Between Clicks and Satisfaction: Study on Multi-Phase User Preferences and Satisfaction for Online News Reading

Hongyu Lu
Beijing National Research Center for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University
Beijing, China
luhy16@mails.tsinghua.edu.cn

Min Zhang*
Beijing National Research Center for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University
Beijing, China
z-m@tsinghua.edu.cn

Shaoping Ma
Beijing National Research Center for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University
Beijing, China
msp@tsinghua.edu.cn

ABSTRACT

Click signal has been widely used for designing and evaluating interactive information systems, which is taken as the indicator of user preference. However, click signal does not capture post-click user experience. Very commonly, the user first clicked an item and then found it is not what he wanted after reading its content, which shows there is a gap between user click and user actual preference. Previous studies on web search have incorporated other user behaviors, such as dwell time, to reduce the gap. Unfortunately, for other scenarios such as recommendation and online news reading, there still lacks a thorough understanding of the relationship between click and user preference, and the corresponding reasons which are the focus of this work. Based on an in-depth laboratory user study of online news reading scenario in the mobile environment, we show that click signal does not align with user preference. Besides, we find that user preference changes frequently, hence preferences in three phases are proposed: *Before-Read Preference*, *After-Read Preference* and *Post-task Preference*. In addition, the statistic analysis shows that the changes are highly related to news quality and the context of user interactions. Meanwhile, many other user behaviors, like viewport time, dwell time, and read speed, are found reflecting user preference in different phases. Furthermore, with the help of various kinds of user behaviors, news quality, and interaction context, we build an effective model to predict whether the user actually likes the clicked news. Finally, we replace binary click signals of traditional click-based evaluation metrics, like Click-Through Rate, with the predicted item-level preference, and significant improvements are achieved in estimating the user's list-level satisfaction. Our work sheds light on the understanding of user click behaviors and provides a method for better estimating user interest and satisfaction. The proposed model could also be helpful to various recommendation tasks in mobile scenarios.

KEYWORDS

User behavior analysis; Item-level preference; Multi-phase user preference; Click-through rate; User satisfaction

*Contact author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210007

ACM Reference Format:
Hongyu Lu, Min Zhang, and Shaoping Ma. 2018. Between Clicks and Satisfaction: Study on Multi-Phase User Preferences and Satisfaction for Online News Reading. In *SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3209978.3210007

1 INTRODUCTION

Learning from user behaviors is a general approach used in online interactive information systems. Clicking on an item has been commonly used as positive implicit feedback of user preference to design systems and evaluate their performance. For example, most recommender systems use click as implicit preference feedback and take Click-Through Rate (CTR) as the optimization target and the online evaluation metric.

Although user click provides implicit information about item-level user preference, it may not represent true preference. On the one hand, besides preference, click behavior may be affected by many other factors, like position [18], trust [34] and presentation [32]. To address these issues, researchers have proposed a number of click models to describe user click behavior in web search scenario [5, 10]. On the other hand, click signal does not capture post-click user experience. User may have clicked an item because of the titlebait, but dislike it after reading its content. For example, a user clicks an item because attracted by its title, "*You Won't Believe What This Guys Does After His Set...*", but turn out to be disappointed with its poor content and result in a negative preference for it. In such cases, using click as positive preference indicator will mislead system in modeling user interest.

Some previous works have realized this issue and have incorporated some other user behaviors to reduce the gap between click and user experience. For example, dwell time (i.e., the time that user spends on a clicked item) has been found well correlated with item-level user satisfaction in information retrieval (IR). Click followed by a long dwell time has traditionally been seen as satisfied click and been successfully used in a number of retrieval applications [11]. Besides dwell time, recent work also tries to learn from other behaviors, like mouse movement [27], scroll information [23], and gaze [1, 25].

However, as for the subjective user experience, it still lacks a thorough understanding of how and why click is not aligned with user preference. In the above example, the user may be interested and expect to like the item when he sees its title, but after reading the news, his/her preference for the item has changed because of the low content quality. Thus, it is necessary to model user preference in different phases.

In this paper, we conduct an in-depth user study in online news reading scenario in the mobile environment, in which we collect

user behavior logs as well as user explicit preference feedback for the news (Section 3). Through comparing the click signal and the item-level user preference, it is observed that click signal are not always aligned with user preference. More than half of the clicked news is disliked by user (Section 4). In addition, we propose user preferences in three phases: the first one is *Before-Read Preference* collected right after users click but before reading the content, the second one is *After-Read Preference* collected right after users finishing reading the content, the third one is *Post-Task Preference* which is context-independently collected after users finishing the task. The statistic analysis of the differences between these multi-phase preferences shows that the changes of preference are highly related to the quality of news and the context of user interactions (Section 5).

Furthermore, we investigate how user behaviors, like viewport time, dwell time and scroll patterns correlate with preference (Section 6). We find that different behaviors represent preference in different phases. For example, the viewport time is correlated with user *Before-Read Preference*, while dwell time is more correlated with user *After-Read Preference*. Based on these observations, by incorporating various kind of user behaviors, news quality, and interaction context information, we build an effective model to predict user's actual preference for the clicked item. Finally, we replace binary click signals to predicted preferences in common click-based online metrics, like CTR, and obtain a significant improvement for estimation of user list-level satisfaction (Section 7).

To sum up, we have made the following contributions:

- To the best of our knowledge, this is the first work which conducts an in-depth user study about the gap between user click and preference in online news reading and recommendation scenarios. We demonstrate that click is not always aligned with user actual preference, and the gap is related to the change of preferences in different phases.
- The analysis on different reading phases show that the news quality and the user interaction context are related to the change of user preferences. Furthermore, different user behaviors, like viewport time, dwell time and read length, are found reflecting preference in different phases.
- Based on these findings, a novel preference prediction model is proposed to predict user actual preferences of the clicked items, in which various user behaviors, news quality, and interaction context are taken into account. Furthermore, a significant improvement of measuring user list-level satisfaction is achieved by incorporating the predicted preference in traditional click-based online metrics, such as CTR.

2 RELATED WORK

In this section, we review two related research directions, the modeling of click behavior and item-level user experience.

2.1 Modeling of Click Behavior

Click behavior has been widely used as essential implicit feedback in interactive information systems. Researchers of IR use click signal to infer document relevance by modeling the relationship between user click behaviors and relevance by click model [10]. Besides, some personalized information filtering systems, like recommender systems, are designed by mining user preference through historical click data [14, 30]. Moreover, click signal is also used for evaluation by calculating several online evaluation metrics, like CTR, UCTR [7] (the binary value representing click) and PLC [4] (the number of clicks divided by the position of the lowest click).

These approaches are based on the assumption that user clicks reflect users' actual experiences, like relevance, preference, and satisfaction for the clicked items. However, previous studies have found that this assumption may be unreliable. Firstly, click behavior is biased by many factors, such as position [18]. Top documents may attract more clicks [19]. Other factors like trust [34] and presentation [32] are also examined. Secondly, click, which usually happens before user examining the content of the items, lacks post-click information [14].

To address the biases of user click behavior, researchers have proposed a number of advanced click models [5, 12], which are designed to eliminate the effects of various biases to obtain an unbiased estimation of result relevance in web search scenarios. Besides, to compensate the missing post-click information, other user behaviors are applied to reduce the gap between click and user actual experiences. Among them, click dwell time has been successfully used in many retrieval applications. A dwell time equaling or exceeding 30 seconds has typically been used to identify clicks with which searchers are satisfied [11]. The correlation between dwell time and user interest is further modeled by document factors (e.g. readability [21], and human factors [33]). Besides dwell time, viewport time is successfully used on mobile devices to infer user interest at the sub-document level [15]. Yixuan et al. [25] examine the correlation between users' eye gaze and user explicit interest and demonstrate the effectiveness of using attention-based behaviors, viewport time and gaze, to predict user interest on mobile.

Previous works mostly focus on modeling user implicit behaviors to bridge the gap between click and user experience, while less effort is made to investigate that how and why the gap exists directly. By leveraging user preference collected in three different phases, we conduct an in-depth analysis of the relationship between click and user preference.

2.2 Modeling of Item-level User Experience

Beside of studying on user behaviors, there are also a few work aim to model user's item-level subjective experience, such as usefulness in web search and preference in recommender systems. Belkin et al. [3] argue that usefulness, which represents users' perceived value of a search result, depends on the scenario and context of accessing the result. Mao et al. [29] also find that usefulness is related to current search task and redundancy with previous documents read by the user. Jiang et al. [17] have collected usefulness judgment not only in situ stage within the context but also post-session context-independent judgment, and find that they are different, which indicates that context will affect user perception of usefulness. Usefulness in these works has been modeled as a one-phase static concept. Through comprehensively inspecting preference in different phases, we model the user preference as a multi-phase dynamic concept and give an in-depth analysis of the change of preference after clicking.

Benefited from modeling user experience, the performances of experience estimation methods are improved by using user behaviors along with various factors. Mao et al.[28] conduct a study to improve the performance of behavior-based usefulness prediction model by incorporating content factors (e.g., the similarity between document and query) and context factors (e.g., the average usefulness of previous pages). In these works, the quality has been less considered. However, from our statistic analysis, we find the quality

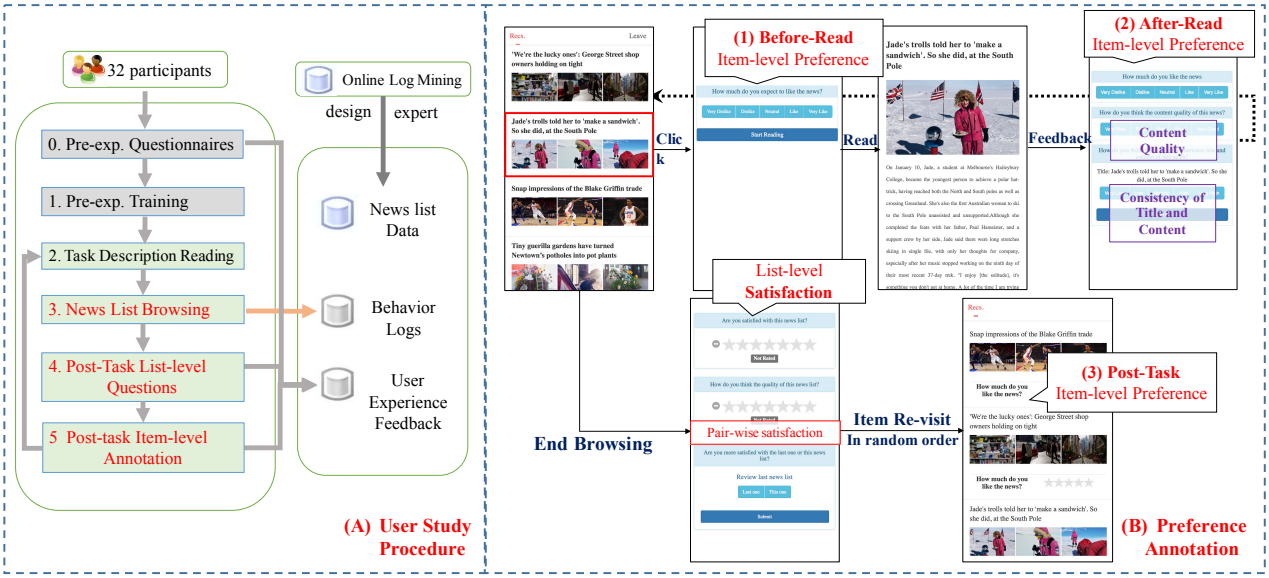

Figure 1: User study procedure. We collect user behavior logs and user experience feedback in the user study(A). Questions for collecting item-level user preference in different phases and list-level satisfaction are injected into user study(B).

of news is highly related to the change of preference and can be applied to boost user preference prediction.

The estimated item-level user experience has been used for the recommendation and online evaluation. Yin et al. [33] develop a model to interpret the dwell time to "pseudo vote," which represents user preference. By incorporating these predicted preferences, traditional recommendation achieves great improvements. As for evaluation in IR, Belkin et al.[2] and Cole et al.[8] propose an idea that replacing relevance-based measurements with usefulness-based ones. Mao et al. [29] demonstrate the improvement of using the usefulness predicted by user behaviors in traditional click-based evaluation metrics. Chen et al. [6] show that online metrics are better aligned with user satisfaction when using item-level gains estimated based on dwell time and mouse hover information. However, the user information need is quite different in different domains, such as IR and recommender systems. Thus document usefulness and user preference may have different correlations with user satisfaction. Therefore, we investigate whether traditional click-based evaluation metrics for recommender systems can be improved by incorporating estimated item-level user preference.

3 USER STUDY METHODOLOGY

In this section, we describe the settings of the user study and the dataset we collected. As mobile phones are becoming the main tools for online reading, all data involved is collected in mobile scenarios.

3.1 Experimental Procedure

We design a laboratory user study to collect user interaction logs and experience feedback simultaneously. To simulate a real online news reading environment, we build an experimental news reading platform (Android application). The user interfaces are similar to the common online news reading applications except for three labeling steps. The main page of the platform is the news browsing page which is a one-column list-style interface. The list contains several news snippets including news title and three thumbnails. The user can scroll to see the full list. When clicking on a snippet, the user jumps to the news full content page. The content pages of news are collected from a commercial news recommender system. To control the variability, the source, publish time, and advertisements are removed from the page. A javascript plugin is injected into both list and content pages to record user's browsing and reading events including scrolling, clicking and page switching.

The news used in the experiment is randomly sampled from the high and middle reading frequency news in the real user logs of a commercial news recommendation website. To control the news quality, we recruit an expert to annotate the overall quality (consider not only the content but also the title) of each news. We randomly sample 165 news with high-quality annotations and then randomly assign to the tasks. To investigate the influence of news quality, we then randomly replace some of the news in the list with low-quality ones. Eleven news reading tasks are generated before the experiment and are same for each participant. Each task contains 15 unique news, and there is no overlap with other tasks.

We recruit participants to read the news using this platform on the same mobile we prepare (1280*720 pixels). The procedure of user study is shown in Figure 1A. Before the experiment (Stage 0), we collect user's preference for five topics, namely social, enterprise, technology, history and sport, which cover all news used in our experiment (1-5 stars). To make sure that every participant is familiar with the experiment procedure, a training task is used for demonstration in pre-experiment training (Stage 1). At the beginning of each task, each participant is asked to read the task description which contains the scenario information and some cautions (Stage 2). After that, the participant is impressed a list of news to browse and read (Stage 3). While no browsing and reading time limits are imposed, he/she can stop at any time. After participants choosing to finish browsing, they are asked to answer some questions about the whole list including satisfaction for the list and pair-wise satisfaction between two adjacent lists (Stage 4). Then, participants are asked to annotate their preference for each news in the list (Stage 5). Note that Stage 2-Stage 5 are same and repeated in each task. Each participant is demanded to complete all of the eleven reading tasks, and the tasks are in a random order.

Table 1: Statistics of user study data

	#users	#tasks	#news	#clicks
Stage 1: for analysis	26	286	4290	1337
Stage 2: Testset for prediction	6	66	990	266

We recruit 32 participants (18 were female) to take part in the user study. The data of the first 26 participants (Stage 1) is used for analysis (see Section 4-6), while the data of the remaining 6 participants (Stage 2) is used as the test set to evaluate the prediction experiment (see Section 7). The descriptive statistics of the dataset in two stages are shown in Table 1

3.2 Multi-Phase Item-level Preferences Annotation

For thoroughly investigating user preference, we inject some questionnaires within the task procedure to collect the user preference for the news in multi-phase.

Before-Read Preference *Q: How much do you expect to like the news?*

Firstly, as soon as the user clicks a news snippet, we ask for his/her expected preference for the news, before he/she see the content. We name this phase as *Before-Read phase* and the preference in this phase as *Before-Read Preference*.

After-Read Preference *Q: How much do you like the news?*

After the user decides to end reading the content of news, we ask a few questions about his/her experience in reading the news. User preference for the news is asked again in the first question. The user perceived content quality and the consistency of title and content are collected subsequently (5 levels each). We name this phase as *After-Read phase* and the preference in this phase as *After-Read Preference*.

Post-Task Preference *Q: How much do you like the news?*

Note that these two item-level preference feedback are only for clicked news, and are collected within the context of the list and user browsing sequence, which may be influenced by the position, surrounding news, and previous read news. To remove these effects, after user finishing the task-level questionnaires, we shuffle all the news shown to the user, not only what he/she clicks but also what he/she not clicks, then ask the user to give his/her actual preference for each news. By doing this, the bias of position and context are all removed. Thus, preferences collected in this phase are considered as ***the users' actual preference*** for each news. We called this phase as *Post-Task phase* and the preference in this phase as *Post-Task Preference*.

Through the user study, user behavior logs, user explicit feedback for item-level preference in different phases, and user satisfaction for each list are collected. The major measures used in this work are summarized in Table 2

4 CLICK SIGNAL V.S. ITEM-LEVEL PREFERENCE

Through the user study, we have collected user preference, as well as user click signals for each news. We first investigate the correlation between user click and user preference for individual items. The user context-independent preference feedback in *Post-Task* phase are used as the ground truth of item-level preferences, to which the implicit feedback based on click signals for each news, will be compared.

Table 2: Description of major measures used in this work. (U:user, E:expert)

	Measures	Labeled by	Stage	Scales
Topic-level Preference	Pre-Exp Topic Pref.	U	0	5
Item-level Preference	*Before-Read Preference*	U	3	5
	After-Read Preference	U	3	5
	Post-Task Preference	U	5	5
Item-level Quality	Overall quality	E	-	2
	Content quality	U	3	5
	Consistency of title and content	U	3	5
Satisfaction	List-level satisfaction	U	4	7
	Pairwise satisfaction of adjacent list (last v.s. current)	U	4	2

4.1 Does click represent preference?

Traditionally, user click signals of items have been used as user's preference in recommender systems and online news reading scenarios. To find out whether click represents preference, we split items into two groups, clicked and not clicked, and show the distribution of user preference in each group (shown in Figure 2a).

We can see a significant difference in user preference between non-clicked and clicked items (t-test, unpaired two-sample, p-value $\ll 0.001$). For non-clicked news, the preference distribution is almost on the low side, nearly 38% non-clicked items are annotated as strong disliked (=1). It indicates that user dislikes most of the non-clicked items. Compared with non-clicked news, user preferences for clicked news are higher in general. However, in the clicked news there is still more than 15% news are annotated as strong disliked, which is contrary to the traditional assumption that users like the clicked news.

We further split the preference into two groups, disliked and liked, as *Post-Task Preference* ≤ 3 and > 3 respectively. By jointly analyzing whether the news is clicked or not and whether the news is liked or not, it is concluded that what user clicked may not be what he/she liked. As shown in Figure 2b, 58% clicked news user find dislike. Although click-or-not signal has a moderate correlation (PCC=0.363, p-value<0.001) with preference, there is still a clear gap between them.

4.2 Why Click and Preference are Different?

To find out why the user dislikes the clicked items, we collect user preferences in several intermediate phases after clicking. By inspecting these multi-phase preferences, we study the gap between click and user final preference.

As described in Section 3, there is two explicit preference feedback collected in the middle phases: *Before-Read Preference* and *After-Read Preference*, along with user unbiased actual preference collected in *Post-Task phase*. The *Before-Read Preference* is collected close to the click event by asking users their expected preference. The *After-Read Preference* is collected right after user read the news, which reflects user immediately preference for the recently-read news. The *Post-Task Preference* is re-rated by the user in a random

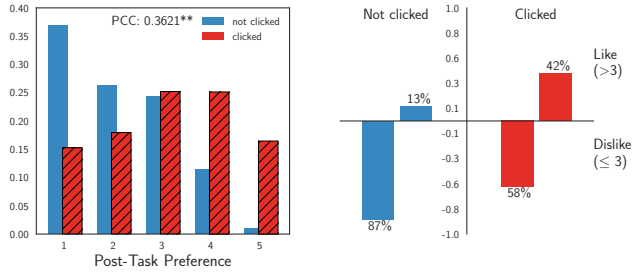

Figure 2: Distribution of *Post-Task* item-level preference in clicked/not clicked news (left). The ratio of user like and dislike in clicked and not clicked news (right).

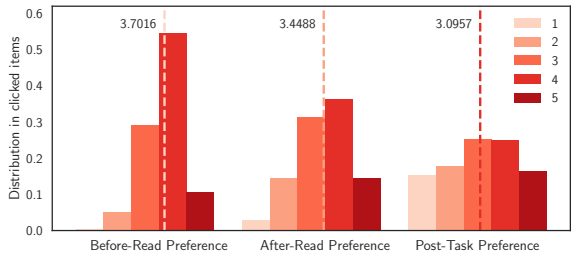

Figure 3: Distributions of the multi-phase item-level preferences of clicked news. The lines indicate the means of the preferences.

order after browsing the list and is regarded as the actual preference because the influence of context information is removed.

This three item-level feedback measure the user preference for an clicked news in three different phases. Through them, we investigate the changes of user preference after clicking. The distribution of multi-phase item-level preferences for clicked news are shown in Figure 3. A significant difference can be seen between these three distributions. The *Before-Read Preference*, which is mostly close to the click, are more concentrate and almost 65% news are user expected to be liked by the user. However, after user read the news, his/her preference may change. We can find that the distribution of *After-Read Preference* is a few more disperse and remain nearly 51% are user annotated as liked. A significant difference between *After-Read Preference* and *Post-Task Preference* which is user context-independently actual preference. The *Post-Task Preference* is more disperse than other intermediate preference. The ratio of strong disliked and strong liked parts in *Post-Task phase* is much higher than those in *Before-Read phase* (31.7% v.s. 11.2%), which indicates that user's preference for a clicked news is more polarized.

From the distribution, we also find the preference declines from *Before-Read phase* to *Post-Task phase*. To inspect this phenomenon, we list the means of preference of different phases in Table 3. The results show that in general, considering all the news user clicked, the preference has a significant decline. Furthermore, we introduce the factor of news quality, which contains three measurements: the user perceived content quality and the title consistency and the expert labeled overall quality. It is clear that user preference for the low-quality news, especially user perceived quality, declines much faster. This phenomenon prompts us that the quality of news may be an essential factor related to the change of preferences.

4.3 Click V.S. Before-Read Preference

In the last section, we show that user preference for the clicked items is not fixed but changing along user further interacting with the item. In this section, we investigate whether user expected preference just after clicking is aligned with the clicked signal.

Table 3: Mean of the item-level preferences along the three phases generally declines. As for low-quality news (based on user perceived content quality, consistency of title and content, and expert labeled overall quality), the decline is greater.

		Before-Read Pref.	After-Read Pref.	Post-Task Pref.
General	all	3.7016	3.4788	3.0957
User perceived: content quality	<=3	3.5017	2.6582	2.1633
User perceived: title consistency	<=3	3.4677	2.7883	2.1613
Expert labeled: overall quality	low	3.5154	2.8567	2.4539

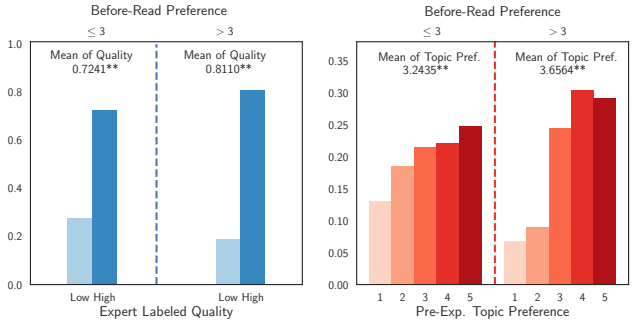

Figure 4: The news expected to be disliked in *Before-Read phase* are of lower quality (left), and more likely from the topic user less liked (left).

From Figure 3, there are more than 65% clicked news expected to be liked by the user. It is surprising to see that there are nearly 35% clicked news user expected to dislike. We manually inspect the clicks with low Before-Read expected preference (*Before-Read Preference*$\leq$ 3). As shown in Figure 4, user expected preference before reading the content is related to the news quality and user prior topic preference.

Through zooming in the news which user clicked but expected to be disliked, we find that the quality of this news is slightly but significantly lower than the news user expect to like. This phenomenon indicates that before user clicking a piece of news, the quality can already be slightly perceived by the user who has read the title and the presented images. Furthermore, the user's prior topic preference of the news which expected to be disliked is also significantly lower.

4.4 Summary on Observations

To sum up, this section analyzes the relationship between user click signal and item-level user preference, and find that:

- Click is not always aligned with user preference, more than half of the clicked items are disliked by the user.
- User preferences for the item in different phases are different, which indicates that preference is not fixed but changing after clicking.

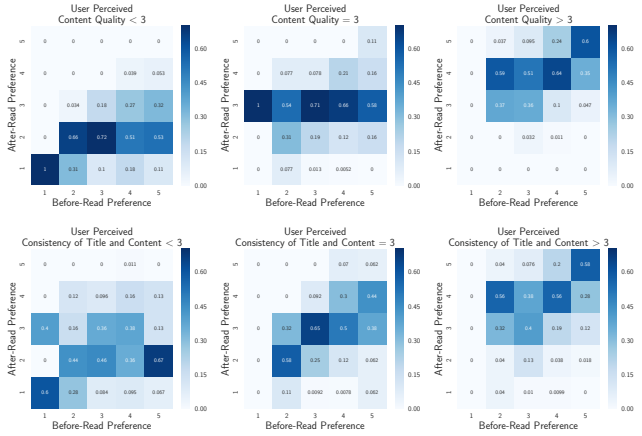

Figure 5: Changes of item-level user preference (from *Before-Read Preference* to *After-Read Preference*). Upper figures split the news by user perceived content quality, while lower figures split news by user perceived consistency of title and content.

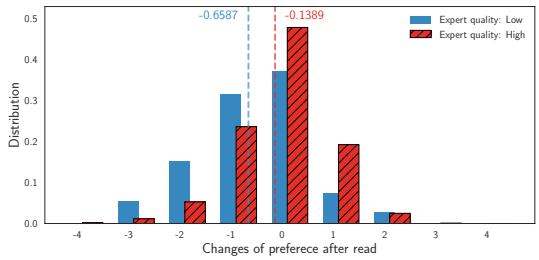

Figure 6: The distribution of the change between *Before-Read Pref.* and *After-Read Pref.* (*After-Read Pref.* minus *Before-Read Pref.*), split by expert labeled overall quality. The lines indicate the means, the difference between them are significant (p-value<0.01)

.

- There is already a gap between click and user expected preference right after clicking.

5 STUDY ON USER PREFERENCE CHANGES IN DIFFERENT PHASES

Beside of the gap between click and user expected preference in *Before-Read* phase, there are still two gaps from *Before-Read Preference* to *After-Read Preference*, and from *After-Read Preference* to *Post-Task Preference*, which will be comprehensively analyzed in this section.

5.1 From Before-Read Preference to After-Read Preference

We are interested in understanding how user preference changes after the user read the content for a clicked news, which is reflected by the difference between user *Before-Read Preference* and *After-Read Preference.*

A significant difference between *Before-Read Preference* and *After-Read Preference* (t-test, paired two-sample, p-value$\ll$0.001) can be seen in Figure 3. For *Before-Read Preference*, people show their preferences in nearly 65% clicked news (rating over 3), while only 51% clicked news are preferred in *After-Read phase.* This indicates a considerable proportion of the news are preferred before read but disliked after read. Meanwhile, the ratio of most dislike (=1) and most like(=5) increases, which indicates that users have a more polarized opinion after read the content.

The consistency between *Before-Read Preference* and *After-Read Preference* is tested by Cohen's weighted kappa k (k=0.2813, linear weighted), and reaches a fair agreement level. We further test the correlation between two preferences using Pearson's correlation coefficient r. A moderate positive correlation is detected, $r(1, 337) = 0.3905$, p-value$\ll$0.001. To further inspect the difference between two preferences, we incorporate two quality factors: 1) the user perceived consistency of title and content; 2) the user perceived content quality. For the first factor, users only read the title of news in *Before-Read* phase, while the *After-Read Preference* is collected after user reading the content. So we investigate whether the consistency of title and content correlates with the change of preference. The user preference for the item of low title consistency ($\leq$ 3) is more likely to change than the item of high title consistency (63.5% v.s. 49.1%). The effect of the consistency is tested by ANOVA [9] (consistency~change or not, p-value$\ll$0.001, one-way).

For the second factor, we inspect if the news content quality relates to the change of user preference. The joint distribution of the *Before-Read Preference* and *After-Read Preference* is plotted in Figure 5-upper. We classify the user perceived content quality into three groups, 1 & 2 as low, 3 as moderate, and 4 & 5 as high content quality. The results show that user preference for the low content quality news is more likely to decline after reading the content (85% news declined). Meanwhile, even users have already shown higher preferences for the high-quality news in *Before-Read Phase*, the preferences in *After-Read Phase* will still increase. It indicates that high content quality news will be helpful in increasing user's preference. We also plot the results of considering the consistency of title and content. A similar trend is found (Figure 5-lower). News with good consistency of title and content will attract more preference too.

Although user perceived quality has a high correlation with the changes of preference from *Before-Read* to *After-Read* phase, it is unrealistic to collect explicit perceived quality from the user. Thus, we test if this correlation still exists when using the expert labeled quality results. The distribution of user preference changes when the news is of low or high quality (expert labeled) is shown in Figure 6).

Compared with high-quality news, user's preference for low-quality news has the higher probability of decline (60% v.s. 55%). And the mean of the vary also shows that user preferences for low-quality news declines more after reading. Besides, the trend is similar to using user perceived quality. These indicate that it is able to use the quality annotated by experts instead of the user perceived quality (which is hard to collect) in practical scenarios.

5.2 From After-Read Preference to Post-Task Preference

In this section, we further investigate whether the preference in *After-Read* phase is consistent with user preference in *Post-Task* phase.

In Figure 3, the results show that there is significant difference between item-level user preference in *After-Read* and *Post-Task* phase (t-test, paired two-sample, p-value$\ll$0.001). Users answer

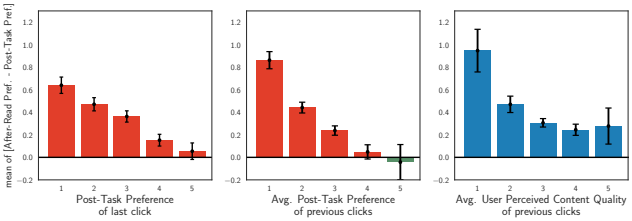

Figure 7: When the user preference for last clicked news (left) is lower, user's *After-Read Preference* for current reading one is significant higher than his/her actual preference. Similar observations are found in the average preference (middle) and the average perceived content quality (right). (the error-bar indicates the standard error).

their *After-Read Preference* for the news in the context of the previous read news, while the *Post-Task Preference*, which is collected after user finish the reading task in a shuffled order, is context-independent. The effect of the context is calculated by *After-Read Preference minus Post-Task Preference*, which can be used to analyze the difference between *After-Read Preference* and *Post-Task Preference*.

Users sequentially click and read the news in the result list, so previous news which user has read may have some effects on user's experience when reading current news, which will result in a misleading preference in *After-Read phase*. We show the changes of *After-Read Preference minus Post-Task Preference* (which shows the differences between *After-Read Preference* and *Post-Task Preference*) along with three context factors in Figure 7.

From the figures, it can be seen that user's preference for the last one or previous clicked news has an impact on user experience of the currently reading news. When the user just read a news he/she does not like, his/her preference for current news in *After-Read phase* is higher than the actual preference. A similar trend appears when the user perceives lower quality news in previous clicks (right), he/she will have an over-high preference for this news in *After-Read phase*. These findings suggest that user interaction context will influence user's immediate preference for the news, which may be inconsistent with his/her actual preference.

5.3 Summary on Observations

In this section, we focus on studying how user preference changes in difference phases, and find that:

- Compared with preference in *Before-Read phase*, user may change his/her preference after reading the content, and the change is related to news quality.
- User *After-Read preference* may be affected by the interaction context, such as user preference for the last or previous read news, and the quality of previous read news.

6 USER BEHAVIOR V.S. MULTI-PHASE PREFERENCES

To better understand user's subjective preference behind observed user behaviors, the correlations between different user behaviors in browsing and reading process and item-level user preference in different phases are investigated.

Table 4: User behavior metrics of the news user liked or disliked in different phases. (*means p-value<0.05, **means p-value<0.01).

	Before-Read Pref.		*After-Read Pref.*		*Post-Task Pref.*	
	<=3	>3	<=3	>3	<=3	>3
viewport time (ms)	6851**	6201**	6415	6439	6522	6293
dwell time (s)	31.33**	38.57**	28.66**	43.72**	30.43**	44.56**
read length (pixel)	6987*	7707*	6113**	8753**	6312**	9068**
read ratio	0.797	0.811	0.776**	0.835**	0.787**	0.833**
read speed (pixel/s)	309.9**	250.5**	296.9**	246.2**	294.8**	237.7**
max scroll interval (s)	5.22*	6.74*	4.58**	7.78**	4.83**	8.15**
direction change times	22.85**	28.25**	19.49**	30.02**	20.38**	34.83**

Table 5: The Pearson correlations between different user behavior metrics and item-level user preference in different phases. (*means p-value<0.05, **means p-value<0.01).

	Before-Read Pref.	After-Read Pref.	Post-Task Pref.
viewport time (ms)	**-0.0635***	-	-0.0515*
dwell time (s)	0.0686*	**0.2797****	0.2611**
read length (pixel)	0.0738**	**0.2770****	0.2534**
read ratio	0.0608*	**0.2693****	0.2142**
read speed (pixel/s)	-0.0725*	-0.0789**	**-0.0924****
max scroll interval (s)	-	**0.0757****	0.0747**
direction change times	0.1070**	0.2249**	**0.2316****

We record user scroll and click events with timestamps in both list and content pages, then calculate several behavior metrics to represent user browsing and reading process.

To investigate whether user behaves differently when he/she likes or dislikes the news in multi-phase, we first separate preference feedback into two parts, like(>3) and dislike(<=3). Then, we compare each behavior metrics of two parts to find out whether there is a significant difference between them. The means of each behavior metrics when the user like or dislike the news in the *Before-Read*, *After-Read*, and *Post-Task* phases are shown in Table 4.

We now zoom in to examine user behaviors in terms of individual metrics. Viewport time represents how long user read a news snippet in the list page. We find that compared with the news user expect to like in *Before-Read phase*, viewport time of the news disliked is longer. This indicates that user may spend more time to make a decision for clicking a news he/she expects to not like so much. When it comes to user preference in *After-Read* and *Post-Task* phases, there is no significant difference in viewport time. The viewport time is more likely reflecting user preference in *Before-Read phase*.

As for dwell time, which represents how long user read the news content, we find a significant difference between user like or dislike a news in all three phases. It is reasonable to see that user may spend more time to read the news he/she likes. And the difference

is larger in *After-Read* and *Post-Task* phases than in *Before-Read phase*. This indicates that dwell time is more likely to reflect user preference after reading its content.

Besides dwell time, we also calculate the read length and read ratio, which indicates how much user reads the news content and are proved to reflect user engagement[24]. Although these metrics are not significantly different between like and dislike conditions in *Before-Read phase*, We find a significant difference in *After-Read* and *Post-Task* phases. The user will read more in the news he/she likes. Moreover, we combine the dwell time and read length to calculate the reading speed, and find a significant difference. For the news user liked, he/she will read slower.

To further inspect user reading behaviors, we analyze two scroll patterns: the max scroll interval, and the times user change his/her scroll directions. The results show that in the news user likes, the max scroll interval which may reflect whether user has ever carefully read some content of the news, is much higher than in the news user dislike, especially in *After-Read* and *Post-Task* phases. Meanwhile, user will change his/her reading direction and will revisit previous content more times in the news he/she likes.

We further analyze the correlation between these behavior metrics and user multi-phase preferences. From the results shown in Table 5, firstly, we find that viewport time is more correlated with user preference in *Before-Read phase*, while dwell time, read length/ratio/speed, and scroll patterns are more correlated with user preference in *After-Read* and *Post-Task* phases. These results indicate that different behaviors may reflect user preference in different phases.

7 PREFERENCE PREDICTION

Previous sections show that the news user clicks may not be the one he/she likes. Further, it finds that the news quality and interaction context are related to the gap between user click and preference. Moreover, several browsing and reading behaviors are found reflecting user preference. Thus, in this section, we attempt to use user various behaviors, along with quality and context to predict user actual preference for a clicked news.

7.1 Experiment Settings

We use user preferences collected in *Post-Task phase* as the ground truth and define two prediction tasks. The first one is a supervised classification task to predict whether a clicked news user likes or not, called Liked-Click prediction. We divide the preference rating into two labels. Rating 4 and 5 are set as liked (41.6% of train set) and the remainder as disliked. The second one is a regression task to predict how much the user prefers a clicked news which is labeled by original preference rating ranging from 1 to 5.

The dataset from the first stage of user study is used for analysis and training, which includes 1337 unique clicked news from 26 participants. While the dataset from the second stage of user study is used for testing, which including 266 unique clicked news from 6 participants. Note that previous analysis does not use the test data, which ensure our evaluation for the prediction model is reliable. All the results reported are of the test set.

7.2 Features & Model

Table 6 summarizes the features extracted from user behavior logs. We categorize these features into three groups: Behavior features (F_b), Context Features (F_c), Quality features (F_q).

Table 6: Features to predict item-level preference

Behavior features F_b	
B1	Viewport time
B2-B3	Dwell time; Normalized dwell time (in user)
B4-B5	Read -length; -ratio
B6	Read speed
B7	Max scroll interval
B8	Direction change times
Context features F_c	
C1-C4	Dwell time; Read-legth /-ratio /-speed of last click
C5-C8	Average dwell time; Read-legth /-ratio /-speed of previous clicks
EQ	Expert labeled quality
Quality features F_q	
Q1	Image num
Q2-Q3	Content / title length
Q4	Stopword num in title
Q5	Similarity of title and content

Behavior features are generated from user browsing logs. These features, such as viewport time, dwell time, read length, and scroll patterns, describe how users interacted with the news in both list and content pages, and are found related to user preference in multi-phase. Beside of user behaviors, the previous analysis shows that the interaction context affects user preference in *After-Read phase*, for example, how user likes the previous clicked news. However, it is hard to collect the preferences for the previous clicked news in real scenarios. We use some user behaviors to represent preference implicitly. As for quality features, we use expert labeled overall quality. Because of the hard collection of news quality, we also generated some content features of the news to represent it, such as the number of images, the length of the title, the number of stop-words in the title, and the similarity between title and content (calculated by cosine similarity based on pre-trained word embedding).

Following with previous literature[22, 29], we use a Gradient Boosting Decision Tree (GBDT) as prediction algorithm, which is able to naturally handle mixed types of features, and has good predictive power.

7.3 Preference Prediction Results

For Liked-Click prediction (binary classification) task, we measure the model performance by precision, recall, and f-measure for the positive class, and overall accuracy.

Two basic baselines are used in the prediction experiment. The first one regards all clicked news as user liked, named as Bin-Click, which is the traditional usage of the click signals. The second one is based on the common approach in the literature[11, 13, 17] as combining dwell time information, named as Sat-Click. The clicks followed by a dwell time of a minimum $t_{threshold}$ seconds are seen as liked clicks. In this study, we set $t_{threshold} = 52$ seconds which has the best discriminatory ability in the training set.

The results are shown in Table 7. Bin-Click reaches the lowest accuracy. The Sat-Click baseline which additionally includes dwell time information performs better than Bin-Click.

As for our prediction models, we sequentially add the feature groups and evaluate whether each group is useful. Only using the behavior features (F_b), our model already performs better than

Table 7: Results for Liked-Click Prediction (classification).

	Precision	Recall	F-measure	Accuracy
Bin-Click	0.4248	1.0000	0.5963	0.4248
Sat-Click	0.5000	0.3008	0.3757	0.5751
① F_b	0.5570	0.4782	0.5133	0.6164
② F_b+F_c	0.5705**	0.4676	0.5128	0.6239**
③ F_b+F_c+F_q	0.5733	0.5007**	0.5338**	0.6293*
④ F_b+F_c+EQ	**0.5818****	**0.5111****	**0.5431****	**0.6358****

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

Table 8: Results for *Post-Task Pref.* Prediction(regression)

	MSE	MAE	PCC
Sat-Click	-	-	0.1400
① F_b	1.2116	0.9099	0.2873
② F_b+F_c	1.1683**	0.8914**	0.3291**
③ F_b+F_c+F_q	1.1587*	0.8890	0.3475**
④ F_b+F_c+EQ	**1.1331****	**0.8789****	**0.3548****

Bin-Click and Sat-Click baselines. As we add more features, the performance of Liked-Click prediction increases, which proves that the context and quality information is useful in preference prediction. Moreover, compared with using some content features to present quality (F_q), directly using expert labeled overall quality (EQ) performs better. This indicates that if we can find more information related to the news quality, the performance of prediction model can still increase.

For preference regression task, we measure the model performance by MSE, MAE and Pearson'r based on the *Post-Task Preference*. Note that the Bin-Click baseline predicts the same for all clicked samples, which cannot be evaluated by these regression metrics. The results are shown in Table 8 and have the same trend as Liked-Click prediction experiment. The model with all features performs best, and replacing quality features with expert labeled quality will further improve the performance.

7.4 Comparison of Click and Predicted Preference with User List-Level Satisfaction

We further demonstrate the validity of preference prediction approach by showing the correlations between evaluation metrics, based on predicted preference labels or commonly used binary click signals, with list-level satisfaction.

We used several general approaches [16, 26] to accumulate the item-level preference indicators to list-level measures, $CG/\#imps$, $CG/\#clicks$, CG/pos_{lc}, and DCG. Cumulative gain (CG) measures the total gain or utility of the list. It is calculated by summing up the item-level measures for all clicks in the list:

$$CG(M) = \sum_{i=1}^{|CS|} M(n_i)$$

Here, $CS = (n_1, n_2, \ldots, n_{|CS|})$ is the click sequence in which each element n_i is a clicked item. $M(n_i)$ is the gain for item n_i. In this section, M can be either binary click-or-not signals (Bin-Click), saitsfied click (Sat-Click) or preference predicted by regression model (Predicted preference). $CG\#imps$ is the average gains per impression, $CG\#clicks$ is the average gains per click, and CG/pos_{lc} is CG divided by the position of the lowest click. When we are using Bin-Click as M, the $CG\#imps$ and CG/pos_{lc} is same as the common used online metrics CTR and PLC [4] respectively. Discounted cumulative gain (DCG) is defined as:

$$DCG(M) = \sum_{i=1}^{|CS|} \frac{M(n_i)}{log_2(i+1)}$$

Based on user list-level satisfaction feedback, we first evaluate how these metrics with different gains M correlate with user list-level satisfaction. Following [6, 20] which says satisfaction judgment may be quite subjective and different users may have different opinions, we regularize the satisfaction scores labelled by each user into Z-scores according to the equation:

$$Z\text{-}score_{ui} = \frac{sat_{ui} - Avg(Sat_u)}{Var(Sat_u)}$$

The results of correlation analysis are shown in Table 9. Replacing binary click signals with predicted preference improves all traditional click-based metrics. Using CG/#imps with predicted preference achieves the best performance.

Different users' understanding of satisfaction feedback may be different. For alleviating the user bias, we also use the explicit pair-wise list-level satisfaction choices collected in the user study to evaluate the efficiency of each metrics. Note that we only collect the pair-wise satisfaction of adjacent lists, therefore we have ten samples per user (60 samples for all users in user study phase 2). Following [6, 31], we use the concordance to measure the performance of different metrics. The concordance is calculated by comparing the relative relation of metrics for two lists with the ground truth labeled by the user.

The results are shown in Table 10 and are similar with the results of correlation analysis. Using predicted preference as the gain of each clicked item achieves the best performance in all metrics.

Table 9: Correlation with list-level satisfaction *L-SAT*. (*means p-value<0.05, **means p-value<0.01)

	Bin-Click	Sat-Click	Predicted pref.
CG/#imps	0.3765**	0.2547	**0.4521****
CG/#clicks	-	-0.0601	**0.3503****
CG/pos_{lc}	0.2784*	0.1625	**0.3856****
DCG	0.2965*	0.2939*	**0.4134****

Table 10: Concordance with list-level pair-wise satisfaction.(*means p-value<0.05, **means p-value<0.01)

	Bin-Click	Sat-Click	Predicted pref.
CG/#imps	0.5667**	0.4000	**0.7333****
CG/#clicks	-	0.4833	**0.6500***
CG/pos_{lc}	0.6500*	0.5000	**0.7333****
DCG	0.6167*	0.5000	**0.6667****

8 CONCLUSIONS AND FUTURE WORKS

In this work, through an in-depth user study of online news reading scenario in the mobile environment, firstly, we find that the click signal is not always aligned with user preference for the item. Moreover, user preference is not fixed and may change after clicking. To our best knowledge, this is the first work that considers the user's subjective preference as a dynamic concept.

Furthermore, the news quality is found to be related to the change of preference from *Before-Read* and *After-Read phases*. For low-quality news, user preference is more likely to decline after user read its content. Meanwhile, user preference in *After-Read phase* is influenced by the user interaction context which is also not aligned with the actual preference.

Besides, several different user behaviors in browsing and reading process are found to reflect preferences in different phases. By using user various behaviors, such as dwell time and read speed, along with the news quality and the user interaction context, a model is successfully built to predict user's actual preference for a clicked news. Furthermore, we replace binary click signal of several click-based metrics with predicted item-level preference and achieve a better estimation of user satisfaction. The conclusion of this work can also be applied to personalized recommendations in the mobile environment.

While we find that the quality of news can be used to improve preference and satisfaction estimation, the news quality used in this work is labeled by expert and is hardly collected in practical applications. We leave the automatic quality estimation based on the content and user behaviors for the future works. Moreover, a new recommendation evaluation framework incorporating item quality to better estimate user satisfaction is also left for future works.

ACKNOWLEDGMENTS

This work is supported by Natural Science Foundation of China (Grant No. 61672311, 61532011). We would like to thank Sogou Inc. for their news recommendation resource articles for our experimental analyses.

REFERENCES

[1] Antti Ajanki, David R Hardoon, Samuel Kaski, Kai Puolamäki, and John Shawe-Taylor. 2009. Can eyes reveal interest? Implicit queries from gaze patterns. *User Modeling and User-Adapted Interaction* 19, 4 (2009), 307–339.

[2] Nicholas J Belkin, Michael Cole, and Ralf Bierig. 2008. Is relevance the right criterion for evaluating interactive information retrieval. In *Proceedings of the ACM SIGIR 2008 Workshop on Beyond Binary Relevance: Preferences, Diversity, and Set-Level Judgments. http://research. microsoft. com/˜ pauben/bbr-workshop.*

[3] Nicholas J Belkin, Michael Cole, and Jingjing Liu. 2009. A model for evaluation of interactive information retrieval. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation.* 7–8.

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

[5] Olivier Chapelle and Ya Zhang. 2009. A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th international conference on World wide web.* ACM, 1–10.

[6] Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. Meta-evaluation of Online and O line Web Search Evaluation Metrics. SIGIR.

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

[8] Michael Cole, Jingjing Liu, NJ Belkin, R Bierig, J Gwizdka, C Liu, J Zhang, and X Zhang. 2009. Usefulness as the criterion for evaluation of interactive information retrieval. *Proc. HCIR* (2009), 1–4.

[9] Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research* 7, Jan (2006), 1–30.

[10] Georges E Dupret and Benjamin Piwowarski. 2008. 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.* ACM, 331–338.

[11] 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.

[12] Fan Guo, Chao Liu, and Yi Min Wang. 2009. Efficient multiple-click models in web search. In *Proceedings of the second acm international conference on web search and data mining.* ACM, 124–131.

[13] Ahmed Hassan. 2012. 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.* ACM, 275–284.

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

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

[16] Jiepu Jiang, Ahmed Hassan Awadallah, Xiaolin Shi, and Ryen W White. 2015. Understanding and predicting graded search satisfaction. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining.* ACM, 57–66.

[17] Jiepu Jiang, Daqing He, and James Allan. 2017. Comparing In Situ and Multidimensional Relevance Judgments. (2017).

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

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

[20] Diane Kelly. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval* 3, 1âĂŤ2 (2009), 1–224.

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

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

[23] Dmitry Lagun, Chih-Hung Hsieh, Dale Webster, and Vidhya Navalpakkam. 2014. Towards better measurement of attention and satisfaction in mobile search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval.* ACM, 113–122.

[24] Dmitry Lagun and Mounia Lalmas. 2016. Understanding user attention and engagement in online news reading. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining.* ACM, 113–122.

[25] Yixuan Li, Pingmei Xu, Dmitry Lagun, and Vidhya Navalpakkam. 2017. Towards Measuring and Inferring User Interest from Gaze. In *Proceedings of the 26th International Conference on World Wide Web Companion.* International World Wide Web Conferences Steering Committee, 525–533.

[26] 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 *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 493–502.

[27] Zeyang Liu, Jiaxin Mao, Chao Wang, Qingyao Ai, Yiqun Liu, and Jian-Yun Nie. 2017. Enhancing click models with mouse movement information. *Information Retrieval Journal* 20, 1 (2017), 53–80.

[28] Jiaxin Mao, Yiqun Liu, Huanbo Luan, Min Zhang, Shaoping Ma, Hengliang Luo, and Yuntao Zhang. 2017. Understanding and Predicting Usefulness Judgment in Web Search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 1169–1172.

[29] Jiaxin Mao, Yiqun Liu, Ke Zhou, Jian Yun Nie, Jingtao Song, Min Zhang, Hengliang Luo, Hengliang Luo, and Hengliang Luo. 2016. When does Relevance Mean Usefulness and User Satisfaction in Web Search?. In *International ACM SIGIR Conference on Research and Development in Information Retrieval.* 463–472.

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

[31] Tetsuya Sakai. 2013. How intuitive are diversified search metrics? Concordance test results for the diversity U-measures. In *Asia Information Retrieval Symposium.* Springer, 13–24.

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

[33] Peifeng Yin, Ping Luo, Wang-Chien Lee, and Min Wang. 2013. 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.* ACM, 989–997.

[34] Yisong Yue, Rajan Patel, and Hein Roehrig. 2010. 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.* ACM, 1011–1018.