Effects of User Negative Experience in Mobile News Streaming

Hongyu Lu
DCST, BNRist, Tsinghua University
Beijing, China
luhy16@mails.tsinghua.edu.cn

Min Zhang*
DCST, BNRist, Tsinghua University
Beijing, China
z-m@tsinghua.edu.cn

Weizhi Ma
DCST, BNRist, Tsinghua University
Beijing, China
mawz14@mails.tsinghua.edu.cn

Ce Wang
Tencent
Beijing, China
cewang@tencent.com

Feng xia
Tencent
Beijing, China
xiafengxia@tencent.com

Yiqun Liu
DCST, BNRist, Tsinghua University
Beijing, China
yiqunliu@tsinghua.edu.cn

Leyu Lin
Tencent
Beijing, China
goshawklin@tencent.com

Shaoping Ma
DCST, BNRist, Tsinghua University
Beijing, China
msp@tsinghua.edu.cn

ABSTRACT

Online news streaming services have been one of the major information acquisition resources for mobile users. In many cases, users click an article but find it cannot satisfy or even annoy them. Intuitively, these ***negative experiences*** will affect users' behaviors and satisfaction, but such effects have not been well understood. In this work, a retrospective analysis is conducted using real users' log data, containing user's explicit feedback of negative experiences, from a commercial news streaming application. Through multiple ***intra-session*** comparison experiments, we find that in current session, users will spend less time reading the content, lose activeness and leave sooner after having negative experiences. Later return and significant changes of user behaviors in the next session are also observed, which demonstrates the existence of ***inter-session*** effects of negative experiences.

Since users' negative experiences are generally implicit, we further investigate the possibility and the approach to automatically identify them. Results show that using changes of both users' *intra-session* and *inter-session* behaviors achieves significant improvement. Besides the effects on user behaviors, we also explore the effects on user satisfaction by incorporating a laboratory user study. Results show that negative experiences reduce user satisfaction in the current session, and the impact will last to the next session. Moreover, we demonstrate users' negative feedback helps on the meta-evaluation of online metrics. Our research has comprehensively analyzed the impacts of users' item-level negative experiences, and shed light on the understanding of user behaviors and satisfaction.

*Contact author

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331247

KEYWORDS

Negative Experience; News Recommendation; User Behavior Modeling; User Satisfaction; Log Analysis

ACM Reference Format:
Hongyu Lu, Min Zhang, Weizhi Ma, Ce Wang, Feng xia, Yiqun Liu, Leyu Lin, and Shaoping Ma. 2019. Effects of User Negative Experience in Mobile News Streaming. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3331184.3331247

1 INTRODUCTION

Online news recommender systems play an important role in meeting users' information needs. Traditionally, click signals and the time spent on reading are widely used as implicit feedback of positive experiences. However, the news a user clicked not always satisfies him/her, sometimes it even makes him/her bored. We name these unpleasant and disagreeable experiences as ***negative experiences***[1]. Such negative experiences are commonly implicit but will have non-trivial influences on user's behaviors and satisfaction, intuitively.

As shown in Figure 1, a user reads an article but experience negatively while browsing the news recommendation list. Then, the user continues to browse and read until he/she decides to leave. After some time, the user returns and starts a new session. In these interaction processes, there remain many questions to answer. Such as, **after the negative experience:**

- will the user click less?
- will the user leave sooner?
- will the user return later after this session?
- will the user change his/her interactions in the next session?

Modeling the effects of negative experiences on user behaviors help us to answer these questions and make better use of interaction behaviors. It also gives us an opportunity to identify users' implicit negative experiences to help provide better recommendations.

In this paper, we focus on providing insights for the understanding of how negative experiences affect users' behaviors and satisfaction, which is less studied in previous work. The impacts of negative experience not only exist in the current session, but also last in

[1]We refer to item-level negative experience in this paper.

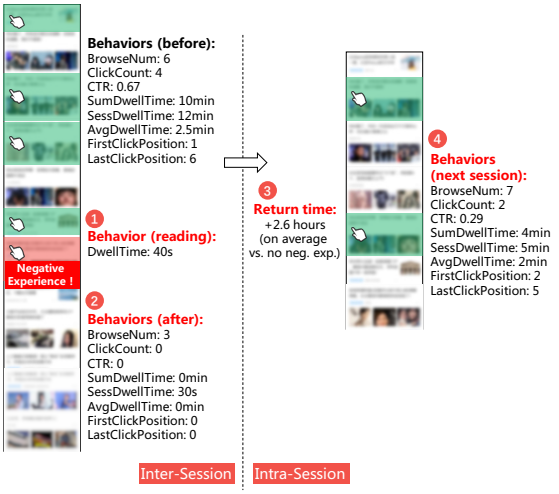

Figure 1: An example: when a user has negative experience, his/her reading behaviors, subsequent behaviors in the current session (intra-session), return time and behaviors in the next session (inter-session) may change.

the next session, which is also studied here. For more comprehensively analysis, we separate the impacts of negative experience into two groups: ***intra-session effects*** and ***inter-session effects***. We conduct analyses using large-scale logs of a commercial newsfeed application. It contains not only the user behaviors, but also some users' explicit feedback of their negative experience.

We aim to answer three research questions in our work:

- **RQ1**: What are the effects of negative experiences on user's behaviors?
- **RQ2**: Can we identify negative experiences using the changes of user interaction behaviors?
- **RQ3**: How negative experiences affect users' satisfaction?

With RQ1, users' reading and subsequent behaviors in current session are inspected to study the *intra-session effects*. Moreover, changes of user behaviors after the current session are examined for *inter-session effects*, including the session return time and the user interactions in the next session. Significant impacts are observed by comparison experiments. To address RQ2, several groups of features are proposed, based on previous observations, to identify users' negative experiences. It is demonstrated that changes of user interactions are helpful for negative experience identification. As for RQ3, a two-step approach, *linking* and *expanding*, is conducted via combining the large-scale log analysis with the laboratory user study. The relationships between user behaviors and satisfaction feedback are discovered in the user study, and further used to expand our previous discovered effects on user behaviors to the effects on user satisfaction.

To sum up, following contributions are made:

- Comprehensive study on the *intra-session* and *inter-session* effects of negative experience are conducted, and six observations on significant impacts are made. To the best of our knowledge, this is the first work studying user's negative experience in online news streaming scenario.
- Both the user behaviors in current document and changes of user subsequent behaviors are proposed and found helpful to identify negative experience. It will be useful to find large-scale implicit negative experiences in real applications and generate better recommendations.
- The influences of negative experiences on user satisfaction in both current and next session are investigated. Furthermore, a criterion based on the negative feedback is proposed to conduct meta-evaluation of online metrics in system logs.

2 RELATED WORK

Since we are studying *how negative experience affects user behaviors and satisfaction*, previous research falls into three directions: the analysis of user negative experience, the analysis of user behaviors, and the analysis of satisfaction.

2.1 User Negative Experience Analysis

In most of previous works, researchers are focusing on user's positive experiences, like user preference, engagement [1, 2] and satisfaction [3], discovering signals to identify and estimate them for both learning and evaluation. In addition, several previous works investigate the negative experiences in different scenarios.

In information search scenario, Pogacar et al. [4] find that search engine results can significantly influence people both positively and negatively. Incorrect search results significantly influence users' decisions and lead users to wrong choices. White et al. [5] show that user are more likely to skip negative results to reach positive in the context of yes-no questions in the medical domain. In general web browsing, Miroglio et al. [6] find that after user install ad blocker to avoid annoying online ads, their engagement increases. And in music discovery scenario, Garcia et al. [3] find that the positive or negative extremes were more correlated to satisfaction than total interactions. Users who are annoyed by at least one track are negatively correlated with user satisfaction. In online news reading scenario, users are quite likely to have negative experiences because of the existence of title-bait and fake news [7]. Lu et al. [8] find that user clicks are not always consistent with user preference, more than 50% clicked news is disliked by the user.

These works indicate that there does exist negative experiences in many scenarios. However, there still lacks the comprehensive analysis of the effects of users' item-level negative experiences, which is the focus topic of our work.

2.2 User Behavior Analysis

User behaviors are widely used in online information systems for ranking, recommendation and evaluation. While the behavior signals contain much information, they are also full of biases.

Click behavior has been widely used as the positive implicit feedback in interactive information systems. Researchers use click signals to infer document relevance [9] in search and user preference [10, 11] in recommender systems. However, click behaviors are found to be biased by many factors, like position [12, 13], trust [14], quality [15], presentation [16], delivery mechanism of the system [17].

Dwell time has also been considered to be correlated to user's positive experience and been widely used in a number of retrieval applications. A dwell time equaling or exceeding 30 seconds has typically been used to identify clicks with which searchers are satisfied [18–21]. The relationship between dwell time and user interest is further modeled with document factors (e.g. readability [22], and human factors [23]). Besides dwell time, viewport time [24] and eye gaze [25, 26] are also used as the implicit positive feedback of user's interest.

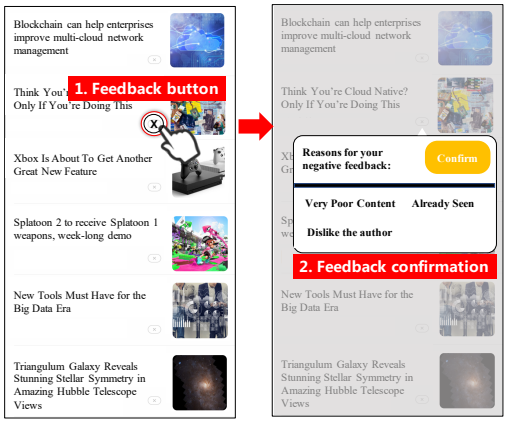

Figure 2: The typical list-style interface (translated version) of the newsfeed system (was used in WeChat). The interactive interface to collect users' feedbacks for negative experience: (left) the non-disturbing controls on the snippets and (right) the popover for feedback confirmation.

Most of the previous works model user's behaviors with user's positive experience and focus on the factors about the current item. In this work, on the one hand, we turn to study the effects of negative experiences. On the other hand, we move further to study the effects in a long-term, not only the behaviors in current item, but also the subsequent interactions in current session (intra-session effects) and next session (inter-session effects).

2.3 User Satisfaction Analysis

User satisfaction has been extensively discussed in the areas of consumer, marketing and psychology research since the mid-1970s [27], and was first proposed by Su et al. [28] in information retrieval system. In information systems, satisfaction is defined as the fullness of user's information need [29, 30].

Many factors that influence user satisfaction have been well studied in previous work [31, 32]. Dan at el. [33] summarize factors related to user satisfaction in search, including the query performance and the task difficulty. In recommender system, many factors, like diversity [34] and serendipity [35] are found to be related to user satisfaction.

Different with previous work, we study the effect of user's item-level negative experience on user satisfaction. Meanwhile, most of the previous works study the satisfaction based on laboratory user study. Through proposed two-step approach, we conduct our analysis in large-scale logs collected from real users in the natural environment.

3 LOG-BASED ANALYSIS METHODOLOGY

In this section, we describe the dataset, the collection of negative experience and the measurements we use in analysis.

3.1 Online News Streaming Scenario

Web users are increasingly using online newsfeed systems, like Google News, Yahoo! News, TopBuzz, "Top Stories" in WeChat, etc., to access information and news, especially on the mobile devices. When a user visits the newsfeed application, the system will recommend a series of news and articles [1] for the user based on his/her interest. A typical interface is shown in Figure 2, the recommended news, with the title and pictures shown, is arranged vertically in the *list page*.

As users scroll to browse the list, the system will continue to load some pieces of news. If a user clicked on any of the news, the system takes the user to the *content page* in which the full content of the news is displayed. At any time, the user can finish reading and return to the *list page* for continuing browsing from previous position. In these processes, the user may decide to leave the system at any time, which means that user can browse any numbers of news in a session.

[1] collectively called "news" in our work

3.2 Negative Feedback Collection

We collect log data in two weeks from the WeChat. It contains not only the content recommended and users' interaction behaviors, but also the users' explicit feedback for their negative experiences.

The feedback of user negative experience is collected by a feedback button (as shown in Figure 2). Similar control is widely used in many online information systems, like *"Fewer stories like this"* button in Google News, and *"Not interested"* button in Movielens. If users experience negatively while reading the news, they can click on the remove ("x") button to provide feedback for this negative experience. The system will not actively disturb users. Although this "non-disturbing" feedback acquisition method gets a small amount of feedback, it does not affect the user's normal interaction behaviors. After clicking the button, the system will ask the user to confirm this feedback, and to give the reasons for negative experience optionally. These options include *Very Poor Content*, *Already Seen*, and *Dislike the author*, etc. In this paper, we mainly focus on studying the impact of the negative experience, the analysis of these reasons remains for future work.

Note that some users are not used to provide negative feedback, and it is unknown that whether they actually feel positive or they do not notice the feedback interface. So taking all the users into analysis may cause biases. Hence, we re-generate the dataset which only contains the interactions from the users who have ever provided negative experience. In total, we collected 106,067 feedbacks in 317,216 sessions from 5,533 users. We further divide the dataset into two parts based on time. The interactions in the first week is used for analysis and model training, the interactions in the second week is used for modeling testing (Section 6).

Besides, all data was observational, anonymous, and analyzed in aggregate. The institutional review board confirmed that the data has no privacy and ethical issues.

3.3 Interaction Measurements

To comprehensively represent user's interaction in a session, we choose to use various measurements which are commonly used as online evaluation metrics in information retrieval scenarios [18, 36–38] and are proved to be correlated with user experiences [39].

- Click-based measurements
  - **ClickCount** - Number of clicks in a session, also known as QCTR [36].
  - **CTR** - Click-Through Rate, which is widely used in evaluating the performance of recommendation.
  - **PLC** - Precision at Lowest Click [36], the number of clicks divided by the position of the lowest click.
  - **SatClickCount, SatCTR, SatClickRatio, DsatClickRatio** - Previous works show that dwell time can be used to identify user item-level satisfaction. The clicks following a long dwell time (e.g. 30s [18, 40]) are considered as satisfied click. We

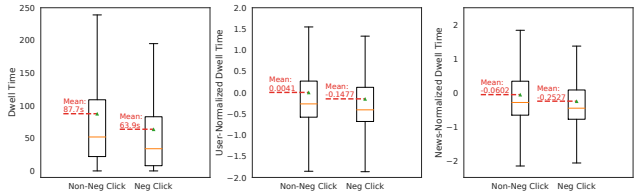

Figure 3: The comparison of dwell time(a), user-normalized dwell time(b) and news-normalized dwell time(c) between the clicks without and with negative experience.

calculate ClickCount and CTR with satisfied clicks [41], and the ratio of satisfied/unsatisfied clicks in all the clicks.

- Time-based measurements
  - **SessionDwellTime** - The total time user spend in browsing the whole session (including reading time).
  - **SumClickDwell, AvgClickDwell, ClickDwellRatio** - The sum and the average user dwell time of clicks in the session, and the ratio of SumClickDwell in SessionDwellTime.
  - **TimeToFirstClick, TimeToLastClick** - The time between the start of session and the first/last click in the session.
  - **BrowsingSpeed** - Considering both SessionDwellTime and BrowseNum, we calculate the speed of user's browsing (number of news per second).
- Position-based measurements
  - **BrowseNum** - Number of impressions user browse in a session. In streaming system, the impressions are loaded as user scrolling.
  - **FirstClickPosition, LastClickPosition** - The position in the recommendation list of the first click and the last click.
  - **MinRR, MaxRR, MeanRR** - The maximum, minimum and mean reciprocal ranks of the clicks respectively.

These measurements are mostly based on user browsing and clicking behaviors as well as the corresponding time and position, which are widely logged in online news recommender systems. In order to cover and com pare with previous work, we have retained some mutual convertible metrics, SatClickRatio vs. DsatClickRatio, FirstClickPosition vs. MaxRR, LastClickPosition vs. MinRR.

To answer $RQ1$, we investigate the effects of negative experiences on user behaviors. We separate the effects into two groups: intra-session effects (Section 4) and inter-session effects (Section 5) based on whether it belongs to the current session.

4 INTRA-SESSION EFFECT OF NEGATIVE EXPERIENCE

In this section, we study the intra-session effects of the negative experience which is reflected by the change of interactions in current news and the change of subsequent interactions in the current session.

4.1 Interactions in Current News

Dwell Time which indicates how long users spend reading the document is widely used to estimate their experience. We now investigate whether it is related to users' negative experiences.

Firstly, we compare users' dwell time(s) of the clicks without (Non-Neg Click) or with (Neg Click) negative experience, shown in Figure 3(a). The average dwell time of Neg Click is 63.9s, significantly shorter than the dwell time of Non-Neg Click (87.7s). The difference is tested by t-test ($pvalue < 0.001$).

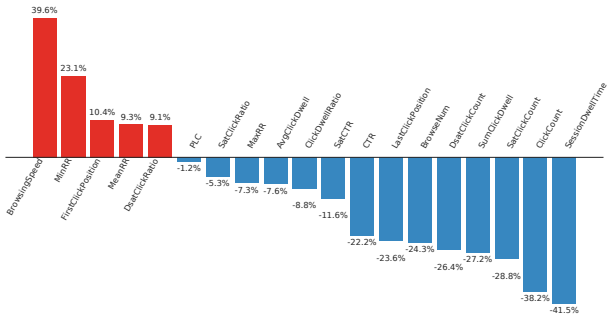

Figure 4: If user click the article at the first position but with negative experience, the following interaction changes. (two-sample t-test, *means p-value<0.05, **means p-value<0.01)

Previous work shows that dwell time is influenced by many factors. Some are related to the user, like the speed of reading and the size of screen, and others are related to its content, like the length of text and the number of images. We eliminate the effects from both user and news by normalizing the dwell time $dt_{u,n}$ within both user (u) and news (n).

$$dt^{user}_{u,n} = \frac{dt_{u,n} - Avg(DT_u)}{Std(DT_u)}\ ,\ dt^{news}_{u,n} = \frac{dt_{u,n} - Avg(DT_n)}{Std(DT_n)}$$

where DT_u and DT_n are all the dwell time of user u and news n respectively.

In between Non-Neg Click and Neg Click, We further compare user-normalized dwell time (Figure 3b) and news-normalized dwell time (Figure 3c) respectively. The results show both normalized dwell time is shorter when user has negative experience. The differences are test by t-test (p-value < 0.001).

Finding #1: Negative experiences lead to shorter reading, indicated by the shorter dwell time.

4.2 Subsequent Interactions in Current Session

As most of the previous work only studies the impact of influencing factors on user behaviors within the current item, the effects on the subsequent behaviors and corresponding analysis methods are less studied. In this section, we study how users change their subsequent behaviors in the current session after they have negative experiences.

User's negative experience for a clicked news may happen after the user has read many other news. These previous interactions may also have effects on the user's subsequent behaviors. To control these effects and leave the effect of negative experience alone, we carefully design two experiments.

4.2.1 *First-Position Experiment.*

Firstly, we introduce an experiment with a strict limitation: we only inspect the clicks occurring at the first position in the recommendation list. In this case, there is no interactions before the click so that we can ignore the influence of previous interactions. Specifically, the comparison is conducted between "click on the first news but with negative experience" (C_{neg}) and "click on the first news and without negative experience" ($C_{non-neg}$).

The difference in behavior measurements after the click (m_{after}) between two conditions are used to measure the change of user

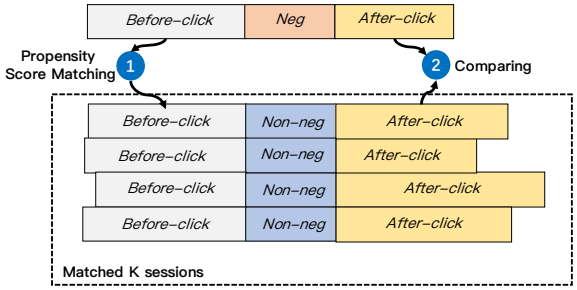

Figure 5: Illustration of the matching experiment. For each click with negative experience, K matched clicks with non-negative experience are found by Propensity Score Matching approach.

interaction after negative experience.

$$Change(m) = \frac{Avg(m_{after}(C_{neg})) - Avg(m_{after}(C_{non-neg}))}{Avg(m_{after}(C_{non-neg}))}$$

where m is a behavior measurement. The change of all the behavior measurements are shown in Figure 4.

Comparing with the non-negative condition, the total time user spend in continuing reading (*SessionDwellTime*) after negative experience has a decline of more than 41.5%. Similarly, the number of news browsed (*BrowseNum*) and clicked (*ClickCount*), as well as the Click-Through-Rate (*CTR*) also have a quite considerable decline. Some other behaviors, like the speed of user browsing (*BrowsingSpeed*), have increased. These changes indicate that user's interest or patients may decline after negative experiences (first-position), and may result in a quick leave and less activeness.

4.2.2 *Matching Experiment.*

Starting from the heuristic first-position analysis, in this section, we further eliminate the first-position limitation and conduct a more general and formal experiment to measure the effect of negative experience on users' subsequent behaviors. To avoid the confounding effects of user's previous interactions, we use the Propensity Score Matching (PSM) [42–44], a widely used matching method to control the confounding effect, to re-generate a paired dataset for analyzing the effects.

As shown in Figure 5, we separate the session into three parts by the clicks: the before-click interaction x_i, the click c_i, and the after-click interaction y_i. The clicks are separated into two groups based on whether the click c_i has negative experience. The experimental group S_{neg} consists of all the clicks with negative experience. For each negative experience in S_{neg}, we find K (K=10 in our work) most similar clicks based on the estimated propensity score to form a controlled group $S_{non-neg}$. We use twenty behavior measurements (listed in Section 3.3) of before-click process, and the Logisitic Regression to estimate the propensity score.

Through PSM, we first build a paired dataset. As shown in Figure 6, the distributions of before-click interactions are very similar between the matched Neg and Non-Neg groups after PSM approach. Based on the matched dataset, we measure the negative experience effects by comparing the average after-click behaviors in Non-Neg and Neg groups, the results are shown in Table 1.

Some behaviors, like *SessionDwellTime*, *TimeToLastClick* and *SumClickDwell*, have significant decline. It indicates that user will spend less time in the following browsing. Some behaviors, like *SatClickCount*, *ClickCount*, and *CTR*, which may represent the activeness of user's following reading, also have significant decline. Other behaviors, like *BrowsingSpeed*, *MeanRR*, *MinRR* and *DsatClickRatio* have some increase. These results confirm the findings of First-Position experiment. Combining the results of two experiments, we have the following finding:

Finding #2: After negative experiences, users lose activeness and leave sooner, indicated by the decreases of some behaviors like *ClickCount* and *SessionDwellTime*.

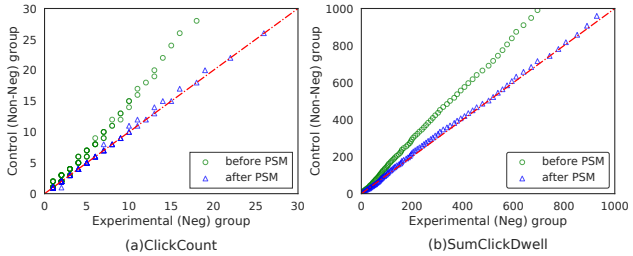

Figure 6: Demonstration of controlling the effects of two example confounders: *ClickCount* before the click, *SumClickDwell* before the click. Use the Q-Q plot to compare the distribution of confounders in Neg and Non-Neg groups.

Table 1: Comparing the interactions after user reading an article without or with negative experience, based on the matched dataset generated by PSM. (two-sample t-test, *means p-value<0.05, **means p-value<0.01)

	Non-Neg [1]	Neg	Δ
SessionDwellTime(s)	**1437.2**	956.8**	-33.4%
TimeToLastClick(s)	**1552.3**	1052.1**	-32.2%
ClickCount	**6.246**	4.572**	-26.8%
SatClickCount	**4.185**	3.177**	-24.1%
SumClickDwell(s)	**544.8**	419.6**	-23.0%
LastClickPosition	**43.53**	33.85**	-22.2%
ImpCount	**43.81**	34.95**	-20.2%
ClickDwellRatio	**0.641**	0.548	-14.6%
TimeToFirstClick(s)	**151.8**	131.6	-13.3%
CTR	**0.225**	0.203**	-9.9%
SatCTR	**0.167**	0.154**	-7.8%
FirstClickPosition	**5.860**	5.577	-4.8%
MaxRR	**0.499**	0.486	-2.6%
SatClickRatio	**0.588**	0.574	-2.3%
AvgClickDwell(s)	**85.79**	83.93	-2.2%
PLC	0.348	**0.349**	0.4%
BrowsingSpeed(num/s)	0.074	**0.076**	3.0%
DsatClickRatio	0.412	**0.426**	3.3%
MeanRR	0.248	**0.268****	7.8%
MinRR	0.158	**0.178****	12.4%

5 INTER-SESSION EFFECT OF NEGATIVE EXPERIENCE

In this section, we proceed with examining the impact of negative experience on user behaviors after the current session, including session return time and interactions in the next session.

[1] Due to commercial reason, we rescale all the values of behavior metrics reported in this paper. However, this should not change any trends or conclusions we observe in this paper.

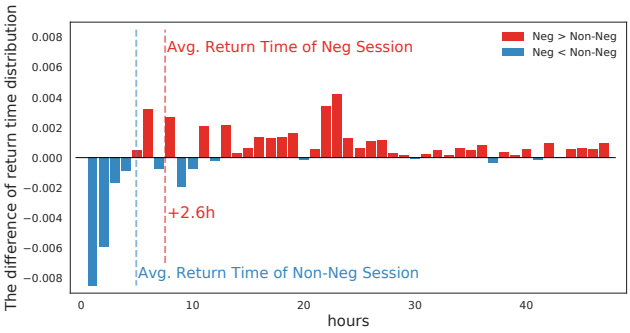

Figure 7: The difference in the distribution of return time between the session with negative experience (*Neg*) and the session without negative experience (*Non-Neg*)

5.1 Session Return Time

We first investigate the effect of negative experience on session return time by answering the question: *Will users return later if they have negative experience in current session?*

How long the user will return is measured by *Session Return time*. If the session is not the user's last visit (the last session is removed in this analysis), its return time can be calculated by the time between the start of the next session $s_{u,i+1}$ and the end of the current session $s_{u,i}$:

$$ReturnTime(s_{u,i}) = StartTime(s_{u,i+1}) - EndTime(s_{u,i})$$

We separate all the session into two groups by whether the session has negative experiences or not, correspondingly named as *Neg* sessions and *Non-Neg* sessions. We calculate the percentage of sessions returned after leaving for i hours ($i = 1, 2, ..., 48$), as the distribution of *Session Return Time*, and compare the distributions of *Neg* and *Non-Neg* session groups.

The differences are shown in Figure 7. After just finishing a session with negative experience, fewer users return within the first five hours. The average return time of *Neg* sessions is higher than the return time of *Non-Neg* sessions. It can conclude that user may return later for about 2.6 hours (two-sample t-test, t=19.54, p-value$\ll$ 0.01, cohen's d=-0.2087) after they have browsed a recommendation list but with negative experience.

Finding #3: After negative experiences, user return later in next session, indicated by the longer session return time.

5.2 Interactions in Next Session

In this section, we investigate whether the effects of negative experiences last to the following next session. We extract session pairs (s_i, s_{i+1}) adjacent in time within each user. Based on whether there exists negative experience (*Neg*) or not (*Non-Neg*), we label the sessions and generate four groups: (*Non-Neg*, *Neg*), (*Neg*, *Neg*), (*Non-Neg*, *Non-Neg*) and (*Neg*, *Non-Neg*). Specifically, we aim to investigate *whether the **user behaviors in the second session** is related to whether the first session has negative experiences.*

We conduct two comparison experiments. The first one is conducted between (*Non-Neg*, *Neg*) and (*Neg*, *Neg*) groups. The differences in behaviors of the second session are calculated by

$$\delta^{neg}(m) = Avg(m(S_{(neg,neg)})) - Avg(m(S_{(nonneg,neg)}))$$

$$S_{(neg,neg)} = \{s_{i+1} : s_i = neg \wedge s_{i+1} = neg\}$$
$$S_{(nonneg,neg)} = \{s_{i+1} : s_i = nonneg \wedge s_{i+1} = neg\}$$

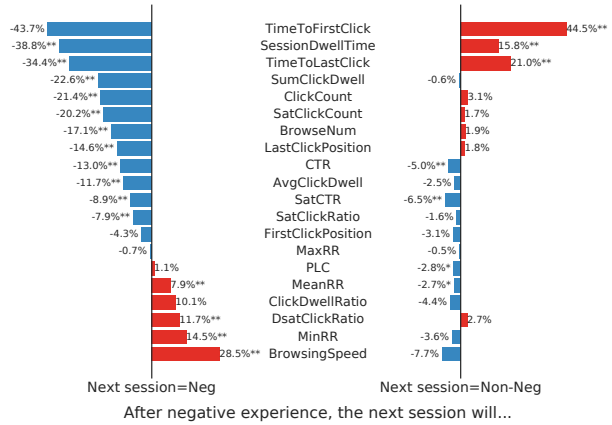

Figure 8: The impacts of negative experience in the next session. User behaviors in the second session is related to whether the first session has negative experience. (Red / blue means the metric increases / decrease after neagtive experience.) (two-sample t-test, *means p-value<0.05, **means p-value<0.01)

The results are shown in Figure 8 (a). We can find that if the first session has negative experience, some main behaviors, like *SessionDwellTime*, *ClickCount*and *CTR*, decrease more than 10%. Some other behaviors, like *DsatClickRatio* and *BrowsingSpeed*, increase more than 10%.

The second comparison is conducted between (*Non-Neg*, *Non-Neg*) and (*Neg*, *Non-Neg*) groups. The differences in behaviors of the second session are calculated by

$$\delta^{nonneg}(m) = Avg(m(S_{(neg,nonneg)})) - Avg(m(S_{(nonneg,nonneg)}))$$

The results are shown in Figure 8 (b). Surprisingly, we find that the differences is almost the opposite of the differences in the first experiment. *TimeToFirstClick*, *SessionDwellTime* and *TimeToLastClick* increased. Most of other behaviors have no significant changes. Part of explanation can be that when a user experience negatively in last session, he/she will reduce his/her expectation. Then, if the current session is good (*Non-Neg*), he/she may be more active and speed more time in the session.

Finding #4: After negative experiences, users behave differently in the next session, indicated by the significant difference of user behaviors in the next session, when they have negative experiences in the current session.

So far, we have given a comprehensive analysis about the effects of negative experience on user behaviors, from the reading process in current news, to following interactions in the current session, the return time and the interactions in next session.

6 NEGATIVE EXPERIENCE IDENTIFICATION

Previous analysis shows that user interacts differently after he/she has a negative experience. In this section, we propose the negative experience identification task, and investigate the research question (RQ2): *Can we identify negative experience using the change of user interactions?*

6.1 Task Definition

The task of identifying user's negative experience is defined as a classification problem: given a click, to predict whether the user

has negative experience. We use user's real negative feedback as the ground truth. The whole two-week dataset is separated into training and testing sets based on time. Note that previous analysis has not used the testing data, which ensure our evaluation for the identification remains reliable. All the following results reported are of the test set.

6.2 Features and Model

Based on our findings from previous analysis, we build various features for identifying negative experience. The features can be categorized into four groups:

F0: Behaviors in current news (Section 4.1). We use the DwellTime ($dt_{u,j}$), the user-normalized DwellTime ($dt^{user}_{u,j}$), and the news-normalized DwellTime ($dt^{news}_{u,j}$) to represent the user reading behaviors in current news.

F1: Change of subsequent interaction in current session (Section 4.2). Users' interactions are represented by twenty behavior measurements m (list in the Section 3.3). Further, the change of interactions in current session are represented by the difference between after-click interactions and before-click interactions.

$$\delta^{curr} m_{u,i,j} = m^{after}_{u,i,j} - m^{before}_{u,i,j}$$

F2: Return Time (Section 5.1). We use the session return time ($rt_{u,i}$) and the user-normalized session return time ($rt^{user}_{u,i}$).

F3: Change of interactions in next session (Section 5.2). It is represented by the difference between the behavior measurements m of next session and the average values of these measurements within the user.

$$\delta^{next} m_{u,i} = m_{u,i+1} - \overline{m_{u,*}}$$

As for the identification model, following previous literature [32, 45], we use a Gradient Boosting Decision Tree (GBDT) as prediction algorithm, which has good predictive power with robustness.

6.3 Evaluation and Results

There are few studies trying to identify user's negative experience, and most of previous work estimating user experience is based on only behaviors within current item [22]. We choose two methods as baseline. The first one is Sat-Click [18, 40, 41], which is the most widely used criterion to identify whether users have positive experience with the click item. The clicks followed by a dwell time less than 30 seconds are seen as unsatisfied (negative) clicks. The second one is the model proposed in Lu [8]. We use the features that can be calculated by the click signals, the dwell time and the impression position.

Considering the negative experience identification as a classification task, we measure the model performance by *Precision*, *Recall*, *F-measure*, and *AUC*.

We sequentially add the feature groups and evaluate whether each group is useful. The results are shown in Table 2. Based on the model which only uses the behavior information in current news ($F0$), we add the feature groups representing the change of user subsequent behaviors within current session ($+F1$), return time ($+F2$) and the change of behaviors in the next session ($+F3$). Using only $F0$ already performs better than the Sat-Click baseline. As we add more feature groups, the performance improves. As we add $F1$ and $F2$, the model performs better than Lu [8] baseline. When we add all groups of features, the model achieves the best performance. It proves that the change of user behaviors is useful for identifying the negative experiences.

Table 2: Results for negative experience identification.

	Precision	Recall	F-measure	AUC
Sat-Click	0.0147	0.4005	0.0283	0.5433
Lu [8]	0.6740	0.2802	0.3958	0.5723
① F_0	0.6231	0.3615	0.4575	0.5714
② F_0+F_1	0.6379	0.3550	0.4562	0.5767
③ F_0+F_1+F_2	0.6499*	0.3941*	0.4906*	0.5909*
④ F_0+F_1+F_2+F_3	**0.7729****	**0.4685****	**0.5834****	**0.6654****

The difference between ②&①, ③&②, ④&③ are tested by paired t-test (*means p-value<0.05, **means p-value<0.01).

Finding #5: Changes of users' subsequent behaviors are useful to identify the negative experiences, indicated by the improvements of model performance when adding more subsequent behaviors.

7 NEGATIVE EXPERIENCE EFFECT ON USER SATISFACTION

Most of the previous work on user satisfaction is conducted based on the explicit feedback of satisfaction. Collecting satisfaction in the real scenario is highly cost and not feasible, thus it is limited in the small-scale laboratory user study.

In this work, to evaluate the effect of negative experience on user's satisfaction (RQ3), we proposed a two-step analysis method. In the first step, we link the user behaviors to user's satisfaction using a laboratory user study data. In the second step, we expand the effects on user behaviors discovered in previous sections to the effects on user satisfaction.

7.1 Link User Behaviors To Satisfaction

User behaviors have been found to be closely related to user satisfaction, and can be used to infer user's satisfaction. Chen et al. [39] have investigated the relationship between user behavior measurements and user satisfaction in general information search. However, in recommendation scenario, the relation between user behaviors and satisfaction is not well studied. In this section, we study the relations based on a laboratory user study.

7.1.1 *User Satisfaction Collection.*

Lu et al. [8] conduct a user study in the same scenario, online news reading in the mobile environment, with the same interface, the list-style recommendation page. In the study, participants are asked to browse several lists of news (15 news each). Their satisfaction for each list (session) is collected by questionnaires after they finish browsing. In total, 32 participants complete 352 browsing tasks. Based on this public user study dataset, we study the relation between user behaviors and satisfaction.

7.1.2 *Relation Between Behaviors and Satisfaction.*

Same as our previous analysis in the log data, we use various user behavior measurements to represent user's interactions. To study the relationship between behaviors and user satisfaction, we apply the correlation analysis. Because user satisfaction feedback may be quite subjective and different users may have different understanding, we normalize the value of satisfaction ($sat_{u,i}$) into zScore according to the equation:

$$zScore_{u,i} = \frac{sat_{u,i} - Avg(Sat_u)}{Std(Sat_u)}$$

where Sat_u represents the set of satisfactions from user u.

Table 3: Pearson's r between online metrics and user satisfaction, including direct Satisfaction feedback and within-user normalized zScore, in the user study. (* means p-value<0.05, ** means p-value<0.01)

	Sat.	zScore
ClickCount	0.398**	0.304**
CTR	0.398**	0.304**
ClickDwellRatio	0.345**	0.297**
SessionDwellTime	0.323**	0.273**
SatClickCount	0.309**	0.270**
SatCTR	0.309**	0.270**
SumClickDwell	0.305**	0.279**
TimeToLastClick	0.273**	0.242**
PLC	0.215**	0.236**
LastClickPosition	0.204**	0.095
SatClickRatio	0.166**	0.181**
MaxRR	0.125*	0.040
AvgClickDwell	0.036	0.115*
MeanRR	-0.079	-0.108*
TimeToFirstClick	-0.132*	-0.067
DsatClickRatio	-0.166**	-0.181**
FirstClickPosition	-0.199**	-0.115*
MinRR	-0.268**	-0.124*
BrowsingSpeed	-0.351**	-0.359**

We calculate the Pearson's r [39, 46] between user behavior measurements and user satisfaction (both absolute values and zScore). The results are shown in table 3. The correlation between these behaviors metrics and user satisfaction proves that there exist two groups of user behavior. The metrics in the first group, like *ClickCount*, *CTR*, *SessionDwellTime*, are positively correlates with user satisfaction, namely SAT-behaviors. The metrics in the other group, like *MeanRR*, *DsatClickRatio*, *FirstClickPosition*, are negatively correlated with user satisfaction, namely DSAT-behaviors. To our best knowledge, this is the first work to study the relationship between online behavior metrics and user satisfaction in mobile news streaming and recommendation scenario.

7.2 Expand the effects on behaviors to satisfaction

In previous sections, we find that negative experience does affect following behaviors in both current session and next session. There exist two opposite effects, for example, after negative experience, the *ClickCount decrease*, while the *BrowsingSpeed* increase. The *ClickCount* have been found to be positively correlated with user satisfaction, while the speed of browsing has been found to be negatively correlated with user satisfaction. Thus, the *ClickCount* decreases may represent the decrease of user satisfaction, as well as the increase of *BrowsingSpeed*.

We summarize the results of previous analysis about the negative effects on user behaviors, shown in Table 4. As the results of *First-Position Analysis* and *Matching Analysis* which represent the negative experience effect on user's interaction in current session, most of SAT-Behaviors decrease and most of DSAT-Behaviors increase. Hence, it can conclude that after negative experience, user satisfaction decreases in current session. As for the results of the next session analysis, we find user satisfaction drops when the next session is also with negative experience (Next-Neg column). However, when the next session is not so bad (without negative experience) (Next-NonNeg column), there exists no consistent result. Some SAT-behaviors increase, like *SessionDwellTime*, while some others decrease, like *CTR*. It suggests that the effects of negative experience on user satisfaction can last to the next session, but quite depends on the performance of next session itself.

Table 4: The relation between online metrics and user experience in several scenarios, including: user study, negative experience effect on current session, negative experience effect on next session. (☻ and ☹ mean positively / negatively related to satisfaction. Down-arrow and up-arrow mean decrease / increase after negative experience. Two types of dark arrows means the results are statistically significant, while two gray ones are not. - means the metric cannot be calculated under this condition.)

	User Study	Curr-FirstPos (Fig.4)	Curr-Match (Tab.1)	Next-Neg (Fig.6a)	Next-NonNeg (Fig.6b)
BrowseNum	-	⬇	⬇	⬇	⇧
ClickCount	☻	⬇	⬇	⬇	⇧
CTR	☻	⬇	⬇	⬇	⬇
ClickDwellRatio	☻	↓	↓	⇧	↓
SessionDwellTime	☻	⬇	⬇	⬇	⇧
SatClickCount	☻	⬇	⬇	⬇	⇧
SatCTR	☻	⬇	⬇	⬇	⬇
SumClickDwell	☻	⬇	⬇	⬇	↓
TimeToLastClick	☻	-	⬇	⬇	⇧
PLC	☻	↓	⇧	⇧	⬇
LastClickPosition	☻	⬇	⬇	⬇	⇧
SatClickRatio	☻	⬇	↓	⬇	↓
MaxRR	☻	⬇	↓	↓	↓
AvgClickDwell	☻	↓	↓	⬇	↓
MeanRR	☹	⇧	⇧	⇧	⬇
TimeToFirstClick	☹	-	↓	↓	⇧
DsatClickRatio	☹	⇧	⇧	⇧	⇧
FirstClickPosition	☹	⇧	↓	⬇	↓
MinRR	☹	⇧	⇧	⇧	↓
BrowsingSpeed	☹	⇧	⇧	⇧	↓

Finding #6: Negative experiences reduce user satisfaction in current session, indicated by the decreases of SAT-Behaviors and the increases of DSAT-Behaviors in current session, and **the impact may last to the next session**.

8 DISCUSSION

8.1 Meta-evaluation of online metrics

Online metrics, calculated based on users' behavior logs, have been adopted to measure how well the system serves real users. To establish the understanding of the effectiveness of these different metrics, researchers proposed meta-evaluation approaches to investigate the relationship between online metrics and actual user satisfaction.

User's satisfaction feedback is required in traditional meta-evaluation approaches, which limits the approaches can only be used in small-scale user studies. In this section, we study whether we can use negative feedback to discover the relation between online metrics and satisfaction in large-scale real logs.

We propose a criterion based on negative experience:

- If a metric **decreases** after negative experiences, we consider it **positively** related with satisfaction.

Table 5: Relationships between behavior-based online metrics and user satisfaction, discovered based on user study (Sat.-Usr. Criterion) and our negative experience experiment in news recommendation scenario (Our Criterion), and previous studies in general search scenario [39] and image search scenario [46]. (☻ and ☹ mean positively / negatively related to satisfaction. ? means findings are not uniform.)

	Sat.-Usr. Criterion	Our Criterion	General Search	Image Search
BrowseNum	-	☻	-	-
ClickCount	☻	☻	-	-
CTR	☻	☻	☹	☻
ClickDwellRatio	☻	?	-	-
SessionDwellTime	☻	☻	☹	☻
SatClickCount	☻	☻	-	-
SatCTR	☻	☻	-	-
SumClickDwell	☻	☻	☹	☻
TimeToLastClick	☻	?	☹	☻
PLC	☻	?	☻	☻
LastClickPosition	☻	☻	-	-
SatClickRatio	☻	☻	-	-
MaxRR	☻	☻	☻	☻
AvgClickDwell	☻	☻	☹	☻
MeanRR	☹	☹	☻	☻
TimeToFirstClick	☹	?	☹	☻
DsatClickRatio	☹	☹	☹	☻
FirstClickPosition	☹	?	☹	-
MinRR	☹	☹	☻	☻
BrowsingSpeed	☹	☹	-	-

- If a metric **increases** after negative experiences, we consider it **negatively** related with satisfaction.

We combine the results of three previous experiments (Seen in Table 4: Curr-FirstPos., Curr-Match., Next-Neg). We summary the consistent results, and further give the meta-evaluation results for user behavior measurements in news reading scenario based on the above criterion, as shown in Table 5 (Our Criterion). In this work, we already have the result of meta-evaluation based on user direct satisfaction feedback in user study (Sat.-User. Criterion). It can be used as ground truth to verify the meta-evaluation criterion based on negative experience.

Through the comparison shown in Table 5, we find that most of the results from our negative feedback based criterion are consistent with the results from the satisfaction feedback based method. This demonstrates that instead of user satisfaction feedback, using negative experience can also do meta-evaluation for behavior-based online metrics. More importantly, instead of small-scale laboratory user study, it can be done by the large-scale log analysis in real environment.

We further compare the performance of behavior-based online metrics in recommendation scenario, in general search scenario [39] and in image search scenario [46]. It shows that some online metrics reflect the opposite satisfaction in different scenario. *CTR*, *SessionDwellTime*, *SumClickDwell*, *AvgClickDwell* and *TimeToLastClick* are negatively correlated with user satisfaction in general search but are positively correlated with user satisfaction in online news reading scenario. *MinRR*, *TimeToFirstClick*, *MeanRR*, *DsatClickRatio* are positively correlated with user satisfaction in image search but are negatively correlated with user satisfaction in online news reading scenario. It indicates that users' information need and behavior patterns may be different in these scenarios.

Table 6: Summary of intra-session and inter-session effects of negative experience on user behaviors and satisfaction.

	Intra-Session (After neg. exp.) (vs. non-neg. exp.)	Inter-Session (If current session has neg. exp.) (vs. has no neg. exp.)
Behaviors	Read shorter Lose activeness Leave sooner	(If next session also has neg. exp.) Lose activeness, and leave sooner. (If next session has no neg. exp.) Browse longer but with lower CTR.
Satisfaction	Decrease	(If next session also has neg. exp.) Decrease

8.2 Intra-session effects vs. inter-session effects

We have studied both *intra-session effects* and *inter-session effects* of negative experiences on user behaviors and user satisfaction. Significant impacts are observed (summarized in Table 6).

As for the *intra-session effects*, we find that after having negative experiences, users spend less time on reading the content (shorter *DwellTime*), loss activeness (less *ClickCount*) and leave session sooner (shorter *SessionDwellTime*), and their satisfaction level decreases.

As for the *inter-session effects*, users return later (longer *SessionReturnTime*) after the negative experience. When moving to the next session, users behave differently in two conditions. Firstly, if users also have negative experiences in the next session, the *inter-session effects* of negative experiences is similar with *intra-session effects*. Secondly, if users do not have negative experience in the next session, they will browse longer (longer *SessionDwellTime*) but has lower *CTR*. It indicates that user negative experiences have *inter-session effects*, but also depend on the user experiences in the next session.

9 CONCLUSIONS AND FUTURE WORK

In this paper, we have presented the first study to investigate the effects of negative experiences on both user behaviors and satisfaction in the news streaming scenario on the mobile device. By inspecting both users' *intra-session* and *inter-session* behaviors after they having negative experience, we study the question, *how negative experience affects users' behaviors*. Four observations are found to demonstrate the significant impacts.

Furthermore, we propose groups of measurements representing the change of user behaviors in different phases, and demonstrate they are helpful to the negative experiences identification. Then, we study the question, *how negative experience affects user satisfaction*, by combining the large-scale log analysis with a laboratory user study, and find negative experience will reduce user satisfaction in the current session, and the impact will last to the next session. Finally, our study shows that the meta-evaluation of online metrics can be done using negative feedbacks. In the future, how to use these explicit or identified negative experiences to help online services, such as personal recommendations, will be studied.

ACKNOWLEDGMENTS

We sincerely thank anonymous reviewers for helpful comments and Prof. dr. Maarten de Rijke for inspiring discussion. This work is supported by Natural Science Foundation of China (Grant No. 61672311, 61532011) and the National Key Research and Development Program of China (2018YFC0831900).

REFERENCES

[1] Nir Grinberg. Identifying modes of user engagement with online news and their relationship to information gain in text. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1745–1754. International World Wide Web Conferences Steering Committee, 2018.

[2] Heather L O'Brien and Elaine G Toms. What is user engagement? a conceptual framework for defining user engagement with technology. *Journal of the American society for Information Science and Technology*, 59(6):938–955, 2008.

[3] Jean Garcia-Gathright, Brian St Thomas, Christine Hosey, Zahra Nazari, and Fernando Diaz. Understanding and evaluating user satisfaction with music discovery. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 55–64. ACM, 2018.

[4] Frances A Pogacar, Amira Ghenai, Mark D Smucker, and Charles LA Clarke. The positive and negative influence of search results on people's decisions about the efficacy of medical treatments. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, pages 209–216. ACM, 2017.

[5] Ryen White. Beliefs and biases in web search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 3–12, New York, NY, USA, 2013. ACM.

[6] Ben Miroglio, David Zeber, Jofish Kaye, and Rebecca Weiss. The effect of ad blocking on user engagement with the web. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 813–821. International World Wide Web Conferences Steering Committee, 2018.

[7] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[8] Hongyu Lu, Min Zhang, and Ma Shaoping. Between clicks and satisfaction: Study on multi-phase user preferences and satisfaction for online news reading. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2018.

[9] Georges E Dupret and Benjamin Piwowarski. A user browsing model to predict search engine click data from past observations. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 331–338. ACM, 2008.

[10] Rong Pan, Yunhong Zhou, Bin Cao, Nathan N Liu, Rajan Lukose, Martin Scholz, and Qiang Yang. One-class collaborative filtering. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 502–511. IEEE, 2008.

[11] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 263–272. Ieee, 2008.

[12] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *ACM SIGIR Forum*, volume 51, pages 4–11. Acm, 2017.

[13] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS)*, 25(2):7, 2007.

[14] Yisong Yue, Rajan Patel, and Hein Roehrig. Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. In *Proceedings of the 19th international conference on World wide web*, pages 1011–1018. ACM, 2010.

[15] Hongyu Lu, Min Zhang, Weizhi Ma, Yunqiu Shao, Yiqun Liu, and Shaoping Ma. Quality effects on user preferences and behaviorsin mobile news streaming. In *The World Wide Web Conference*, pages 1187–1197. ACM, 2019.

[16] Chao Wang, Yiqun Liu, Min Zhang, Shaoping Ma, Meihong Zheng, Jing Qian, and Kuo Zhang. Incorporating vertical results into search click models. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 503–512. ACM, 2013.

[17] Jimmy Lin, Salman Mohammed, Royal Sequiera, and Luchen Tan. Update delivery mechanisms for prospective information needs: An analysis of attention in mobile users. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 785–794, New York, NY, USA, 2018. ACM.

[18] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems (TOIS)*, 23(2):147–168, 2005.

[19] Ahmed Hassan and Ryen W White. Personalized models of search satisfaction. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2009–2018. ACM, 2013.

[20] Youngho Kim, Ahmed Hassan, Ryen W White, and Imed Zitouni. Comparing client and server dwell time estimates for click-level satisfaction prediction. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 895–898. ACM, 2014.

[21] Ahmed Hassan. A semi-supervised approach to modeling web search satisfaction. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 275–284. ACM, 2012.

[22] Youngho Kim, Ahmed Hassan, Ryen W White, and Imed Zitouni. Modeling dwell time to predict click-level satisfaction. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 193–202. ACM, 2014.

[23] Peifeng Yin, Ping Luo, Wang-Chien Lee, and Min Wang. Silence is also evidence: interpreting dwell time for recommendation from psychological perspective. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 989–997. ACM, 2013.

[24] Jeff Huang and Abdigani Diriye. Web user interaction mining from touch-enabled mobile devices. In *HCIR workshop*, 2012.

[25] Qian Zhao, Shuo Chang, F Maxwell Harper, and Joseph A Konstan. Gaze prediction for recommender systems. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 131–138. ACM, 2016.

[26] Yixuan Li, Pingmei Xu, Dmitry Lagun, and Vidhya Navalpakkam. Towards measuring and inferring user interest from gaze. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 525–533. International World Wide Web Conferences Steering Committee, 2017.

[27] Thorsten Hennig-Thurau and Alexander Klee. The impact of customer satisfaction and relationship quality on customer retention: A critical reassessment and model development. *Psychology & marketing*, 14(8):737–764, 1997.

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

[29] Diane Kelly et al. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval*, 3(1–2):1–224, 2009.

[30] Henry A Feild, James Allan, and Rosie Jones. Predicting searcher frustration. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 34–41. ACM, 2010.

[31] Azzah Al-Maskari and Mark Sanderson. A review of factors influencing user satisfaction in information retrieval. *J. Am. Soc. Inf. Sci. Technol.*, 61(5):859–868, May 2010.

[32] Jiaxin Mao, Yiqun Liu, Ke Zhou, Jian Yun Nie, Jingtao Song, Min Zhang, Hengliang Luo, Hengliang Luo, and Hengliang Luo. When does relevance mean usefulness and user satisfaction in web search? In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 463–472, 2016.

[33] Ovidiu Dan and Brian D Davison. Measuring and predicting search engine users' satisfaction. *ACM Computing Surveys (CSUR)*, 49(1):18, 2016.

[34] Martijn C. Willemsen, Mark P. Graus, and Bart P. Knijnenburg. Understanding the role of latent feature diversification on choice difficulty and satisfaction. *User Modeling and User-Adapted Interaction*, 26(4):347–389, Oct 2016.

[35] Saúl Vargas. Novelty and diversity enhancement and evaluation in recommender systems and information retrieval. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 1281–1281, New York, NY, USA, 2014. ACM.

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

[37] Aleksandr Chuklin, Pavel Serdyukov, and Maarten De Rijke. Click model-based information retrieval metrics. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 493–502. ACM, 2013.

[38] Filip Radlinski, Madhu Kurup, and Thorsten Joachims. How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 43–52. ACM, 2008.

[39] Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. Meta-evaluation of online and o line web search evaluation metrics. SIGIR, 2017.

[40] Ahmed Hassan, Xiaolin Shi, Nick Craswell, and Bill Ramsey. Beyond clicks: query reformulation as a predictor of search satisfaction. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 2019–2028. ACM, 2013.

[41] Masrour Zoghi, Tomáš Tunys, Lihong Li, Damien Jose, Junyan Chen, Chun Ming Chin, and Maarten de Rijke. Click-based hot fixes for underperforming torso queries. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 195–204. ACM, 2016.

[42] Kun Kuang, Peng Cui, Bo Li, Meng Jiang, Shiqiang Yang, and Fei Wang. Treatment effect estimation with data-driven variable decomposition, 2017.

[43] Kun Kuang, Meng Jiang, Peng Cui, and Shiqiang Yang. Steering social media promotions with effective strategies. In *ICDM*, pages 985–990, 2016.

[44] Ben Miroglio, David Zeber, Jofish Kaye, and Rebecca Weiss. The effect of ad blocking on user engagement with the web. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 813–821, Republic and Canton of Geneva, Switzerland, 2018. International World Wide Web Conferences Steering Committee.

[45] Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anastasakos. Predicting user satisfaction with intelligent assistants. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 45–54. ACM, 2016.

[46] Fan Zhang, Ke Zhou, Yunqiu Shao, Cheng Luo, Min Zhang, and Shaoping Ma. How well do offline and online evaluation metrics measure user satisfaction in web image search? In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 615–624. ACM, 2018.