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


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

Besides 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
design a laboratory user study to collect user interaction logs. 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 (1280x720 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.

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
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

### Table 1: Statistics of user study data

<table>
<thead>
<tr>
<th>Stage 1: for analysis</th>
<th>#users</th>
<th>#tasks</th>
<th>#news</th>
<th>#clicks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>26</td>
<td>286</td>
<td>4290</td>
<td>1337</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stage 2: Testset for prediction</th>
<th>#users</th>
<th>#tasks</th>
<th>#news</th>
<th>#clicks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6</td>
<td>66</td>
<td>990</td>
<td>266</td>
</tr>
</tbody>
</table>

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

| 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      |
| Item-level Preference     | After-Read Preference | U     | 3      | 5      |
| Item-level Preference     | Post-Task Preference | U     | 5      | 5      |
| Item-level Quality        | Overall quality | E     | -      | 2      |
| Item-level Quality        | Content quality | U     | 3      | 5      |
| Item-level Quality        | Consistency of title and content | U     | 3     | 5      |
| Item-level Satisfaction   | List-level satisfaction | U    | 4     | 7      |
| Item-level Satisfaction   | Pairwise satisfaction of adjacent list (last v.s. current) | U    | 4     | 2      |

the implicit feedback based on click signals for each news, will be compared.

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 < 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
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.

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 ≤ 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.

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.

<table>
<thead>
<tr>
<th></th>
<th>Before-Read Pref.</th>
<th>After-Read Pref.</th>
<th>Post-Task Pref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>all</td>
<td>3.7016</td>
<td>3.4788</td>
</tr>
<tr>
<td>User perceived:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>content quality &lt;3</td>
<td>3.5017</td>
<td>2.6582</td>
<td>2.1633</td>
</tr>
<tr>
<td>title consistency &lt;3</td>
<td>3.4677</td>
<td>2.7883</td>
<td>2.1613</td>
</tr>
<tr>
<td>Expert labeled:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>overall quality low</td>
<td>3.5154</td>
<td>2.8567</td>
<td>2.4539</td>
</tr>
</tbody>
</table>

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.

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).
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<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 $\kappa$ ($k=0.2813$, 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<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<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<0.001). Users answer...
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 different 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.

![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).](image)

<p>| Table 4: User behavior metrics of the news user liked or disliked in different phases. (* means p-value&lt;0.05, ** means p-value&lt;0.01). |
|---------------------------------|-----------------|-----------------|-----------------|</p>
<table>
<thead>
<tr>
<th></th>
<th>Before-Read Pref.</th>
<th>After-Read Pref.</th>
<th>Post-Task Pref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>viewport time (ms)</td>
<td>6851**</td>
<td>6201**</td>
<td>6415</td>
</tr>
<tr>
<td>dwell time (s)</td>
<td>31.33**</td>
<td>38.57**</td>
<td>28.66**</td>
</tr>
<tr>
<td>read length (pixel)</td>
<td>6987**</td>
<td>7707**</td>
<td>6113**</td>
</tr>
<tr>
<td>read ratio</td>
<td>0.797</td>
<td>0.811</td>
<td>0.776**</td>
</tr>
<tr>
<td>read speed (pixel/s)</td>
<td>309.9**</td>
<td>250.5**</td>
<td>296.9**</td>
</tr>
<tr>
<td>max scroll interval (s)</td>
<td>5.22*</td>
<td>6.74*</td>
<td>4.58**</td>
</tr>
<tr>
<td>direction change times</td>
<td>22.85**</td>
<td>28.25**</td>
<td>19.49**</td>
</tr>
</tbody>
</table>

<p>| Table 5: The Pearson correlations between different user behavior metrics and item-level user preference in different phases. (* means p-value&lt;0.05, ** means p-value&lt;0.01). |
|---------------------------------|-----------------|-----------------|-----------------|</p>
<table>
<thead>
<tr>
<th></th>
<th>Before-Read Pref.</th>
<th>After-Read Pref.</th>
<th>Post-Task Pref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>viewport time (ms)</td>
<td>-0.0635*</td>
<td>-0.0515*</td>
<td>-0.0515*</td>
</tr>
<tr>
<td>dwell time (s)</td>
<td>0.0686</td>
<td>0.2797**</td>
<td>0.2611**</td>
</tr>
<tr>
<td>read length (pixel)</td>
<td>0.0738**</td>
<td>0.2770**</td>
<td>0.2534**</td>
</tr>
<tr>
<td>read ratio</td>
<td>0.0608</td>
<td>0.2693**</td>
<td>0.2142**</td>
</tr>
<tr>
<td>read speed (pixel/s)</td>
<td>-0.0725*</td>
<td>-0.0789**</td>
<td>-0.0924**</td>
</tr>
<tr>
<td>max scroll interval (s)</td>
<td>-</td>
<td>0.0757**</td>
<td>0.0747**</td>
</tr>
<tr>
<td>direction change times</td>
<td>0.1070**</td>
<td>0.2249**</td>
<td>0.2316**</td>
</tr>
</tbody>
</table>

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 ensures 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

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

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 $t_{\text{threshold}}$ seconds are seen as liked clicks. In this study, we set $t_{\text{threshold}} = 52$ 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).

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bin-Click</td>
<td>0.4248</td>
<td>1.0000</td>
<td>0.5963</td>
<td>0.4248</td>
</tr>
<tr>
<td>Sat-Click</td>
<td>0.5000</td>
<td>0.3008</td>
<td>0.3757</td>
<td>0.3751</td>
</tr>
<tr>
<td>① $F_b$</td>
<td>0.5570</td>
<td>0.4782</td>
<td>0.5133</td>
<td>0.6164</td>
</tr>
<tr>
<td>② $F_b+F_c$</td>
<td>0.5705**</td>
<td>0.4676</td>
<td>0.5128</td>
<td>0.6239**</td>
</tr>
<tr>
<td>③ $F_b+F_c+F_g$</td>
<td>0.5733</td>
<td>0.5000**</td>
<td>0.5338**</td>
<td>0.6293*</td>
</tr>
<tr>
<td>④ $F_b+F_c+EQ$</td>
<td>0.5818**</td>
<td>0.5111**</td>
<td>0.5431**</td>
<td>0.6358**</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th></th>
<th>MSE</th>
<th>MAE</th>
<th>PCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sat-Click</td>
<td>-</td>
<td>-</td>
<td>0.1400</td>
</tr>
<tr>
<td>① $F_b$</td>
<td>1.2116</td>
<td>0.9999</td>
<td>0.2873</td>
</tr>
<tr>
<td>② $F_b+F_c$</td>
<td>1.1683**</td>
<td>0.8914**</td>
<td>0.3291**</td>
</tr>
<tr>
<td>③ $F_b+F_c+F_g$</td>
<td>1.1587*</td>
<td>0.8989</td>
<td>0.3475**</td>
</tr>
<tr>
<td>④ $F_b+F_c+EQ$</td>
<td>1.1331**</td>
<td>0.8789**</td>
<td>0.3548**</td>
</tr>
</tbody>
</table>

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 ($F_g$), 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’s 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 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:

$$CG(M) = \frac{1}{\log_2(l+1)} \sum_{i=1}^{[CS]} M(n_i)$$

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), satisfied 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:

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

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:

$$Z-score_{ui} = \frac{sati - Ave(sati)}{Var(sati)}$$

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.

### 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 recommendation resource articles for our experimental analyses.

REFERENCES