuggets for a session. The gain of a result will be discounted if its subtopic has been covered. These two metrics are more like diversity metrics for which relevance of results regarding different nuggets/subtopics are required. In this paper, we do not consider diversity and leave this for future work.

Instead of extending query-level metrics with another dimension for session search, Jiang and Allan [13] compare different methods to aggregate the nDCG scores of individual queries composing a search session. They show that the metric sDCG/q, which takes the cost factor into account, significantly correlates with user-rated performance. In addition, nDCG of the last query may have a stronger influence on users' search experience compared with other queries in a session.

From the perspective of the user model, all of these metrics proposed for session search are still based on the cascade hypothesis [9]. However, Liu et al. [19] investigate the effectiveness of this assumption for session search and find that user satisfaction in a session is highly correlated with the most recently issued queries. Their recent work [20] is most relevant to our paper. They investigate the influence of cognitive effects on users' satisfaction in terms of a whole search session and find that the recency effect has a stronger influence on users' satisfaction in a session. It indicates that users' impressions of later queries receive greater weight in forming their satisfaction feedback. Our study differs from their work on considering the recency effect in session search by combining users' examination process and cognitive process, rather than only modifying the weighting functions for the scores of the queries. We also compare the differences in users' examination process and cognitive process when considering tasks of different difficulty and complexity. This work bridge the gap between the metrics based on user examination model and user satisfaction by introducing user cognitive model for a whole search session.

3 RECENCY-AWARE SESSION-BASED METRICS

Before introducing the recency effect into the design of session search evaluation metrics, we first take a view of the framework of existing metrics.

3.1 A Framework of Existing Metrics

Most search effectiveness metrics can be calculated with a generic framework [37] as the inner-product between a utility vector $\vec{g}$ and a discount vector $\vec{d}$:

$$f(q) = \sum_{n=1}^{N} g_n(q) \cdot d_n \tag{1}$$

where $g_n(q)$ denotes the gain users obtain from the n-th result returned by the system given the query q. It is usually measured by relevance judgement. The discount factor d_n for the n-th result is usually estimated by the probability that the result is examined by users. N is the number of results.

Recently, Lipani et al. [17] generalize Equation 1 for session search as follows:

$$f(s) = \sum_{m=1}^{M} \sum_{n=1}^{N} g_{m,n}(q_m) \cdot d_{m,n} \tag{2}$$

where M is the number of queries. $g_{m,n}(q_m)$ and $d_{m,n}$ are respectively the gain and the discount for the n-th result returned by the system given the m-th query in the session.

Given this framework, the differences between sDCG and sRBP mainly come from the definitions of their discount functions:

$$\begin{aligned} d_{m,n}(sDCG) &= \frac{1}{(1+\log_{b_r} n)(1+\log_{b_q} m)} \\ d_{m,n}(sRBP) &= \left(\frac{p-bp}{1-bp}\right)^{m-1} (bp)^{n-1} \end{aligned} \tag{3}$$

For sDCG, b_r and b_q are the logarithm base parameters for the rank and query discount. These two parameters are set to model different search behaviors: larger values are used to model more patient users who are willing to examine more results and submit more queries, respectively. Note that there are variant logarithmic forms of the discount function for sDCG in previous works [12, 13, 17, 29]. In this paper, we adopt the form used in [29] to ensure $d_{1,1} = 1$, which means that users will examine the first result of the first query all the time. For sRBP, b and p are the balance and persistence parameters. As shown in [17], $\alpha = b \cdot p$ and $\beta = (1-b) \cdot p$ are the probabilities that users examine the next result in the current query and issue a new query after examining a result, respectively.

3.2 Incorporating the Recency Effect

As mentioned in the introduction, Carterette [4] suggests that model-based measures are generally composed of a browsing model, a document utility model, and a utility accumulation model. From this perspective, g in Equation 2 measures the utility of a result, while d is determined by users' browsing model. As for the utility accumulation model, the sum of the expected utility of each result is computed to measure the performance of a session. However, Liu et al. [20] indicate that users' impressions of more recent queries should receive greater weight in forming their satisfaction perceptions. Inspired by their work, we assume that the utility accumulation model should be affected by the recency effect.

To incorporate the recency effect into the utility accumulation model, we describe the process of forming user satisfaction in a search session. As shown in Figure 1, a user starts searching by issuing an initial query and gets results returned by the system. After examining a result, the user will decide whether to examine the next result or reformulate a new query until the user ends this search session. That is the interactive process between a user and a system described by the browsing model. For example, sRBP assumes that the user will examine the next result and reformulate a new query with constant probabilities. Besides the browsing model, each time the user examine a search result, he/she will assess the utility of the result, which derives from the document utility model. Most metrics (e.g. nDCG [11], RBP [25], TBG [28], etc.) use relevance

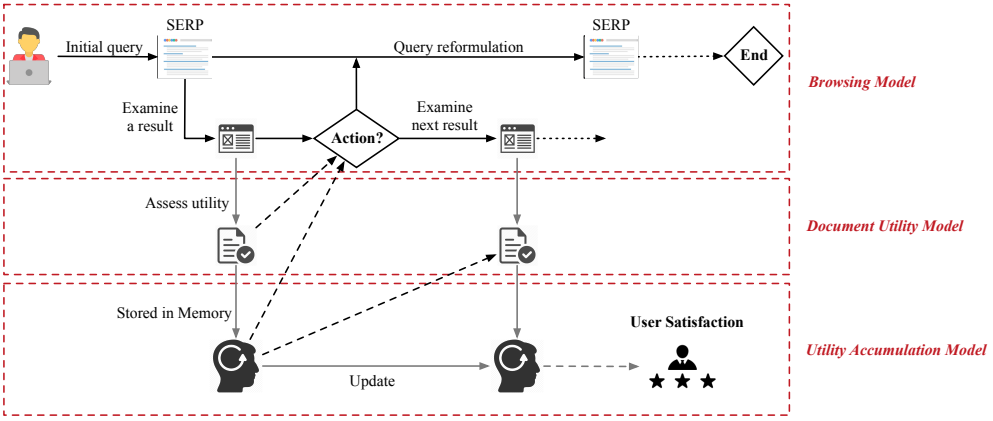

Figure 1: An illustration of the process of forming user satisfaction in a search session.

as a substitute for utility. For simplicity, we also use relevance to model document utility in this paper. In addition, before feeling satisfied, the user will store the results he/she has encountered in his/her memory and update it as the search session proceeds. A utility accumulation model describes how users update the memory and aggregate the utility of different results. Existing metrics usually add the utility of all the results to get the overall utility of the session. However, we assume that this utility accumulation model should be constructed based on the user's cognitive process. Psychological researchers have found that there is a recency effect for short-term memory [2], where the last items presented tend to be the best recalled. Liu et al. [20] also find that the recency effect has a stronger influence on user's session satisfaction. Therefore, in this paper, we focus on the recency effect, and incorporate it by considering the decay of users' memory of results in utility accumulation model.

To formalize this framework with the recency effect, we generalize Equation 2 as follows:

$$f(s) = \sum_{m=1}^{M} \sum_{n=1}^{N} g_{m,n}(q_m) \cdot d_{m,n} \cdot mem_{m,n} \tag{4}$$

where $mem_{m,n}$ is users' memorized weight of the n-th result returned by the system given the m-th query in session s when forming satisfaction. For simplicity, in this paper, we assume that the memory is only related to the distance between the m-th query and the last query, which can be denoted as mem_m. The design of mem_m will be discussed in Section 5. We call the metrics defined by Equation 4 Recency-aware Session-based Metrics (RSMs), denoting session-based metrics which take recency effect into consideration. Note that in Figure 1 we show that the user browsing model may be affected by how users access utility of a result and what have been stored in their memory. The discussion about these influences on the user browsing model is outside the scope of this paper.

4 FIELD STUDY

To verify the effectiveness of proposed RSMs in a more realistic view, we conducted a field study involving 30 participants for one month. This field study is similar to the field studies in [10] and [31], which were proposed to overcome the limitations of lab studies and large-scale log analysis. He and Yilmaz [10] conducted a field study during which they collected data of 21 participants for 5 days to investigate the relationship between user behavior and task characteristics in Web search. Our study involved 30 participants for one month, mainly focusing on the influence of user behavior and task characteristics on user satisfaction. Wu et al. [31] conducted a study quite similar to ours, while their study was prepared specifically for image search scenario. To emphasize on users' examination process and cognitive process in search sessions, we collected additional feedbacks from users about the query reformulation process, which were not considered in previous field studies.

4.1 Procedure

The procedure of our field study consisted of three stages, following [10] and [31].

Introduction Stage. In this stage, we instructed the participants about the requirements of our field study. They were invited to our laboratory with their own laptops so that we can install a browser extension on their laptops. This browser extension could record their daily Web search activities, while the participants can opt to turn it on and off anytime. They were asked to fill in a pre-experiment questionnaire to collect demographic information and sign an agreement showing that they consented to data collection in our field study. After a detailed introduction of our study procedure, two training sessions were provided to make the participants familiar with our experimental platform and understand some concepts better in this study. After the training process, they were told to use their laptops to search for daily purposes as usual.

Data Collection Stage. This stage lasted for about one month. With the permission of the participants, their daily Web search activities would be recorded automatically. They were required to review their past search queries and provide feedbacks at their convenience. However, to ensure that the participants had distinct impressions on their search events, each query and its corresponding log entries would be reserved for at most two days if not being reviewed in time. When the participants reviewed their search queries, they were allowed to remove any queries that they were

Table 1: Search feedback information collected in the field study.

	Attribute	Description	Value
Task	Background	Please describe the time, location and your intention for this search task.	open-ended question
	Satisfaction	Were you satisfied with the process of searching for completing this task?	(0) unsatisfied - (4) very satisfied
	Success	How much useful information have you found for this task?	(0) not any - (4) all you want
	Difficulty	How do you feel about the difficulty of this task to find relevant information?	(0) easy - (4) extremely difficult
Query	Expectation	What information did you expect to find for this query?	open-ended question
	Relation	What is the relation between this query B and the last query A?	(0) initial query; (1) substitute/(2) add/(3) delete terms for the same topic; (4) B is a subtopic of A; (5) B and A are two subtopics of a same topic; (6) B is a new topic related to A.
	Satisfaction	Were you satisfied with the search results for this query?	(0) unsatisfied - (4) very satisfied
	Reason for Leaving	Why do you reformulate this query or end your search?	(A) have found enough information; (B) come up with a better query; (C) cannot find useful information; (D) other reason___
Result	Usefulness	For each result, how do you rate its usefulness for completing this search task?	(0) useless - (4) highly useful

not willing to share with us. Once a query was removed by the participants, all related log entries would be removed from the recorded data collection automatically. Therefore, allowing the participants to remove search logs would not have impact on sessions log completeness.

Summarization Stage. After one month for data collection, the participants were informed about the finishing of this field study. They were paid according to their contributions: about $5 for participating our field study and $0.15 for each valid search query log they provided. Finally, we collected their experiences and suggestions for our field study with a post-experiment questionnaire. Most participants were satisfied with the design of our field study and felt free for providing search logs and feedbacks because they were allowed to remove any logs that they were not willing to share.

4.2 Search Behavior Log

To record participants' daily Web search activities, we developed a Chrome extension which could record related information when some specific events were triggered by browser operations. This extension could be installed on different kinds of chrome-based Web browsers, which the participants were familiar to use.

When the participants started to search with a general Web search engine, the queries they issued were recorded. These events could be triggered by issuing a query or clicking query suggestions. Besides the queries, we collected abundant information for each page the participants browsed, including search engine result pages (SERPs) and landing pages. Specifically, we recorded the URLs and HTML contents of the pages. We also recorded participants' mouse activities such as mouse movement, scrolling, and click events. In addition, each search log was associated with a timestamp, from which we could get the dwell time on the SERPs and landing pages.

4.3 Search Feedback

To collect user feedbacks for their search sessions, we also developed an annotation platform for the participants to review their search logs and provide feedbacks. To submit valid feedback for a search task, the participant needed to go through the following steps.

Review Logs. In this step, participants' search logs were shown with issued queries in the log review panel. They could review the SERPs and click events by clicking the corresponding queries. If there were some queries they would not like to share with us, they could remove them freely.

Identity Search Tasks. To provide feedbacks for search tasks, the participants first needed to identify which queries belong to the same task. Different from previous study [15], we required the participants to segment the whole query logs into search tasks themselves, rather than identify task boundaries with a 30-minutes gap between two logged activities. We believe that tasks identified by themselves will consist of queries related to the same information needs. In this paper, note that "search task" is referred to as "search session" and we will use them interchangeably, depending on the context.

Collecting Detailed Feedbacks. After identifying search tasks, the participants should provide feedback for each task. The queries that participants identified for each task, as well as the corresponding SERPs and click events, would be shown to participants. As shown in Table 1, we collected both task-level, query-level and result-level feedbacks from the participants. Compared to previous field study [31], we focused more on users' query-level feedbacks. For each query, we asked the participants to describe information they expected to find and the reason why they stopped searching with this query. Further, they should choose from several possible options for the relationship between the current query and the previous one, based on whether these queries belong to the same subtopic. Finally, we also collected user satisfaction feedback for each query. With these query-level feedbacks from users, we can better understand how users' search states and intents change from one query to another query and how they feel about those changes.

4.4 Participants and Collected Data

To recruit participants who are familiar with the usage of search engines and search for daily purposes frequently, we first posted a related questionnaire on social network platforms. Then we invited 30 participants, 13 females and 17 males, to take part in our field study. The ages of participants range from 18 to 41. These participants include 13 undergraduates, 10 graduates and 7 employees coming from different universities and companies.

Through the field study, after filtering some invalid search sessions (part of user behaviors or page contents were not recorded successfully), we constructed a dataset which contains 1,169 sessions and 3,875 queries in total. For each search session, the participants on average submitted 3.3 queries and clicked 3.1 unique

results. All the collected data used in this paper is available online for academic research.[2]

5 EXPERIMENTS

We conduct a series of experiments to investigate the effectiveness of RSMs compared with existing session-based metrics. In particular, we try to answer the following three research questions:

- **RQ1:** How do our proposed RSMs perform compared to sDCG and sRBP?
- **RQ2:** How is the performance of RSMs affected by different settings of model parameters of the metrics?
- **RQ3:** What is the influence of cognitive factors on performance of RSMs?

5.1 Datasets

As shown in Section 4, we construct a field study based session search dataset. Besides this self-constructed dataset, we also use a public available search behavior dataset from an existing user study [14]. Taking user satisfaction as the ground truth, we can compare the performance of different session-based metrics by computing the correlation between metrics and user satisfaction. For the field study dataset, we find that participants examined more than one SERP for only 60 (about 1.5%) queries. For simplicity, we filter out these queries and their corresponding sessions in our experiments. In addition, for the user study dataset [14], we find that there are no results returned for the first two queries in session #22. Therefore, we also remove this session in our experiments. To summarize, Table 2 shows the statistics of these two datasets we utilize in our experiments.

5.2 Instantiations of RSMs

In Section 3, we propose the framework of Recency-aware Session-based Metrics (RSMs). In our experiments, we drive two instantiations of RSMs based on sDCG and sRBP, given definitions of g, d and mem in Equation 4.

Gain. Following previous works, we measure the gain of a result based on its relevance or usefulness. Mao et al. [23] have concluded that the metrics based on usefulness have a better correlation with user satisfaction compared to relevance. For the user study dataset, we map a 3-level graded relevance score $r_{m,n}$ to a gain in {0.0, 0.5, 1.0} similar to [1]. While for our field study dataset, we map a 5-level graded usefulness score $r_{m,n}$, which is collected from participants rather than external assessors, to a gain in {0.0, 0.25, 0.5, 0.75, 1.0}.

Discount. Our instantiations of RSMs in this paper are based on sDCG and sRBP, so we use their discount functions shown in Equation 3.

Memory. To incorporate the recency effect, we consider the decay of users' memory of results in utility accumulation model of metrics. In this paper, we assume that once a user issues a new query to a system, his/her impressions of previous queries will decay. Therefore, users' memory of a query is related to the distance between this query and the last query, which can be expressed as follows:

$$mem_m = FF(M-m) = e^{-\lambda(M-m)} \tag{5}$$

[2] http://www.thuir.cn/tiangong-ss-fsd/

Table 2: Statistics of the datasets used in our experiments.

	Field Study	User Study [14]
# sessions	1,124	79
# queries	3,535	383
# results per SERP	~10	9
language	Chinese	English

where FF denotes the forgetting function that decay the memory with the sequence of queries. Following a recent work [16] which models the recency effect in recommender systems, we adopt an exponential function to model the forgetting function in this paper. λ is a parameter that reflects the rate at which users forget. If $\lambda = 0$, $mem_m = 1$, which means that a user will remember all the results he/she has encountered in previous queries. The RSMs will then return to their original forms which do not take the recency effect into consideration. Discussion about the design of better forgetting functions is not the focus of this paper and left for future work.

Given the above definitions of three components of RSMs, we can derive two instantiations of RSMs as follows:

$$\begin{aligned} RS-DCG &= \sum_{m=1}^{M} e^{-\lambda(M-m)} \sum_{n=1}^{N} g(r_{m,n}, q_m) \cdot d_{m,n}(sDCG) \\ RS-RBP &= \sum_{m=1}^{M} e^{-\lambda(M-m)} \sum_{n=1}^{N} g(r_{m,n}, q_m) \cdot d_{m,n}(sRBP) \end{aligned} \tag{6}$$

where $g(r_{m,n}, q_m)$ maps the relevance or usefulness score of the n-th result in q_m to a gain.

5.3 Baseline Metrics

To show the influence of incorporating recency effect, we use sDCG and sRBP as baselines. Metrics like ERR and MAP are not taken into account because they are query-level metrics. In addition, sDCG/q and Last-DCG are two metrics which perform best in [13]. sDCG/q takes the cost factor into account, while Last-DCG only considers DCG score of the last query. We also compare our RS-DCG and RS-RBP with these two metrics. Further, compared with the recency effect, there is a competing hypothesis that the best query in a session determines users' perceived satisfaction. Therefore, Best-DCG and Best-RBP are also compared as baselines.

6 RESULTS

6.1 The Overall Performance of RSMs

We first compare the performance of RSMs with respect to baseline metrics (**RQ1**) by computing their Spearman's correlations with user satisfaction on both two datasets shown in Table 2.

Note that all the above metrics have parameters. Among these parameters, b_r and b_q are the logarithm base parameters for the rank and query discounts in the user browsing model of sDCG. Similarly, b and p are the balance and persistence parameters of sRBP. Besides, RS-DCG and RS-RBP have a new parameter λ to control the rate of forgetting.

As discussed in [17], the user browsing model behind an evaluation metric should provide an accurate prediction of actual user

behavior. Following their work, we first try to find the optimal values of parameters related to user browsing models, which are b_r and b_q for DCG-based metrics, and b and p for RBP-based metrics, respectively. Based on the discount functions of metrics, $d_{m,n}(sDCG)$ and $d_{m,n}(sRBP)$, we can compute the probabilities that results of different ranks and queries are examined by users. Meanwhile, assuming that all the results with higher ranks than the last clicked position are examined by users [1, 17], we can compute the observed probabilities that users examine results and issue queries based on the observed sessions. To compare the probabilities of examining results derived from the user browsing models of metrics and the observed user behavior, we compute the Total Squared Error (TSE) of probabilities over all the ranks and queries. We perform a grid search on the TSE measure with step of 0.1 to find the optimal values. For DCG-based metrics, we search the values of b_r and b_q in range (1.0, 5.0][3]. For RBP-based metrics, the values of b and p are searched from 0 to 1.

After finding the best values of parameters related to user browsing models, metrics without the recency effect are determined by these optimal parameters. For RS-DCG and RS-RBP, they have an additional parameter λ. To tune this parameter, we also perform a grid search for λ to optimize the Spearman's correlation between the values of metrics and the feedbacks of user satisfaction in search sessions. We search the values for λ in range [0, 5] with step of 0.1. Considering the fairness of comparison with the baselines, besides optimizing baseline metrics to fit user behavior, we also compare the performance of baseline metrics by optimizing them to fit user satisfaction. We adopt a 5-fold cross-validation method. Each time one fold of data is retained for comparing the performance of different metrics, and the other four folds of data are used to find the optimal value of parameters. We repeat 5-fold cross-validation method ten times by randomly split the data into five folds and report the average Spearman's correlation between each metric and user satisfaction.

The results of Spearman's correlations between different metrics and user satisfaction on both two datasets are shown in Table 3. Note that for the field study dataset, the statistical tests of significance (two-tailed Student's t-test) are based on repeating 5-fold cross-validation method ten times. We report the average Spearman's correlations over all the folds in repetitions. However, for the user study dataset which involves only 79 sessions, we do not apply a cross-validation method and significance tests on this dataset. The parameters are tuned over all sessions in the user study dataset and we only report the final results.

From table 3 we can see that RS-DCG performs better than baseline metrics in both two datasets, no matter baseline metrics are optimized to fit user behavior or user satisfaction. However, RS-RBP does not perform as well as sRBP/q in field study dataset or Last-RBP in user study dataset when baseline metrics are optimized to fit user satisfaction. Although RS-RBP does not correlates best with user satisfaction, we should note that the primary strength of RSMs compared with baseline metrics is bridging the gap between user examination model and user satisfaction. For example, if parameters of sRBP/q are optimized to fit user satisfaction, the average TSE between examination probabilities of user model and observed user behavior is 0.447, which is significantly larger than that of RS-RBP (TSE = 0.112). It means that sRBP/q cannot well characterize user behavior and user satisfaction simultaneously. Considering RS-RBP also performs well in satisfaction (Spearman's correlation is 0.530), we think RSMs can characterize user behavior and user satisfaction simultaneously by incorporating the recency effect into utility accumulation model of metrics. We also find that the recency effect significantly performs better than the hypothesis that the best query in a session determines user satisfaction (Best-DCG and Best-RBP). In addition, although parameter λ is optimized by using user satisfaction in our experiments, further discussion (see Section 6.2 and 6.3) show that λ is relatively stable and can be roughly estimated by some task attributes without user behavior. It makes RSMs more applicable than baseline metrics since it is more difficult to collect user satisfaction than user behavior.

[3]We also tried more values larger than 5, but the results showed that they did not perform as well as the values in (1.0, 5.0].

Table 3: Spearman's correlations between session-based metrics and user satisfaction on both two datasets. UB means User Behavior while US means User Satisfaction. Bold font indicates the strongest correlation in each block. * indicates the difference is significant at p < 0.001 level with a Bonferroni correction involving all tests performed on the same dataset, comparing to the RSM in each block.**

	Field Study		User Study	
	fit UB	fit US	fit UB	fit US
sDCG	0.262***	0.321***	0.019	0.221
sDCG/q	0.519***	0.535***	0.305	0.343
Last-DCG	0.458***	0.456***	0.332	0.340
Best-DCG	0.323***	0.359***	0.222	0.229
RS-DCG	**0.548**		**0.356**	
sRBP	0.215***	0.392***	0.081	0.238
sRBP/q	0.530	**0.538**	0.323	0.346
Last-RBP	0.455***	0.458***	0.371	**0.372**
Best-RBP	0.287***	0.374***	0.260	0.260
RS-RBP	0.530		0.345	

6.2 The Effects of Parameters

To investigate how the performance of RSMs are affected by different settings of parameters (**RQ2**), we make a sensitivity analysis of RSMs. First we investigate to what extent the parameters are stable when we apply 5-fold cross-validation method on the field study dataset. Table 4 shows the mean and standard deviation of each parameter optimized on each fold (5-fold cross-validation with ten repetitions). From this table, we can find that the variances of parameters describing user browsing models (b_r and b_q for RS-DCG, b and p for RS-RBP) on different folds are very small. It seems that the parameter λ has a larger standard deviation ($\sigma = 0.35$ for RS-DCG and $\sigma = 0.42$ for RS-RBP). We plot curves of the exponential forgetting function with different values of λ in Figure 2. We can see that when λ is large (e.g. $\lambda \geq 1$), the values of forgetting functions with different parameters are very close. It suggests that the standard deviation of λ in Table 4 is relatively small considering the variation of the forgetting function.

Table 4: The mean and standard deviation of each parameter optimized on each fold.

	RS-DCG			RS-RBP		
	b_r	b_q	λ	b	p	λ
Mean	1.30	1.30	2.03	0.60	0.80	1.97
σ	0.00	0.00	0.35	0.00	0.00	0.42

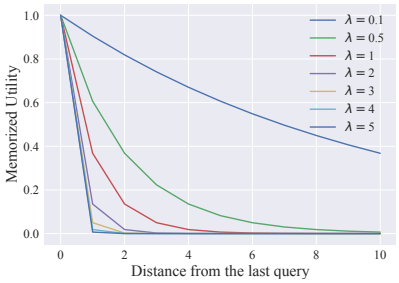

Figure 2: Curves of the exponential forgetting function with different values of λ.

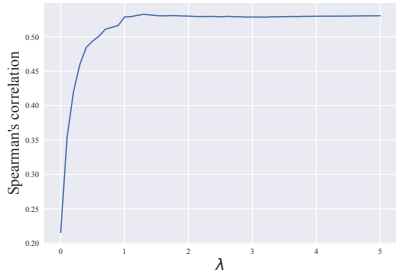

Figure 3: Sensitivity analysis of parameter λ for RS-RBP.

To further analyze how the performance of RSMs changes with the value of λ, especially in the range where the optimal values of λ on different folds lie, we make a sensitivity analysis of RSMs. Take RS-RBP as an example, we show the sensitivity of parameter λ of RS-RBP by fixing both b and p to their optimal values. We depict the Spearman's correlation between RS-RBP and user satisfaction when considering different values of λ in Figure 3. It shows that Spearman's correlation between RS-RBP and user satisfaction increases significantly in the beginning as λ increases from 0 to 1. However, when λ is larger than 1, the Spearman's correlation becomes stable, which may be caused by the small change of the forgetting function shown in Figure 2.

In summary, the sensitivity analysis based on Table 4 and Figure 3 indicates the stability of the learned parameters of RSMs.

6.3 The Influence of Cognitive Factors

Finally, we analyze the influence of two cognitive factors, task difficulty and complexity, on the model parameters and performance of RSMs (**RQ3**). In our experiments, task difficulty describes users' feeling about the difficulty to find relevant information, which was collected as a 5-level graded feedback (see Table 1: Difficulty) from users in our field study. We denote tasks with difficulty score of 0 or 1 as Low-Difficulty (**LD**) tasks, while the remaining tasks are denoted as High-Difficulty (**HD**) tasks. In total, we get 804 LD task sessions and 320 HD task sessions from the field study dataset. Different from task difficulty, task complexity is related to activities and information sources required to complete the task [18]. To measure task complexity of a session, we check the relations of all the adjacent queries in this session, which were collected as feedbacks from users given seven options (see Table 1:Relation) in the field study. If all the relations are selected from {0, 1, 2, 3}, which means that there is only one topic/subtopic through the whole search session, the session is called a Low-Complexity (**LC**) task session. In contrast, once there is a relation between two adjacent queries selected from {4, 5, 6}, which means that this session involves multiple topics/subtopics, the session containing these queries is denoted as a High-Complexity (**HC**) task session. Through this measurement of task complexity, finally we get 667 LC task sessions and 457 HC task sessions from the field study dataset. Since the user study dataset do not contain these abundant information such as the relation between adjacent queries, we conduct experiments for **RQ3** only on the field study dataset.

First we compare the performance of metrics based on tasks with different difficulties and complexities by performing similar correlation analysis as done for **RQ1**. Table 5 displays the Spearman's correlations between session-based metrics and user satisfaction on tasks with different difficulties and complexities of the field study dataset. Note that here the baseline metrics are optimized to fit user satisfaction. The results reveal some interesting findings:

For LD and LC tasks, RS-DCG and RS-RBP have the strongest correlations with user satisfaction compared with baseline metrics. While for HD and HC tasks, sRBP/q performs better than RS-RBP. There are some possible reasons for the results. For task difficulty, on the one hand, when users search for HD tasks, they are more focused, thus have deeper impressions for what they have encountered in search sessions. On the other hand, it may be difficult for users to issue an effective query at the beginning of HD tasks. They have to learn how to reformulate more effective queries in the first few query rounds within a session. Therefore, the contributions of previous queries are as important as subsequent queries for users. For task complexity, when a task is complex, it involves multiple topics or subtopics, which makes the number of queries become an important factor to evaluate the performance. Given the above reasons, it is intuitive that sRBP/q performs best in HD and HC tasks because it takes the cost factor into consideration. In addition, here sRBP/q is optimized to fit user satisfaction. As we have discussed in Section 6.1, it cannot well characterize user behavior and user satisfaction simultaneously. We can also find that RS-DCG performs better than baseline metrics in all situations. In general, our proposed RSMs are robust and perform well with the strength of characterizing user behavior and user satisfaction simultaneously.

Then we compare the optimal parameters of RS-DCG and RS-RBP learned on different tasks to show the influence of task difficulty and complexity. We assume that task difficulty and complexity not only change users' browsing behavior, but also affect users'

Table 5: Spearman's correlations between session-based metrics and user satisfaction on tasks with different difficulties and complexities of the field study dataset. Bold font indicates the strongest correlation of its column in each block. * indicates the difference is significant at p < 0.001 level with a Bonferroni correction involving all tests performed on the same dataset, comparing to the RSM in each block.**

	LD	HD	LC	HC
sDCG	0.208***	0.522***	0.435***	0.252***
sDCG/q	0.364***	0.552***	0.516***	0.453
Last-DCG	0.336***	0.506***	0.464***	0.397***
Best-DCG	0.263***	0.522***	0.424***	0.331***
RS-DCG	**0.391**	**0.574**	**0.540**	**0.459**
sRBP	0.297***	0.536***	0.446***	0.297***
sRBP/q	0.373	**0.598**	0.525	**0.455*****
Last-RBP	0.349***	0.498***	0.468***	0.403
Best-RBP	0.318***	0.538	0.405***	0.340***
RS-RBP	**0.381**	0.576	**0.528**	0.418

Table 6: The mean of each parameter optimized on each fold of tasks with different difficulties and complexities.

	RS-DCG			RS-RBP		
	b_r	b_q	λ	b	p	λ
LD	1.30	1.20	2.45	0.60	0.70	1.11
HD	1.33	1.62	0.16	0.60	0.90	0.15
LC	1.30	1.11	2.33	0.76	0.60	2.11
HC	1.31	1.79	0.94	0.54	0.90	1.13

memory of their searching process. The results are shown in Table 6. In terms of the examination parameters of user browsing models, the probabilities that users examine the next result (described by b_r for RS-DCG and $\alpha = b \cdot p$ for RS-RBP) in different tasks are quite similar. However, in tasks of high difficulty or high complexity, users are more likely to issue a new query (described by b_q for RS-DCG and $\beta = (1 - b) \cdot p$ for RS-RBP). It requires more queries for users to find useful information in HD tasks or cover multiple topics/subtopics in HC tasks. As for the parameter λ of forgetting functions, we can also find a difference between different tasks. Referring to task difficulty, λ is large for LD tasks while small for HD tasks. For task complexity, we can also find that λ is large for LC tasks while small for HC tasks. Note that λ is a parameter that reflects the rate at which users forget. The different values of λ on different tasks indicate that the influence of the recency effect is affected by cognitive factors. For HD and HC tasks, users pay more attention to previous queries in a session probably because cognitive load is larger in HD and HC tasks for users.

To summarize, the analysis in this section indicates that users' examination and cognitive behaviors vary among tasks with different difficulties and complexities, which means we should use different model parameters for different types of search tasks.

7 CONCLUSION

Constructing session-based evaluation metrics is essential for evaluating and improving the performance of search engines in a realistic search scenario. In this paper, we propose novel session-based metrics, Recency-aware Session-based Metrics (RSMs), by incorporating the recency effect to characterize users' cognitive process in search sessions. Regarding user satisfaction for a search session as the golden standard in search performance evaluation, we show the effectiveness of RSMs based on both self-constructed and public available search user behavior datasets. The self-constructed dataset comes from a field study where we collect daily search logs and explicit feedbacks of 30 participants for one month. Comparing the correlations between different session-based metrics and user satisfaction, we find that our proposed RSMs generally have stronger correlations with user satisfaction than existing metrics when characterizing user behavior and user satisfaction simultaneously. We further make a sensitivity analysis of RSMs parameters. The results indicate our proposed RSMs are stable in terms of the learned parameters. Considering the features of tasks, we investigate the influence of task difficulty and task complexity on the performance and model parameters of RSMs. Although the cost factor is important to evaluate user satisfaction for HD tasks and HC tasks, we find that our proposed RSMs have stronger and robust correlations than existing metrics with user satisfaction across different tasks. It is also found that users' examination and cognitive behaviors vary among tasks with different difficulties and complexities. For examination behaviors, users are more likely to issue a new query in HD tasks and HC tasks. In spite of more queries issued in HD tasks and HC tasks, for cognitive behaviors, users focus more on previous queries in HD tasks and HC tasks. In particular, what users have learnt in previous queries makes a contribution to their more efficient search process in subsequent queries. These results suggest that different parameters should be considered for different types of search tasks.

Our work suggests the importance and effectiveness of incorporating cognitive effects besides examination process or browsing process in search evaluation. As we only incorporate the forgetting function to model the recency effect for sRBP and sDCG in this study, the framework can be further adopted to augment other evaluation metrics. Furthermore, it can be extended to more sophisticated search scenarios. For example, taking the influence of memory on assessment of utility of results into consideration, we can change the definition of $g_{m,n}$ in Equation 4. We can also design better functions for $d_{m,n}$ to model how the user browsing model is affected by user's assessment of utility and user's memory.

Nevertheless, our work has a few limitations which we would like to address for the future work. (1) We incorporate the recency effect with forgetting functions into the design of session-based evaluation metrics. In the future work, we plan to explore more sophisticated functions to model the recency effect. It is also essential to consider some other cognitive effects. For example, Tversky and Kahneman [30] studied the anchoring effect, which indicated that assessments made by participants were influenced by the standard of an initial comparative judgment. Similarly, in a search session, the initial query may become an anchor, thus affecting the user's perception of the subsequent queries. (2) Our RSMs do not consider

information nuggets or subtopics, which are utilized in Expected Utility [33] and Cube Test [22]. Given that task complexity has an influence on the performance and parameters of RSMs, we would like to enhance RSMs by inspecting users' cognitive process from a fine-grained perspective in the future. The recency effect may have different influence on users' impression of the results and queries related to different subtopics respectively. We can consider subtopics for RSMs and compare them to Expected Utility and Cube Test. (3) We focus on the correlation between metrics and user satisfaction on a single search engine in this work. To further validate the effectiveness of our metrics, we will compare the performance of different search systems with RSMs. The evaluation results can be further compared with preference tests of different systems. To achieve this goal, we think future research needs to take the dynamics of session search (users may reformulate different queries given different results returned by different systems) into consideration. Building a simulation-based evaluation framework [5, 36] is a possible way to address this problem.

REFERENCES

[1] Leif Azzopardi, Paul Thomas, and Nick Craswell. 2018. Measuring the utility of search engine result pages: an information foraging based measure. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 605–614.

[2] AD Baddeley. 1968. Prior recall of newly learned items and the recency effect in free recall. *Canadian Journal of Psychology/Revue canadienne de psychologie* 22, 3 (1968), 157.

[3] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2015. User variability and IR system evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 625–634.

[4] Ben Carterette. 2011. System effectiveness, user models, and user utility: a conceptual framework for investigation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 903–912.

[5] Ben Carterette, Ashraf Bah, and Mustafa Zengin. 2015. Dynamic test collections for retrieval evaluation. In *Proceedings of the 2015 international conference on the theory of information retrieval*. ACM, 91–100.

[6] Ben Carterette, Evangelos Kanoulas, Mark Hall, and Paul Clough. 2014. *Overview of the TREC 2014 session track.* Technical Report. DELAWARE UNIV NEWARK DEPT OF COMPUTER AND INFORMATION SCIENCES.

[7] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 621–630.

[8] Cyril Cleverdon, Jack Mills, and Michael Keen. 1966. ASLIB Cranfield Research Project: factors determining the performance of indexing systems. (1966).

[9] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*. ACM, 87–94.

[10] Jiyin He and Emine Yilmaz. 2017. User behaviour and task characteristics: A field study of daily information behaviour. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. ACM, 67–76.

[11] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[12] Kalervo Järvelin, Susan L Price, Lois ML Delcambre, and Marianne Lykke Nielsen. 2008. Discounted cumulated gain based evaluation of multiple-query IR sessions. In *European Conference on Information Retrieval*. Springer, 4–15.

[13] Jiepu Jiang and James Allan. 2016. Correlation between system and user metrics in a session. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*. ACM, 285–288.

[14] 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 *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 607–616.

[15] Rosie Jones and Kristina Lisa Klinkner. 2008. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 699–708.

[16] Santiago Larrain, Christoph Trattner, Denis Parra, Eduardo Graells-Garrido, and Kjetil Nørvåg. 2015. Good times bad times: A study on recency effects in collaborative filtering for social tagging. In *Proceedings of the 9th ACM Conference on Recommender Systems*. ACM, 269–272.

[17] Aldo Lipani, Ben Carterette, and Emine Yilmaz. 2019. From a User Model for Query Sessions to Session Rank Biased Precision (sRBP). In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*. ACM, 109–116.

[18] Jingjing Liu, Michael J Cole, Chang Liu, Ralf Bierig, Jacek Gwizdka, Nicholas J Belkin, Jun Zhang, and Xiangmin Zhang. 2010. Search behaviors in different task types. In *Proceedings of the 10th annual joint conference on Digital libraries*. ACM, 69–78.

[19] Mengyang Liu, Yiqun Liu, Jiaxin Mao, Cheng Luo, and Shaoping Ma. 2018. Towards designing better session search evaluation metrics. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 1121–1124.

[20] Mengyang Liu, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Investigating Cognitive Effects in Session-level Search User Satisfaction. KDD.

[21] Cheng Luo, Yiqun Liu, Tetsuya Sakai, Fan Zhang, Min Zhang, and Shaoping Ma. 2017. Evaluating mobile search with height-biased gain. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 435–444.

[22] Jiyun Luo, Christopher Wing, Hui Yang, and Marti Hearst. 2013. The water filling model and the cube test: multi-dimensional evaluation for professional search. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, 709–714.

[23] 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 *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 463–472.

[24] Alistair Moffat, Paul Thomas, and Falk Scholer. 2013. Users versus models: What observation tells us about effectiveness metrics. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, 659–668.

[25] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)* 27, 1 (2008), 2.

[26] Tetsuya Sakai and Zhicheng Dou. 2013. Summaries, ranked retrieval and sessions: A unified framework for information access evaluation. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 473–482.

[27] Mark Sanderson et al. 2010. Test collection based evaluation of information retrieval systems. *Foundations and Trends® in Information Retrieval* 4, 4 (2010), 247–375.

[28] Mark D Smucker and Charles LA Clarke. 2012. Time-based calibration of effectiveness measures. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 95–104.

[29] Zhiwen Tang and Grace Hui Yang. 2017. Investigating per topic upper bound for session search evaluation. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. ACM, 185–192.

[30] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *science* 185, 4157 (1974), 1124–1131.

[31] Zhijing Wu, Yiqun Liu, Qianfan Zhang, Kailu Wu, Min Zhang, and Shaoping Ma. 2019. The influence of image search intents on user behavior and satisfaction. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 645–653.

[32] Grace Hui Yang and Ian Soboroff. 2016. TREC 2016 Dynamic Domain Track Overview.. In *TREC*.

[33] Yiming Yang and Abhimanyu Lad. 2009. Modeling expected utility of multi-session information distillation. In *Conference on the Theory of Information Retrieval*. Springer, 164–175.

[34] Emine Yilmaz, Milad Shokouhi, Nick Craswell, and Stephen Robertson. 2010. Expected browsing utility for web search evaluation. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 1561–1564.

[35] Fan Zhang, Yiqun Liu, Xin Li, Min Zhang, Yinghui Xu, and Shaoping Ma. 2017. Evaluating web search with a bejeweled player model. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 425–434.

[36] Yinan Zhang, Xueqing Liu, and ChengXiang Zhai. 2017. Information retrieval evaluation as search simulation: A general formal framework for ir evaluation. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. ACM, 193–200.

[37] Yuye Zhang, Laurence AF Park, and Alistair Moffat. 2010. Click-based evidence for decaying weight distributions in search effectiveness metrics. *Information Retrieval* 13, 1 (2010), 46–69.