Comparing In Situ and Multidimensional Relevance Judgments

Jiepu Jiang
Center for Intelligent Information Retrieval, University of Massachusetts Amherst
jpjiang@cs.umass.edu

Daqing He
School of Computing and Information, University of Pittsburgh
dah44@pitt.edu

James Allan
Center for Intelligent Information Retrieval, University of Massachusetts Amherst
allan@cs.umass.edu

ABSTRACT

To address concerns of TREC-style relevance judgments, we explore two improvements. The first one seeks to make relevance judgments contextual, collecting in situ feedback of users in an interactive search session and embracing usefulness as the primary judgment criterion. The second one collects multidimensional assessments to complement relevance or usefulness judgments, with four distinct alternative aspects examined in this paper—novelty, understandability, reliability, and effort.

We evaluate different types of judgments by correlating them with six user experience measures collected from a lab user study. Results show that switching from TREC-style relevance criteria to usefulness is fruitful, but in situ judgments do not exhibit clear benefits over the judgments collected without context. In contrast, combining relevance or usefulness with the four alternative judgments consistently improves the correlation with user experience measures, suggesting future IR systems should adopt multi-aspect search result judgments in development and evaluation.

We further examine implicit feedback techniques for predicting these judgments. We find that click dwell time, a popular indicator of search result quality, is able to predict some but not all dimensions of the judgments. We enrich the current implicit feedback methods using post-click user interaction in a search session and achieve better prediction for all six dimensions of judgments.

KEYWORDS

Relevance judgment; search experience; implicit feedback.

ACM Reference format:
Jiepu Jiang, Daqing He, and James Allan. 2017. Comparing In Situ and Multidimensional Relevance Judgments. In *Proceedings of SIGIR '17, Shinjuku, Tokyo, Japan, August 07–11, 2017,* 10 pages.
DOI: 10.1145/3077136.3080840

1 INTRODUCTION

Test collection-based IR evaluation relies on human assessments of search result quality. The most popular method is the Cranfield-style relevance judgments [9], such as the approach used in TREC [10], where assessors (usually trained experts) judge a preassigned set of search results one after another using criteria that focus on topical relevance. This method had achieved great success but also attracted criticism such as focusing solely on topical relevance and ignoring real users' perceptions of the usefulness of results in a particular search context. We examine two directions to improve this status quo.

One direction is to incorporate context into assessments. That is, the value of a search result depends on the scenario and context of accessing the result. Belkin et al. [5] proposed to evaluate interactive search systems by the usefulness of each interaction for accomplishing a search task. We can apply this model to search result judgments—to assess the usefulness of a click (the perceived usefulness of a clicked result). This intrinsically requires us to switch from relevance to usefulness as the primary judgment criteria, and to collect in situ judgments to take into account the particular time and context of accessing a search result.

Two recent efforts [25, 36] examined this direction. Kim et al. [25] collected users' in situ feedback of clicked results after they had finished examining the results. However, they restricted the in situ feedback to "thumbs-up" or "thumbs-down". Mao et al. [36] asked users to assess the usefulness of the clicked results after a search session without considering the particular context. Both studies reported improved correlations with search experience measures comparing to TREC-style relevance judgments *by external assessors*. However, neither study excluded the influence of the difference between searchers and external assessors on relevance judgments.

The other direction is to use a combination of multiple aspects of judgments. Many previous studies tried to complement relevance with seemingly reasonable dimensions, such as novelty [6, 55], understandability [41, 56], credibility [39, 46, 51, 53], readability [42, 49], effort [20, 50, 54], freshness [11], etc. Multidimensional judgments are also popular approaches used in user-centric evaluation models [19, 27, 52]. However, most previous IR studies had only examined one particular alternative dimension to relevance, and they had not verified the value of multidimensional judgments by correlating with user experience measures.

We evaluate and compare these two directions. We collected users' search result judgments from six dimensions (relevance, usefulness, novelty, understandability, reliability, and effort) in two settings—an in situ one that happened right after users had finished examining a clicked search result (called *in situ judgments*), and a context-independent one collected after a search session (called *post-session judgments*). We evaluate the two types of judgments on six dimensions by correlating with six search experience measures collected from a laboratory user study. We also examined implicit feedback methods for predicting these judgments.

We examine the following questions in the rest of this article:

- Do in situ judgments better correlate with search experience measures than context-independent (post-session) ones? Do multiple

SIGIR '17, Shinjuku, Tokyo, Japan

DOI: 10.1145/3077136.3080840

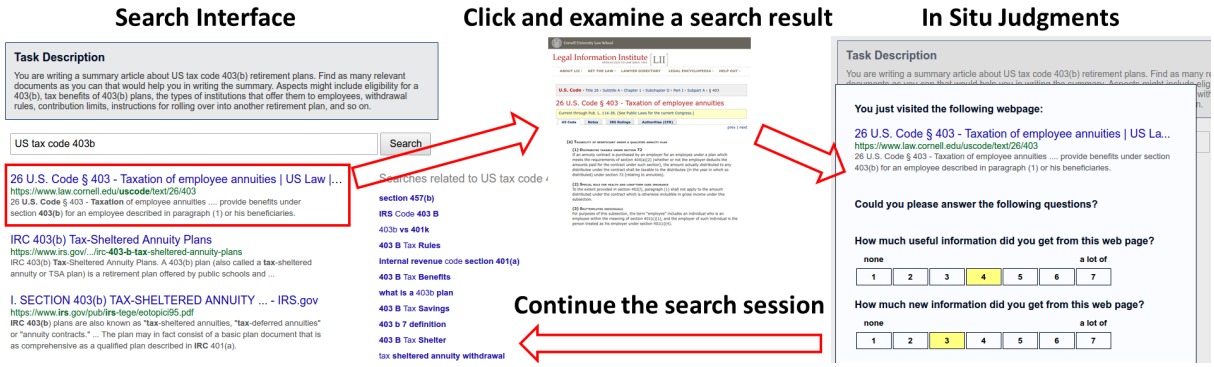

Figure 1: A screenshot of the search interface and the in situ judgments interface.

dimensions of judgments help relevance/usefulness judgments better correlate with search experience measures? Which dimensions of judgments should we collect to improve a particular user experience measure? Section 3 seeks answers to these questions.

- Can we effectively predict different search result judgments using implicit feedback signals? Section 4 and Section 5 examine techniques for addressing this challenge.

2 USER STUDY

We designed a user study to collect search result judgments. The user study asked participants to work on different tasks in an experimental search system. We recorded users' search behavior and collected their in situ and post-session search result judgments.

2.1 Experiment Design

The user study employed a 2×2 within-subject design to balance different types of search tasks. The tasks come from the TREC session tracks [7] and were categorized into four types by the targeted task product and goal based on Li and Belkin's faceted classification framework [28]. The targeted task product is either *factual* (to locate facts) or *intellectual* (to enhance the user's understanding of a topic). The goal of a task is either *specific* (clear and fully developed) or *amorphous* (an ill-defined or unclear goal that may evolve along with the user's exploration).

We divided participants into groups of four. Participants in the same group worked on the same four tasks (one task for each type) but using a different sequence (rotated using a Latin square). We assigned different tasks to different groups to increase task diversity.

For each task, the participants went through two stages:

- **Search stage (10 minutes)**. The participants performed an interactive search session to address the task. They could submit and reformulate any queries and click on any search results. After clicking on a result's link, the participants switched to the result webpage in a new browser tab. When they had finished examining the result and turned back to the SERP, the participant needed to provide in situ judgments on the clicked results before they could resume the search session. Figure 1 shows the screenshots of the search interface and the in situ judgments.
- **Judgment stage (about 10 minutes)**. The participants rated their search experience in the session and finished post-session judgments on each result they visited in the session. Section 2.2 introduces details of the in situ and post-session judgments.

As Figure 1 shows, the interface of the experimental system is similar to popular web search engines. The system redirected users' queries to Google and returned filtered Google search results. The system only showed the "10-blue links", vertical search results (except image verticals), and related queries. Other SERP elements were removed to simplify the user study. The system displayed results in the same way they would appear on Google. The main difference between our system and Google in SERP design was that our system showed task description on the top of a SERP (to help participants recall task requirements) and we showed related searches on the right side of a SERP.

The participants spent about 100 minutes to finish an experiment. First, they worked on a training task (including all the steps) for 10 minutes. Then, they worked on four formal tasks, spending about 20 minutes on each task. We required the participants to take a 5-minute break after two formal tasks to reduce fatigue.

2.2 Collecting Search Result Judgments

We collected search result judgments in two different scenarios:

- **In situ judgments** – participants assessed a clicked result when they had finished examining it and turned back to the SERP.
- **Post-session judgments** – the judgments collected after a search session (in the judgment stage).

The in situ judgments measure the participants' perceptions of the clicked result at (roughly) the same time and contexts they visit the result. The approach is similar to Kim et al. [25], except that we adopted different measures to assess search results. In the search stage, we instructed the participants to examine results as they would normally do when using a search engine in their daily lives. For example, they did not need to fully read a result and they could abandon examining. Particularly, they were instructed that during the in situ judgments, they should not revisit the result for the purpose of answering the judgment questions (and we did not offer a link for revisiting in the in situ judgment interface). This is to ensure that the in situ judgments only measure participants' perceptions of the latest click activity.

The post-session judgments resemble the TREC-style relevance judgments, where the assessors judge results without a particular search context and in a random order—they are asked to judge a set of results one after another in detail. In our post-session judgments, the assessors are real searchers. We asked them to judge the set of results they visited in the session. We instructed them to examine

Table 1: Questions for collecting search result judgments and users' search experience.

Search Result Judgments	Question & Options
Topical Relevance (TRel)	How relevant is this webpage? • *Key* (3): this page or site is dedicated to the topic; authoritative and comprehensive; it is worthy of being a top result. • *Highly Relevant* (2): the content of this page provides substantial information on the topic. • *Relevant* (1): the content of this page provides some information on the topic, which may be minimal. • *Not Relevant* or *Spam* (0).
Usefulness (Usef)	**In Situ**: How much useful information did you get from this web page? From 1 (none) to 7 (a lot of). **Post-session**: How much useful information did this web page provide for the task? From 1 (none) to 7 (a lot of).
Novelty (Nov)	How much new information did you get from this web page? From 1 (none) to 7 (a lot of).
Effort (Effort)	How much effort did you spend on this web page? From 1 (none) to 7 (a lot of).
Understandability (Under)	How difficult was it for you to follow the content of this web page? From 1 (very difficult) to 7 (very easy).
Reliability (Relia)	How trustworthy is the information in this web page? From 1 (not at all trustworthy) to 7 (very trustworthy).

Search Experience Measures	Question & Options
Satisfaction (Sat)	How satisfied were you with your search experience? From 1 (very unsatisfied) to 7 (very satisfied).
Frustration (Frus)	How frustrated were you with this task? From 1 (not frustrated) to 7 (very frustrated).
System Helpfulness (Help)	How well did the system help you in this task? From 1 (very badly) to 7 (very well).
Goal Success (Succ)	How well did you fulfill the goal of this task? From 1 (very badly) to 7 (very well).
Session Effort (S.Eff)	How much effort did this task take? From 1 (minimum) to 7 (a lot of).
Difficulty (Diff)	How diffcult was this task? From 1 (very easy) to 7 (very difficult).

the results in a better detail in the post-session judgments. The system also required participants to revisit each clicked result and spend at least 30 seconds to judge a result.

We collected users' in situ and post-session judgments of six different measures. Table 1 shows the detailed questions and options.

- **TREC relevance (TRel)** – the de facto standard of relevance judgments due to the popularity of TREC test collections. We collected TRel using the criteria of the latest TREC web track [10]. As Table 1 shows, the criteria focus on topical relevance. We excluded the relevance level *Nav* (the correct homepage of a navigational query) from the original TREC criteria because our search tasks do not include navigational search.
- **Usefulness (Usef)** – Following Belkin et al.'s model [5] and Mao et al.'s study [36], we collected users' perceptions regarding the usefulness of the clicked results.
- **Novelty (Nov)** – Novelty was often assessed algorithmically in previous studies based on sub-topic or "nugget" level relevance judgments [8, 40, 43, 55]. In contrast, we collect users' explicit novelty judgments.
- **Understandability (Under)** – the easiness of understanding the content of the result. Recent studies incorporated understandability into search result ranking [41] and evaluation [56].
- **Reliability (Relia)** – the reliability, credibility, and trustworthy of the information presented in the result [39, 46, 51] (here we do not distinguish the three constructs).
- **Effort** – Yilmaz et al. [54] and Verma et al. [50] examined effort as a dimension of evaluating search result.

The following table summarizes the measures collected in in situ and post-session judgments. We only collected TRel in post-session judgments because the TREC criteria do not consider context. We only collected Nov and Effort during in situ judgments because the participants of a pilot study reported confusions assessing the two measures twice. In the rest of this paper, we will use .i and .p suffixes to denote in situ and post-session judgments, respectively. For example, Usef.i denotes users' in situ usefulness judgments.

	In Situ (.i)	Post-session (.p)
TREC relevance (TRel)		✓
Usefulness (Usef)	✓	✓
Novelty (Nov)	✓	
Effort (Effort)	✓	
Understandability (Under)	✓	✓
Reliability (Relia)	✓	✓

Except for TRel, we collected judgments using a 7-point Likert scale, because a previous study [48] showed that assessors approximate the optimal level of confidence when using a 7-point scale for relevance judgments. TRel used a different scale so that it is consistent with the TREC web track (as a representative example of the state-of-the-art relevance judgment methods).

2.3 Search Experience Measures

In the judgment stage, participants rated their search experience in a session. We collected six representative user experience measures used in previous studies of information retrieval and recommender systems—satisfaction (Sat) [17, 21, 26, 35, 36, 45], goal success (Succ) [1, 18], frustration (Frus) [12, 13], task difficulty (Diff) [4, 15, 29, 31, 32], the helpfulness of the system (Help) [19] and the total effort spent (S.Eff) [27]. Table 1 includes the questions.

2.4 Rationale of Experiment Design

The way we balance different types of tasks is similar to previous studies [22, 24, 30, 33, 36]. However, we acknowledge that the selected tasks cannot cover all varieties. It is also worth noting that the TREC session track tasks [7] are more complex than regular web search requests such as navigational search.

Our study aims to collect both in situ judgments and user behaviors related to the clicked results. This poses challenges to the experiment design. On the one hand, we hope to collect accurate in situ judgments, which often requires multi-item measurements [27, 52]. On the other hand, interrupting participants for in situ judgments breaks the flow of search session and can affect their subsequent search behaviors. To balance between the two purposes,

Table 2: Spearman's correlation (ρ) matrix of different judgments for the 727 unique clicks.

		In Situ Judgments					Post-session Judgments		
		Usef.i	Novelty	Effort	Under.i	Relia.i	TRel	Usef.p	Under.p
In Situ	Novelty	0.67	-	-	-	-	-	-	-
	Effort	0.22	0.24	-	-	-	-	-	-
	Understandability	0.20	0.14	−0.45	-	-	-	-	-
	Reliability	0.42	0.37	0.05	0.26	-	-	-	-
Post-session	Topical Relevance	0.63	0.46	0.16	0.14	0.42	-	-	-
	Usefulness	0.72	0.52	0.16	0.18	0.43	0.83	-	-
	Understandability	0.20	0.18	−0.36	0.68	0.29	0.18	0.24	-
	Reliability	0.43	0.38	0.04	0.22	0.82	0.48	0.51	0.31

The reported values are estimated from 1000 bootstrap samples (we used stratified sampling to balance user and task dependency).

we made a few compromises in experiment design, e.g., we only collected six popular dimensions of judgments, and we simply used one question to measure each dimension.

While examining search behaviors, we excluded the time spent on answering in situ judgments from dwell time. On average the participants spent 57.1 seconds on a clicked result and 12.1 seconds to answer the five in situ judgment questions.

2.5 Collected Data

We recruited 28 participants (16 are female) through fliers posted on the campuses of two universities in the United States. We required participants to be English native speakers to exclude the influence of language fluency on relevance judgments [16]. All the participants were undergraduate or graduate students studying different fields. They were reimbursed $15 per hour.

We collected 112 sessions by 28 participants on 28 tasks. Each participant worked on four unique tasks and each task was performed by four unique users. In total, we collected 537 queries (4.8 per session) and 736 clicks (6.6 per session) on 727 unique session-URL pairs (9 cases of revisiting). We exclude the 9 cases of revisiting from the analysis (about 1% of the data) to simply the analysis.

3 IN SITU VS. POST-SESSION JUDGMENTS

3.1 Correlation of Different Judgments

Table 2 reports the correlation of different judgments, which are generally consistent with previous studies. For example, relevance and usefulness positively correlate with novelty and reliability [52], understandability negatively correlates with effort [50], etc. We examined the relationship of the judgments in another article [23].

Note that Mao et al. [36] reported a weak correlation (0.332) of *searchers*' post-session usefulness judgments and *external assessors*' relevance judgments. However, Table 2 shows that `TRel` and `Usef.p` are strongly correlated ($\rho = 0.83$) when both of them are assessed by searchers. This suggests that the low correlation reported by Mao et al. [36] may be mostly due to the disparity between searchers and external assessors, rather than the difference between using relevance or usefulness as the judgment criteria.

3.2 Correlating with User Experience

We evaluate different search result judgments by correlating with (regressing) users' search experience measures in a session. This is based on the assumption that the "quality" of the clicked results in a session can influence users' search experience in that session—thus, a reasonable search result judgment (assumed to indicate certain "quality"), or a reasonable set of judgments, should also correlate with users' search experience in a session.

3.2.1 Regression Analysis. We use multilevel regression analysis to examine the relationship between the judgments of the clicked results and users' search experience in a session. The dependent variables (DVs) are each of the six search experience measures. The independent variables (IVs) include the statistics of judgments regarding the clicked results in a session (such as the mean, maximum, and minimum ratings). For `TRel`, `Usef.i`, and `Usef.p`, we include the mean, maximum, and minimum ratings of the clicked results in a session as IVs in the regression analysis. For other search result judgments, we only include the maximum and minimum ratings of the clicked results as IVs. This is because the mean ratings of the other measures often highly correlate with those of `TRel` and `Usef`, causing multicollinearity issues for regression analysis.

For each user experience measure (the DV), we examine six different models that include different judgments as IVs.

- **Unidimensional** & **Context-independent** – Model ① and ② only include context-independent search result judgments from a single dimension—Model ① includes the statistics of `TRel` and Model ② includes those of `Usef.p`.
- **Unidimensional** & **In Situ** – Model ③ includes in situ judgments from a single dimension (the statistics of `Usef.i`) as IVs.
- **Multidimensional** & **Context-independent** – Model ④ and ⑤ extend Model ① and ② to include other dimensions of judgments (the statistics of `Under.p`, `Relia.p`, `Nov`, and `Effort`). Note that Model ④ and ⑤ include two in situ judgments (`Nov` and `Effort`) because we did not collect post-session judgments on these two dimensions (as discussed in § 2.2).
- **Multidimensional** & **In Situ** – Model ⑥ extends Model ③ to include other dimensions of judgments (the statistics of `Under.i`, `Relia.i`, `Nov`, and `Effort`).

	Context-independent	In Situ
Unidimensional	① `TRel only` ② `Usef.p only`	③ `Usef.i only`
Multidimensional	④ `TRel + others` ⑤ `Usef.p + others`	⑥ `Usef + others`

All six models also include the same set of control variables, including: gender (*Male* or *Female*), age (four levels; 0 for *18–24*, 1 for *25–30*, 2 for *31–40*, and 3 for *Over 40*), highest degree obtained or expected (*Undergraduate* or *Graduate*), the expertise of using web search engines (SE Expertise) rated using a Likert scale from 1

Table 3: The adjusted R^2 of different regression models.

Models	Sat	Frus	Succ	S.Eff	Help	Diff
Base (control only)	0.12	0.06	0.11	0.06	0.11	0.03
① TRel	0.23	0.09	0.18	0.10	0.16	0.09
② Usef.p	0.25	0.15	0.36	0.17	0.18	0.18
③ Usef.i	0.29	0.14	0.35	0.16	0.22	0.22
④ TRel + others	0.31	0.25	0.33	0.33	0.31	0.21
④ vs. ①	**	**	**	**	**	**
⑤ Usef.p + others	0.30	0.26	0.42	0.37	0.31	0.26
⑤ vs. ②	**	**	**	**	**	**
⑥ Usef.i + others	0.30	0.27	0.34	0.37	0.27	0.33
⑥ vs. ③		**		**	*	**

* and ** indicate $p < 0.05$ and $p < 0.01$ by F-test.

(*very badly*) to 5 (*very well*), task product and goal, user's familiarity with the topic of the task (Topic Familiarity) rated using a Likert scale from 1 (*very unfamiliar*) to 7 (*very familiar*), and the number of clicks (# clicks) and queries (# queries) in the session.

We examine multicollinearity between variables using variance inflation factor (VIF). The IVs of all models satisfy VIF < 4, the commonly suggested threshold (4–10) for concerns of multicollinearity issues [37]. Table 3 reports the adjusted R^2 of the six models for regressing the six dimensions of search experience.

3.2.2 TREC Relevance vs. Usefulness. We first compare TREC relevance criteria (TRel) and post-session usefulness judgments (Usef.p). This is a revisit of Mao et al.'s study [36], which compared searchers' usefulness judgments and external assessors' relevance judgments. Here we collected both judgments from real searchers, removing the influence caused by the difference between searchers and external annotators in relevance judgments. The regression analysis suggest that switching from TREC relevance to usefulness is fruitful, consistently enhancing the ability of the regression models to correlate with user experience (by adjusted R^2).

Models ① and ② include the mean, maximum, and minimum TRel or Usef.p ratings of the clicked results. Model ② consistently explains the six search experience measures better than Model ① (by adjusted R^2). We note that usefulness (Usef.p) seems to be particularly better than TREC relevance (TRel) in terms of correlating with goal success (Succ), with adjusted $R^2 = 0.36$ vs 0.18.

Models ④ and ⑤ further include other dimensions of judgments as IVs. This helps compare TRel and Usef.p judgments with other search result judgments as controls. Still, we consistently observe that Model ⑤ explains the six search experience measures better than or as well as model ④. These results verify that usefulness is indeed a better criteria of relevance judgments than TREC-style relevance (in terms of correlating with users' search experience).

3.2.3 In Situ vs. Context-independent (Post-session) Judgments. We further compare in situ and post-session judgments in both unidimensional and multidimensional settings. Results suggest in situ usefulness judgments have better correlations with a few (but not all) user experience measures than post-session usefulness judgments. However, after combining search result judgments from different dimensions, in situ judgments show limited advantages over post-session ones.

Models ③ and ② include the mean, maximum, and minimum Usef.i or Usef.p ratings of the clicked results as IVs. Results show Model ③ explains satisfaction (Sat), helpfulness (Help), and task difficulty (Diff) slightly better than Model ②, with about 0.04 difference in adjusted R^2.

We further compare in situ and post-session judgments in a multidimensional setting, using a combination of Usef.p/Usef.i and other four judgments as IVs (Models ⑤ and ⑥). Results show that the post-session multidimensional model (⑤) better correlates with search success (Succ) than the in situ one (adjusted R^2 0.42 vs 0.34), but the latter also better correlates with task difficulty (adjusted R^2 0.26 vs. 0.21). Overall, no evidence suggests either model is *consistently* better than another in terms of correlating with users' search experience measures.

Even though Model ③ (Usef.i only) performs slightly better than Model ② (Usef.p only), results suggest limited advantages of in situ judgments over post-session ones in terms of correlating with search experience measures. We suspect a possible reason is that a 10-minute session is not long enough to trigger sufficient differences between in situ and post-session judgments. Although we expect to observe a greater difference between in situ and post-session judgments in longer sessions, we believe a substantial proportion of web search sessions are no longer than 10 minutes, which may not benefit much from in situ judgments. In addition, it also requires a more complex experiment design to collect in situ judgments.

3.2.4 Unidimensional vs. Multidimensional Judgments. We further compare models using a combination of multiple aspects of judgments (Models ④, ⑤, and ⑥) with those using a single dimension (Models ①, ②, and ③). Results suggest that it is almost always helpful (enhancing the correlation with most of the six search experience measures significantly) to complement either relevance or usefulness with the alternative dimensions.

Models ④ and ⑤ explain all six dimensions of search experience measures significantly better than Models ① and ②, suggesting that multidimensional judgments are almost always helpful for TREC-style relevance judgments (TRel) and post-session usefulness judgments (Usef.p). We also note that in situ usefulness judgments (Usef.i) worked particularly well for correlating with users' satisfaction (Sat) and goal success (Succ), such that combining with more dimensions of judgments adds little to the model.

Results demonstrate that multidimensional search result judgments are helpful, complementing unidimensional judgments and yielding better correlation with search experience measures. This also suggests the advantages of multidimensional search result judgments over the in situ one—the former can consistently improve relevance/usefulness to better correlate with almost all user experience measures, while the latter shows limited advantages.

3.3 Which Dimensions To Judge?

A crucial issue of information retrieval is deciding which criteria to use to rank search results. We come to initial answers by looking into the standardized coefficients (β) of Model ⑤ (Table 4) as an example due to its superiority over other models. The standardized coefficient β stands for the magnitude of change in the DV (relative to its standard deviation) caused by one-unit change in the IV (relative to the IV's standard deviation) while other variables being equal. The coefficients of the model indicate how changes in the "quality" of the clicked results will (theoretically) affect users' search experience in a session. Table 4 suggests that:

Table 4: Multilevel regression: standardized coefficients (β) of independent variables for Model ⑤ – `Usef.p + others`.

Independent Variables	DV: session-level search experience					
	Sat	Frus	Succ	S.Eff	Help	Diff
Gender: *Male*	0.10	0.19	0.09	0.06	0.00	0.16
Age	−0.05	0.00	−0.03	−0.02	−0.11	0.00
Degree: *Graduate*	−0.03	0.05	0.01	0.13	−0.08	0.23
SE Expertise	0.12	0.04	0.08	0.01	0.12	−0.00
Product: *Factual*	0.02	−0.02	−0.06	−0.09	−0.00	0.02
Goal: *Specific*	0.02	−0.07	0.04	0.05	0.07	0.01
Topic Familiarity	0.10	−0.23	0.17	−0.20	0.19	−0.24
# clicks	0.20	−0.12	0.17	−0.07	0.18	−0.01
# queries	−0.36	0.17	−0.25	0.16	−0.35	−0.02
† Usef.p (mean)	0.23	−0.38	0.36	−0.36	0.08	−0.43
† Usef.p (max)	0.16	0.09	0.22	−0.07	0.11	−0.08
† Usef.p (min)	0.01	0.19	−0.04	0.19	0.01	0.18
† Nov (max)	0.24	−0.10	0.18	−0.09	0.25	−0.11
† Nov (min)	−0.01	−0.20	−0.08	−0.06	0.07	−0.00
† Under.p (max)	0.09	−0.27	0.30	−0.15	0.14	−0.22
† Under.p (min)	0.16	−0.08	0.14	−0.26	0.29	−0.27
† Relia.p (max)	−0.13	−0.08	0.01	0.05	−0.03	0.06
† Relia.p (min)	0.06	0.01	−0.05	0.08	−0.07	0.04
† Effort (max)	−0.12	0.16	0.08	0.28	−0.13	0.02
† Effort (min)	0.21	0.04	0.12	0.01	0.25	0.02
Adjusted R^2	0.30	0.26	0.42	0.37	0.31	0.26

Light and dark shadings indicate $p < 0.05$ and 0.01, respectively.

- To enhance user **satisfaction**, a search system should present useful and novel results—both Usef.p (mean) and Nov (max) show significant positive effects on Sat in Model ⑤.
- To reduce user **frustration**, a search system should offer results that are useful and easy-to-understand—both Usef.p (mean) and Under.p (max) show significant negative effects on Frus.
- To help users **success**fully reach the goal (Succ), a search system should retrieve useful, novel, and easy-to-understand results—Usef.p (mean), Nov (max), and Under.p (max) show significant positive effects on Succ.
- To reduce the **total effort** of a search session, the system should retrieve easy-to-understand results and avoid those requiring too much effort—Under.p (min) shows a significant negative effect on S.Eff and Effort (max) shows a positive one.
- To better help users in a session (enhance the **helpfulness** of the system), a system should retrieve novel and easy-to-understand results—both Nov (max) and Under.p (max) show significant positive effects on Help.
- To reduce the perceived **task difficulty**, we need to retrieve useful and easy-to-understand results—both Usef.p (mean) and Under.p (min) show significant negative effects on Diff.

The coefficients suggest that the mean usefulness of the clicked results is helpful for explaining all six search experience measures (has statistically significant coefficients). In addition, novelty, understandability, and effort also significantly relate to many different search experience measures, suggesting they are useful complements to usefulness in search result judgments. In contrast, reliability shows no significant effect on any of the six user experience measures in Model ⑤. However, we suspect this is because the top-ranked results returned by Google are mostly reliable ones, which makes reliability a less important judgment measure among the clicked results.

Table 5: Statistics of the absolute difference of two users' ratings on the same results ($|\Delta|$).

	$\|\Delta\|$ mean (SD)	$\|\Delta\| = 0$	$\|\Delta\| \leq 1$	$\|\Delta\| \leq 2$
Usef.i	1.55 (1.45)	25.9%	58.2%	79.1%
Effort	1.52 (1.25)	22.9%	57.7%	76.1%
Nov	1.60 (1.47)	26.4%	54.2%	78.1%
Relia.i	1.23 (1.21)	31.3%	67.2%	86.1%
Under.i	1.18 (1.23)	34.8%	68.2%	88.1%
TRel	0.63 (0.68)	48.3%	89.1%	100.0%
Usef.p	1.53 (1.54)	29.9%	60.7%	77.6%
Relia.p	1.38 (1.35)	30.8%	62.2%	81.1%
Under.p	1.08 (1.31)	38.8%	76.6%	90.5%

3.4 Variability of Judgments

We further examine the variability of judgments among different searchers, because in many practical scenarios we may have to train and evaluate retrieval systems based on relevance judgments made by external assessors. We suspect different users may have a greater degree of inconsistencies in their in situ judgments than their post-session ones (due to the contextual nature of the former). However, results do not support this conjecture well.

We examine the absolute difference of two users' ratings on the same result. Table 5 reports the mean absolute difference and the distribution. The mean absolute difference for in situ and post-session usefulness judgments (Usef.i and Usef.p) are very close (1.55 vs. 1.53). The mean absolute difference of post-session reliability judgments (Relia.p) is slightly higher than that for in situ ones (Relia.i) (1.38 vs. 1.23), but that for post-session understandability judgments (Under.p) is also slightly lower than the in situ ones (Under.i, 1.08 vs. 1.18). Overall, no evidence suggests that either in situ or post-session judgments is more or less consistent than the other across different users.

Further, we note that different users' reliability and understandability judgments seem more consistent than those for usefulness, effort, and novelty judgments, regardless of performed in an in situ setting or a post-session one. This suggests that usefulness, effort, and novelty judgments may suffer from inter-rate consistency by a greater extent, while inter-rate agreement is less likely a concern for understandability and reliability judgments. However, since users judged TRel by a different scale, it remains unclear how do the other five judgments compare with standard TREC relevance judgments in terms of inter-rate consistency.

3.5 Summary

To sum up, this section discloses both opportunities and challenges for future search result judgments.

- **Opportunity** – Since a combination of multidimensional judgments explains user experience measures better than using relevance or usefulness alone, we expect that an appropriate ranking of search results by multiple criteria may potentially yield better user experience as well. The results in Table 4 also help select ranking criteria according to a targeted user experience measure.
- **Challenge** – Extending current judgments from a single dimension to multiple aspects largely increases the cost of judgments. This is a crucial issue for the scalability of multidimensional judgments. The following sections address this concern by predicting multidimensional judgments using implicit feedback techniques.

Table 6: Implicit feedback features and their correlation with different search result quality measures.

			Pearson's r with search result judgments					
	Click Dwell Time Features	**Note**	**TRel**	**Usef.p**	**Nov**	**Effort**	**Under.p**	**Relia.p**
T1	Click dwell time (log).		0.38	0.43	0.41	0.36	0.12	0.34
T2	$(t-\mu)/\sigma$. t is the result's dwell time; μ is average click dwell time; σ is the standard deviation of click dwell time. T3-5 are based on personalized versions of μ and σ.	all clicks	0.31	0.34	0.30	0.32	0.06	0.24
T3		by user	0.31	0.36	0.38	0.32	0.09	0.24
T4		by task	0.31	0.35	0.30	0.32	0.06	0.24
T5		by length	0.29	0.33	0.29	0.31	0.06	0.24
	Follow-up Query Features		**TRel**	**Usef.p**	**Nov**	**Effort**	**Under.p**	**Relia.p**
Q1	The number of terms in the next query found in the URL/title/body of the result.	URL	−0.04	0.03	−0.01	−0.02	−0.02	0.03
Q2		title	−0.03	−0.00	−0.00	0.01	−0.00	−0.02
Q3		body	−0.03	−0.03	0.10	0.06	0.03	−0.02
Q4	The percentage of terms in the next query found in the URL/title/body of the result.	URL	0.02	0.09	0.02	−0.04	0.01	0.08
Q5		title	0.07	0.10	0.06	−0.00	0.05	0.07
Q6		body	0.17	0.18	0.21	0.04	0.13	0.18
Q7	The number of newly added query terms in the next query reformulation found in the URL/title/body of the result.	URL	0.03	−0.01	−0.03	−0.06	0.02	0.02
Q8		title	0.07	0.04	0.00	−0.07	0.01	−0.00
Q9		body	0.07	0.07	0.13	−0.04	0.04	−0.02
Q10	The number of removed query terms in the next query reformulation found in the URL/title/body of the result.	URL	0.01	0.01	−0.00	−0.07	−0.04	−0.12
Q11		title	0.06	0.09	0.06	−0.09	0.03	−0.06
Q12		body	0.08	0.07	0.06	0.01	−0.09	−0.04
Q13	The mean/max/min log likelihood scores between the full content of the result and follow-up queries.	mean	0.22	0.23	0.22	−0.03	0.19	0.23
Q14		max	0.22	0.23	0.21	−0.03	0.17	0.18
Q15		min	0.15	0.19	0.19	−0.01	0.17	0.19
	Follow-up Click Features		**TRel**	**Usef.p**	**Nov**	**Effort**	**Under.p**	**Relia.p**
C1	The mean/max/min similarity between the title of the result and the titles of clicked results in follow-up searches.	mean	0.04	0.05	0.12	0.02	0.04	0.01
C2		max	0.05	0.04	0.09	0.00	0.06	0.00
C3		min	0.06	0.08	0.10	0.01	−0.01	0.03
C4	The mean/max/min similarity between the snippet of the result and the snippets of clicked results in follow-up searches.	mean	−0.00	0.01	0.01	0.04	0.02	0.03
C5		max	−0.04	−0.06	−0.06	−0.05	0.01	−0.04
C6		min	0.08	0.11	0.10	0.07	0.06	0.09
C7	The mean/max/min similarity between the full content of the result and the full contents of SAT clicks (dwell time > 30s) in follow-up searches.	mean	0.19	0.23	0.09	0.01	−0.02	0.10
C8		max	0.20	0.20	0.05	−0.00	0.00	0.08
C9		min	0.12	0.17	0.07	−0.00	−0.01	0.07
C10	The mean/max/min similarity between the title of the result and the titles of skipped results in follow-up searches.	mean	0.09	0.11	0.11	0.02	0.05	0.05
C11		max	0.11	0.09	0.06	−0.01	0.05	0.02
C12		min	0.09	0.13	0.13	−0.05	0.00	0.05
C13	The mean/max/min similarity between the snippet of the result and the snippets of skipped results in follow-up searches.	mean	0.01	0.03	0.03	0.11	−0.05	0.03
C14		max	0.02	0.01	−0.01	−0.00	0.02	−0.02
C15		min	0.10	0.12	0.12	0.13	−0.05	0.11

Light and dark shadings indicate the correlation is significant at 0.05 and 0.01 levels, respectively.

4 PREDICTION

This section introduces our techniques for predicting multidimensional judgments of clicked results from search logs. We model the prediction task as a regression problem—the input is features related to a target click, the output is the predicted judgment score of the clicked result. We use gradient boosted regression trees (GBRT) for prediction. Table 6 lists the prediction features. Due to the limited space, we only report results for predicting TRel and the judgments included in Model ⑤—Usef.p, Nov, Effort, Under.p, and Relia.p. However, the described approach can also effectively predict other search result judgments as well.

4.1 Click Dwell Time Features

Click dwell time (T1) is one of the most widely used implicit feedback measure. As Table 6 shows, T1 does not correlate much with understandability, but it still has 0.3–0.4 correlations (significant at 0.01 level) with other measures.

T2–T5 measure the deviation of a click's dwell time from the mean dwell time (μ) of a group of clicks (normalized by the standard deviation σ). T2 computes μ and σ based on all clicks in the training sets. T3 is based on clicks by the same user. T4 is based on clicks in sessions with the same task type. T5 is based on clicks on documents with similar length (we divide the clicked results into ten bins by length and compute μ and σ of a click based on its bin).

4.2 Follow-up Query Features

Follow-up query features are based on the intuition that a clicked result may influence follow-up query reformulation in a session. Thus, we can infer the quality of a click from queries issued after the clicked result in the same session.

Q1–Q6 match the terms in the immediate follow-up query with the target click. Q7–Q12 match the newly added and removed terms in the immediate follow-up query reformulation with the target click. Q13–Q15 match the target click with all follow-up queries.

Many of the follow-up query features (such as Q6 and Q13–Q15) have significant correlations with the search result quality measures, confirming that the intuition is reasonable. We also note that Q6 and Q13–Q15 have stronger correlations with understandability than click dwell time features.

4.3 Follow-up Click Features

Similar to follow-up query features, we may also infer the quality of a target click based on follow-up clicks in a session.

C1–C6 measure the similarity between the target click and follow-up clicks. C7–C9 measure the similarity with follow-up satisfactory (SAT) clicks. C10–C15 measure the similarity with follow-up skipped results (unclicked results ranked higher than a clicked result). Some features have significant correlations with the search result quality measures, suggesting they may be useful predictors.

4.4 Prior-to-click Features (Baseline)

Prior-to-click features include the existing techniques that predict search result quality measures using information available before users clicking on the result. In this paper, they serve as the baseline for the implicit feedback features. We include a full list of prior-to-click features in an online appendix[1].

We incorporate different prior-to-click features for predicting different measures. The shared features for all six measures include the rank of the result by Google search, ad hoc search models (QL, BM25, DFR [3], and SDM [38]), and session search models [14, 47]. The unique features for predicting each measure are:

- TRel – a subset of LETOR features [34].
- Usef.p – a subset of LETOR features [34] and a subset of the usefulness features by Mao et al. [36] that do not rely on post-click information.
- Nov – the similarity of the click with previous clicks and higher ranked results in the same SERP (motivated by previous work on novelty-based search result diversification [6, 43, 44, 55]).
- Effort – Yilmaz et al. [54] and Verma et al. [50].
- Under.p – Palotti et al. [41, 42].
- Relia.p – Olteanu et al. [39] and Wawer et al. [51].

Our prior-to-click features are representatives of the state-of-the-art techniques for predicting each dimension of judgments without using implicit feedback. However, we did not include features that we do not have the resource to calculate, which include link structure based features and social media popularity features such as Twitter mention. Note this may reduce the effectiveness of predicting reliability since the excluded features take about 1/3 of the features by Olteanu et al. [39] and Wawer et al. [51].

5 EVALUATION

5.1 Experiment Settings

We evaluate prediction (regression) by the Pearson's correlation between the predicted values and actual judgments (prediction correlation) and the root mean square error (RMSE) of the predicted values. Note that the RMSE for predicting different measures is not comparable—first, TREC relevance ranges from 0–3 while others from 1–7; second, their distributions vary a lot. Here we only report prediction correlation for its easy interpretability. The results of RMSE is highly consistent with that using prediction correlation.

The dataset for evaluation includes multidimensional judgments on the 727 unique clicked results. We use 10-fold cross validation for evaluation (using eight folds for training, one for validation, and one for testing). We randomly shuffle the dataset 10 times and apply 10-fold cross-validation for each random shuffling of the whole dataset—this generates prediction results on $10 \times 10 = 100$ test folds in total (note that we are not using a 100-fold cross validation). We report the mean and standard deviation (SD) of prediction correlation on the test folds. We note that the prediction correlation reported in this section is different from and cannot be compared with the correlation in Table 6, which are computed for the whole dataset without cross validation.

5.2 Click Dwell Time Features

Current techniques for inferring search result quality from logs rely on click dwell time. Results (① in Table 7) suggest the click dwell time features work reasonably well for predicting usefulness, novelty, and effort, but they have difficulties inferring the understandability and reliability of results.

The click dwell time features (①) are effective predictors for usefulness, novelty, and effort. For these three measures, the predicted values have about 0.3–0.4 mean Pearson's correlation with the actual judgments, which is comparable to that for predicting TREC relevance (mean $r = 0.35$). However, the click dwell time features perform much worse for predicting understandability and reliability. On average the predicted and actual judgments have only 0.10 and 0.22 correlation, suggesting it is necessary to incorporate new implicit feedback signals.

5.3 Follow-up Query and Click Features

We extend click dwell time to include signals from follow-up search activities. Results suggest the new features are helpful.

The follow-up query (②) and click features (③) alone have limited prediction capability. However, combining them with the click dwell time features (④) consistently produces better prediction than using click dwell time features alone (①): except for effort, the prediction correlation for the other five measures using feature set ④ is significantly better than that for click dwell time features (①). This indicates that follow-up queries and clicks indeed provide useful implicit feedback that are complementary to click dwell time.

The follow-up query and click features are particularly helpful for predicting reliability. Combining them with the click dwell time features improves the mean correlation of prediction from 0.22 to 0.36. The new features are also helpful for predicting TREC relevance and usefulness as well. This partly confirms our intuition—the quality of a clicked result may influence follow-up search activities, making it possible to infer the quality of a clicked result based on what happened afterward in the session.

The new features also improved the mean prediction correlation for understandability from 0.10 to 0.20. However, we note the combination of all implicit feedback features still does not work well for predicting understandability (mean $r = 0.20$). This suggests that, compared with other judgments, it is more challenging to predict understandability based on the implicit feedback information.

[1] http://ciir.cs.umass.edu/downloads/mdrel/

Table 7: The effectiveness of different features for predicting multidimensional search result judgments.

		Mean (SD) Pearson's r between true and predicted judgments over the test folds					
	Features	**TRel**	**Usef.p**	**Nov**	**Effort**	**Under.p**	**Relia.p**
①	Click Dwell Time	0.35 (0.11)	0.40 (0.11)	0.42 (0.11)	0.31 (0.10)	0.10 (0.14)	0.22 (0.13)
②	Follow-up Query	0.19 (0.11)	0.17 (0.14)	0.13 (0.13)	0.12 (0.11)	0.14 (0.12)	0.19 (0.12)
③	Follow-up Click	0.15 (0.12)	0.20 (0.11)	0.11 (0.12)	0.14 (0.11)	0.11 (0.12)	0.17 (0.12)
④	All (①+②+③)	**0.39** (0.09)	**0.46** (0.08)	**0.45** (0.09)	**0.33** (0.11)	**0.20** (0.13)	**0.36** (0.12)
	① vs. ④	**	**	*		**	**
⑤	Prior-to-click	0.36 (0.10)	0.29 (0.10)	0.28 (0.11)	0.13 (0.12)	0.20 (0.14)	0.18 (0.13)
	④ vs. ⑤	**	**	**	**		**
⑥	All+Prior-to-click	**0.45** (0.08)	**0.49** (0.09)	**0.47** (0.09)	**0.39** (0.09)	**0.26** (0.12)	**0.40** (0.11)
	⑤ vs. ⑥	**	**	**	**	**	**

* and ** indicate the difference is statistically significant at 0.05 and 0.01 levels by two-tail paired t -test.

5.4 Comparing to Prior-to-click Features

An important application of implicit feedback techniques is to infer relevance labels from search logs. Aggregating inferred relevance labels or implicit feedback signals from past search logs may help rank search results in the future [2]. We examine whether or not implicit feedback techniques can serve a similar purpose for multidimensional judgments.

The combination of the implicit feedback features and the prior-to-click features (⑥) generated significantly better prediction results on all the six judgments than using the prior-to-click features alone (⑤). This suggests that the implicit feedback features are indeed helpful and complementary to the prior-to-click features for predicting these judgments. We also note that the improvements in mean prediction correlation can be as large as over 0.2 (such as for predicting reliability and effort). However, even combining the two sets of features still cannot adequately predict understandability (mean $r = 0.26$).

6 DISCUSSION AND CONCLUSION

A crucial issue of information retrieval is deciding which criteria to use to rank search results. We compared two seemingly reasonable directions for improving current TREC-style relevance judgments. One direction is to collect in situ search result judgments. The other one is to complement a single dimension of judgments (such as relevance or usefulness) by combining with other aspects. We found that the latter direction seems more effective and versatile—using a combination of different dimensions of judgments, we can almost always improve correlation with user experience measures.

We envision future search engines should rank results by multiple aspects. We also offered initial suggestions on which criteria to adopt and when to adopt them. We further examined and improved implicit feedback techniques for predicting multiple judgments, addressing the scalability concern of applying multidimensional judgments in real web search applications.

Our study makes the following contributions:

- We evaluated and compared in situ usefulness judgments with regular relevance/usefulness judgments by searchers. We show that using usefulness as the judgment criteria is fruitful, but in situ judgments do not show clear benefits over regular ones.
- We evaluate multidimensional search result judgments considering four alternative aspects other than relevance/usefulness. We show that multidimensional judgments better correlate with user experience measures than using relevance/usefulness judgments alone. We also note that multidimensional judgments is a better direction for improving TREC-style relevance judgments.
- Our study also discloses the connections between different user experience measures and various dimensions of search result judgments. This offers practical suggestions for system design, such as the appropriate dimensions to judge search results for the purpose of improving a particular user experience measure.
- We successfully generalize implicit feedback signals to include follow-up searches and clicks in a search session to help click dwell time better predict multidimensional judgments. To the best of our knowledge, we are also the first to examine the effectiveness of implicit feedback approaches for predicting novelty, understandability, reliability, and effort.

Our work also sheds lights on a few critical areas for exploration in the future:

An important line of future work is to provide more accurate criteria for search result ranking and evaluation. Based on a regression analysis, we have already offered initial suggestions on what criteria to use and when to use them, as discussed in Section 3.3. We note that, with a sufficiently large dataset, one can possibly learn a prediction model for search experience measures by taking multidimensional judgments of results as input. Such a model can further address issues such as what are the proper weights to put on different aspects when ranking search results. It may also solve the discrepancy between offline evaluation measures and user experience measures, and ultimately serve as a better objective function for training ranking models.

Another important application is to perform multidimensional ranking of search results based on implicit feedback signals and other information. We have already demonstrated that implicit feedback approaches can infer judgments of usefulness, novelty, effort, and reliability with reasonable accuracy comparing to those for relevance labels. Aggregating such inferred judgments from past search logs may serve as useful features for performing multidimensional search result ranking in the future. However, we also note that our current technique needs to be improved to better infer understandability of results from search logs.

We do admit certain limitations in our current study. First, our analysis and experiments are solely based on data collected from one laboratory user study, which is limited in both scale and representativeness. We suggest that further studies employ larger

datasets to verify our findings. Second, it is worth noting that our way of collecting in situ judgments influenced users' natural search behaviors. We observed in our log that users spent on average 12.1 seconds to finish the in situ judgments. Thus some particular user behavior patterns may vary when applied to another scenario (without interrupting users for in situ judgments). Third, we also note that we only collected search result judgments for the clicked results, while it remains unclear to which extent the findings can be generalized to the unclicked ones. Last but not least, the collected post-session judgments are more or less influenced by the search session and the in situ judgments (although we meant to collect context-independent judgments such as to compare with contextual ones). It is also worth noting that our post-session judgments are not fully representative of the existing TREC-style approach.

7 ACKNOWLEDGMENT

This work was supported in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

REFERENCES

[1] M. Ageev, Q. Guo, D. Lagun, and E. Agichtein. Find it if you can: A game for modeling different types of web search success using interaction data. In *SIGIR '11*, pages 345–354, 2011.
[2] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *SIGIR '06*, pages 19–26, 2006.
[3] G. Amati and C. J. Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, 2002.
[4] J. Arguello. Predicting search task difficulty. In *ECIR '14*, pages 88–99, 2014.
[5] N. J. Belkin, M. J. Cole, and J. Liu. A model for evaluation of interactive information retrieval. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, 2009.
[6] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98*, pages 335–336, 1998.
[7] B. Carterette, P. Clough, M. Hall, E. Kanoulas, and M. Sanderson. Evaluating retrieval over sessions: The TREC session track 2011-2014. In *SIGIR '16*, pages 685–688, 2016.
[8] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR '08*, pages 659–666, 2008.
[9] C. W. Cleverdon. The evaluation of systems used in information retrieval. In *Proceedings of the International Conference on Scientific Information*, pages 687–698, 1959.
[10] K. Collins-Thompson, C. Macdonald, P. Bennett, F. Diaz, and E. Voorhees. TREC 2014 web track overview. In *TREC 2014*, 2014.
[11] N. Dai, M. Shokouhi, and B. D. Davison. Learning to rank for freshness and relevance. In *SIGIR '11*, pages 95–104, 2011.
[12] H. A. Feild and J. Allan. Modeling searcher frustration. In *HCIR '09*, pages 5–8, 2009.
[13] H. A. Feild, J. Allan, and R. Jones. Predicting searcher frustration. In *SIGIR '10*, pages 34–41, 2010.
[14] D. Guan, S. Zhang, and H. Yang. Utilizing query change for session search. In *SIGIR '13*, pages 453–462, 2013.
[15] J. Gwizdka. Revisiting search task difficulty: Behavioral and individual difference measures. In *ASIS&T '08*, 2008.
[16] P. Hansen and J. Karlgren. Effects of foreign language and task scenario on relevance assessment. *J. Doc.*, 61(5):623–639, 2005.
[17] A. Hassan. A semi-supervised approach to modeling web search satisfaction. In *SIGIR '12*, pages 275–284, 2012.
[18] A. Hassan, R. Jones, and K. L. Klinkner. Beyond DCG: User behavior as a predictor of a successful search. In *WSDM '10*, pages 221–230, 2010.
[19] R. Hu and P. Pu. A study on user perception of personality-based recommender systems. In *UMAP '10*, pages 291–302, 2010.
[20] J. Jiang and J. Allan. Adaptive effort for search evaluation metrics. In *ECIR '16*, pages 187–199, 2016.
[21] J. Jiang, A. Hassan Awadallah, X. Shi, and R. W. White. Understanding and predicting graded search satisfaction. In *WSDM '15*, pages 57–66, 2015.
[22] J. Jiang, D. He, and J. Allan. Searching, browsing, and clicking in a search session: Changes in user behavior by task and over time. In *SIGIR '14*, pages 607–616, 2014.
[23] J. Jiang, D. He, D. Kelly, and J. Allan. Understanding ephemeral state of relevance. In *CHIIR '17*, pages 137–146, 2017.
[24] D. Kelly, J. Arguello, A. Edwards, and W.-c. Wu. Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework. In *ICTIR '15*, pages 101–110, 2015.
[25] J. Y. Kim, J. Teevan, and N. Craswell. Explicit in situ user feedback for web search results. In *SIGIR '16*, pages 829–832, 2016.
[26] J. Kiseleva, E. Crestan, R. Brigo, and R. Dittel. Modelling and detecting changes in user satisfaction. In *CIKM '14*, pages 1449–1458, 2014.
[27] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):441–504, 2012.
[28] Y. Li and N. J. Belkin. A faceted approach to conceptualizing tasks in information seeking. *Inf. Process. Manage.*, 44(6):1822–1837, 2008.
[29] C. Liu, J. Liu, and N. J. Belkin. Predicting search task difficulty at different search stages. In *CIKM '14*, pages 569–578, 2014.
[30] J. Liu, J. Gwizdka, C. Liu, and N. J. Belkin. Predicting task difficulty for different task types. In *ASIS&T '10*, 2010.
[31] J. Liu, C. Liu, M. Cole, N. J. Belkin, and X. Zhang. Exploring and predicting search task difficulty. In *CIKM '12*, pages 1313–1322, 2012.
[32] J. Liu, C. Liu, J. Gwizdka, and N. J. Belkin. Can search systems detect users' task difficulty?: Some behavioral signals. In *SIGIR '10*, pages 845–846, 2010.
[33] J. Liu, C. Liu, X. Yuan, and N. J. Belkin. Understanding searchers' perception of task difficulty: Relationships with task type. In *ASIS&T '11*, 2011.
[34] T.-Y. Liu, J. Xu, T. Qin, W. Xiong, and H. Li. LETOR: Benchmark dataset for research on learning to rank for information retrieval. In *SIGIR 2007 workshop on learning to rank for information retrieval*, pages 3–10, 2007.
[35] Y. Liu, Y. Chen, J. Tang, J. Sun, M. Zhang, S. Ma, and X. Zhu. Different users, different opinions: Predicting search satisfaction with mouse movement information. In *SIGIR '15*, pages 493–502, 2015.
[36] J. Mao, Y. Liu, K. Zhou, J.-Y. Nie, J. Song, M. Zhang, S. Ma, J. Sun, and H. Luo. When does relevance mean usefulness and user satisfaction in web search? In *SIGIR '16*, pages 463–472, 2016.
[37] S. Menard. *Applied Logistic Regression Analysis*. Sage, 1997.
[38] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *SIGIR '05*, pages 472–479, 2005.
[39] A. Olteanu, S. Peshterliev, X. Liu, and K. Aberer. Web credibility: Features exploration and credibility prediction. In *ECIR '13*, pages 557–568, 2013.
[40] P. Over. The TREC interactive track: An annotated bibliography. *Inf. Process. Manage.*, 37(3):369–381, 2001.
[41] J. Palotti, L. Goeuriot, G. Zuccon, and A. Hanbury. Ranking health web pages with relevance and understandability. In *SIGIR '16*, pages 965–968, 2016.
[42] J. Palotti, G. Zuccon, and A. Hanbury. The influence of pre-processing on the estimation of readability of web documents. In *CIKM '15*, pages 1763–1766, 2015.
[43] D. Rafiei, K. Bharat, and A. Shukla. Diversifying web search results. In *WWW '10*, pages 781–790, 2010.
[44] R. L. Santos, C. Macdonald, and I. Ounis. On the role of novelty for search result diversification. *Inf. Retr.*, 15(5):478–502, 2012.
[45] A. Schuth, K. Hofmann, and F. Radlinski. Predicting search satisfaction metrics with interleaved comparisons. In *SIGIR '15*, pages 463–472, 2015.
[46] J. Schwarz and M. Morris. Augmenting web pages and search results to support credibility assessment. In *CHI '11*, pages 1245–1254, 2011.
[47] X. Shen, B. Tan, and C. Zhai. Context-sensitive information retrieval using implicit feedback. In *SIGIR '05*, pages 43–50, 2005.
[48] R. Tang, W. M. Shaw, Jr., and J. L. Vevea. Towards the identification of the optimal number of relevance categories. *J. Am. Soc. Inf. Sci.*, 50(3):254–264, 1999.
[49] J. van Doorn, D. Odijk, D. M. Roijers, and M. de Rijke. Balancing relevance criteria through multi-objective optimization. In *SIGIR '16*, pages 769–772, 2016.
[50] M. Verma, E. Yilmaz, and N. Craswell. On obtaining effort based judgements for information retrieval. In *WSDM '16*, pages 277–286, 2016.
[51] A. Wawer, R. Nielek, and A. Wierzbicki. Predicting webpage credibility using linguistic features. In *WWW '14 Companion*, pages 1135–1140, 2014.
[52] Y. Xu and Z. Chen. Relevance judgment: What do information users consider beyond topicality? *J. Am. Soc. Inf. Sci. Technol.*, 57(7):961–973, 2006.
[53] Y. Yamamoto and K. Tanaka. Enhancing credibility judgment of web search results. In *CHI '11*, pages 1235–1244, 2011.
[54] E. Yilmaz, M. Verma, N. Craswell, F. Radlinski, and P. Bailey. Relevance and effort: An analysis of document utility. In *CIKM '14*, pages 91–100, 2014.
[55] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *SIGIR '03*, pages 10–17, 2003.
[56] G. Zuccon. Understandability biased evaluation for information retrieval. In *ECIR '16*, pages 280–292, 2016.