Quality Effects on User Preferences and Behaviors in Mobile News Streaming

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

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

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

Yunqiu Shao
BNRist, DCST, Tsinghua University
Beijing, China
shaoyunqiu14@gmail.com

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

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

ABSTRACT

User behaviors are widely used as implicit feedbacks of user preferences in personalized information systems. In previous works and online applications, the user's click signals are used as positive feedback for ranking, recommendation, evaluation, etc. However, when users click on a piece of low-quality news, they are more likely to have negative experiences and different reading behaviors. Hence, the ignorance of the quality effects of news may lead to the misinterpretation of user behaviors as well as consequence studies. To address these issues, we conducted an in-depth user study in mobile news streaming scenario to investigate whether and how the quality of news may affect user preferences and user behaviors. Firstly, we verify that quality does affect user preferences, and low-quality news results in a lower preference. We further find that this effect varies with both interaction phases and user's interest in the topic of the news. Secondly, we inspect how users interact with low-quality news. Surprisingly, we find that users are more likely to click on low-quality news because of its high title persuasion. Moreover, users will read less and slower with fewer revisits and examinations while reading the low-quality news.

Based on these quality effects we have discovered, we propose the Preference Behavior Quality (PBQ) probability model which incorporates the quality into traditional behavior-only implicit feedback. The significant improvement demonstrates that incorporating quality can help build implicit feedback. Since the importance and difficulty in collecting news quality, we further investigate how to identify it automatically. Based on point-wise and pair-wise distinguishing experiments, we show that user behaviors, especially reading ratio and dwell time, have high ability to identify news quality. Our research has comprehensively analyzed the effects of quality on user preferences and behaviors, and raised the awareness of item quality in interpreting user behaviors and estimating user preferences.

*Contact author

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313751

CCS CONCEPTS

• **Information systems → Evaluation of retrieval results**; **Users and interactive retrieval**;

KEYWORDS

Online news reading; Quality effect; User behavior analysis; User item-level preference; Implicit feedback

ACM Reference Format:
Hongyu Lu, Min Zhang, Weizhi Ma, Yunqiu Shao, Yiqun Liu, and Shaoping Ma. 2019. Quality Effects on User Preferences and Behaviors in Mobile News Streaming. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3308558.3313751

1 INTRODUCTION

User preference has been widely used as a critical concept in personalized information system [13, 28]. Yet, it is difficult to acquire due to its subjective nature. The most direct way is to ask users to give explicit feedback, for example, the rating in e-commerce websites [30, 38]. However, because of the difficulty and the potential biases [24] in its collection in real-world applications, commercial information systems have exploited implicit feedback signals derived from user behaviors. Most commonly, users' clicking on the item have been used as implicit feedback of document relevance [1, 16] and user experiences [19]. In addition to the click signal, dwell time (i.e. the time that user spends on a clicked item) has also been found well correlated with document relevance [35], item-level satisfaction [9, 12, 20] and user preferences [6]. A click followed by a long dwell time has traditionally been seen as satisfied click and been widely used in a number of retrieval applications [9, 11]. Other behaviors, like mouse movement [27], scroll information [22], and gaze [2, 25] are also investigated for inferring user experiences.

However, user behaviors have been proved to be informative but noisy. Click behavior is found to be affected by many factors, like position [16], trust [37], and presentation [34]. Moreover, dwell time is also found related to the content length [36], readability [21]and search task [18].

The credibility of the behaviors may also be related to the item quality. For example, as shown in Figure 1, two pieces of news with different levels of quality are both clicked by the user. The low-quality news is more likely to be disliked by the user after reading its poor content. As a result, although user clicking on both news,

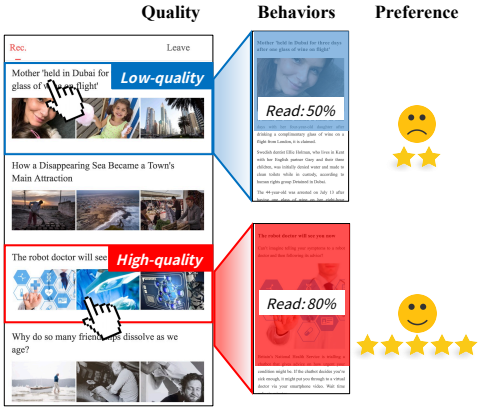


Figure 1: User clicking on items may have different behaviors (e.g. reading ratio) and different preferences because of the different levels of quality.

his/her preference is significantly different because of the quality. Thus, traditional implicit feedback may yield imprecise user preference modelling because of the ignorance of the quality effect. To the best of our knowledge, there still lacks comprehensive understanding of how item quality affects user preference and user behaviors, which is important for modeling user preference and utilizing user behaviors. In this paper, we study the quality effects and aim to address the following questions:

- **RQ1**: Does quality affect user preferences? If yes, how?
- **RQ2**: Does quality affect user behaviors during the browsing and reading process? If yes, how?
- **RQ3**: Can incorporating quality help build implicit feedback?
- **RQ4**: Can we identify quality based on user behavior signals?

We conduct an in-depth user study in online news reading scenario on the mobile device. At the beginning of the study, we collected user's interest in the topics. After that, users are asked to read several lists which contains the news of different levels of quality. In this process, we collected their behaviors as well as explicit feedback for experiences, including the perception of quality and the preferences for the news.

As for RQ1, by comparing user preferences for low-quality and high-quality news, we find that lower quality leads to a lower user preference. From this, we further analyze the degree of quality effect, measured by the difference of user preferences between low-quality and high-quality news, and find it is related to the interaction phases and user's interest for the news topic. On the one hand, from before-read phase to after-read phase, the quality effects increase. On the other hand, the quality effect is much larger when the user is more interested in the news topic.

Besides the effects on user preferences, we further investigate the quality effects on user behaviors (RQ2). From the comparison of click behaviors, we find that users have higher probability to click the low-quality news. Part of the explanation may be the higher title persuasion of the low-quality news. This effect is more distinct when a user has higher interest in the topic of the news. Furthermore, we find that a user's reading behaviors after clicking are also influenced by the news quality. Users will read less in the low-quality news, as shown by the reduction of *dwell time* and *reading ratio*. They will also have fewer revisits and examinations. These quality effects promote us to re-think the implicit feedback (RQ3).

Traditional implicit feedback relies on the relation between user preferences and user behavior. Because of our findings about quality effects on user preferences and behaviors, we further take the quality and its effects into consideration, and propose a probability model to estimate user preference. The out-performance of the proposed model demonstrates that it is helpful to incorporate quality into implicit feedback building.

Considering the usefulness but the hard annotation of quality, it is valuable to automatically identify news quality (RQ4). We examine the point-wise and pair-wise distinguishing ability of each behavior, and confirm the possibility of using user behavior signals, like reading ratio and dwell time, to identify the quality of news.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 introduces the methodology of our user study and collected measurements. Section 4 studies the quality effects on user preferences (RQ1). The analysis of the quality effect on user behaviors (RQ2) is given in Section 5. Section 6 investigates the usage of quality for improving the implicit feedback building (RQ3). Section 7 discusses the possibility of utilizing behaviors to identify quality (RQ4). Finally, Section 8 summarizes the conclusion and future work.

2 RELATED WORK

In this section, we review two related research directions, the modeling of user item-level experiences and the analysis of user behaviors.

2.1 User Item-Level Experience Modeling

Modeling user item-level experiences is a fundamental work for implicit feedback building and performance evaluation in information system. Some previous works have paid attention to this direction, such as modeling user perceived usefulness in web search and user preferences in recommender systems.

Belkin et al. [4] argue that usefulness, which represents users' perceived value of a search result, depend 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. [15] have collected usefulness judgment not only in situ stage but also post-session context-independent judgment, and find that they are different, indicating that context will affect user perception of usefulness.

Compared with the usefulness, user preference is more personalized and subjective, making it hard to model it. Lu et al. [28] model user preference as a dynamic concept varying in different interaction phases, and found that the change of preference after user reads the news content is related to the news quality. Their work provides a good starting point for modeling user preferences with item quality, but more in-depth researches still remains to be studied. We go further to analyze the quality effect on user preferences jointly with interaction phases and user topic interest, examining how the effect varies in different interaction phases and when users have different levels of interest on the news topic.

2.2 User Behavior Analysis

Learning from user behaviors is a general approach used in online interactive information systems. Researchers of Information Retrieval use click signals to infer document relevance [8] under the assumption that relevant documents attract more clicks. Personalized information filtering systems, like recommender systems, are also designed by mining user preferences through historical click data [13, 32].

However, user click behavior has been found to be biased because of the influence caused by some factors, for example, position [16]. Documents on the top may attract more user clicks [17]. Other factors like trust [16], result attractiveness [37] and presentation [34] are also examined. To address the biases of user click behavior, researchers have proposed a number of advanced click models [5, 10], which are designed to eliminate the effects of various biases to obtain an unbiased estimation of result relevance in web search scenarios.

Besides click, some other behaviors are also examined and incorporated to model user 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 [9]. The correlation between dwell time and user interest is further modeled by document factors (e.g. readability [21], genres [3], and human factors [36]). Besides dwell time, viewport time is successfully used on mobile devices to infer user interest at the sub-document level [14]. Li et al. [25] examine the correlation between users' eye gaze and user explicit interest, and demonstrate the effectiveness of using attention-based behaviors, like viewport time and gaze, to predict user interest. Lin et al. [26] find that the delivery mechanism of the system greatly affects user information consumption behaviors. Ben et al. [31] investigate the causal effects of ad blocking on user long-term behaviors, like active time spent in the browser and the number of page. Lu et al. [28] investigate the influence of user preferences on user behaviors in news reading scenario and in mobile environment.

While many factors related to user behaviors have been investigated, the item quality is less studied. Based on a carefully designed user experiment, we studied the directly effects of news quality on user browsing and reading behaviors with other factors controlled, like position, topic and user preferences.

3 USER STUDY METHODOLOGY

To measure the effects of news quality on user preferences and behaviors, we designed a laboratory user study in which we varied the news quality with other factors controlled and inspect whether user experiences and behaviors become different.

3.1 Experimental Setup

In this section, we introduce the settings of our experiment, including how we sample the news data and annotate their quality as low-quality and high-quality, how we generate news lists with other factors controlled, like position and topic and what are the interfaces of the experimental platform.

3.1.1 Data.

The news used in the user experiment is sampled from the real log of *Sougou NewsFeed*, a popular commercial newsfeed service on mobile device. We firstly choose five most popular topics: *social*, *entertainment*, *technology*, *history*, and *sport*. Then we randomly sample one hundred pieces of news from each topic. By doing this, we cover a variety of topics and ensure a balance between them to eliminate the possible bias.

After that, a human annotation is conducted to measure the quality of these pieces of news. We recruit three experts to label the overall quality (binary-scales: low-quality or high-quality) by discussion according to the following four properties.

Authenticity: High authenticity means that this piece of news is authentic or has high credibility [33]. On the contrary, if the content is imaginary or exaggerated, the news is seen of low authenticity.

Value: A piece of news is of low-quality value when its content is full of problems with being vulgar, violent, bloody and pornographic.

Expression: The expression is high-quality when the statement is objective and precise, and the information is rich but not redundant.

Headline: Low-quality headlines have one of such problems: information is incomplete or fake; the expression is exaggerating or vulgar and inconsistent with the content.

The validity of the overall quality label, namely *Expert Labeled Quality (EQ)*, are tested by comparing it with the average perceived quality annotated by the experiment participants. (EQ vs. User perceived content quality Fleiss's k=0.5017; EQ vs. User perceived title consistency k=0.4737).

3.1.2 News Lists.

Based on the binary overall quality labelled by experts, the news are classified into two groups, low-quality and high-quality news. We generate sixteen news lists (fifteen news each) containing both low-quality and high-quality news as tasks in advance.

Previous works show that user behaviors are affected by many factors, like position [16]. To control the impact of position, we take the Latin square experimental design principles to assign the positions of low and high-quality news in the lists. Meanwhile, we assigned the same number of news from five topics in each task list to control the effect of news topic. The careful design ensures that the low-quality and high-quality news have the same position and topic distribution.

3.1.3 Online Platform.

To simulate a real online news reading environment, we build an experimental platform which is similar to the common used newsfeed website. There are two types of pages, one is the list page for the user to browse the news list (Figure 2 left and right pictures), the other is the news content page for users to read the news content (Figure 2 middle picture). A JavaScript plugin is injected into both pages to record user's scrolling, clicking, and page switching behaviors.

3.2 Experimental Procedure

In this section, we describe the procedure of the experiment and the collection of user experiences in multi-phases.

3.2.1 Procedure.

Before the experiment, we first collect user's interest of the five topics, namely *Topic Interest* (5-ratings each). To get familiar with

Figure 2: Collect user experiences in different phases: *Before-Read, After-Read,* and *Post-Task Phase.*

the experiment platform and procedure, the participants need to complete an example task for training. After that, users are asked to finish four tasks.

At the beginning of each task, users need to read the task description carefully. Then, a list of 15 pieces of news is shown to the user for browsing and reading. As no browsing and reading time limitations are imposed, users can finish the browsing task at any time.

3.2.2 Multi-Phase Experience Measurement.

Within the task procedure, we add several questionnaires to collect user experiences in three phases: Before-Read, After-Read and Post-Task, as shown in Figure 2.

Before-Read Phase

- **Q1:** How much do you expect to like or dislike reading this piece of news? (5-point Likert scale, from very dislike to very like)

Firstly, as soon as a user clicks a piece of news, before showing its content, we ask the user about his/her expected preference for the news. This preference is named as *Before-Read Preference*.

After-Read Phase

- **Q2:** How much do you like reading this piece of news? (5-point Likert scale, from very dislike to very like)
- **Q3:** What do you think of the content quality of this piece of news? (5-point Likert scale, from very poor to very good)
- **Q4:** What do you think of the consistency between title and content of the news? (5-point Likert scale, from very low to very high)

After a user reads the news content and deciding to leave (note that there is no constraints on user reading process, users can leave at anytime as he/she wish), we ask the user several questions for his/her experiences, including *After-Read Preference* (Q2), *User Perceived Content Quality* (Q3) and *User Perceived Title Consistency* (Q4).

Post-Task Phase

- **Q5:** How much do you like reading this piece of news? (5-point Likert scale, from very dislike to very like)

After users finish browsing, we randomly shuffle and re-display the news that they have seen (right subplot in Figure 2. "Seen" means that the news has appeared on the user's viewport (including the news not clicked). We ask users to re-annotate his/her preference on each news, named *Post-Task Preference*.

We finally recruit 32 participants, including 18 females. Across 128 tasks finished, 1920 impressions (including 576 low-quality news) are shown, and 631 clicks (including 209 clicks on the low-quality news) are collected.

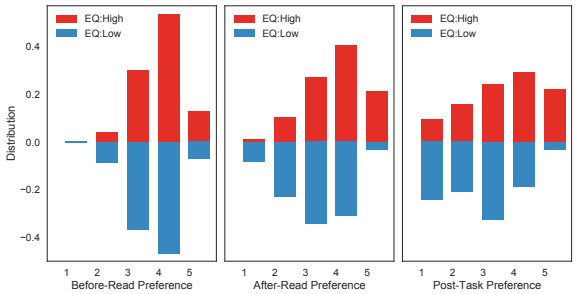


Figure 3: When the news quality is different (EQ:High vs. EQ:Low), the distributions of user preference in multi-phases: Before-Read (left), After-Read (center) and Post-Task (right).

Table 1: The Pearson correlations between news quality and user preference in multi-phases. The quality is measured by both expert (EQ) and user (User Perceived Content Quality, UQ-C; User Perceived Title Consistency, UQ-T).

	EQ	UQ-C	UQ-T
Before-Read Preference	0.1480	0.3267	0.2964
After-Read Preference	0.3292	0.8170	0.5874
Post-Task Preference	0.3077	0.6992	0.5652

4 QUALITY EFFECTS ON USER PREFERENCE

Based on the quality measurements labeled by both expert and users, and the explicit user preference feedback collected in the experiment, we now investigate the effect of quality on user preference.

4.1 Different Interaction Phases

To begin with, we compare the distribution of user preferences for low-quality and high-quality news, as shown in Figure 3. Generally, significant difference is found in user preference in all the three interaction phases (Before-Read Preference, p<0.01, d =0.32; After-Read Preference, p<0.01, d =0.74; Post-Task Preference, p<0.01, d =0.69). The results suggest that news quality has significant effects on user preference in all three phases. To be more specific, we find the degrees of the differences vary in different phases. It promotes us to jointly analyze the quality effects with the interaction phases.

We separate the news into different groups based on the quality measurement, not only the expert labeled quality (binary-scale, two groups), but also the user perceived content quality (5-scales, five groups) and the user perceived consistency between title and content (5-scales, five groups). For each group, we calculate the mean of user preferences in three phases (*Before-Read, After-Read, Post-Task Phases*). The results are shown in Figure 4.

Firstly, we can see that user preferences for the high-quality news are more likely to stay or increase (from Before-Read to After-Read and to Post-Task, shown as mark-1 in the figure 4), and vice versa, user preferences for the low-quality news continually drop along user reading. This finding is consistent with Lu's work [28]. Starting from this point, we further analyze how the quality effects

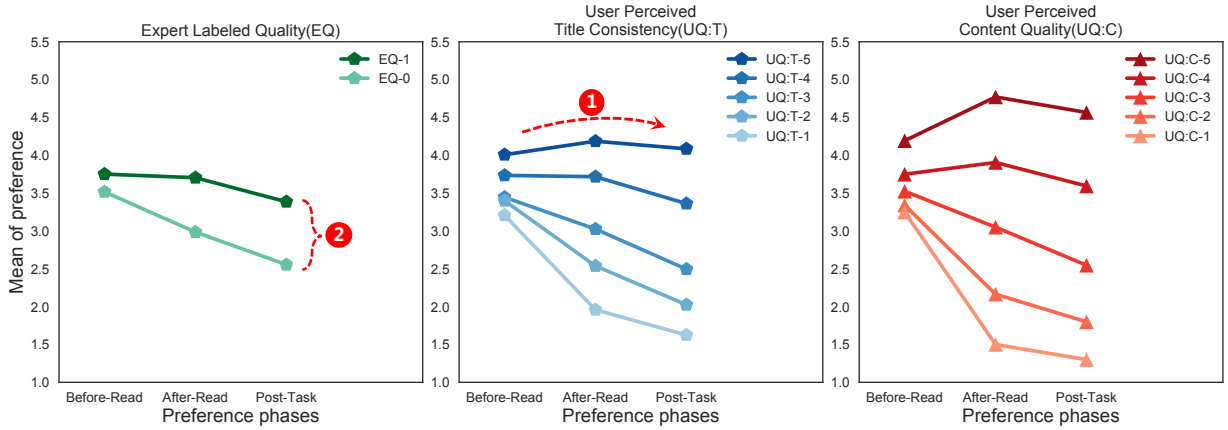


Figure 4: Means of user preference (in three phases) for the news of different levels of quality. The quality is measured by Expert Labeled Quality (left), User Perceived Title Consistency (center), and User Perceived Content Quality (right). The level of quality is distinguished by the shade of color.

Table 2: Quality effects, measured by the difference (Δ, Cohen's d) of user preferences between low-quality and high-quality news, when user has different topic interests (TI).

	Before-Read Preference				After-Read Preference				Post-Task Preference			
	EQ:Low	EQ:High	Δ	d	EQ:Low	EQ:High	Δ	d	EQ:Low	EQ:High	Δ	d
TI=Min	3.610	3.689	+0.079	0.114	3.170	3.597	+0.427	0.447	2.627	3.176	+0.549	0.454
TI=Mid	3.465	3.630	+0.166	0.223	2.831	3.674	**+0.843**	**0.870**	2.563	3.442	+0.879	0.751
TI=Max	3.494	3.897	**+0.403**	**0.532**	2.987	3.806	+0.819	0.838	2.494	3.491	**+0.997**	**0.801**

varies in different phases. The degree of quality effect is seen as the gap of preferences between low-quality and high-quality news (larger gap means higher degree, shown as mark-2 in the figure 4). In Before-Read Phase, user preferences for the news of low-quality and high-quality are already different, but more slightly than in After-Read and Post-Task Phases. It may be due to that user only knows a little part of the news before reading its content, thus quality is less perceived and has less effect on user preferences.

Meanwhile, although the trend of quality effect is similar when using different quality measurements, user perceived quality has a larger effect than objective expert labeled quality. It may because that user's perception of quality effect may be personalized and related to other subjective factors like user's interest on the topic of the news, which motivates us to further investigate the quality effect with user topic interest in next section.

In addition, the effect of user-perceived content quality is larger than the title consistency. By applying the Pearson's r, we further directly compare the correlation between user multi-phase preferences and different quality measurements. The results, shown in Table 1, indicate that the correlation between quality and preference is largest in After-Read Phase, and content quality is more closely related to user preference than title consistency and expert labelled quality.

4.2 Different Levels of Topic Interests

Beside of interaction phases, we further investigate whether the quality effects on user preferences are related with user's interest of the news topic. As mentioned in Section 3.2.1, users' interests on all the five topics are collected at the beginning of the user study, by asking users to rate each topic. To avoid users' different understandings of topic interest ratings, we separate the five topics within each user into three groups. The topics with the highest user interest are set as "Max" group (not necessarily one topic), and vice versa, the topics with lowest user interest are set as "Min" group, and the other topics are set as the "Mid" group.

The means of user preferences for low-quality and high-quality news with different levels of topic interest are shown in Table 2. We further measure the quality effect on user preferences by calculating Δ (preference of high-quality news − preference of low-quality news) and Cohen's d. On the one hand, the quality effect is lowest in *Before-Read phase* and highest in *Post-Task phase*, which is same as previous analysis.

On the other hand, the quality effect is highly related to the topic interest. When user has low topic interest (Min), the difference of preferences between low-quality and high-quality news is small. On the contrary, when user has higher topic interest (Mid and Max), the quality effect is much larger. It can be interpreted as the user has less quality sensitiveness for the news of his/her less interested topics. Moreover, although topic interest is generally positively correlated with user preferences, when the news is of low quality, higher topic interest leads to lower preferences. It can be explained by the less user tolerance of quality for his/her highly preferred topics.

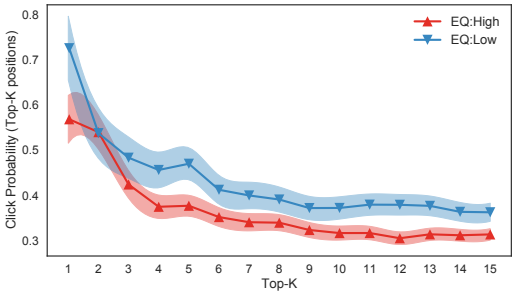


Figure 5: Click Probability of the news *up to position k* conditioned by the news quality. The low-quality news attracts more clicks.

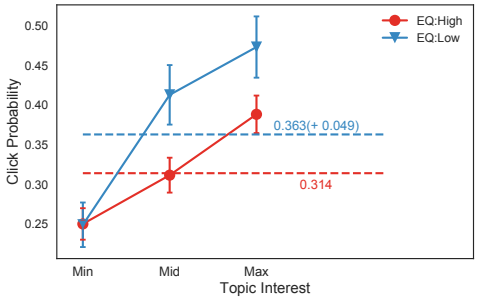


Figure 6: Click probability of the news with different levels of quality and user topic interests.

5 QUALITY EFFECTS ON USER BEHAVIORS

In this section, we focus on investigating whether users behave differently when interacting with low-quality and high-quality news. Users' scroll and click events with timestamps are recorded in user browsing (list page) and reading (content page) processes, and are used to build several frequently used behavior measurements.

5.1 Click Behavior

Click behavior has been widely used as implicit feedback for recommendation, ranking and evaluation. In this section, we investigate the quality effects on user's click behavior, measured by the probability of users clicking the news. If there is a dependency with the quality, any interpretation of clicks as implicit preference feedback should be relative to the quality.

We first evaluate the probability for users clicking on low-quality or high-quality news. The conditional probability shows that if the news is of high quality, it will be clicked with a probability of 0.3140 ($P(Click|EQ = 1)$), vice versa, the click probability for low-quality news is 0.3628 ($P(Click|EQ = 0)$). The difference is slightly but significant according to the independent t-test, p<0.05, d=0.10. It implies that low-quality news is more likely to be clicked than high-quality news.

Incorporating the position of the news into analysis, we calculate the click probabilities conditioned by the quality for top-k impressions (K ranging from 1 to 15) are shown in Figure 5. By separating different levels of quality, we can find that the news of low quality has higher click probability than high-quality news in all the Top-K positions.

To further verify the reliability of the finding that low-quality news has a higher probability of being clicked, we test whether it exists in real system. Note that the news used in our work is sampled from *Sougou Newsfeed*. We further collect one week's log data about these piece of news (from October 16, 2017, to October 22, 2017, more than 1.5 thousands interactions per news).

Based on the quality labeled by expert (EQ) and the CTR calculated in the log for each news, we are able to compare the average CTR of low-quality and high-quality news. The result of independent t-test shows that there is a significant difference of CTR for low-quality news (0.1539) and high-quality news (0.0835), p<0.01. The observation is similar with the result in user study, and proves that the news quality does affect user click behavior and low-quality news is more likely to be clicked.

5.1.1 *Title Persuasion.*

Table 3: The contextual effect of quality on click probability. When the quality of last impression news (lEQ) varies, how is the click probability of current (c) impression news?

	$lEQ = low$		$lEQ = high$	
$P(Click\|\ lEQ)$	0.3507		0.2898	
$P(Click\|\ lEQ, cEQ)$	$cEQ = 0$	$cEQ = 1$	$cEQ = 0$	$cEQ = 1$
	0.4000	0.3108	0.2838	0.2917

To further study why low-quality news attracts more clicks, we conduct a supplementary annotation for the persuasion of the title. The *title persuasion* is defined as the extent that user is seduced to click the news (4-scales). We hired external assessors from JD crowd source platform and make sure each piece of news was annotated by 3 different assessors (Fleiss' $k = 0.4259$, reach moderate agreement). We use majority voting to get the final score.

Based on the annotation, we compare the title persuasion of low and high-quality news and find low-quality news has higher persuasion than high-quality news (2.16 vs. 1.61, p<0.01). It is reasonable that low-quality news usually uses literary methods, like the exaggeration, to generate a more compelling title. Thus, user may click on low-quality news because of its high title persuasion.

5.1.2 *Topic Interest.*

We further study how quality effects on user's clicking decision change when user has different levels of topic interest. We show the click probability conditioned by the news quality and topic interest ($P(Click|Q, TI)$), in Figure 6.

Beside of the separate effects of quality and topic interest, there also exists an interaction effect. When topic interest is low (Min), user has almost the same click probability for the news of low or high quality. However, when topic interest is high (Mid, Max), the click probability of low-quality news exceed high-quality news. The reason is that when the news is of low topic interest, the user may be not easily aware of its quality, so the click decision is less affected.

5.1.3 *Contextual Effect.*

As the user interacts with news in the list context, the quality of the news around may have contextual effects. In the mobile environment, most of users scroll to browse the list, examine news one-by-one. So we investigate how the quality of the last impression affects user click probability for the current news.

We calculate the probability of clicking on current news conditioned by the quality of the last impression. The results, shown in

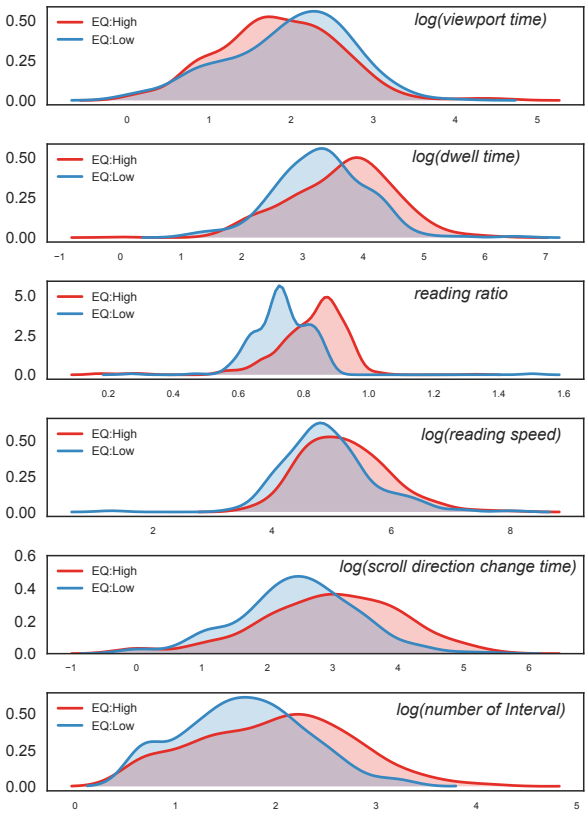


Figure 7: The probability density function (estimated by kernel density estimation) of behaviors when users are reading the news of different qualities (EQ:High vs. EQ:Low)

Table 3, indicate that user has a higher click probability when the quality of the last impression is low (0.3570 vs. 0.2898). We further control the quality of current news. Results show that no matter what the quality of current news is, the click probability is higher when the last impression is of low quality. It may be caused by the saliency attention mechanism, the saliency and click probability of a piece of news may depend on the context news.

5.1.4 *Summary.*

From above analysis, we can conclude that user click behavior is affected by news quality. Users have a higher probability of clicking the low-quality news. It promotes us that we should take quality into account when using user click signals to build implicit feedback. User topic interest is found to be related to the quality effect on click probability. The quality effect is lower when user has lower interest in the topic. Beside of affecting the click decision of current news, news quality also has effects on context news, like the next news in the recommendation list.

5.2 Reading Behaviors

Beside of the click behavior, in this section, we investigate how news quality affects user's behaviors in the reading process. First of all, we directly compare the distributions of user behavior measurements when the news is of high-quality and low-quality, as

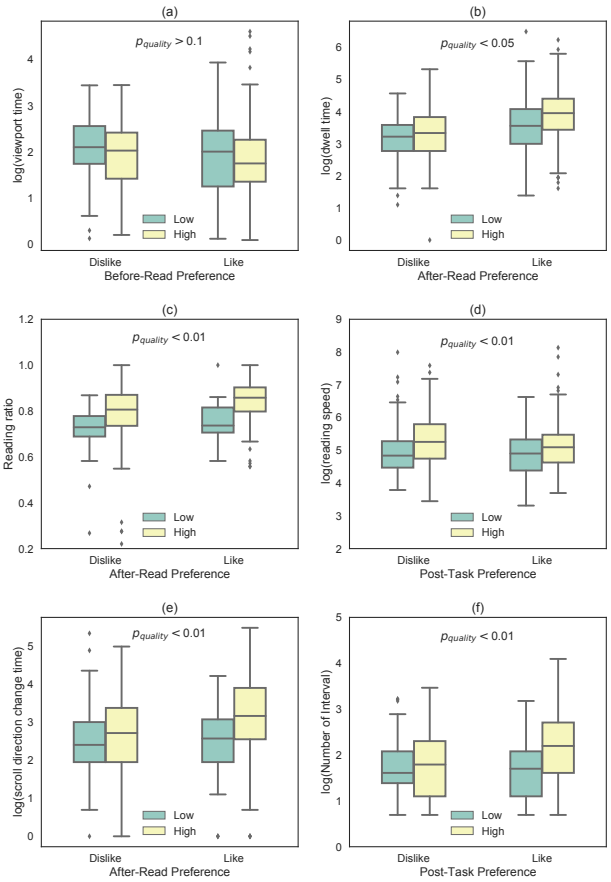


Figure 8: User behaviors for the news of different levels of quality (EQ: low, high) with different user preferences (Dislike(≤3), Like(>3)). The p-value displayed in each figure represents the main effect of quality in two-way ANOVA (with preferences).

shown in Figure 7. By doing this, we get an overview of news quality effects on behaviors.

In last section, we demonstrates that the news quality affects user's preferences, while previous work [28] shows that user preferences influence user behaviors. Hence, to analyze the impact of news quality on user behaviors, we further control the influence comes from user preferences.

For each behavior, we choose the preference in its most closely related phase based on the behavior-preference correlations, marked by the y-label in Figure 8. For example, the most closely related preference for dwell time is *After-Read Preference*, while for viewport time is *Before-Read Preference* and for direction change times is *Post-Task Preference*. If the preference less than or equal to 3 (≤ 3), we consider it as *dislike*, other levels of preference are considered as *like*. Based on both analysis approaches, we study the quality effects on each reading behavior:

Viewport time.

Viewport time, reflecting how long user reads the news snippet in the list browsing page, may be related to the user click decision process. Because the data for viewport time does not meet the

normality assumption, we use log-transformation, $log(x+1)$, before statistic analysis. In general, the result of independent t-test suggests there is a slightly but significant difference of viewport time between low-quality and high-quality news (p<0.05, d=0.17). We further take the effect of user *Before-Read Preference*, which is proved mostly related to viewport time, into account. The result of the analysis of variance (ANOVA) [7] considering both quality and preferences shows that there is no independently significant effect of quality on the viewport time (two-way ANOVA, $F(1,929) = 4.47$, p>0.1). We then separate the news into dislike and like groups, and compare the means of viewport time between low-quality and high-quality news within each group. The result, shown in Figure 8 (a), indicates that user spends a longer time reading the snippet of low-quality news. While the difference is not significant (t-test, $p_{dislike}$>0.1, d=0.13 and p_{like}>0.1, d=0.11), such a longer time is plausible: user can perceive part of the quality by reading the title and images. Thus he/she needs longer time to make the click decision for low-quality news.

Dwell time.

Dwell time, indicating how long user reads the content of the news, has been found highly related to user post-click experiences. Same as viewport time, we also use log-transformation to normalize the data. Generally, there is a significant effect due to the news quality (t-test, p<0.01, d=0.42). By jointly analyzing with user preferences in *After-Read Phase*, the independently significant effect of quality is confirmed by two-way ANOVA ($F(1,929) = 4.50$, p<0.05). With preferences controlled, we find that higher quality leads to higher dwell time, especially when the preferences is high (see Figure 8 (b)). Independent t-test is applied within both dislike and like groups. Results indicate that there is a significant difference when user likes the news, $p_{dislike}$=0.13, d=0.17 and p_{like}<0.01, d=0.36.

Reading ratio.

Dwell time has been proved to be affected by many factors, like content length and readability. For eliminating the effect of content length, we normalize the absolute read length by the whole length [23], named as reading ratio which directly reflect how deep user read. In general, user's reading ratio for low-quality and high-quality news is significantly different (t-test, p<0.01, d=0.85).

With user *After-Read Preference* controlled, the statistic results show that quality has still a significant effect on user reading ratio (two-way ANOVA, $F(1,929) = 62.78$, p<0.01). Low-quality news has the lower reading ratio, which indicates that user will leave earlier when reading the low-quality news (see Figure 8(c)). t-test is applied. Results indicate that there is a significant difference in both dislike and like groups, $p_{dislike}$<0.01, d=0.49 and p_{like}<0.01, d=1.19.

Reading speed.

Combining dwell time and reading length (pixel), we calculated user reading speed (reading length/dwell time). As shown in Figure 8 (d), user reading slower when the quality is lower. The result is supported by both two-way ANOVA with *Post-Task Preference* ($F(1,929) = 27.6$, p<0.01) and t-test within each preference group ($p_{dislike}$<0.01, d=0.49 and p_{like}<0.01, d=0.54)

Scroll direction change times.

When a user is reading the news content, he/she may change his/her reading direction to revisit some previous information. The number of the user changing his/her scroll direction indicates how often the user revisits. Considering only the news quality, there is a significant difference reported by t-test, p<0.01, d=0.49. With further investigation by controlling user *After-Read Preference*, we can conclude that quality effect still exists (two-way ANOVA, $F(1,929) = 19.09$, p<0.01) and user revisits more when the news quality is high, especially when the user like the news. (t-test, $p_{dislike}$ =0.10, d=0.19 and p_{like}<0.01, d=0.69)

Number of interval.

In the content page, user read the news by scrolling, scroll intervals may represent user's examinations. Therefore, the number of user scroll interval indicates how many times user carefully examines some information along reading the news. In general and with user preferences controlled, the quality has a significant effect on the number of intervals (t-test, p<0.01, d=0.43; two-way ANOVA, $F(1,929) = 9.05$, p<0.01). The result shown in Figure 8 indicates that when the user likes the news, higher quality leads to more scroll intervals (t-test, $p_{dislike}$=0.25, d=0.14 and p_{like}<0.01, d=0.39).

Summary.

In this section, we give the detailed analysis of the quality effects on user's reading behaviors. We find that when users read low-quality news, they may:

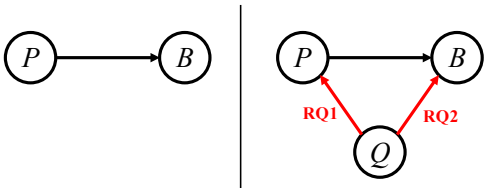
- spend less time reading (dwell time decrease)
- leave earlier (reading ratio decrease)
- read slower (reading speed decrease)
- have fewer revisits (scroll direction change times decrease)
- have fewer careful examinations (number of interval decrease)

The existence of quality effects on user behaviors promotes us to take quality into consideration when using user reading behaviors as implicit feedback. In next session, we further examine it through the probability model.

6 CAN INCORPORATING QUALITY HELP IN BUILDING IMPLICIT FEEDBACK?

To collect user's preferences, information systems have exploited implicit feedback signals derived from user behaviors. For example, the Satisfied-Click, which considers the click following a long dwell time as positive preference feedback, is widely used in today's systems. These approaches are relied on the assumption that user preferences affects user behaviors (e.g. dwell time). While no other factors are considered, the graphical model of this assumption can be shown as Figure 9(a), namely Preference Behavior (PB) model here.

Based on our previous findings about the quality effect on user preferences and user behavior, we argue that quality need to be taken into consideration when building implicit feedback. By studying RQ1, we find that the quality affects user preferences, so we add a relation between quality and user preference. Besides, by studying RQ2, we also conclude that news quality does have an effect on user behavior and the effect is independent of the effect of user preferences. Therefore, we further add a relation between news quality and user behavior. Note that although topic interest



(a) PB-model (b) PBQ-model

Figure 9: The graph models of traditional implicit feedback which use Behavior to estimate Preference, namely PB-model (a), and implicit feedback which incorporates Quality (b). *P*: preference; *B*: behavior; *Q*: quality.

Table 4: The performance of PB and PBQ with different behavior metrics in estimating user preference.(* represents p-value<0.05, ** represents p-value<0.01)

Behavior metric	AUC(PB)	AUC(PBQ)	p	cohens' d
viewport time	0.5775	**0.6249**	**	1.25
dwell time	0.6225[1]	**0.6526**	**	0.88
reading ratio	0.6382	**0.6486**		0.23
reading speed	0.4490	**0.6142**	**	3.32
direction change times	0.5904	**0.6477**	**	1.17
number of interval	0.6111	**0.6709**	**	1.33

[1] Sat-click, the widely used implicit feedback, can be interpreted as dwell time-based PB-model.

is found to influence the quality effect, we do not include it in the implicit feedback building approach because it is subjective and hardly collected in real applications. Figure 9(b) summarizes the proposed relations among news quality, user preferences and user behaviors, namely Preference Behavior Quality (PBQ) model.

Among the three variables, user behavior and news quality are observable by system logging or expert annotation. Based on the conditional dependence in the graphical model, we can infer user preferences when given the behavior and quality. Specifically, for PB-model, when observing a user behavior signal B (e.g. dwell time), we can infer the probability if the user likes the news ($P = 1$):

$$P(P=1|B) = \frac{P(B|P=1)P(P=1)}{\sum_{i\in\{0,1\}} P(B|P=i)P(P=i)}$$

As for PBQ-model which further taking the news quality Q (low or high) into consideration, the probability is calculated by:

$$P(P=1|B,Q) = \frac{P(B|P=1,Q)P(P=1,Q)}{\sum_{i\in\{0,1\}} P(B|P=i,Q)P(P=i,Q)}$$

The Satisfied-Click (with dwell time (DT) threshold =30s) can be interpreted as PB-model with $P(DT <= 30s|P=1) = 0, P(DT > 30s|P=0) = 0$. While most of user behavior signals conform the normal distribution (some need log-transformation at first), we assume that the probability of user behavior B given user preference or news quality follows normal distribution to get a better robustness:

$$P(B|P) \sim N(\mu_p, \sigma_p^2) \text{ or } P(B|P,Q) \sim N(\mu_{pq}, \sigma_{pq}^2)$$

Note that there is no constraint for what the behavior is, the PB-model and PBQ-model can be built by different behavior signals, like viewport time, dwell time, reading ratio, reading speed, direction change times, and the number of interval.

To investigate whether incorporating quality helps in building implicit feedback of user preferences, we compare the PB-model and PBQ-model with the performance of estimating whether a user likes a clicked news. We choose *Post-Task Preference* as the ground truth of user preference, dislike (Post-Task Preference<=3) and like (Post-Task Preference>3), because it is collected in a random order after user browsing the whole list, which eliminate the potential biases from position and interaction order.

We use 10-fold cross-validation to evaluate the model performance. The parameters of the probability model are estimated using training set in each fold, and the performance is measured by AUC in testing set. The results are shown in Table 4. Firstly, while PB-model performs best when using *reading ratio* and *dwell time*, similar with our traditional usage, the PBQ-model performs best when using a less used user behavior metric: *number of interval*. Secondly, the PBQ-model significantly outperforms the PB-model when using all the behavior signals, especially *reading ratio*. It proves that incorporating news quality does help in building implicit feedback of user preferences.

7 CAN WE IDENTIFY THE NEWS QUALITY BASED ON USER BEHAVIOR?

From above analyses, we conclude that quality is useful for estimating user preferences. Thus, how to get news quality is a valuable research topic. In this work, we labelled news quality by external assessors' annotation. It is reliable but high cost and hard to be applied in real applications. Another way is to predict the quality of news automatically. Traditional prediction models are mostly based on the content information.

In this section, We conduct two identification analysis to examine the ability of each single behavior for distinguishing the news quality.

7.1 Point-Wise Distinguishing Ability

To begin with, we evaluate the performance of using absolute value of behavior metric to identify the news quality in an interaction (a user read a piece of news), namely point-wise distinguishing ability:

Given a threshold t_b and a direction coefficient α, for an interaction i with the behavior b_i, the inferred quality of the news is calculated by:

$$\hat{q}_{\alpha,t_b}(i) = I[\alpha(b_i - t_b) > 0]$$

where α indicates whether the identification criteria is above the threshold ($\alpha = 1$) or below the threshold ($\alpha = -1$), and I is a indicator function. Then, the point-wise distinguishing ability D_{point} of a behavior b is defined as the highest accuracy that can be achieved in click interaction set C, when changing the direction α and threshold t_b:

$$D_{point}(b) = \max_{\alpha,t_b} \frac{\sum_{i\in C} I[\hat{q}_{\alpha,t_b}(i) = q_i]}{n(C)}$$

The quality labelled by the expert (EQ) is used as the ground truth (q_i). For each behavior, we evaluate its point-wise distinguishing ability based on all the click interactions in user study.

Table 5: The performance of different behaviors in identifying quality, measured by point-wise distinguishing ability (D_{point}) and pair-wise distinguishing ability (D_{pair}).

	Expert Quality				UQ-C		UQ-T	
	D_{point}	α	D_{pair}	α	D_{pair}	α	D_{pair}	α
viewport time	0.6703	-	0.5850	-	0.5470	-	0.5300	-
dwell time	0.6751	+	0.6650	+	0.6940	+	**0.6350**	+
reading ratio	**0.7084**	+	**0.8010**	+	**0.702**	+	0.6270	+
reading speed	0.6799	+	0.6210	+	0.5240	-	0.5300	-
direction change times	0.6688	+	0.6590	+	0.6300	+	0.5650	+
number of interval	0.6719	+	0.5174	+	0.5547	+	0.5106	+

+ Positive relative relation. - negative relative relation.

7.2 Pair-Wise Distinguishing Ability

The different reading habits lead to user biases in user behavior metrics, for example, some users usually read longer no matter the news quality. To eliminate these user biases, we design an pair-wise evaluation approach to measure the ability of a behavior metric to distinguishing the quality. The pair-wise distinguishing ability is calculated by comparing the relative relation between the behavior metric and the news quality.

Specifically, consider an example set S of news pairs [$<n_1, n_2>$,... $<n_i, n_j>$], with the behavior metric b and news quality q. We first define the relative relation between behavior and quality of a news pair $<n_i, n_j>$:

$$r_\alpha(n_i, n_j) = I\big[\ \alpha I[b(n_i) < b(n_j)] = I[q(n_i) < q(n_j)]\ \big]$$

where α is the direction coefficient and I is the indicator function. Then the pair-wise distinguishing ability is defined as:

$$D_{pair}(b) = \max_{\alpha \in \{-1,1\}} \frac{\sum_{<n_i,n_j> \in S} r_\alpha(n_i, n_j)}{n(S)}$$

To eliminate the user bias, we generate the evaluation pairs within users, which means the n_i and n_j is from the same user. Beside of the expert labelled quality, we also use the user perceived quality as ground truth for news quality q, including content quality (UQ-C) and title consistency (UQ-T).

7.3 Results

The results of both point-wise and pair-wise distinguishing experiments are shown in Table 5. As for the objective quality labelled experts (EQ), among these behaviors, *reading ratio* achieves the highest point-wise distinguishing ability (0.7084) with threshold $t_b = 0.74$ and direction $r = 1$. It means that whether users read more than 74% of the news content can be used as an indicator for the high quality news. *Reading ratio* also achieves the highest pair-wise distinguishing ability (0.8010). It means that if a user reading more in news i than another news j, the quality of news i is more likely higher than the quality of news j. As for the perceived quality labelled by the user, UQ-C and UQ-T the *dwell time* and *reading ratio* performs better in distinguishing pair-wise quality.

To summary, based on the proposed point-wise and pair-wise distinguishing ability measurements, we investigate whether user behaviors can be used to identify news quality. The results demonstrate that user behaviors, especially *reading ratio* and *dwell time*, have highly ability to distinguish news quality. Starting from here, the powerful behavior-based quality identification models are promoted to be developed in the future.

8 CONCLUSION

In this work, through an in-depth user experiment, we investigate the quality effects on user item-level preference and user behaviors. To begin with, we verify that the quality does affect user preferences, and further find that the quality effect varies at different phases and it is more closely related to the user's interest for the news topic.

As for the quality effects on user behaviors, we firstly show that user click behavior is significantly affected by the quality. Specifically, low-quality news attracts more clicks especially when the user has higher interest in news topic. It can be interpreted by the higher title persuasion of low-quality news. Besides, a piece of low-quality news will affect not only the click probability of current news but also the news following it. Secondly, we find that a user behaves significantly different when interacting with low-quality news. A user will read less and slowly, with fewer revisits and fewer examinations.

Based on our finding that quality affects user preferences and behaviors, we proposed the PBQ-model which incorporates quality effects into the traditional behavior-only implicit feedback. The significant improvement in estimating user preferences demonstrates that considering quality is useful for building accurate implicit feedback. Because of the usefulness of quality, how to automatically identify quality is a valuable research topic. Our study shows the possibility of using user behaviors to identify news quality, such as *reading ratio* and *dwell time*.

The study on designing the model using content and behaviors information to predict the quality is left for future work. Beside of individual-level preference, the quality effect on user list-level satisfaction is also left for further study.

ACKNOWLEDGMENTS

We thank Prof. dr. Maarten de Rijke and Prof. Fang Yang for providing very useful comments for this paper. This work is supported by Natural Science Foundation of China (Grant No. 61672311, 61532011) and The National Key Research and Development Program of China (2018YFC0831900), and is partly supported by the Tsinghua-Sogou Tiangong Institute for Intelligent Computing.

REFERENCES

[1] Eugene Agichtein, Eric Brill, and Susan Dumais. 2006. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 19–26.

[2] 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.

[3] Ioannis Arapakis, Mounia Lalmas, B Barla Cambazoglu, Mari-Carmen Marcos, and Joemon M Jose. 2014. User engagement in online N ews: Under the scope of sentiment, interest, affect, and gaze. *Journal of the Association for Information Science and Technology* 65, 10 (2014), 1988–2005.

[4] 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.

[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] Mark Claypool, Phong Le, Makoto Wased, and David Brown. 2001. Implicit interest indicators. In *Proceedings of the 6th international conference on Intelligent user interfaces*. ACM, 33–40.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

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

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

[16] 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.

[17] 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.

[18] Diane Kelly and Nicholas J Belkin. 2004. Display time as implicit feedback: understanding task effects. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 377–384.

[19] Diane Kelly and Jaime Teevan. 2003. Implicit feedback for inferring user preference: a bibliography. In *Acm Sigir Forum*, Vol. 37. ACM, 18–28.

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

[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] 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.

[23] Dmitry Lagun, Donal McMahon, and Vidhya Navalpakkam. 2016. Understanding mobile searcher attention with rich ad formats. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 599–608.

[24] Gael Lederrey and Robert West. 2018. When Sheep Shop: Measuring Herding Effects in Product Ratings with Natural Experiments. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 793–802.

[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] Jimmy Lin, Salman Mohammed, Royal Sequiera, and Luchen Tan. 2018. Update Delivery Mechanisms for Prospective Information Needs: An Analysis of Attention in Mobile Users. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. ACM, New York, NY, USA, 785–794. https://doi.org/10.1145/3209978.3210018

[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] Hongyu Lu, Min Zhang, and Ma Shaoping. 2018. Between Clicks and Satisfaction: Study on Multi-Phase User Preferences and Satisfaction for Online News Reading. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

[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] Benjamin M Marlin and Richard S Zemel. 2009. Collaborative prediction and ranking with non-random missing data. In *Proceedings of the third ACM conference on Recommender systems*. ACM, 5–12.

[31] Ben Miroglio, David Zeber, Jofish Kaye, and Rebecca Weiss. 2018. The Effect of Ad Blocking on User Engagement with the Web. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 813–821.

[32] 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.

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

[34] 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.

[35] Ryen W White and Diane Kelly. 2006. A study on the effects of personalization and task information on implicit feedback performance. In *Proceedings of the 15th ACM international conference on Information and knowledge management*. ACM, 297–306.

[36] 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.

[37] 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.

[38] Xiaoying Zhang, Junzhou Zhao, and John Lui. 2017. Modeling the assimilation-contrast effects in online product rating systems: Debiasing and recommendations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 98–106.