Investigating Cognitive Effects in Session-level Search User Satisfaction

Mengyang Liu, Jiaxin Mao, Yiqun Liu*, 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 100084, China
yiqunliu@tsinghua.edu.cn

ABSTRACT

User satisfaction is an important variable in Web search evaluation studies and has received more and more attention in recent years. Many studies regard user satisfaction as the ground truth for designing better evaluation metrics. However, most of the existing studies focus on designing Cranfield-like evaluation metrics to reflect user satisfaction at query-level. As information need becomes more and more complex, users often need multiple queries and multi-round search interactions to complete a search task (e.g. exploratory search). In those cases, how to characterize the user's satisfaction during a search session still remains to be investigated. In this paper, we collect a dataset through a laboratory study in which users need to complete some complex search tasks. With the help of hierarchical linear models (HLM), we try to reveal how user's query-level and session-level satisfaction are affected by different cognitive effects. A number of interesting findings are made. At query level, we found that although the relevance of top-ranked documents have important impacts (primacy effect), the average/maximum of perceived usefulness of clicked documents is a much better sign of user satisfaction. At session level, perceived satisfaction for a particular query is also affected by the other queries in the same session (anchor effect or expectation effect). We also found that session-level satisfaction correlates mostly with the last query in the session (recency effect). The findings will help us design better session-level user behavior models and corresponding evaluation metrics.

KEYWORDS

Session Search; User Satisfaction; Search Evaluation

ACM Reference Format:
Mengyang Liu, Jiaxin Mao, Yiqun Liu*, Min Zhang, Shaoping Ma. 2019. Investigating Cognitive Effects in Session-level Search User Satisfaction. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292500.3330981

*Corresponding author.

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08.
https://doi.org/10.1145/3292500.3330981

1 INTRODUCTION

Search evaluation is one of the major concerns in information retrieval (IR) studies. The traditional search evaluation method, referred to as Cranfield paradigm, plays an important role in the development of a large number of IR systems. Most existing evaluation metrics (e.g., RBP [27], ERR [5], etc.) are designed for the result lists of a single query. Previous studies [1, 25] have showed that these query-level evaluation metrics have a strong correlation with users' satisfaction. However, while search request becomes more and more complex, there are many scenarios in which multiple queries and multi-round search interactions are needed (e.g. exploratory search). Under this scenario, the evaluation metric designed for single-query interactions may be not enough to reflect users' satisfaction in those complex sessions. Therefore, how to design better session-level evaluation metrics has received more and more attention in recent years.

Existing query-level evaluation metrics are designed based on the cascade hypothesis [6] which assumes that the users view search results from top to bottom and their attention will gradually decay. Many existing session-level evaluation metrics (e.g. Session-based DCG (sDCG) [14] and Expected Utility (EU) [34]) also follows the same paradigm. Although the cascade hypothesis makes sense for the evaluation of a single query, whether it is valid for session-level evaluation still remains under-investigated. Some previous studies [14, 21] suggested that the performance of the last query or the average performance of all queries within a session have a stronger correlation with the session performance.

To design a better session-level evaluation metrics, we need to consider different *cognitive effects*, that may influence user's satisfaction with the whole search sessions, and properly incorporate them into the evaluation metric. For example, the metrics with a decaying weighting function emphasize the *primacy effect* [32] that the initial documents examined by users are more influential for their satisfaction. On the other hand, the metrics with an increasing weighting function emphasize the *recency effect* [3] that the last-examined documents have a greater impact on their satisfaction. In addition, queries are not completely independent of each other in a session, therefore, we should also consider the impact of the previous issued query when evaluating the latter query. Specifically, the initial query may become an anchor (*anchoring effect* [29]) or generate additional expectation (*expectation effect* [4]) for the user's perception of the subsequent queries.

So in this study, we investigate the interaction effect between queries and whether the primacy effect or the recency effect is more important for query-level and session-level evaluation. Particularly, we try to answer the following three research questions:

- **RQ1** How does the cascade assumption perform in characterizing query-level and session-level satisfaction?
- **RQ2** How query-level satisfaction is affected by other queries in the same session?
- **RQ3** How to design better session-level evaluation metrics with the findings in RQ1 and RQ2.

To answer these research questions, we conducted a laboratory user study to construct a dataset containing 675 search sessions of complex search tasks. In these tasks, we collected interaction logs and explicit satisfaction feedback from users. We also collected the relevance annotations of documents from a third-party crowdsourcing platform. With this constructed dataset[1], we investigated the cognitive effects that may determines user's session-level satisfaction and how to design better session-level evaluation metrics.

The remainder of this paper is organized as follows. Section 2 reviews some related work. Section 3 describes the experimental settings of user study and data annotation method. In Section 4 and Section 5, we present data analysis to address **RQ1** and **RQ2**. Regarding **RQ3**, in Section6 we propose a framework in which new session-level evaluation metrics can be defined. Finally, we give our discussions and conclusions in Section 7.

2 RELATED WORK

2.1 Search evaluation

Evaluation is one of the most important research problems in the field of information retrieval (IR) related studies. The traditional search evaluation method, referred to as Cranfield paradigm, is mainly based on corpus, fixed queryset, relevance judgment of "query-document" pairs, and evaluation metrics.

2.1.1 Query-level evaluation. A lot of query-level evaluation metrics have been proposed based on different insights about users' search behavior, such as Normalized Discounted Cumulative Gain (NDCG) [13], Expected Reciprocal Rank (ERR) [5], Rank-biased Precision (RBP) [27], Time-biased Gain (TBG) [28] and etc.

Moffat et al. [26] concluded that there are user behavior models behind different evaluation metrics. For example, RBP [27] is a query-level metric which assumes that the users will browse from top to bottom and end the current browsing with a certain probability. Maskari et al. [1] showed that the evaluation metrics have a strong correlation with users' satisfaction, and a combination of measures can better evaluate the effectiveness of IR systems.

There are also some researchers arguing that the relevance annotation does not take into account the interaction between the results, which may lead to certain differences between the evaluation of the results and actual users' feelings. For example, Mao et al. [25] find that the measures based on usefulness rather than relevance annotation has a better correlation with user satisfaction. So that we would like to further investigate how to design new evaluation metrics based on these different measures (e.g., relevance, usefulness, etc.) to better characterize user satisfaction.

2.1.2 Session-level evaluation. As search tasks become more complex, users often submit multiple queries in one session. In this condition, query-level evaluation metrics will not be suitable for session-level evaluation. Some recent studies have focused on session-level evaluation methods and proposed some session-level evaluation metrics, such as Session-based DCG (sDCG) [14], Expected Utility (EU) [34], and Cube Test (CT) [24].

sDCG [14] is a extended version of the Discounted Cumulative Gain (DCG) [13], it assumes that the documents at lower position and retrieved by later query are less likely to be read by users, and therefore, have a weaker influence on session-level satisfaction. EU [34] takes into account the contribution of fine-grained information nuggets and the corresponding importance of each nugget. It also considers the novelty and efforts of the results. The gain of a result will be discounted if the same nugget has been encountered in previous results. Similar to the EU, CT [24] also takes into account the information nuggets and its importance. The gain of a point will get more discount if this point has appeared for multiple times.

However, these three metrics are built on the cascade hypothesis [6] which assumes that the user browses search results from top to bottom and the user's attention will gradually decay during the browsing process. Although this assumption holds for a single query, whether it is suitable for session-level evaluation has not been verified. The metrics that have a decaying weighting function imply that the initial documents and queries are more influential for the overall session-level satisfaction. However, whether this assumption is consistent with the real user's perception has not been verified by empirical studies. Therefore, we would like to investigate whether the cascade hypothesis still holds at session level and which kind of evaluation metric is more consistent with the behavior and satisfaction judgment of real users.

2.2 User satisfaction

Besides the traditional system-oriented evaluation methods, i.e. Cranfield paradigm, more and more studies begin to pay attention to user-oriented evaluation methods. Many studies [9, 12, 22] proposed methods to evaluate search engines from the perspective of real users. Researchers have focused on modeling users' subjective feelings with various document features (relevance, usefulness, etc.) [11, 21, 25] and users' implicit feedback signals (click, hover, scroll, etc.) [7, 8, 10].

User satisfaction is an important concept in this line of research, it measures users' subjective feelings about their interactions with the system and can be understood as the fulfillment of a specified information requirement [18]. Ali et al. [2] has mentioned that a more realistic evaluation of system performance can be made, if actual users can provide the explicit judgments.

The relationship between user satisfaction and user's behavior has been widely investigated. Kim et al. [19] found that the click-level satisfaction can be predicted with click dwell-time. Wang et al. [31] proposed a model in which user's action-level satisfaction was considered as a latent factor that affects the session-level satisfaction. Xu et al. [33] found that user's duration of completing a search task is negatively correlated with satisfaction. Liu et al. [23] extracted users' mouse movement information on search result pages and proposed an effective method to predict user satisfaction.

There are also many studies focusing on investigating the relationship between user satisfaction and search system's outcomes. Maskari et al. [1] found that user satisfaction is strongly correlated

[1] http://www.thuir.cn/KDD19-UserStudyDataset/

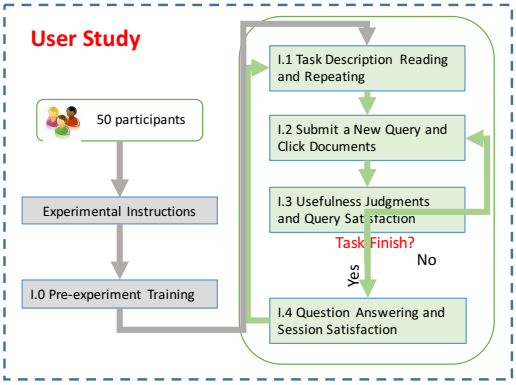

Figure 1: User study procedure.

with some evaluation metrics such as CG and DCG. Huffman and Hochster [11] found a strong correlation between session-level satisfaction and some simple relevance metrics. Jiang et al. [15] proposed the concept of graded search satisfaction and observed a strong correlation between satisfaction and average search outcome per effort. Jiang et al. [16] compared user's feedback in two experimental settings, an in situ one and a context-independent one. Mao et al. [25] found that users' usefulness feedback reflects users' satisfaction better than relevance. They compared a series of evaluation metrics based on user's click sequence, but they did not investigate whether these metrics were suitable for session evaluation.

In this study, we mainly focus on investigating the cognitive effects that influence users satisfaction in complex search task environments. More specifically, we are interested in what factors will affect the user's experiences with a query or a document, and how these fine-grained experiences contribute to session-level satisfaction. We believe addressing these questions can support the design a session-level evaluation metric that can better characterize user satisfaction.

3 DATA COLLECTION

To investigate cognitive effects that influence users' query-level and session-level satisfaction, we conducted a laboratory user study (see Figure 1). We collect the four kinds of measures in the user study: (1) User's click dwell time; (2) Document-level usefulness feedback; (3) Query-level satisfaction feedback; (4) Session-level satisfaction feedback. In addition, we conduct a crowdsourcing study to collected the query-document relevance of all documents.

3.1 Main user study

In our main user study, we recruited 50 undergraduate students via email and poster on campus. 24 participants were female, and the other 26 participants were male. The ages of participants ranged from 18 to 27. All the participants were familiar with the basic usage of web search engines and most of them used search engines every day. Each participant needed to complete 9 tasks which were selected from the topics of TREC Session Track. We made some modifications to the original TREC task descriptions so that these search tasks could satisfy the following criteria. First, the task should be easily interpreted by all participants so that they will have a clear search target. Second, the task should not be a trivial one, since we mainly focused on search sessions with multiple queries.

An experimental search engine system, which had a similar user interface as the commercial web search engine, was developed for the user study. The system had a common user interface and enabled the users to click multiple documents and reformulate the queries. There was no limit to the initial query so that the users could organize their query terms in the way they were used to. When users submited queries to this system, it would crawl corresponding results from a major Chinese commercial search engine. The organic results of each query would be stored in our system when the query was submitted for the first time so we could make sure that the participants submited the same query would see the same search engine results page (SERP). We injected a javascript plugin into the system to collect the users' search interactions including query reformulation, click, scrolling, tab switching, and mouse movement.

We made sure that each participant understood the experimental process through a pre-experiment training task. After the training stage, each participant was asked to perform 9 tasks in a random order. As shown in Figure 1, the main experiment consisted of four stages:

(I-1) In the first stage, the participant should read and memorize the task description on an initial page, and he/she was asked to repeat the task description without viewing it to ensure that he/she has remembered it.

(I-2) Next, the participant could submit a query and clicked on the results to collect information as they usually do with commercial search engines.

(I-3) After finishing the current query, she was asked to mark whether each document was useful for her at an evaluation page (0: not at all, 1: somewhat, 2: fairly, 3: very useful). He/she was also asked to give a 5-level graded satisfaction feedback on this query in this stage. If he/she wanted to find more information, he/she could go back to step (I-2) and submit a new query. He/she could end the search whenever he/she thought enough information had been found, or he/she could find no more useful information.

(I-4) Finally, the participant was required to give a search answer to a question related to the search task. Finally, the participant was further required to give an overall 5-level graded satisfaction feedback for the whole search session of the task.

3.2 Crowdsourcing annotation

Relevance assessment is very important in the field of IR evaluation, it is usually provided by human judges or annotators [30]. Crowdsourcing has been widely used for obtaining annotations for IR system development and evaluation [20]. In this work, we collected the relevance assessment of all the documents in our user study with a popular Chinese crowdsourcing platform.

During our crowdsourcing tasks, every crowd worker was provided with a "query-document" pair each time. The crowd workers needed to read the issued query, and the content of clicked documents. Then they were required to give the corresponding relevance score according to the following rating criteria [17]:

- **Rating 0:** The page is not relevant or a spam page.
- **Rating 1:** The page only provides minimal information about the query.

- **Rating 2:** The page provides substantial information about the query.
- **Rating 3:** The page is dedicated to the query, it is worthy of being a top result in a web search engine.

Through the crowdsourcing platform, we collected the relevance labels for 10,246 pages.

4 QUERY-LEVEL ANALYSIS

Since user satisfaction is users' subjective feelings about their interactions with the system, it may be influenced by a lot of factors. In this section, we first investigate how to characterize user's satisfaction at query-level. We use the hierarchical linear modeling (HLM) to fit user's query-level to investigate such effects. We also examine the relationship between users' satisfaction and a range of metrics at query-level. We find that the performance of these metrics will be different when they are computed based on different measures. In addition, we investigate the interaction effect of user satisfaction between adjacent queries in a session. Result shows that the users' perception of the initial query will have an impact on their satisfaction of subsequent query.

4.1 Modeling query-level satisfaction

In our user study, we collected users' query-level satisfaction after they finish each query. To investigate the impact of the sequential information of documents list on the user's query-level satisfaction, we use the hierarchical linear models (HLM) to fit the user's query-level satisfaction because it allows us to analyze the impact of document order and list length.

Intuitively, documents at different rank in a list may have different contributions to user's satisfaction. So the regression model of user's query-level satisfaction can be expressed as Equation 1.

$$SAT_{query} = \beta + \sum_{r=1}^{L} w_r \cdot rel_r \tag{1}$$

The Equation 1 is called the Level-1 model. SAT_{query} represents user's satisfaction with a query, rel_r represents the relevance of the r^{th} document, β and w_r are regression coefficients where w_r can be understood as the weight of the r^{th} document. Considering the document list length is not uniform and the w_r may be affected by the document position, we use two Level-2 regression models are constructed to represent the regression coefficients observed in Equation 1.

$$\beta = a_0 + a_1 \cdot L \tag{2}$$

$$w_r = b_0 + b_1 \cdot r + b_2 \cdot r^2 \tag{3}$$

Equation 2 can be understood as the intercept in Equation 1 and the intercept is a linear function of document list length (L). a_0 represents a general intercept and a_1 represents whether the intercept will change according to the list length. Equation 3 represents the regression coefficients of the documents, it is modeled as a function of document position (r). Considering that most previous metrics suggest a curvilinear decaying weighting function, so we add a second-order quadratic term of document position to Equation 3. We can know if there is an order effect and what kind of order effect it is according to the significance and fitted value of b_1 and b_2.

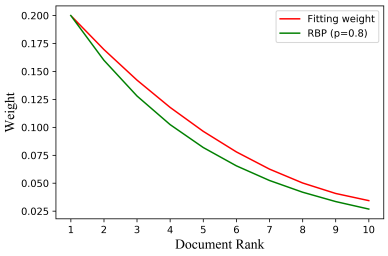

Figure 2: Comparison of document weight between our fitted model and RBP (p=0.8).

4.1.1 Based on ranking list. To examine whether the SERP has an order effect on user's query-level satisfaction. We first fit the hierarchical linear model by using the relevance scores and the ranking list on the SERP. The coefficients of the model are shown in the Table 1.

Table 1: Model fitting for query-level satisfaction based on the SERP. (*p<0.05. **p<0.01. ***p<0.001)

	a_0	a_1	b_0	b_1	b_2
Coefficient	2.102***	-0.063*	0.233***	-0.035***	0.002***
SE	0.248	0.027	0.030	0.009	0.001

We can see that both the b_1 and b_2 are significant at p<0.001, which implies that there exists an order effect on user satisfaction. To visualize order effect, we plot the trend of w_r of our fitted model in Figure 2. For comparison, we also plot the weights of a traditional metric $RBP_{0.8}$ in the same figure. We can see that the two curves are very close. The result also implies that there is a primacy effect on user's query-level satisfaction because the top-ranked documents have a larger influence on the SAT_{query}.

4.1.2 Based on Click Sequence. Similarly, to examine whether the users' click sequence has an order effect on their satisfaction. We fit the same hierarchical linear model based on the relevance scores and the click sequence. The coefficients of this model are shown in the Table 2.

Table 2: Model fitting for query-level satisfaction based on the click sequence. (*p<0.05. **p<0.01. ***p<0.001)

	a_0	a_1	b_0	b_1	b_2
Coefficient	3.504***	-0.451***	0.235***	-0.009	0.001
SE	0.107	0.057	0.035	0.016	0.001

We can see that both the b_1 and b_2 are not significant, this implies the order effect does not exist when the calculation is based on the click sequence. The results illustrate that there is no need to consider the primacy effect when a metric is calculated based on the click sequence. So that primacy effect which has been captured in the ranking list may only be due to that the document position has a effect on the user's click.

Table 3: Correlation of different metrics with query-level satisfaction (All correlations are significant at p < 0.001).

Metrics	Click Sequence			SERP
	$Usefulness$	$DwellTime$	$Relevance$	$Relevance$
CG	0.536	0.269	0.279	0.347
DCG	0.684	0.310	0.351	0.376
$RBP_{0.8}$	0.668	0.297	0.331	**0.381**
ERR	0.689	0.312	0.364	0.199
Min	0.631	0.292	0.343	0.176
$Mean$	**0.824**	**0.357**	**0.439**	0.362
Max	0.818	0.344	0.419	0.308

4.2 Correlation with query-level metrics

As mentioned in Section 3, we obtained the user's usefulness feedback and the click dwell time of each clicked document. We also obtained the query-document relevance annotation of all documents through the crowdsourcing platform. In this section, we investigate the relationship between users' query-level satisfaction and metrics based on these three measures ($usefulness$, $dwell\ time$, $relevance$).

We calculated seven query-level metrics (CG, DCG, RBP, ERR, Min, $Mean$, Max) based on these three measures. As shown in Equation(4)-(8), $DL = (d_1, d_2, ..., d_{|DL|})$ represents the document list in which each element d_r is the r^{th} document, s_r is the measure score of d_r and we use $(2^{s_r} - 1)$ to represent the gain of it. Considering that the scale of other measures is from 0 to 3 except for dwell time, to adapt to the calculation of the these evaluation metrics, we take the logarithm of dwell time and then map it to the interval of [0,3] according to the max-min method. We also use another mapping method which makes the distribution of dwell time the same with the distribution of usefulness. Since the results obtained by the two mappings are almost the same, we only report the results of the first mapping method.

$$CG = \sum_{r=1}^{|DL|} gain(d_r) = \sum_{r=1}^{|DL|} 2^{s_r} - 1 \tag{4}$$

$$DCG = \sum_{r=1}^{|DL|} \frac{gain(d_r)}{log_2(r+1)} = \sum_{r=1}^{|DL|} \frac{2^{s_r} - 1}{log_2(r+1)} \tag{5}$$

$$RBP = (1-p) \sum_{r=1}^{|DL|} gain(d_r) \cdot p^{r-1} = (1-p) \sum_{r=1}^{|DL|} (2^{s_r} - 1) \cdot p^{r-1} \tag{6}$$

$$ERR = \sum_{r=1}^{|DL|} \frac{1}{r} \prod_{i=1}^{r-1} (1 - R_i) R_r = \sum_{r=1}^{|DL|} \frac{1}{r} \prod_{i=1}^{r-1} (1 - \frac{2^{s_i} - 1}{2^3}) \frac{2^{s_i} - 1}{2^3} \tag{7}$$

$$Min/Mean/Max = (min/mean/max)(u_1, u_2, ..., u_{|DL|}) \tag{8}$$

The Pearson's correlation coefficient between query satisfaction and these metrics are shown in Table 3. The metrics calculated based on not only the click sequence but also the ranking list on the SERP. All metrics have significant correlations with query satisfaction. Comparing the performance of the same metric under different measures, we can see that the metrics based on usefulness perform the best. Similarly, comparing the seven metrics under the same measure. We can see that the $Mean$ performs the best when the calculation is based on the click sequence while the $RBP_{0.8}$ performs the best when the calculation is based on the ranking list. The $Mean$ calculated based on usefulness has the strongest correlation ($r = 0.824$) with query satisfaction, which suggests that the mean usefulness of clicked documents can best reflect user's query-level satisfaction. We can also see that the second best alternative metric is different given different conditions. When the calculation is based on the click sequence, the second best metric is Max. Differently, when the calculation is based on the ranking list, the second best metric is DCG.

This result illustrates that the traditional evaluation metrics which are calculated based on the ranking list should have a decaying weighting function, just because the user's click is affected by a primacy effect. This result is consistent with the cascade assumption. However, these metrics have limited ability to reflect user's satisfaction. On the other hand, there is no need to consider the primacy effect if we have known the user's click sequence. At this time, the metric of $Mean$ or Max can better characterize user's satisfaction.

4.3 Interaction effect between queries

In the previous section, we have found that the metrics calculated based on users' explicit feedback have better correlation with their satisfaction. One important reason is that the users' subjective perception is not directly decided by the objective measure because of the context information. Considering that many previous works have investigated the impact between documents, we only focus on the interaction effect at query-level in this work. Specifically, we want to investigate whether the user's satisfaction with the second query will be affected by his/her satisfaction perception of the initial query.

Firstly, we conducted a case analysis. We collected some examples in which the users issued the same query at the second query position. Considering that the usefulness and satisfaction are users' subjective perception of a query. If there is an impact between the queries, then the usefulness and query satisfaction will be affected at the same time. This can be reflected by what we have mentioned in the previous section that the user's usefulness have a strong correlation with satisfaction. Therefore, we try to analyze the influence of satisfaction between adjacent queries based on the relevance of the click sequence. Table 4 shows two comparison examples, including the initial query and the second query, the satisfaction of the query, and the mean relevance of clicked documents.

In the first example, user A had clicked some relevant documents (*mean relevance* = 4.0) in the first query and it was reasonable that he/she felt very satisfied with the query (*satisfaction* = 5). Although the documents that he/she clicked in the second query were also very relevant (*mean relevance* = 4.0), his/her satisfaction with the second query is decreased (*satisfaction* = 4). The documents that user B clicked in the first query were not very relevant (*mean relevance* = 2.6) so that his/she satisfaction with this query was not very high (*satisfaction* = 3). Differently, the clicked documents in the second query were very relevant (*mean relevance* = 4.0), and he/she is very satisfied with the second query (*satisfaction* = 5). We can know from this example that the user's satisfaction with the second query may be affected not only by the document relevance of the query but also by his/her satisfaction perception of the initial query. For the same query which is in the second query position, if

Table 4: Two examples of the comparison of different users' satisfaction perception with the same query at the second query position.

Example #1				
		Query	Satisfaction	Mean Relevance
User A	Initial query	smoking cessation advantage	5	4.0
	Second query	smoking cessation side-effect	4	4.0
User B	Initial query	smoking cessation	3	2.6
	Second query	smoking cessation side-effect	5	4.0
Example #2				
		Query	Satisfaction	Mean Relevance
User C	Initial query	google glass	3	3.3
	Second query	google glass price	4	4.0
User D	Initial query	google glass	4	3.2
	Second query	google glass price	5	3.5

the initial query is good enough, the user's satisfaction perception of the second query may decrease; if the initial query is not very satisfying, the user may get higher satisfaction on the second query. These results indicate that users' satisfaction perception may be affected by their expectation on the second query.

For the second example, the initial query and second query submitted by user C and user D are the same, but their clicked documents are not exactly the same. For both of the two users, the mean relevance of clicked documents in the second query is higher than the mean relevance of clicked documents in the initial query. Both users are more satisfied with the second query than the initial one. Comparing the two users in each query we can see that, although the mean relevance of clicked documents of user C is higher than user D (*mean relevance:* 3.3 vs. 3.2 and 4.0 vs. 3.5), the satisfaction of user C is lower (*satisfaction:* 3 vs. 4 and 4 vs. 5). It seems that there is a difference in absolute satisfaction between users, but the relative satisfaction perception of users is less affected by user factors. At this time, the satisfaction of users may be more affected by the anchor effect.

Therefore, the user's satisfaction perception of a query may be affected by many factors. Although the user's perception of relative difference in satisfaction between queries is almost consistent. The satisfaction perception criteria of each user may be different and the satisfaction of the initial query in a session will cause an anchor effect or expectation effect on the satisfaction perception of the subsequent query. To investigate which factor has a stronger impact on user's satisfaction perception, we regard it as a regression problem and evaluate the effectiveness of the regression models in terms of the correlations between the model predictions and user's satisfaction. Specifically, we use the query satisfaction and mean relevance of the initial query as features to predict the second query satisfaction and analyze the impact in terms of regression weight. To avoid the user bias that some users prefer to give high scores or low scores, we perform a user-level z-score processing on the user's satisfaction. We perform a 5 fold cross-validation to evaluate the performance of the regression model, the results were shown in Table 5.

Table 5: The regression weight of features and prediction results for different model.

	Mean Relevance (Q2)	Satisfaction (Q1)	PCC	MSE
Model 1	0.445	-	0.475	0.745
Model 2	0.411	0.226	0.530	0.692

We can see that when we only use the mean relevance of the second query as a feature to predict the satisfaction of the second query, the correlation between the prediction and the user satisfaction is 0.475. If we add the satisfaction of the initial query as a feature to predict the satisfaction of the second query, the corresponding correlation will increase to 0.530. This indicates that the satisfaction of the initial query does have an impact on the satisfaction of the second query. Regarding the specific impact, we can see that the coefficient of mean relevance is 0.411 and the coefficient of satisfaction of the initial query is 0.226. So that the satisfaction of the initial query will have a positive impact on the satisfaction of the second query, which suggests that the satisfaction perception of the second query is mainly affected by the anchor effect of the satisfaction of initial query.

5 SESSION-LEVEL ANALYSIS

In Section 4, we have found that user's query-level satisfaction is affected by some cognitive effects and the cascade assumption is applicable for query-level evaluation in some cases. In this section, we try to investigate what factors will influence a user's session-level satisfaction and whether the cascade assumption is still applicable for session-level evaluation.

5.1 Modeling session-level satisfaction

Intuitively, a user's session satisfaction comes from the contribution of each query (e.g. query satisfaction). So that we again use the hierarchical linear models to fit the users' session-level satisfaction based on their query-level satisfaction. And the corresponding Level-1 model can be expressed as Equation 9.

$$SAT_{session} = \beta + \sum_{r=1}^{L} w_r \cdot sat_r \tag{9}$$

$SAT_{session}$ represents user's satisfaction with a session, sat_r represents the satisfaction perception of the r^th query, β and w_r are regression coefficients where w_r can be understood as the weight of the r^{th} query. Considering the session lengths are not uniform, β and w_r should adapt to the session length (L). Moreover, w_r may also be affected by the order of a query because of the order effect. So we construct the Level-2 model as expressed in Equation 10 and Equation 11.

$$\beta = a_0 + a_1 \cdot L \tag{10}$$

$$w_r = b_0 + b_1 \cdot r + b_2 \cdot L \tag{11}$$

Equation 10 models the β in Equation 9 as a linear function of session length (L). In the equation, a_0 represents a general intercept and a_1 represents whether the intercept will change as the session length changes. Equation 11 models the regression coefficients of the queries as a linear function of session length (L) and query order (r). The significance test of the b_1 can reflect whether there is an order effect. We use the sessions in a different range of session length to fit the model, Table 6 reports the fitting results.

Table 6: Model fitting for session-level satisfaction based on query satisfaction. (*p<0.05. **p<0.01. ***p<0.001)

Session		a_0	a_1	b_0	b_1	b_2
$L \leq 3$	RC	3.108***	-0.790***	0.403***	0.236***	-0.196***
	SE	0.273	0.128	0.054	0.034	0.025
$L \leq 4$	RC	3.861***	-0.935***	0.222***	0.194***	-0.112***
	SE	0.195	0.088	0.032	0.023	0.016
$L \leq 5$	RC	3.938***	-0.785***	0.204***	0.100***	-0.056***
	SE	0.164	0.068	0.025	0.017	0.010
All	RC	4.050***	-0.560***	0.149***	0.034***	-0.017***
	SE	0.113	0.039	0.013	0.008	0.004

We can see from Table 6 that all the fitting coefficients are significant. This indicates that we have constructed an effective model to capture the user's session satisfaction. We can see that b_1 is a positive number, this means that the recent query has a higher w_r in a session. So that the user's session satisfaction is affected by the recency effect of query level satisfaction. b_2 is a positive number indicating that w_r will decrease as the session length increases, this implies that it is necessary to normalize the w_r. This result indicates that the user's attention is increasing during a session so that the cascade assumption is not suitable for characterizing session-level satisfaction.

5.2 Correlation with session-level metrics

In the last section, we find that the users' query satisfaction has a recency effect on their session satisfaction. However, this result conflicts with existing session-level metrics which usually have a decaying weighting function which emphasize the primacy effect. Therefore, we would like to investigate which kind of metric can better characterize session satisfaction when the calculation is based on different measures (usefulness, dwell time or relevance) and sequences (click sequence or SERP).

As shown in Table 7, we investigate the following five types of weighting functions which emphasize different ordering effect to weight query-level measures:

- **Decreasing weight:** the earlier queries have higher weight.
- **Increasing weight:** the later queries have higher weight.
- **Equal weight:** all queries have the same weight.
- **Middle low:** the earlier and later queries have higher weight.
- **Middle high:** the middle queries have higher weight.

The session metrics can be calculated with Equation(12) in which s_i represents the i^{th} query metric score. The metric is normalized so that it can adapt to different session lengths.

$$M = \frac{\sum_{r=1}^{N} w_r \cdot s_r}{\sum_{r=1}^{N} w_r} \tag{12}$$

Table 8 shows the Pearson's correlation coefficient between users' session satisfaction and session-level metrics consisting of different combination of query-level metrics and session weighting function.

Table 7: The query's weight at the r^{th} query position of different session weighting functions (N is the query number of a session).

Metrics	$w_r(0 < r \leq N/2)$	$w_r(N/2 < r \leq N)$
Decreasing_weight	$1/r$	$1/r$
Increasing_weight	r	r
Equal_weight	1	1
Middle_low	$1/r$	$1/(N+1-r)$
Middle_high	r	$N+1-r$

The first subtable shows the result of weighting the query-level satisfaction directly using different weighting functions. We can see that the metric with decreasing weights has the lowest correlation coefficient with session satisfaction ($r = 0.644$) while the metric with increasing weight has the strongest correlation ($r = 0.765$). The second subtable shows the result that when the calculation of the query-level metric is based on users' usefulness feedback. We can see that *Mean* performs the best while *CG* performs the worst among the seven query-level metrics. *Increasing_weight* performs the best while *Decreasing_weight* performs the worst among the five session-level weighting function. The metric which combines the *Mean* of usefulness and an increasing weighting function has the highest correlation with users' session satisfaction ($r = 0.638$). Since usefulness and satisfaction are directly obtained from user's feedback, these results suggest that it is the recency effect but not the primacy effect that has a stronger influence on user's session-level satisfaction perception.

The third and forth subtable shows the result that when the calculation of the query-level metric is based on relevance annotations. We can see that whether the calculation is based on the click sequence or based on the SERP, the *Middle_low* performs the best among the five session-level weighting function. As for the best performing query-level metric, *Mean* performs the best when the calculation is based on click sequence and *DCG* performs the best when the calculation is based on the SERP. The metric performance based on the click sequences ($r = 0.319$) is slightly higher than the metric performance based on the SERP ($r = 0.290$). This tells us that when designing relevance-based evaluation metrics, we should consider both the primacy effect and recency effect. And if we know the users' click sequence, we can better characterize their session-level satisfaction. Compared with the performance of metric based on usefulness, we see that the performance of metric based on relevance is much lower. This reflects the limitations of relevance measure in characterising user satisfaction, and users' usefulness perception can better reflect their session satisfactions.

The performance when the calculation of metric is based on user's dwell time is shown in the last subtable. The result shows that the session-level metric has a higher correlation with user's session satisfaction when the query-level metric is *Min*. The metric which combines the *Min* of dwell time and the *Middle_high* weighting function has the highest correlation with users' session satisfaction, however, there is only a weak correlation has been achieved ($r = 0.214$). However, there is not a big difference between these five session weighting functions. Therefore, it is difficult to reflect user's

Table 8: Correlation of different combination of query-level metrics and session weighting functions with session-level satisfaction. The darker shadings indicate the strongest correlation under specific measures.

Query-level Metrics		Session Weighting Function				
		$Middle\_low$	$Middle\_high$	$Equal_w$	$Decreasing_w$	$Increasing_w$
Query Satisfaction		0.732	0.696	0.724	0.644	0.765
Usefulness Based	CG	0.331	0.333	0.336	0.313	0.344
	DCG	0.452	0.446	0.454	0.420	0.468
	$RBP_{0.8}$	0.444	0.438	0.445	0.415	0.456
	ERR	0.468	0.450	0.467	0.431	0.472
	Min	0.487	0.480	0.494	0.454	0.497
	$Mean$	0.614	0.596	0.617	0.551	0.638
	Max	0.532	0.513	0.532	0.463	0.560
Relevance Based (Click)	CG	0.113	0.092	0.104	0.121	0.090
	DCG	0.173	0.152	0.165	0.173	0.154
	$RBP_{0.8}$	0.156	0.135	0.147	0.158	0.136
	ERR	0.222	0.207	0.218	0.210	0.210
	Min	0.311	0.289	0.307	0.285	0.302
	$Mean$	0.319	0.285	0.309	0.281	0.314
	Max	0.247	0.205	0.232	0.212	0.240
Relevance Based (SERP)	CG	0.289	0.250	0.276	0.277	0.264
	DCG	0.290	0.254	0.277	0.278	0.266
	$RBP_{0.8}$	0.284	0.248	0.270	0.273	0.260
	ERR	0.135	0.133	0.133	0.132	0.125
	Min	0.124	0.123	0.126	0.129	0.114
	$Mean$	0.263	0.213	0.243	0.235	0.247
	Max	0.138	0.118	0.122	0.114	0.155
Dwell Time Based	CG	0.049	0.074	0.062	0.057	0.059
	DCG	0.072	0.101	0.088	0.077	0.088
	$RBP_{0.8}$	0.080	0.104	0.093	0.081	0.094
	ERR	0.063	0.083	0.074	0.068	0.073
	Min	0.202	0.214	0.212	0.201	0.207
	$Mean$	0.173	0.197	0.189	0.172	0.187
	Max	0.092	0.126	0.112	0.094	0.115

session-level satisfaction by using dwell time with our proposed method.

6 A FRAMEWORK FOR SESSION-LEVEL METRICS

In the previous sections, we have presented that user's query-level and session-level satisfaction are affected by different cognitive effects. In this section, we try to propose a two-step framework to capture user's session-level satisfaction based on these findings. The first step is to estimate query-level satisfaction and the second step is to estimate session-level satisfaction based on the query-level satisfaction estimation.

$$\begin{cases} M_{q_1} = g_p(l_1) \\ M_{q_n} = g_p(l_n) + g_e(l_n, l_1) + g_a(M_{q_1}) \end{cases} \tag{13}$$

As shown in Equation 13, M_{q_n} represents the metric score of the n^{th} query, it comes from the contribution of three parts. The $g_p(l_n)$ represents the contribution of the result list (l_n) of the current query in which we should consider the primacy effect. The $g_e(l_n, l_1)$ is a function of the result list of the initial query and the current query, it represents the contribution of expectation effect. The $g_a(M_{q_1})$ is affected by the user's satisfaction perception of the initial query, it represents the contribution of the anchoring effect. Existing query-level metrics only consider the primacy effect so that they only include the first part $g_p(l_n)$. But when evaluating a query in a session, we need to consider the interaction effect since the queries are not independent. We have mentioned in Section 4.3 that we can better characterize the query satisfaction if considering the anchoring effect and the expectation effect. However, due to the limited space of this paper, we only focus on how to design a session-level metric based on the query-level metrics, the investigation of designing better query-level metrics is left to future work.

Table 9: The correlation of user satisfaction with evaluation metrics under different λ.

λ	0.2	0.3	0.4	0.5	0.6	0.7
PCC	0.762	0.774	**0.777**	0.775	0.768	0.760

$$\begin{cases} M_{s_1} = M_{q_1} \\ M_{s_n} = (1 - w_n) \cdot M_{s_{n-1}} + w_n \cdot M_{q_n} \end{cases} \tag{14}$$

Equation 14 shows the method of obtaining the session-level evaluation metric, which is calculated based on the result of the query-level metrics. M_{s_n} represents the session metric score of the previous n queries in the session, the contribution of each query is controlled by the parameter w_n. This kind of form ensures that the metric is normalized. To verify the validity of the framework, we directly use the query satisfaction feedback from the user as the score of the query metrics and use the following function for w_n.

$$w_n = \frac{1}{n^\lambda} \tag{15}$$

As shown in Equation 15, w_n is simply chosen as a function of n in this paper. In the future work, more complex functions, such as the function that considers the gain and cost of the search, can be used to model w_n. When λ is equal to 1, the weight of all queries is the same. When λ is larger than 1, the weight of the subsequent query will become smaller. So the metric will emphasize the primacy effect. When λ is less than 1, the weight of the subsequent query will be larger which emphasizes the recency effect. When λ is equal to 0, only the last query contributes to the session metric.

In the previous section, we have found that the user's session-level satisfaction is affected by the recency effect, so λ should be

less than 1. We test 5 different λ shown in Table 9, and compute the PCCs between the M_{s_n} and the user's session-level satisfaction feefback. We found that when $\lambda = 0.4$, the evaluation metric has a strong correlation of $r = 0.777$ with the session-level satisfaction. Based on this framework, it is easy to derive new session-level evaluation metric from the query-level.

7 CONCLUSION

Understanding what factors will influence user satisfaction in the whole search session and how to design session-level evaluation metrics accordingly is crucial for improving the search engine, especially in supporting complex search tasks. To achieve this goal, we conducted a laboratory user study to collect real users' feedback during their search process. Furthermore, we analyze our collected data in the user study to address the three RQs.

We find that the query-level metrics have different performance under different conditions. When the calculation is based on the SERP, the metrics which emphasize the primacy effect (e.g. RBP) perform better. However, when the calculation is based on the click sequence, the *Mean* performs better than the other metrics. The metrics based on usefulness perform better than the metrics based on relevance or dwell time. We show that user's satisfaction on the initial query in a session will cause an anchor effect or expectation effect on the satisfaction on the subsequent queries. The satisfaction on the second query is mainly affected by the anchor effect of the initial query. The user's perception of the relative difference in satisfaction between queries is almost consistent. We find that users' experience with the later queries has a greater impact on the session-level satisfaction feedback, which proves that the recency effect has a stronger influence on user's session-level satisfaction than the primacy effect. On the other hand, if we want to characterize user's session-level satisfaction with the document relevance annotations, we need to consider both the recency effect and primacy effect. We compare different combinations of query-level metrics and session-level weighting functions. Results show that the *Increasing_weight* session weighting function is more suitable for the metrics based on usefulness-based measures. However, if the metrics are computed with the external relevance annotations of documents, the *Middle_low* session weighting function will be more appropriate.

From our results, we can see that traditional evaluation metrics may not be suitable for characterizing user satisfaction because they ignore the recency effect in users' perception of session-level satisfaction. Furthermore, the experiment results suggest that a metric for session-level satisfaction should meet the following criteria: (1) The SERP will cause a primacy effect on users' query satisfaction; (2) The average and maximum usefulness perception of documents in a query is the most important; (3) The initial query has an anchoring effect on the perception of the subsequent query; (4) The session weighting function should be normalized to adapt to different session length; (5) The recency effect has a stronger influence on user's session satisfaction.

8 ACKNOWLEDGMENTS

This work is supported by Natural Science Foundation of China (Grant No. 61622208, 61732008, 61532011) and the National Key Research and Development Program of China (2018YFC0831700).

REFERENCES

[1] Azzah Al-Maskari, Mark Sanderson, and Paul D. Clough. 2007. The relationship between IR effectiveness measures and user satisfaction. In *SIGIR'07*.
[2] Rashid Ali and MM Sufyan Beg. 2011. An overview of Web search evaluation methods. *Computers & Electrical Engineering* 37, 6 (2011), 835–848.
[3] 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.
[4] Ernest R Cadotte, Robert B Woodruff, and Roger L Jenkins. 1987. Expectations and norms in models of consumer satisfaction. *Journal of marketing Research* (1987), 305–314.
[5] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *CIKM'09*.
[6] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *WSDM'08*.
[7] Henry A Feild, James Allan, and Rosie Jones. 2010. Predicting searcher frustration. In *SIGIR'10*.
[8] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems (TOIS)* 23, 2 (2005), 147–168.
[9] Michael Gordon and Praveen Pathak. 1999. Finding information on the World Wide Web: the retrieval effectiveness of search engines. *Information Processing & Management* 35, 2 (1999), 141–180.
[10] Ahmed Hassan, Ryen W White, Susan T Dumais, and Yi-Min Wang. 2014. Struggling or exploring?: disambiguating long search sessions. In *WSDM'14*.
[11] Scott B Huffman and Michael Hochster. 2007. How well does result relevance predict session satisfaction?. In *SIGIR'07*.
[12] Jaana Kekalainen Kalervo Jarvelin. 2003. User-oriented evaluation methods for information retrieval: A case study based on conceptual models for query expansion. *Exploring artificial intelligence in the new millennium* (2003), 355.
[13] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.
[14] Kalervo Järvelin, Susan L Price, Lois ML Delcambre, and Marianne Lykke Nielsen. 2008. Discounted cumulated gain based evaluation of multiple-query IR sessions. In *ECIR'08*.
[15] Jiepu Jiang, Ahmed Hassan Awadallah, Xiaolin Shi, and Ryen W. White. 2015. Understanding and Predicting Graded Search Satisfaction. In *WSDM'15*.
[16] Jiepu Jiang, Daqing He, and James Allan. 2017. Comparing In Situ and Multidimensional Relevance Judgments. In *SIGIR'17*.
[17] Jiepu Jiang, Daqing He, Diane Kelly, and James Allan. 2017. Understanding ephemeral state of relevance. In *CHIIR'17*.
[18] Diane Kelly. 2009. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends in Information Retrieval* 3, 1-2 (2009), 1–224. https://doi.org/10.1561/1500000012
[19] Youngho Kim, Ahmed Hassan Awadallah, Ryen W. White, and Imed Zitouni. 2014. Modeling dwell time to predict click-level satisfaction. In *WSDM'14*.
[20] Matthew Lease and Emine Yilmaz. 2012. Crowdsourcing for information retrieval. In *SIGIR'12*.
[21] Mengyang Liu, Yiqun Liu, Jiaxin Mao, Cheng Luo, and Shaoping Ma. 2018. Towards Designing Better Session Search Evaluation Metrics. In *SIGIR'18*.
[22] Mengyang Liu, Yiqun Liu, Jiaxin Mao, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. "Satisfaction with Failure" or "Unsatisfied Success": Investigating the Relationship between Search Success and User Satisfaction. In *WWW'18*.
[23] 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*.
[24] Jiyun Luo, Christopher Wing, Hui Yang, and Marti A. Hearst. 2013. The water filling model and the cube test: multi-dimensional evaluation for professional search. In *CIKM'13*.
[25] 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*.
[26] Alistair Moffat, Paul Thomas, and Falk Scholer. 2013. Users versus models: What observation tells us about effectiveness metrics. In *CIKM'13*.
[27] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)* 27, 1 (2008), 2.
[28] Mark D Smucker and Charles LA Clarke. 2012. Time-based calibration of effectiveness measures. In *SIGIR'12*.
[29] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *science* 185, 4157 (1974), 1124–1131.
[30] Ellen M Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information processing & management* 36, 5 (2000), 697–716.
[31] 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*.
[32] Nancy C Waugh and Donald A Norman. 1965. Primary memory. *Psychological review* 72, 2 (1965), 89.
[33] Ya Xu and David Mease. 2009. Evaluating web search using task completion time. In *SIGIR'09*.
[34] Yiming Yang and Abhimanyu Lad. 2009. Modeling expected utility of multi-session information distillation. In *ITCIR'09*.