Standing in Your Shoes: External Assessments for Personalized Recommender Systems

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

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

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

Maarten de Rijke
University of Amsterdam
& Ahold Delhaize
Amsterdam, The Netherlands
M.deRijke@uva.nl

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

The evaluation of recommender systems relies on user preference data, which is difficult to acquire directly because of its subjective nature. Current recommender systems widely utilize users' historical interactions as implicit or explicit feedback, but such data usually suffers from various types of bias. Little work has been done on collecting and understanding user's personal preferences via third-party annotations.

External assessments, that is, annotations made by assessors who are not the systems' users, have been widely used in information search scenarios. *Is it possible to use external assessments to construct user preference labels?* This paper presents the first attempt to incorporate external assessments into preference labeling and recommendation evaluation. The aim is to verify the possibility and reliability of external assessments for personalized recommender systems. We collect both users' real preferences and assessors' estimated preferences through a multi-role, multi-session user study. By investigating the inter-assessor agreement and user-assessor consistency, we demonstrate the reasonable stability and high accuracy of external preference assessments. Furthermore, we investigate the usage of external assessments in system evaluation. A higher degree of consistency with users' online feedback is observed, even better than traditional history-based online evaluation.

Our findings show that external assessments can be used for assessing user preference labels and evaluating systems in personalized recommendation scenarios.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; **Evaluation of retrieval results**; *Test collections*.

*Corresponding author.

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462916

KEYWORDS

Recommender System; Offline Evaluation; Preference Assessment

ACM Reference Format:
Hongyu Lu, Weizhi Ma, Min Zhang, Maarten de Rijke, Yiqun Liu, and Shaoping Ma. 2021. Standing in Your Shoes: External Assessments for Personalized Recommender Systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3404835.3462916

1 INTRODUCTION

User preference, as a critical ingredient of recommendation, is personalized, subjective, and even implicit [26]. This creates challenges for evaluating recommender systems. Traditionally, the collection of user preference labels has mainly relied on users' explicit and implicit feedback extracted from their historical interactions, such as ratings, clicks and dwell time [20, 29, 38]. However, historical feedback is known to be biased by various confounding factors, such as position [17, 18] and popularity [2], bring creates challenges for the evaluation of recommender systems.

So-called *external assessments*, that is, annotations made by external people (outside the system users themselves), are considered very difficult to use for such personalized user preference judgments [23]. In information search scenarios, external assessments have been widely, and successfully, used to build relevance labels for decades, as is evidenced by the Cranfield evaluation methodology and TREC-like benchmarking activities [6]. However, compared to a document's relevance in search, a user's preferences in a recommendation scenario are even more subjective and personalized [26]. There is little previous work trying to conduct the recommendation experiments via external assessments.

In this paper, we systematically examine the possibility and reliability of external preference assessments. As illustrated in Figure 1, users' personalized interests are hidden, but partially reflected in their historical interactions with recommender systems, e.g., through ratings or reviews on movie platforms. When presented with such preference records, external assessors can perceive users' implicit interests and estimate their preference labels for candidate items and fine-grained attributes. This process is similar to how humans perform peer recommendations in the real world, indicating

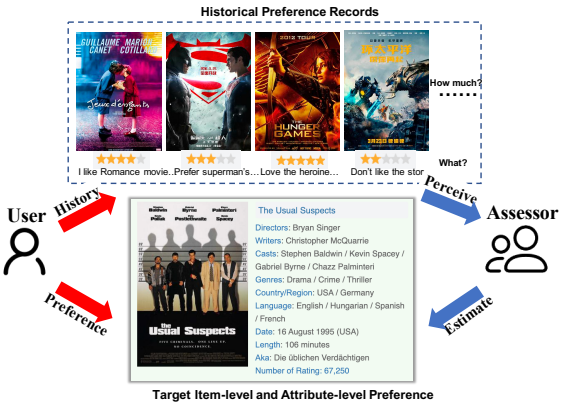


Figure 1: The main idea of our study. The assessor perceives the user interest based on the observation of user's preference history and estimates user preferences for candidate items and fine-grained attributes.

the potential of reliable external preference assessments.

We revisit the intuition that subjective preference is difficult to assess, by systematically investigating the reliability and consistency of external preference assessments, as well as their usability in system evaluations. Specifically, we aim to answer the following research questions:

(RQ1) Can a user's personalized preference be externally assessed? Are the results of external preference assessments consistent across different assessors? (*inter-assessor agreement*)

(RQ2) How accurate is the external preference assessment? (*user-assessor consistency*)

(RQ3) Can external preference assessment be used in recommender system evaluation? (*assessment-based evaluation*)

We conduct an in-depth, multi-role, multi-session laboratory user study in which participants are separated into two groups, called users and assessors, respectively. In the user-part of the study, the user's *historical preference records* are first collected. Based on those records, candidate items are built by *pooling* with various recommenders. After being exposed to these candidate items, the user's real preference labels are further collected. In the parallel assessor-part, the assessors are given the user's historical preference records, then asked to determine the user's interests and estimate the preference on the same set of items.

For the first research question, *inter-assessor agreement*, the agreement on point-wise and pair-wise preferences between multiple assessors is examined. To investigate the second research question, we measure the performance in terms of the gap between the external assessments and a user's real preferences. In addition, we investigate whether external assessors can correctly label a user's preferences on fine-grained item attributes. For the third research question, we examine the ability and strength of using external preference assessments to conduct system evaluation. By comparing with system evaluation results based on users' online feedback, we observe highly consistent results, outperforming traditional history-based offline system evaluation.

Our comprehensive analyses based on our user study data lead us to the interesting conclusion that *user preference can be assessed by external assessors with a moderate degree of agreement and accuracy, comparable to the performance of external assessments of relevance in search.* These encouraging findings suggest a new direction, i.e., to build user preference data and evaluate systems through external assessments. The external assessments do not rely on users and are not limited by the interacted items, and, hence, they are of great value for mitigating key issues faced by current recommender systems, such as bias in traditional history-based training and evaluation, and difficulties in evaluating the performance on new or inactive users, items and algorithms.

To summarize, our main contributions are the following:

(1) To the best of our knowledge, ours is the first work to systematically study the external assessment of users' personalized and subjective preferences. Our findings reveal new directions for preference labeling and evaluation of recommender systems.

(2) We demonstrate the possibility and reliability of external assessments of user preferences through comprehensive analyses on both inter-assessor agreement and user-assessor consistency.

(3) The usability and strength of incorporating external assessments into evaluation are demonstrated through a high degree of consistency with the online evaluation results.

2 RELATED WORK

Preference modeling in recommender systems. Finding items that match users' preferences is the main target of recommender systems. Hence, modeling user preferences forms the basis of recommendations. Recommender systems typically record historical information from a user's past interactions, such as ratings, clicks, reviews, and purchases [14, 21, 22, 34]. These records implicitly reflect a user's preferences [12, 25] and are widely used as implicit feedback [14, 32]. However, historical interaction feedback always suffers from confounding biases, such as position bias [17, 18], trust bias [17], result attractiveness [39], selection bias [30], positivity bias [16], presentation bias [37], and exposure bias [2, 11]. To address these issues, past research has explored a number of approaches, such as using advanced click models [4, 5, 10] to eliminate the effects of the position at which an item is presented, incorporating more post-click behavior [9, 15, 24], designing unbiased exploration strategies [13], and learning algorithms [19].

As current usage of historical feedback for measuring user preference suffers from these problems and may mislead systems by incorrectly modeling a user's preferences, we ask whether there could be other ways to gather users' preferences to supply current feedback, i.e., *external assessments*. If external assessors can reliably and accurately annotate users' preferences, it is possible to mitigate the biases listed above in training and evaluation. In this work we present a first and fundamental step, examining the possibility and reliability of external preference assessment.

External assessments in information retrieval. External assessment is the practice of collecting labels, such as image labels, document relevance labels, etc., by annotations from external peoples, and are commonly applied for human computation tasks [33] and used in information retrieval scenarios. A dominant way of experimentally evaluating information retrieval (IR) systems, widely known as the Cranfield or TREC-like paradigm, relies on the assessment of document relevance [7] by recruited human assessors,

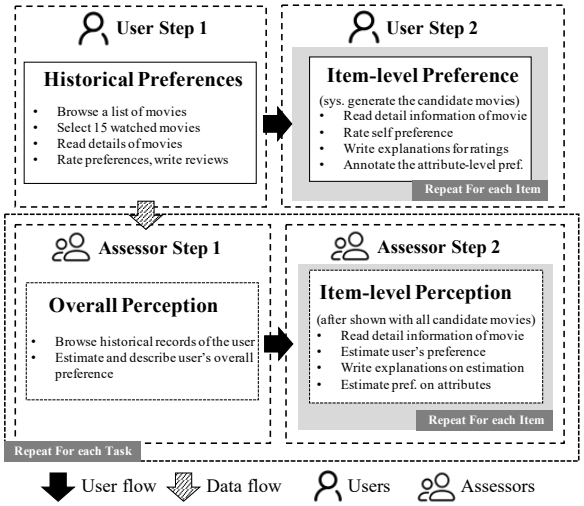


Figure 2: Procedure of our two-part user study: user-part and assessor-part.

such as experts or crowdsourced workers. Using instructions describing the underlying information needs and queries, assessors are asked to label the relevance of candidate documents [1, 27]. The underlying assumption is that relevance is an objective factor that can be perceived consistently among humans. However, such an objective target limits the system to perform similarly to all users, and may not be suitable for personalized scenarios that rely on subjective experiences. Mao et al. [28] find that usefulness, the user's subjective perception of document utility, is better correlated to user satisfaction.

As for user preferences in recommendation scenarios, Krishnan et al. [23] compare estimations for user preferences by human workers given historical ratings of the particular user, with the MovieLens system performance. Results indicate that the recommender system performs better than human annotators. Organisciak et al. [31] further investigate two task designs for collecting user preference annotations, taste-matching and taste-grokking, and demonstrate that when given a user's historical profiles, crowd workers can predict their preferences and the external assessors' prediction performance depends on the task scenarios.

The publications listed above show the promise of crowdsourcing user preferences in domains with a lack of explicit or implicit feedback from users. But we still lack a systematic examination of the external assessment of user preferences, including agreement among assessors and alignment with users themselves. Another important and less studied question concerns the usability and advantages of preference assessment in recommendation evaluation. All of these are matters that we address in this paper.

3 USER STUDY SETUP

In this section, we describe the design of our user study. As our participants have two roles, namely *user* and *assessor*, the study is divided into corresponding two parts. See Figure 2 for an overview.

3.1 User Part: Collecting Actual Preferences

The participants in this part are called *users*, and they are asked to complete two tasks in order.

Historical preference collection. To collect users' interests, we follow the preference elicitation approach proposed in previous work [23, 31]. In particular, users are presented with movies ordered by: $\log(pop_i) \times ent_i$, with the poster, title, attributes (e.g., directors, writers, actors, region) and plot synopsis shown. Here, pop_i denotes the popularity of item i, and ent_i denotes the entropy of all ratings for item i. To avoid biases, some information, such as average ratings and reviews are not displayed. The user is asked to browse the movies and select the ones he/she has watched until 15 movies have been collected. Then, for each movie selected (watched), the user is asked to rate his/her preference and to write a sentence explaining the rating. These collected historical preference records ([item, rating, review]×15) are of the same format as the Movielens dataset and can be directly used for training and recommendations.

After this session, for each user u, based on his/her historical ratings, we generate a set of recommendation movies by a *pooling* approach based on six classic recommenders (see Section 3.3 for details). These recommended movies will be used in later parts of the user study.

Self preference labeling. The second task starts after the *historical preference collection* and is aimed at collecting user's self-feedback for preference labels on the recommended items. Users are shown candidate items in a *random* order, with the same details as in the first task. After browsing the details, users are asked about their experience and preference, as shown in Figure 3 (left). Users' real (self-reported) preference for the item is collected ("Do you like this movie?"), along with an explanation. Moreover, we also collect the users' fine-grained preferences on the item's attributes using a 3-point scale (negative, neutral, positive). The collected users' self-feedback for their preferences at both the item level and the attribute level are used as self-preference labeling to measure the accuracy of external assessments.

3.2 Assessor Part: Collecting External Assessments of Preferences

The assessor-part of the user study starts after the user-part, the participants, namely (external) *assessors*, are disjoint from the *user* group. After a pre-experiment questionnaire including demographics and expertise, each assessor is *randomly* assigned with three target users as tasks. The steps within each task are the same: (i) overall interest perception, and (ii) item preference assessment.

Overall interest perception. First, the external assessor is shown the user's historical rating records collected in the user-part (historical preference collection), including the detailed information, user's ratings, and reviews for each movie. After browsing the records, assessors are asked to write their perception of the user's overall interest. This phase aims to guide the assessor to understand the user's preference.

Item preference assessment. This is the main step of the study, in which we collect assessor's annotations for the user's preference on the experimental items. The assessors are shown the movies in *random* order, and are asked to assess the user's preference on this movie, as shown in Figure 3 (right).

Figure 3: The main interface (translated) used in the User-Part for item preference collection (left), and in the Assessor-Part for item preference assessment (right).

The first question concerns the user's item-level preference ("Do you think he/she will like this movie?") plus an explanation. The second question asks for the user's attribute-level preferences on 3-point scale, the same as in the user-part. We refer to the collected annotations as (external) *preference assessments*, and they include both item-level and attribute-level annotations.

3.3 Experimental Details

Candidate item generation: Pooling. As the movie scenario is widely used in recommender systems research, a large number of the recommendation algorithms are designed and evaluated in this scenario. We also choose this scenario to conduct our study. Movielens, probably the most widely used dataset for recommender systems research, is selected for our study so as to allow for solid generalizability of our findings. To acquire precise recommendations and include the most recent movies, we use the *Movielens Latest* (9/2018) dataset, which has 27 million user-item interactions and the newest movies (up to 9/2018).

Candidate item generation. To generate a pool of candidate items used in preference labeling and assessment, we choose the *pooling* method, which is widely used in research and development in search to extract a sample of documents to be assessed for relevance [35], but is not yet applied for recommender systems, to the best of our knowledge. Here we utilize a pooling method to achieve more diverse candidate items, and make evaluation more fair and less biased. A total of six recommenders are used to build the item pool: *Random* (for controlling), *Pop*, *UserKnn*, *ItemKnn*, *BiasedMF* and *BPR*. These recommenders cover the most classical algorithms for rating prediction and ranking task, and are implemented by Lenskit [8] with well-tuned parameters.[1] The user's historical records collected in the first session (15 per user) are randomly split into two groups, 10 ratings named *hist/train* and 5 ratings named *hist/test*. The split results are fixed during the user study and the subsequent analyses. The whole Movielens dataset and the *hist/train* data are used to train the basic models for the target user. Considering the participants' burden and fatigue, we build the pool with depth = 3, i.e., the top-3 recommendations of each basic recommender are included.

Participants and data statistics. We recruit 16 user participants (8 are female) via email and online social networks for the user-part of the study. Most of the participants have a great interest for watching movies, 43.75% watch more than 10 movies yearly. We further recruit 19 assessor participants (different from the user participants, 14 are female). Each user and assessor is rewarded with US$20 as an incentive. After filtering the task data with errors due to network transmission and data logging, we end up with user preference labels for 284 ⟨user, item⟩ pairs and a total of 3,908 ⟨user, item, attribute⟩ labels, and 870 assessor preference annotations for these ⟨user, item⟩ pairs, and a total of 11,099 attribute-level annotations. The user study platform, setup, and collected data are released alongside this paper for future research.

4 CAN USER PREFERENCE BE ASSESSED?

4.1 Inter-Assessor Agreement

Analyzing the agreement and disagreement among assessors plays a crucial role in validating the possibility and stability of external assessment tasks. However, the question whether different assessors can achieve consistent assessments on the personalized and subjective user preference has so far remained unanswered.

As described in Section 3.2, for each target user, we *randomly* assign *three* assessors to conduct external assessments. Given the user's historical preference records, assessors perceive and annotate

[1] UserKnn (nnbrs = 30), ItemKnn (nnbrs = 30), BiasedMF (# factors = 40), BPR (# factors = 64).

Table 1: Summary of different statistics measuring inter-assessor agreement. Generally, for assessing preferences, the agreement among assessors achieves a moderate level.

Statistic	Value
<user, item>	
#samples	852
Krippendorff's α	0.425
Individual assessor ($\Delta(\alpha)$)	-0.037∼+0.043
Percentage agreement (5-scales)	0.3764
Percentage agreement (2-scales)	0.6777
Pearson's r	0.4150 (p<0.01)
Pearson's r (normalized pref.)	0.5030 (p<0.01)
⟨user, itemA, itemB⟩ pairwise	
#samples	2,383
Concordance	0.5921
Concordance (filter equal cases)	0.6968

the user's preference on the pooled recommendations. In total, at the item-level, for each ⟨user, item⟩ pair, three assessors' annotations are collected. Based on this data, we examine the agreement by multiple measurements used in the literature [1, 3, 36]:

- *Percentage agreement.* This metric counts the cases that received the same annotations by assessors, and divides the number by the total number of cases. The measure directly reflects the agreement among assessors.
- *Krippendorff's α.* This metric is calculated by looking at the overall distribution of assessors regardless of which assessors produced the judgments.
- *Pearson's correlation r.* This coefficient measures the linear correlation between two variables.
- *Concordance.* This metric measures the agreement of two assessors on pairwise relative relations.

The measurement results are summarized in Table 1.

Pointwise agreement (item-level). First, we consider the overall pointwise agreement among assessors, i.e., how different assessors agree when judging the same task (⟨user, item⟩ pair). Specifically, 19 assessors annotate the preference for a total of 870 ⟨user, item⟩ pairs. By using Krippendorff's α, we measure the overall agreement and obtain a value of 0.425, which can be seen as a reasonably moderate level. The degree of agreement will be discussed by comparing it to relevance assessments, at the end of this section.

To have an intuitive understanding, we measure the *percentage agreement* and obtain a value of 37.6% (5-scale), slightly lower but close to the 39% value of relevance assessment reported by Alonso and Mizzaro [1]. This finding is very encouraging as subjective preference can get a comparable level of inter-assessor agreement as relevance. To inspect the distinction between dislike and like, we convert the assessments into binary scale (≤ 3 as dislike, > 3 as like), and obtain a higher agreement, i.e., of 67.8%.

Pairwise agreement (item-level). To mitigate the rating scale bias of different assessors [27], we examine the agreement based on the pairwise relative relation. Specifically, for each ⟨user, itemA, itemB⟩ triple produced by an assessor, the original 5-scale annotations are converted into 3-scale (A < B, A = B, and A > B). For pairwise percentage agreement, we find that 59.2% of the pairs are consistent among different assessors, increasing to 69.7% if we only consider the non-neutral (A < B and A > B) cases. This high value of pairwise agreement further indicates the reasonable agreement of preference assessment.

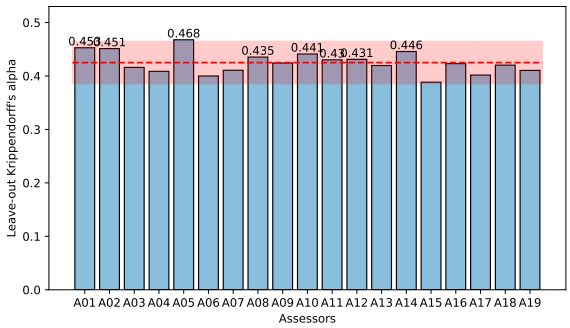


Figure 4: Krippendorff's α (ordinal) scores of overall preference assessments when leaving out results from one specific assessor. There is no outliers ($\pm 2\sigma$, outside the red area), indicating all the assessors did a reasonable job.

Variance of agreement among assessors. Besides overall agreement, we are also interested in variance, i.e., whether individual assessors agree with others. To this end, we conduct a leave-one-out experiment among assessors to study the impact of one assessor's assessments on the overall agreement. Specifically, each time we mask the assessments from one assessor, and compute the overall Krippendorff's α. If α increases, it means that his/her assessments undermine the overall agreement and hence has less agreement with other assessors. Figure 4 shows the leave-one-out α for each assessor, and compares to the overall value (dotted line). We see that there are small number of assessors slightly hurt the agreement, but there are no outliers. This indicates that all assessors perform reasonably, and the variance due to different assessors is within a moderate range.

Comparison to the literature and summary. To the best of our knowledge, ours is the first work to study the agreement for external assessment on subjective user preferences. We compare the agreement degree to the well-studied relevance assessment in the IR literature. Voorhees [36] measured the agreement among three assessors on the TREC-4 topics using the *overlap* metric (defined as the size of the intersection of the relevant document sets divided by the size of the union of the relevant document sets), and obtained 0.426 for two assessors and 0.301 for three assessors. By measuring the same measure on our preference assessment data, we get a value of 0.4728 for two assessors and 0.3252 for three assessors. Carterette et al. [3] examined and found using a 5-point relevance scale that trained assessors achieve a 43% percentage agreement, which increases to 69% for a binary scale. As for the 5-rating scale that we use, we obtain a lower agreement of 37.6%, but this increases to a comparable 67.7% when using a binary scale.

To summarize, we find that different external assessors can achieve a reasonably moderate agreement on the assessment of

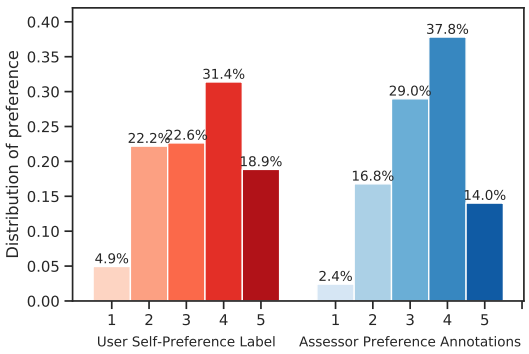


Figure 5: The distribution of assessor's item-level perceptions (red bars) and user's actual preference (blue bars).

personalized and subjective preferences. This observation is an encouraging finding on the possibility and stability of externally assessing user preferences in a recommendation scenario.

4.2 User-Assessor Consistency

In this section, based on the users' self-feedback on their own preferences, we are able to further examine the accuracy of the assessments by determining the user-assessor consistency. The results via multiple metrics are summarized in Table 2. We examine the consistency for the original 5-rating scale, and binary values converted ($\leq$ 3:0, $>$ 3:1).

Distribution. We first inspect the distribution of both assessor's assessments and user's self-feedback of actual preference, shown in Figure 5. Generally, the two distributions are very similar (Kullback-Leibler divergence $k = 0.0363$, high similarity). Compared to user preference, the assessors give slightly higher ratings (mean: 3.44 vs. 3.37). The preference assessment is more concentrated ($\sigma^2 = 1.0076$), while the users themselves give more decentralized ratings ($\sigma^2 = 1.3504$), indicating that assessors tend to rate more narrowly than users.

Individual accuracy. We further inspect the relationships between *individual* assessment and user self-preference at the item-level. Specifically, we group the ⟨user, item⟩ pairs based on the user preference ratings, and show the distribution of received assessments. Results are shown in Figure 6(a), which can also be seen as a transition matrix, e.g., the 0.41 at the self = 4, assess = 4 means that 41% of the items liked by the users are accurately assessed as like. Generally, the high values are close to the diagonal, intuitively reflecting the reasonable accuracy of preference assessments. The values derive the diagonal reflects the errors of assessments. We can observe that assessors are more likely to overestimate the preference, consistent to the finding of distribution analysis. Moreover, the errors are higher when user preference is low, indicating that user's negative preference, i.e., disliked items, are relatively harder to assess.

The results of the *percentage agreement* are shown in Figure 6(b); 32.9% of the ⟨user, item⟩ pairs are exactly matched between the assessments and user preference, similar to the agreement among assessors. If the constraints are eased to small adjacent errors (within one rating), the agreement increases to 77.8%.

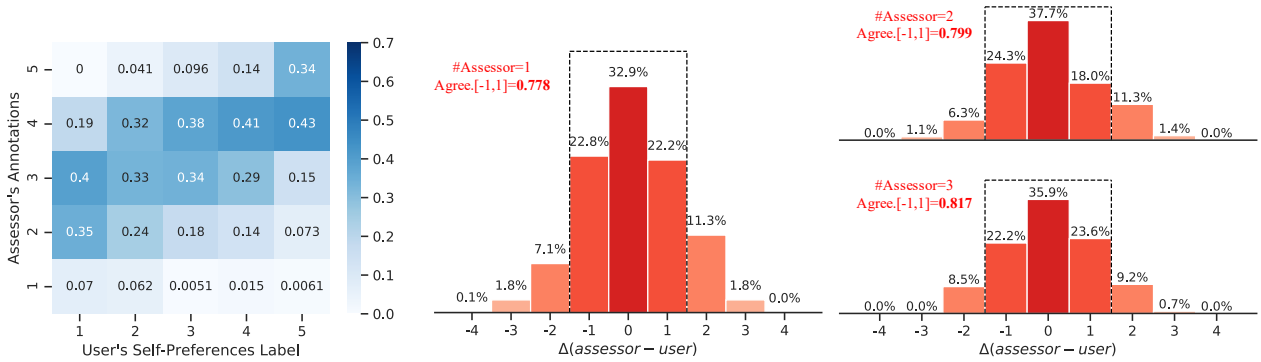
Table 2: Summary of different statistics measuring the consistency between external assessments and users' actual preference.

Stats	Five-scale	Binary
⟨user, item⟩ [user, assessor-1, assessor-2, ...]		
#samples	284	284
KL divergence	0.0363	0.0005
Krippendorff's α	0.393	0.296
Individual assessor ($\Delta(\alpha)$)	-0.032~0.017	-0.034~0.029
Percentage agreement	32.9%;[-1, 1]:77.8%	62.07%
Pearson's r	0.3552**	0.2414**
RMSE (as rating prediction)	1.2387	–
Accuracy (as classification)	–	0.6207
⟨user, item⟩ [user, assessor (aggregated)]		
α(average)	0.3933	–
α(majority vote)	–	0.2899
Percentage agreement (avg.)	36.3%;[-1,1]:82.0%	63.7%
Percentage agreement (maj.)	–	64.4%
⟨user, itemA, itemB⟩: pairwise relative		
#samples	2,383	2,383
Krippendorff's α	0.426	0.337
Concordance	0.5111	0.5115
Concordance (filter equal cases)	0.5387	0.4425

Aggregated accuracy. To achieve more reliable results, repeat and aggregation processes are widely utilized in relevance assessment [1]. Recall that in our user study, for each target user and preference, we collect repeated assessments from three different assessors. Based on this data, we examine to which extent aggregation can help the assessments for user preference. We show the change in percentage agreement when merging more repeat assessments in Figure 6(c). We observe that aggregation indeed improves the accuracy of assessment for user preference, from the individual (32.9%, soft match 77.8%) to 3-aggregated (35.9%, soft match 81.7%).

Accuracy on attribute-level preferences. User preferences on more fine-grained item attributes, such as genres and actors in our movie scenario, category and price in an e-commerce scenario, are of great value for many recommendation tasks, such as user profiling and explanation, etc., but are hardly collected. In the user study, we let external assessors annotate such fine-grained attribute-level preference. The labels of user's real preferences on the attributes are also collected. The assessments and user's self-feedback for attributes are collected via 3-scale feedback (negative, neutral, positive). On average, 20.2% (2.85 per item) attributes are labeled as positive and 4.8% (0.62 per item) attributes are labeled by users, leading to a very skewed distribution, while the ratios of assessors' assessment are positive (20.9%), negative (7.6%).

In order to examine the accuracy and gain an intuitive understanding of its degree, we measure the accuracy by three metrics, *precision*, *recall* and *Jaccard*, and compare its performance to the *Random* method (based on the general probability of *user* preferences and averaged by 100 repeated experiments) and a heuristic personalized method $AvgR_u$ (based on the user's average rating for the attribute: $>$ 3 = pos, $\leq$ 3 = neg, no occurrence = neutral). The



(a) Joint distribution of assessment and self preference label

(b) Distribution of gap between assessment and self preference

(c) Assessment consistency when aggregate different number of assessors

Figure 6: Comparison of assessments and users' actual self-preference.

results are shown in Table 3. We observe that assessors perform much better than the compared method, indicating the ability of assessors to perceive user's preferences on attributes. Compared to negative attributes, the positive ones are easier to perceive and assess (with both higher precision and recall).

Comparison to the literature and summary. We further compare to literature in information search scenario. Voorhees [36] measures the consistency between assessors to the author of the target topics (as primary assessor) by the *overlap*, and obtains values in the 0.4–0.5 range. By computing the same measure between assessors and users, we obtain the value 0.4378. Alonso and Mizzaro [1] investigate assessments given by crowdsourcing workers and achieve 68% agreement to the TREC assessments (binary scale). Comparing assessors and users, we obtain a similar result (62.1%), slightly lower but comparable to the level of relevance judgments. Query-document relevance is a more objective property than personalized user-item preferences. Therefore, it is encouraging that external preference assessment can achieve comparable levels of agreement and accuracy as relevance assessments.

In summary, by measuring the *inter-assessor agreement* and *user-assessor consistency*, we find that external assessors can perceive user's preferences on items, with a moderate-level agreement and accuracy. By aggregating multiple annotations, the accuracy can be further improved. These findings suggest the possibility of using external assessments to gather labels of user preference, in a stable and accurate manner. The finding that assessors can assess user preference not only on items but also on attributes has important implications. By only using historical offline records, we may not know about a user's interest concerning attributes, while through external assessments, we may be able to collect these fine-grained preference labels, and further improve recommendation applications, e.g., user profiling, explanation evaluation, etc.

5 USING ASSESSMENT IN RECOMMENDER SYSTEM EVALUATION

In this section we study whether external preference assessments can be used in the offline evaluation of recommender systems.

Table 3: Accuracy of attribute-level preference assessments.

		Precision	Recall	Jaccard
Positive	random	0.2051	0.1815	0.0935
	AvgR_u	0.3488	0.3987	0.2019
	Assessor	**0.4585**	**0.4568**	**0.2543**
Negative	random	0.0432	0.0783	0.0273
	AvgR_u	0.0500	0.0292	0.0165
	Assessor	**0.2484**	**0.3226**	**0.1522**
Neutral	random	0.6778	0.7443	0.5482
	AvgR_u	0.5829	0.6581	0.4451
	Assessor	**0.7531**	**0.7455**	**0.5874**

5.1 Multi-Source Labels

Traditional offline approaches to recommender system evaluation rely on users' historical interactions (as labels). They treat all missing value items as negative, leading to underestimation and biases of the evaluation results for new systems. Online evaluation can truly capture the degree of satisfaction of users with the recommender system and collect their feedback as online labels, but it is costly and faces risks of harming the user experience. We consider external assessments as an alternative to help recommender system evaluation. To validate this assumption, we design and conduct three evaluation experiments based on our user study data.

As described in Section 3.2, for each participant (user) in our user study, we have collected feedback for his/her preference from three different sources: (1) user's preference on historical watched items; (2) user's preference on recommended items; and (3) assessor's assessment for the user's preference on the same recommended items. Recall that the items in the study are collected by *pooling* based on the *random* and five classic recommenders, *Pop*, *UserKnn*, *ItemKnn*, *BiasedMF* and *BPR*. It is important to note that through the pooling method, we collect full information about the users' preferences on the recommended items of each algorithm. Based on the full information, we can evaluate the systems without the impact of missing value issue as we discussed.

Based on these multi-source labels, we first build several different datasets, as summarized in Table 4, for simulating the evaluation experiments. The Movielens dataset is used as basic data for training

Table 4: Datasets from multiple sources for the users (history, self-preference, external assessment).

Movielens	
ml/train	ml/test (leave-5-out)
Historical Preference Label (hist)	
hist/train(×10)	hist/test(×5)
Self Preference Label (self)	
self/all (×18)	self/pop(×3), itemknn(×3), biasemf...
External Assessment (assess)	
assess/all (×18)	assess/pop(×3), itemknn(×3), biasedmf

Table 5: Experimental settings for recommender system training and evaluation. ("B.O." is short for "based on.")

	Train	Test
Dataset Settings		
B.O. User History	ml+hist/train	hist/test
B.O. User self-preference	ml+hist/train	self
B.O. Assessment	ml+hist/train	assess
	negative samples	metrics
Evaluation Settings		
Rank (Full)	all other items	nDCG
Rank (Sample)	sampled other items	nDCG
Rating Prediction	no, only test samples	RMSE

the recommender models; users' historical preferences collected in the user-part of the user study, namely *hist*, are randomly split into two groups, 10 ratings for training (*hist/train*) and 5 ratings for testing (*hist/test*); users' explicit feedback on their actual preference labels on the pooled recommendations are referred to as *self*; and external assessors' assessments of users' preferences on the same recommendations are referred to as *assess*. Using these datasets, we further design and conduct evaluation experiments to address our research question.

5.2 Evaluation Based on External Assessment

To demonstrate the validity of using external assessments for evaluation, we design evaluation experiments by comparing the system evaluation results based on user's actual preference feedback, or based on external assessments. The settings of our experiments are shown in Table 5. The training set is the same across experiments, the only difference is the test set, including *self*, *assess*, *hist/test*. Based on the *self* test set, we simulate an online evaluation. We also simulate a traditional offline evaluation by using the left-out *hist/test* set which includes only user ratings on some watched items (as with most traditional recommendation evaluation setups).

The systems to evaluate include *Random*, *Pop*, *UserKnn*, *ItemKnn*, *BiasedMF* and *BPR*, which are used in the user study (as described in Section 3.3). To cover a diverse of evaluation settings and facilitate the generalizability of our findings, we choose three widely-used evaluation tasks and options:

- *Ranking task, full-ranking setting* (treating the missing value as zero, i.e., disliked), measured using nDCG (@all, grade gains)
- *Ranking task, negative sampling setting* (for each user, sample the non-interacted items as negative items), measured using nDCG (@all, grade gains).
- *Rating prediction task*, measured using RMSE. In this experiment, we only compare the rating prediction methods: UserKnn, ItemKnn, and BiasedMF.

The results of our experiments are shown in Figure 7. Based on the user's self-feedback of the preference labels for the pooled recommendations, we evaluate the performance of each recommender (shown in Figure 7(a)). Notice that we are not aiming to find the best performing recommendation algorithm, hence the observations about the performance ranking are only used to compare the evaluation settings.

Instead of using the users' preference labels on pooled items (Figure 7(a)), which are infeasible to collect in a real scenario, an evaluation using external preference assessments is conducted. The results are shown in the Figure 7(b). We see that both results are very similar to each other, leading to the same relative rank of the recommendation algorithms, though the absolute metric values are different. Through the bootstrapping method, we measure the agreement of system rankings evaluated based on user preference and external assessor's assessments in terms of Kendall's τ (for the ranking-full setting: $\tau = 0.769$; for the rank-sampling setting: $\tau = 0.627$; for the rating prediction setting: $\tau = 0.874$). The Kendall τ scores are consistently high across different evaluation settings (rank and rating prediction, full and sampling).

We also compare the results of the traditional history-based offline evaluation method, which suffers from the missing-values issue, as shown in Figure 7(c). Generally, the rankings of systems are significantly different from the evaluation results based on full-information of user's real preference labels. This observation confirms the gap between traditional offline evaluation results and online performance. The consistency between using external assessments and user's real preference labels reveals the potential of the proposed external assessment-based offline evaluation method.

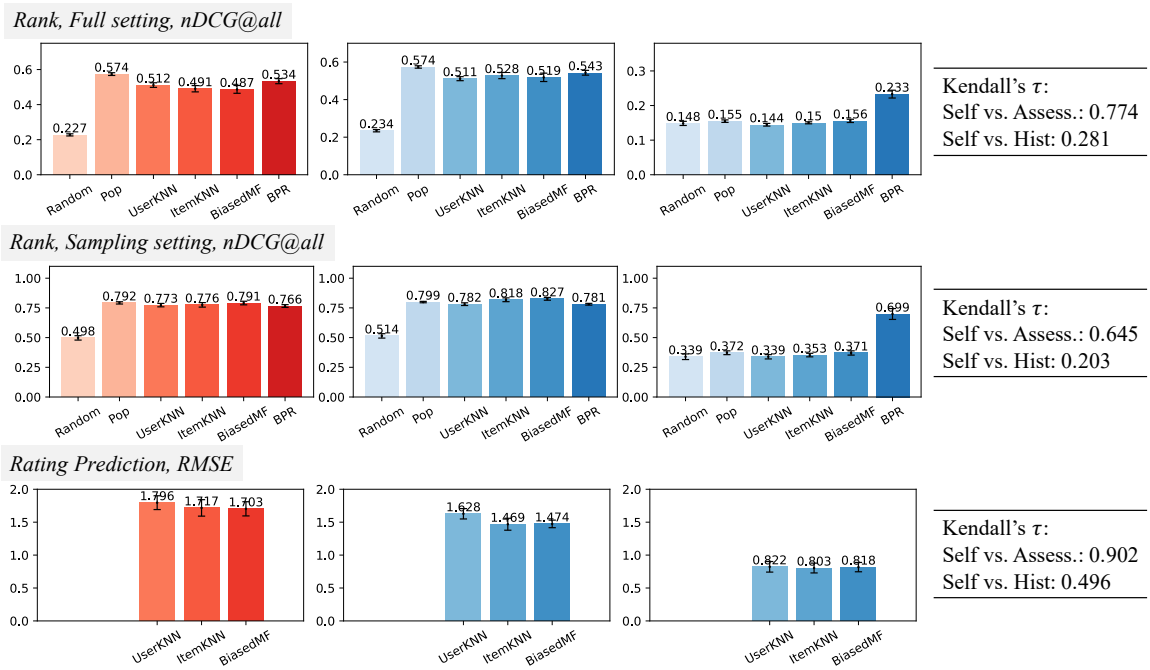
6 DISCUSSION

Strength of preference assessments. We have demonstrated the potential of using external assessments of user preferences in a personalized recommendation scenario. Here, we summarize the strengths of this method for preference labeling and evaluation.

First, external assessment has no constraint on the labeled items. Previous preference feedback relies on the users' historical interactions, which means that only the items that have been interacted can have a label and be included for training and evaluation, while external assessment can be applied to items that have not been interacted with. Second, external assessors can not only annotate item-level preferences, but also more fine-grained attribute-level preferences. This has great implications for many recommendation tasks. Moreover, it suggests a potential way of incorporating human knowledge into recommender system building and optimization.

Limitations. Our study is only conducted in the movie recommendation scenario to seek a match with a large proportion of previous recommender system research. As an intrinsic human cognitive activity, preference perception as well as our findings and methodologies could be generalized to other domains.

As the research about assessment-based recommendation evaluation is still at its beginning and needs exploration, we conduct an



(a) Evaluation based on user self-preference labels

(b) Evaluation based on external assessments

(c) Evaluation based on user history (simulate offline)

Figure 7: Evaluation results based on different labels: (a) users' *real preference labels* for the pooled recommendations; (b) the external *assessors' assessments*; (c) users' *historical preference records*. Rows present different evaluation tasks. Notice the high consistency between users and assessors in system rankings.

in-depth user study (average two hours per participants) in a laboratory environment. Although the number of participants (32 in total) is reasonable compared to previous user studies [23, 31], we believe that conducting large-scale, more in-depth and controllable studies or directly investigating a large-scale industrial recommender system is a natural and valuable but challenging next step.

7 CONCLUSION AND FUTURE WORK

Through an in-depth user study and systematic analyses, we have investigated whether users' personalized preferences can be assessed by external assessors. In general, we find that different external assessors can achieve a consistent assessment of users' preferences, reaching a reasonably moderate agreement. Moreover, external preference assessments can accurately match users' real preferences, and the accuracy can be further improved by aggregating multiple assessments. Based on the findings, we conclude that user preferences can be assessed with a reasonable consistency and accuracy. Furthermore, we have examined whether the external preference assessment results can help the system evaluation. Our results show that system rankings based on external preference assessments are consistent with those based on users' actual preferences, more than those based on user history, indicating its potential and strength.

To the best of our knowledge, this is the first work that systematically examines the usability and reliability of external assessments of user preferences in a recommendation scenario. Our findings represent a step towards a new assessment-based offline evaluation methodology for recommender systems. Along this direction, there are many valuable directions for future research, such as (i) improving the accuracy of external preference assessments; (ii) improving both training and evaluation of recommender systems by incorporating *small-size* external assessment data; and (iii) applying external assessments in recommendation tasks, e.g., for explanation generation and evaluation.

REPRODUCIBILITY

To facilitate reproducibility of the results in this paper, we are sharing the code and data used in this paper at https://github.com/luhongyu/Preference-Assessment.

ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China (2018YFC0831900), the Natural Science Foundation of China (Grant No. 62002191, 61672311, 61532011), and Tsinghua University Guoqiang Research Institute. This study is also funded by the China Postdoctoral Science Foundation (2020M670339) and Dr Weizhi Ma is supported by the Shuimu Tsinghua Scholar Program. This research was (partially) funded by the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Omar Alonso and Stefano Mizzaro. 2012. Using crowdsourcing for TREC relevance assessment. *Inf. Process. Manag.* 48, 6 (2012), 1053–1066. https://doi.org/10.1016/j.ipm.2012.01.004

[2] Rocío Cañamares and Pablo Castells. 2018. Should I Follow the Crowd?: A Probabilistic Analysis of the Effectiveness of Popularity in Recommender Systems. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '18)*. ACM, New York, NY, USA, 415–424. https://doi.org/10.1145/3209978.3210014

[3] Ben Carterette, Paul N Bennett, David Maxwell Chickering, and Susan T Dumais. 2008. Here or There: Preference Judgments for Relevance. In *European Conference on Information Retrieval*. Springer, 16–27.

[4] 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, WWW 2009, Madrid, Spain, April 20-24, 2009*. ACM, 1–10. https://doi.org/10.1145/1526709.1526711

[5] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. 2016. Click Models for Web Search and their Applications to IR: WSDM 2016 Tutorial. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, February 22-25, 2016*. ACM, 689–690. https://doi.org/10.1145/2835776.2855113

[6] Cyril Cleverdon. 1967. The Cranfield Tests on Index Language Devices. In *Aslib Proceedings*. MCB UP Ltd.

[7] 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 (SIGIR '08)*. Association for Computing Machinery, New York, NY, USA, 331–338. https://doi.org/10.1145/1390334.1390392

[8] Michael D. Ekstrand. 2018. The LKPY Package for Recommender Systems Experiments: Next-Generation Tools and Lessons Learned from the LensKit Project. *arXiv preprint arXiv:1809.03125* abs/1809.03125 (2018). arXiv:1809.03125 http://arxiv.org/abs/1809.03125

[9] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan T. Dumais, and Thomas White. 2005. Evaluating implicit measures to improve web search. *ACM Trans. Inf. Syst.* 23, 2 (2005), 147–168. https://doi.org/10.1145/1059981.1059982

[10] Fan Guo, Chao Liu, and Yi Min Wang. 2009. Efficient multiple-click models in web search. In *Proceedings of the Second International Conference on Web Search and Web Data Mining, WSDM 2009, Barcelona, Spain, February 9-11, 2009*. ACM, 124–131. https://doi.org/10.1145/1498759.1498818

[11] Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Tat-Seng Chua. 2016. Fast Matrix Factorization for Online Recommendation with Implicit Feedback. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*. ACM, 549–558. https://doi.org/10.1145/2911451.2911489

[12] Wendell Hicken, Frode Holm, James Clune, and Marc Campbell. 2005. Music Recommendation system and method. (February~17 2005). US Patent App. 10/917,865.

[13] Katja Hofmann, Shimon Whiteson, and Maarten de Rijke. 2011. Balancing Exploration and Exploitation in Learning to Rank Online. In *Advances in Information Retrieval - 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings (Lecture Notes in Computer Science)*, Vol. 6611. Springer, 251–263. https://doi.org/10.1007/978-3-642-20161-5_25

[14] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*. IEEE Computer Society, 263–272. https://doi.org/10.1109/ICDM.2008.22

[15] Jeff Huang and Abdigani Diriye. 2012. Web User Interaction Mining from Touch-enabled Mobile Devices. In *HCIR Workshop*.

[16] Jin Huang, Harrie Oosterhuis, Maarten de Rijke, and Herke van Hoof. 2020. Keeping Dataset Biases out of the Simulation: A Debiased Simulator for Reinforcement Learning based Recommender Systems. In *RecSys 2020: Fourteenth ACM Conference on Recommender Systems, Virtual Event, Brazil, September 22-26, 2020*. ACM, 190–199. https://doi.org/10.1145/3383313.3412252

[17] Thorsten Joachims, Laura A. Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017. Accurately Interpreting Clickthrough Data as Implicit Feedback. *SIGIR Forum* 51, 1 (2017), 4–11. https://doi.org/10.1145/3130332.3130334

[18] Thorsten Joachims, Laura A. 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 Trans. Inf. Syst.* 25, 2 (2007), 7. https://doi.org/10.1145/1229179.1229181

[19] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2018. Unbiased Learning-to-Rank with Biased Feedback. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*. ijcai.org, 5284–5288. https://doi.org/10.24963/ijcai.2018/738

[20] Youngho Kim, Ahmed Hassan Awadallah, Ryen W. White, and Imed Zitouni. 2014. Modeling dwell time to predict click-level satisfaction. In *Seventh ACM International Conference on Web Search and Data Mining, WSDM 2014, New York, NY, USA, February 24-28, 2014*. ACM, 193–202. https://doi.org/10.1145/2556195.2556220

[21] Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*. ACM, 426–434. https://doi.org/10.1145/1401890.1401944

[22] Yehuda Koren, Robert M. Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (2009), 30–37. https://doi.org/10.1109/MC.2009.263

[23] Vinod Krishnan, Pradeep Kumar Narayanashetty, Mukesh Nathan, Richard T. Davies, and Joseph A. Konstan. 2008. Who predicts better?: results from an online study comparing humans and an online recommender system. In *Proceedings of the 2008 ACM Conference on Recommender Systems, RecSys 2008, Lausanne, Switzerland, October 23-25, 2008*. ACM, 211–218. https://doi.org/10.1145/1454008.1454042

[24] 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, Perth, Australia, April 3-7, 2017*. ACM, 525–533. https://doi.org/10.1145/3041021.3054182

[25] Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. 2010. Personalized news recommendation based on click behavior. In *Proceedings of the 15th International Conference on Intelligent User Interfaces, IUI 2010, Hong Kong, China, February 7-10, 2010*. ACM, 31–40. https://doi.org/10.1145/1719970.1719976

[26] Hongyu Lu, Min Zhang, and Shaoping Ma. 2018. Between Clicks and Satisfaction: Study on Multi-Phase User Preferences and Satisfaction for Online News Reading. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*. ACM, 435–444. https://doi.org/10.1145/3209978.3210007

[27] Eddy Maddalena, Stefano Mizzaro, Falk Scholer, and Andrew Turpin. 2017. On Crowdsourcing Relevance Magnitudes for Information Retrieval Evaluation. *ACM Trans. Inf. Syst.* 35, 3 (2017), 19:1–19:32. https://doi.org/10.1145/3002172

[28] Jiaxin Mao, Yiqun Liu, Huan-Bo Luan, Min Zhang, Shaoping Ma, Hengliang Luo, and Yuntao Zhang. 2017. Understanding and Predicting Usefulness Judgment in Web Search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*. ACM, 1169–1172. https://doi.org/10.1145/3077136.3080750

[29] Douglas W Oard, Jinmook Kim, et al. 1998. Implicit feedback for recommender systems. In *Proceedings of the AAAI workshop on recommender systems*, Vol. 83. WoUongong.

[30] Harrie Oosterhuis and Maarten de Rijke. 2021. Unifying Online and Counterfactual Learning to Rank: A Novel Counterfactual Estimator that Effectively Utilizes Online Interventions. In *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021*. ACM, 463–471. https://doi.org/10.1145/3437963.3441794

[31] Peter Organisciak, Jaime Teevan, Susan T. Dumais, Robert C. Miller, and Adam Tauman Kalai. 2014. A Crowd of Your Own: Crowdsourcing for On-Demand Personalization. In *Proceedings of the Seconf AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2014, November 2-4, 2014, Pittsburgh, Pennsylvania, USA*. AAAI. http://www.aaai.org/ocs/index.php/HCOMP/HCOMP14/paper/view/8972

[32] Rong Pan, Yunhong Zhou, Bin Cao, Nathan Nan Liu, Rajan M. Lukose, Martin Scholz, and Qiang Yang. 2008. One-Class Collaborative Filtering. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*. IEEE Computer Society, 502–511. https://doi.org/10.1109/ICDM.2008.16

[33] Alexander J. Quinn and Benjamin B. Bederson. 2011. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the International Conference on Human Factors in Computing Systems, CHI 2011, Vancouver, BC, Canada, May 7-12, 2011*. ACM, 1403–1412. https://doi.org/10.1145/1978942.1979148

[34] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009*. AUAI Press, 452–461.

[35] Nicola Stokes. 2006. *TREC: Experiment and Evaluation in Information Retrieval* Ellen M. Voorhees and Donna K. Harman (editors) (National Institute of Standards and Technology), Cambridge, MA: The MIT Press (Digital libraries and electronic publishing series, edited by William Y. Arms), 2005, x+462 pp; hardbound, ISBN 0-262-22073-3. *Comput. Linguistics* 32, 4 (2006), 563–567. https://doi.org/10.1162/coli.2006.32.4.563

[36] Ellen M. Voorhees. 1998. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*. ACM, 315–323. https://doi.org/10.1145/290941.291017

[37] Chao Wang, Yiqun Liu, Min Zhang, Shaoping Ma, Meihong Zheng, Jing Qian, and Kuo Zhang. 2013. Incorporating vertical results into search click models. In *The 36th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '13, Dublin, Ireland - July 28 - August 01, 2013*. ACM,

503–512. https://doi.org/10.1145/2484028.2484036

[38] Peifeng Yin, Ping Luo, Wang-Chien Lee, and Min Wang. 2013. Silence is also evidence: interpreting dwell time for recommendation from psychological perspective. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*. ACM, 989–997. https://doi.org/10.1145/2487575.2487663

[39] 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, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*. ACM, 1011–1018. https://doi.org/10.1145/1772690.1772793