When does Relevance Mean Usefulness and User Satisfaction in Web Search?

Jiaxin Mao†, Yiqun Liu†, Ke Zhou⋆, Jian-Yun Nie#, Jingtao Song†, Min Zhang†, Shaoping Ma†, Jiashen Sun‡, Hengliang Luo‡
†Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science & Technology, Tsinghua University, Beijing, China
⋆Yahoo! Research, London, U.K.
#Université de Montréal
‡Samsung R&D Institute China - Beijing
yiqunliu@tsinghua.edu.cn

ABSTRACT

Relevance is a fundamental concept in information retrieval (IR) studies. It is however often observed that relevance as annotated by secondary assessors may not necessarily mean usefulness and satisfaction perceived by users. In this study, we confirm the difference by a laboratory study in which we collect relevance annotations by external assessors, usefulness and user satisfaction information by users, for a set of search tasks. We also find that a measure based on usefulness rather than relevance annotated has a better correlation with user satisfaction. However, we show that external assessors are capable of annotating usefulness when provided with more search context information. In addition, we also show that it is possible to generate automatically usefulness labels when some training data is available. Our findings explain why traditional system-centric evaluation metrics are not well aligned with user satisfaction and suggest that a usefulness-based evaluation method can be defined to better reflect the quality of search systems perceived by the users.

Keywords

Relevance; Usefulness; User satisfaction; Evaluation

1. INTRODUCTION

Relevance, which "expresses a criterion for assessing effectiveness in retrieval of information, or of objects potentially conveying information" [37], is a central concept in IR and plays an important role in search engine evaluation. However, this notion involves multiple aspects. In the traditional system evaluation paradigm [10, 43], in order to compare the performances of different search systems, we typically rely on a test collection that consists of a document corpus, a set of predefined statements of information needs, and a set of *relevance judgements*. Based on the relevance judgements of query-document pairs, evaluation metrics, such as MAP, NDCG [21], and ERR [7], are computed for the ranked lists returned by different systems. Each of these measures is defined according to a different user model, which describes how the user interacts with the ranked list [33], and links the document-level relevance judgments with an estimation of the query-level user satisfaction [1, 28].

SIGIR '16, July 17-21, 2016, Pisa, Italy
© 2016 ACM. ISBN 978-1-4503-4069-4/16/07…$15.00
DOI: http://dx.doi.org/10.1145/2911451.2911507

Conceptually, the relevance judgements are expected to represent users' opinions about whether the retrieved documents are relevant and meet users' information needs [43, 44] and should be made by the users themselves. However in practice, it is usually hard to collect relevance feedbacks directly from actual search users, especially in Web search. We therefore ask external (secondary) to make the relevance judgements instead. In this case, there is a high risk that the collected relevance judgments may not necessarily reflect the user-perceived *usefulness* of retrieved documents. This is due to several reasons. On the one hand, in general, the assessors do not originate the information needs themselves and thus may not fully understand what the user actually wants. It has been indeed questioned whether the search intent can always be captured by the assessors [42]. On the other hand, conventionally the relevance judgments are made in a much simplified environment in which the assessor is asked to judge the relevance relation between each query-document pair independently. The assessor does not have access to much contextual information that may affect relevance judgment such as the queries the user issued previously in the session, the documents examined or clicked by the user, and so on. In addition, the assessor is only provided with a single short query, which may hardly describe accurately the user's information need. In reality, the Web search engine users often issue multiple queries in a search session [38], especially for exploratory and struggling search tasks [19]. It has been well documented that there are dependency and redundancy among the result documents [5, 9] . When all these contextual factors are ignored, it is very difficult for the assessor to put herself in the shoe of the user to make correct relevance judgment.

The lack of contextual information and accurate description of the information need often leads the assessor to limit herself to judge the *topical* aspect of relevance, and therefore, different from the highly *situational*, potentially *subjective*, user-perceive usefulness. The difference can be easily observed when the relevance judgment of the assessor is compared to that of the user. Table 1 shows a search session collected in our experimental study in which we collect relevance judgments of the user and the assessor (see Section 3 for more details). Given a search task, a user issued two queries and viewed several results. For the first query, `baggage restrictions`, the user clicked on two results in order. The contents of these two documents are very similar and both are topically related to the query. The assessor judged both document to be "highly relevant". However, the user judged the first document to be more useful than the second. This may be due to the fact that the second result does not contain much novel information after reading the first one. As for the second query, the user clicked on the result titled `The Best Way to Pack a Suitcase`. From the assessor's point of view, this document is not so relevant to the query `carry-on baggage liquids`, but the

Table 1: An example session showing the difference between user's feedbacks on usefulness and assessor's relevance labels.

Search Task:	
You are going to US by air, so you want to know what restrictions there are for both checked and carry-on baggage during air travel.	
Query Logs:	
Query #1	`baggage restrictions`
Click #1	Checked baggage policy - American Airlines Relevance: 4(Highly) Usefulness: 3(Fairly)
Click #2	Air Canada - BaggageInformation Relevance: 4(Highly) Usefulness: 2(Somewhat)
Query #2	`carry-on baggage liquids`
... **Click #3**	The Best Way to Pack a Suitcase Relevance: 2(Somewhat) Usefulness: 4(Very)

user finds it very useful when he or she is preparing for an air trip (this is part of the task specification missing in the short query). In these examples, we clearly spot that the usefulness of a document is dependent on previously read contents and on the accurate specification of the search task, and therefore, is different from its topical relevance. In this paper, the difference on relevance judgments between the user and the assessors will be further analyzed.

A number of existing studies have noticed the differences between users' and external assessors' relevance labels. Vakkari and Sormunen found that the relevance criterion of some users is more liberal than that used by TREC assessors [41]. Al-Maskari et al. [2] also compared the differences in relevance labeling process between users and TREC assessors, and observed that various factors, such as the number of retrieved relevant documents and the ranking of relevant documents (i.e. context of the current document), contribute to the differences. Although these previous studies show that users' judgements are different from the assessors', they do not attempt to proposed a new way to evaluate systems to better correspond to user's perception. Yilmaz et al. [47] compared document usefulness for users (called utility in their paper) with relevance annotation by assessors, which is in line with our work. They come to an interesting conclusion that some of the differences between user's usefulness and assessor's relevance are caused by the amount of effort required to find the relevant information in a document. However, they used dwell time as a sign of usefulness; while in our work, users' explicit feedback information is collected, which is expected to be more reliable than implicit behavior signals. We also investigate more thoroughly the reasons besides user effort that lead to users and assessors' differences. The idea of replacing relevance-based measurements with usefulness-based ones is also proposed by Belkin et al. [3] and Cole et al. [13]. The possibilities of adopting usefulness in evaluation of interactive information retrieval systems are also discussed in their work. However, the idea has not been implemented and no experimental study has been carried out so far on realistic data.

In this paper, we examine the relationship between relevance, usefulness and user satisfaction in a realistic Web search setting. In particular, we will design a protocol to collect data [1] containing both (1) user's explicit feedbacks on document usefulness and user satisfaction, which will be considered as ground truth; and (2) external assessor's relevance judgments. Based on the collected data, we examine the following research questions:

RQ1 What is the difference between user's perceived usefulness and the external assessor's relevance annotation?

RQ2 How do document's usefulness and relevance correlate with user's satisfaction?

We examine these two questions to study whether it is possible to evaluate systems in terms of usefulness rather than topical relevance. However, in a practical Web search setting, it is impossible to ask users to provide explicit feedback on usefulness. We have to resort to an alternative approach. This motivates us to examine the following two additional research questions:

RQ3 Can we rely on external assessors to make reliable and valid assessments for the document-level usefulness?

RQ4 Can we automatically generate usefulness labels based on user behavior and search context features?

Regarding **RQ3** and **RQ4**, we propose two approaches that can collect usefulness labels in practical Web search settings. The first approach relies on manual labeling by external assessors. We study if showing search task and search context information to assessors enables them to estimate user-perceived usefulness. The results show that this is fairly possible. The second approach goes a step further by utilizing machine learning techniques based on user behavior data to automatically generate usefulness labels. We show that this approach is feasible when a small amount of training data is available. Such an automatic usefulness labeling approach can help save the tedious work of manual labeling.

By answering these research questions, we aim to propose a new evaluation framework in which usefulness, instead of the current simplified relevance, is used. With manually labeled or automatically generated usefulness labels, evaluation metrics in this new framework are expected to better correlate with users' feelings of satisfaction in search tasks, which will be confirmed in our experiments. We do not hope to fully replace the current practice of relevance judgment with usefulness assessment in all situations. Instead, we hope to show that in certain circumstances where usefulness information can be collected or deduced, evaluation based on usefulness assessment can better reflect users' opinions.

The rest of this paper is organized as follows: Related studies are discussed in Section 2. In Section 3, we describe the experiment design and the data collecting procedure. In Section 4, we compare user's usefulness feedback with assessor's relevance annotation to answer **RQ1**. In Section 5, we characterize the relationship between document-level measures and user satisfaction, and answer **RQ2**. To answer **RQ3** and **RQ4**, we propose and test two approaches for acquiring usefulness labels, manually or automatically, in Section 6. Finally we draw conclusions and provide future work directions in Section 7.

2. RELATED WORK

Our work is related to a broad range of IR evaluation studies, as relevance sits at the core of the system-centric evaluation paradigm, and usefulness and satisfaction are key concepts in the user-centric evaluation of Web search engines.

In the traditional system-centric Cranfield-style evaluation [10, 43], most evaluation metrics are based on an implicit user model describing how the user interacts with a SERP [33]. They assess and summarize the effectiveness at *query* level. Recent studies extend the Cranfield-style evaluation paradigms to (1) cope with the redundancy and diversity of documents [9, 35]; (2) assess the overall effectiveness in a search session [6, 22, 25].

On the other hand, the user-centric evaluation draws more and more attention along with the emergence of Web search engines. It has been argued for a long time that instead of relevance, usefulness (or utility) should be used as a measure of retrieval effectiveness [13, 14]. Using the user behavior information that can be implicitly collected at a large scale, the utility or usefulness of a document (sometimes referred to as click satisfaction or intrinsic relevance) are estimated [4, 8, 24, 46]. These studies are based on the assumption that there are correlations between user behaviors and usefulness of documents. We will further investigate these correlations in Section 6.2. Our work is complementary to the existing work in the following ways: (1) instead of relying on natural log data from a search engine, we collected explicit usefulness feedbacks as well as comprehensive user behavior and search context information in a laboratory user study, which are expected to be more reliable and make it possible

[1] This dataset and the detailed instructions used to construct the dataset are publicly available on https://github.com/THUIR/UsefulnessUserStudyData

Table 2: Descriptions of major measures used in this work.

Concepts	Measures	Descriptions
Relevance	Relevance annotation (R)	4-level graded relevance annotations made by external assessors in Stage II.1 (see Figure 1 and Section 3.2).
Usefulness	Usefulness feedbacks (U_u)	4-level graded usefulness feedbacks collected in Stage I.4 (see Figure 1 and Section 3.2). We use them as the ground truth labels for usefulness.
	Usefulness annotation (U_a)	4-level graded usefulness annotations made by external assessors reviewing augmented search logs in Stage II.2 (see Figure 1 and Section 3.2).
	Usefulness prediction or predicted usefulness (U_{Q}, $U_{\mathrm{Q+S}}$, etc.)	Predicted usefulness labels. We utilize a machine learning method and different combinations of features to predict usefulness (See Section 6.2).
Query-level Satisfaction	Query-level satisfaction feedbacks ($QSAT_u$)	5-level graded satisfaction feedbacks for each issued query collected in Stage I.4 (see Figure 1 and Section 3.2).
	Query-level satisfaction annotation ($QSAT_a$)	5-level graded satisfaction annotations made by external assessors in Stage II.2 (see Figure 1 and Section 3.2).
Task-level Satisfaction	Task-level satisfaction feedbacks ($TSAT_u$)	5-level graded satisfaction feedbacks for each search task collected in Stage I.4 (see Figure 1 and Section 3.2). .
	Task-level satisfaction annotation ($TSAT_a$)	5-level graded satisfaction annotations made by external assessors in Stage II.2 (see Figure 1 and Section 3.2).

to investigate the correlations between different notions. (2) Different from the studies that focus on predicting user satisfaction and search success at query- or task-level in Web search [18, 23, 32, 34], our primary goal in this study is to exploit the possible correlations between *document*-level measures and *query*- and *task*-level user satisfaction so that (1) we can understand the relationship between document utility and user satisfaction; and (2) we can derive appropriate usefulness measures.

Some existing studies investigated the relation between the system-centric and user-centric evaluation by comparing the system-centric evaluation metrics, usually based on relevance, with user performance, satisfaction, or preference [1, 20, 28, 36, 40]. In a recent work, Yilmaz et al. [47] compared document utility with document relevance. They showed that the required effort plays an important role in the degree of document utility perceived by a real search user. Our work also extensively compares document usefulness with document relevance, and extends their work as follows: (1) instead of relying on the dwell time as a surrogate for utility, we collect users' explicit feedbacks of usefulness; (2) in addition to the required effort and dwell time, we also consider other factors, such as the the current search task and the redundancy with previous documents read by the user, as they have been shown to influence document usefulness perceived by the user.

Usefulness and satisfaction are both subjective. Therefore, our work is also indirectly related to the field of personalized search [15, 39]. As personalized search aims to take into account the diverse and volatile information needs from different users, to make our study more reliable, we control this variability by designing predefined and clearly stated search tasks for the participants, which differentiate our work from personalized search studies.

3. DATA COLLECTION

As shown in Figure 1, the data collection procedure consists of two parts: I. User Study and II. Data Annotation. The first part is collected in a laboratory environment. We collected users' behavior logs and their explicit feedbacks for both usefulness and satisfaction. In the second step, we hired external assessors to generate corresponding relevance annotations. To investigate **RQ3** and **RQ4**, we also asked the assessors to provide their usefulness and satisfaction annotations. We use these feedbacks and annotations as *measures* for relevance, usefulness or satisfaction. Table 2 provides a summary of these measures.

3.1 User Study Design

Table 3: Examples of search tasks. The TREC topic indexes are given in parentheses ().

Init. Query	Description
baggage restrictions	You are going to US by air, so you want to know what restrictions there are for both checked and carry-on baggage during air travel. (2010-7)
long-term care insurance	You just learned about the existence of long-term care insurance and want to know about it: costs / premiums, companies that offer it, types of policies, people's opinion about long term care insurance; what are the differences between long term care and health insurance? (2013-8)
quit smoking	Your friend would like to quit smoking. You would like to provide him with relevant information about: the different ways to quit smoking, benefits of quitting smoking, second effects of quitting smoking. (2013-12)

We conducted a laboratory user study to collect search logs and user feedbacks. Each participant was asked to complete 12 search tasks using an experiment search engine system. Compared with collecting data from real search logs [23, 34], or by browser plugins [16, 45], the laboratory user study had a smaller scale, but enabled us to fully control the variabilities in search tasks and information needs as well as to collect explicitly the information needed.

To simulate a real Web search environment, we built an experimental search engine that can access the open Web. As shown in Figure 1(I.3), this experimental search engine has an interface similar to common Web search engines, and supports query reformulation and pagination. When the user issues a query, or clicks a pagination link, the experimental search system will forward the request to a commercial search engine in real time, and retrieve the corresponding search engine result page (SERP). To control the variability in presentation styles, all query suggestions, ads, sponsor search results, and vertical results in the retrieved SERP are removed, only the remaining organic results are returned to the user. We also store these organic results in our system, not only for further annotation and analysis, but also to make sure that if another participant issues the same query, he or she will be shown the same SERP. A javascript plugin was injected into the returned SERP to log users' search behaviors including query reformulation, click, scrolling, tab switching and mouse movement.

12 search tasks were selected from the topics of TREC Session Track 2010-2014[2]. Several criteria were considered when selecting the search tasks. Firstly, a search task should be clearly stated so that different participants could interpret the task description in the same way. Secondly, the difficulty and complexity of a search task should be appropriate. The search task should be neither too time-consuming, nor so easy that only requires one query and a few clicks on top results to complete. A pilot experiment was conducted to test whether these criteria were met. Based on the result of the pilot experiment, we made necessary modifications to the original TREC task descriptions to adjust the difficulties and complexities. We further provided an initial query for each search task. While providing initial query might threaten the ecological validity of our experiment, it can make sure that all participants will see the same initial SERP for each search task, and thus effectively prevent potential topic drifts. Table 3 shows some examples of the selected search tasks.

We recruited 29 undergraduate students, via emails and online social networks, to take part in the user study. 15 participants were female and 14 were male. The ages of participants range from 18 to 26. The distribution of their major is: 15 in engineering, 10 in humanities and social sciences, and 4 in designs and arts. All the participants were familiar with basic usage of Web search engines, and most of them reported using Web search engines daily.

Each participant was asked to complete all of the 12 search tasks in a random order. As shown in Figure 1(I), to make sure that every participant was familiar with the experiment procedure,

[2] http://trec.nist.gov/data/session.html

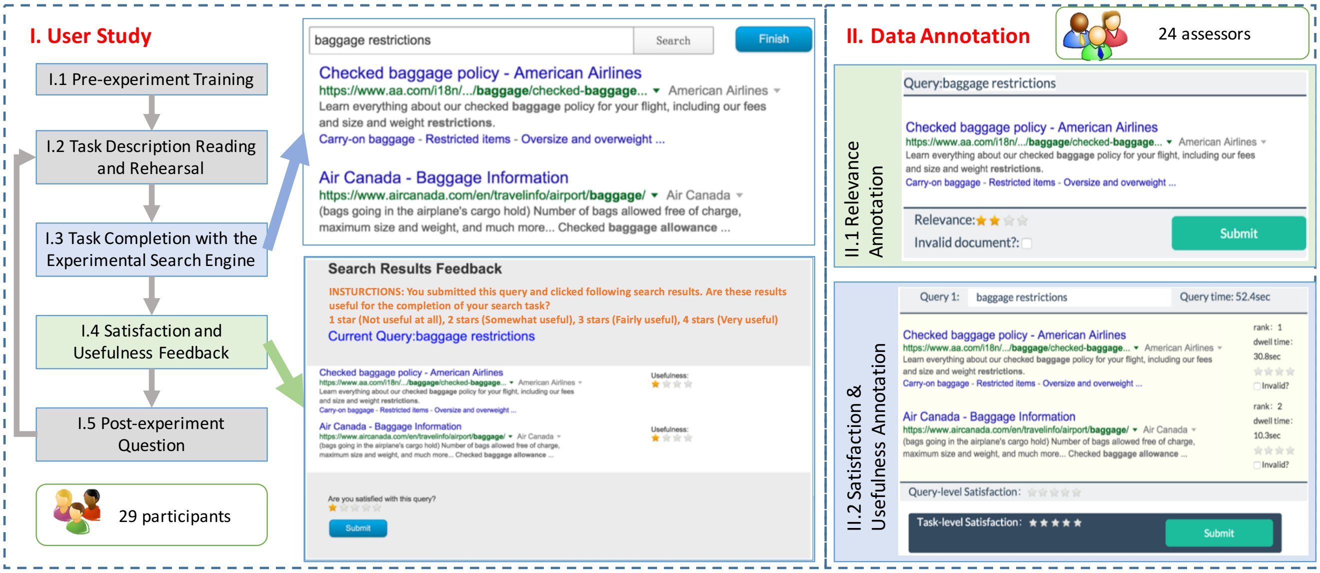

Figure 1: Data collection procedure. With enrolled participants, we collected behavior logs and feedback data in I. User Study. With hired external assessors, we generated relevance, usefulness and satisfaction annotation data in II. Data Annotation.

an example task was used for demonstration in the Pre-experiment Training stage (I.1). For each search task, the participant had to go through 4 different stages (I.2-I.5). Firstly, the participant should read and memorize the task description (note that the complete task description is provided to the participant). After that, s/he was required to re-input the task description without viewing it again during searching (I.2). Then s/he would be redirected to the SERP of the initial query, and start completing the search task (I.3). The participant could click on the results and submit new queries freely, just like using a normal Web search engine. While no task time limits were imposed, s/he could stop searching and click the finish button when s/he thought the task was completed, or no more helpful information would be found. After the task completion stage, the participant was required to review the search process and provide explicit feedbacks (I.4). Figure 1(I.4) shows the interface for collecting usefulness and query-level satisfaction feedbacks. We used a 4-level graded usefulness feedback (U_u, 1: not useful at all; 2: somewhat useful; 3: fairly useful; 4: very useful) since we aim to compare it against 4-level graded relevance annotation [26]. We used a 5-level graded *query-level satisfaction* feedback. The 5-level satisfaction scale and instructions are in accordance with to those introduced by Liu et al. [32]. We only collected usefulness feedbacks for documents that were clicked by that particular participant in the task completion stage. After reviewing all issued queries, the participant would further submit a 5-level *task-level satisfaction* feedback ($TSAT_u$). The explicit feedback stage was immediately after, but did not interfere with, the search process. We believe such an experiment design could collect most accurate feedbacks while introduce a minimal interference to users' search behavior. A question answering stage was put at the end of each search task (I.5). The participant must answer a question related to the search task (e.g. "Please provide three suggestions for quitting smoking." for the task "quit smoking") in voice. In the pilot test we found that the voice question answering introduces much less cognitive cost than requiring the participants to use keyboard to input the answers, and it can effectively ensure that the participants indeed put some effort in finishing the search task.

3.2 Data Annotation

After collecting the search behavior logs and user feedbacks in the user study, we hired external assessors to generate (1) relevance annotations (R) for all the documents that were clicked by users or shown in the top 5 positions of a SERP; (2) usefulness annotations (U_a) for all clicked documents; (3) query-level satisfaction annotations ($QSAT_a$) for all issued queries; and (4) task-level satisfactions ($TSAT_a$) for all search sessions.

Annotation Instructions:
Search Task: You are going to US by air, so you want to know what restrictions there are for both checked and carry-on baggage during air travel.
The left part shows the issued queries and clicked documents when a user is doing the search task via a search engine, you need to complete the following 3-step annotation:
STEP1: Annotate the usefulness of each clicked document for accomplishing the search task:
1 star: Not useful at all;
2 stars: Somewhat useful;
3 stars: Fairly useful;
4 stars: Very useful.
STEP2: Annotate query-level satisfaction for each query
(1 star: Most unsatisfied - 5 stars: Most satisfied)
STEP3: Finally, please annotate the task-level satisfaction
(1 star: Most unsatisfied - 5 stars: Most satisfied)
Completed units/all units : 0/29

Figure 2: Annotation instructions shown to assessors.

Figure 1(II.1) shows the interface for relevance annotation. A relevance annotation unit consists of a query-document pair. For each unit, we showed the short query and the snippet of the document, in a single page, to the assessors. The assessors were required to click and examine the document and make a 4-level graded relevance judgment (1: irrelevant; 2: somewhat relevant; 3: fairly relevant; 4: highly relevant). The relevance scale and annotation instructions are similar to those introduced by Kekäläinen et al. [26] and are also consistent with the current practice in Web search. Some documents were not accessible during the annotation process because the page had been removed or deleted. So we asked the assessor to check the `Invalid document?` checkbox when s/he could not access the document.

The annotation unit of usefulness and satisfaction annotation is a search session in which a participant completed a single search task (with full task description). As shown in Figure 1(II.2), for each annotation unit, we showed an augmented search log, along with the instruction and the search task description (see Figure 2), to the assessor. All queries and clicked documents in the log were presented in the same order as when the participant issued and clicked them in the user study. To imitate the search process and reproduce the search context, the assessors were instructed to

Table 4: Statistics of behavior logs.

#tasks	#participants	#sessions	#queries	#clicks
9	25	225	935	1,512

Table 5: Statistics of annotation data.

	R_{nc}	R_c	U_a	$QSAT_a$	$TSAT_a$
#Annotations	1,944	1,161	1,512	935	225
Weighted κ	0.344	0.413	0.530	0.535	0.274

inspect the search session and judge the document-level usefulness, query-level satisfaction and task-level satisfaction sequentially. Similar to the laboratory user study, we used the same 4-level graded scale for usefulness, and 5-level graded scale for satisfaction. We also showed behavioral informations including the query dwell time, click dwell time and the ranks of clicked documents to the assessors. The above annotation approaches are consistent with the existing studies [23, 29, 32] on user satisfaction.

24 assessors were enrolled in the data annotation tasks. They were all graduate, or senior undergraduate students. We *randomly* assigned 9 of them to complete the relevance annotation task, and 15 of them the usefulness and satisfaction annotation task.

3.3 Quality Control and Data Filtering

To make sure the data annotations are *reliable*, we ensured that each unit was judged by at least 3 different assessors. As the annotations are *ordinal*, we applied Cohen's Weighted κ [11] to assess the inter-assessor agreements. This requires a weight matrix W to indicate how severe a disagreement is. We chose to use the difference on the ordinal scale as the values in W.

After a careful inspection of the annotation data, we filtered out three search tasks: one search task which contained a considerable number of invalid documents; and two other search tasks for which there were many documents with commercial intents, and the assessors had difficulties in determining whether they were spam or not. While these tasks represent real search situations, we judge that the collected judgments are not reliable enough to serve as ground truth, so they are discarded. We also examined the search log collected in the user study, and removed the data generated by 4 participants who did not put sufficient effort in search tasks. They completed search tasks in a significantly shorter time than other participants, and gave very vague answers in the question answering stage.

Summary

Through the user study, data annotation and filtering, we collected user behavior logs, users' explicit feedbacks for usefulness and satisfaction, and a set of corresponding annotation data from external assessors. The statistics of the behavior logs are shown in Table 4. The number of collected relevance, usefulness and satisfaction annotations are shown in Table 5. We separate the assessor's relevance annotations R into two groups: R_c and R_{nc}. R_c are the relevance annotations for clicked documents, which will be compared with usefulness measures. R_{nc} are the relevance annotations for the documents that were among the top 5 results of a query, but never clicked by a user. We also list the average Weighted κ for each kind of annotations. According to Landis et al. [30] [3] , fair inter-assessor agreements between assessors are reached for R_{nc} and $TSAT_a$, and moderate agreements are reached for R_c, U_a, and $QSAT_a$, which indicates the annotation data are of reasonable quality.

4. USEFULNESS V.S. RELEVANCE

Based on the data collected, we first investigate the difference and relationship between assessor's relevance and user's usefulness to answer **RQ1**. In this work, we use usefulness feedbacks (U_u)

[3]Landis et al. [30] characterize κ values < 0 as no agreement, $0-0.20$ as slight, $0.21-0.40$ as fair, $0.41-0.60$ as moderate, $0.61-0.80$ as substantial, and $0.81-1$ as almost perfect agreement.

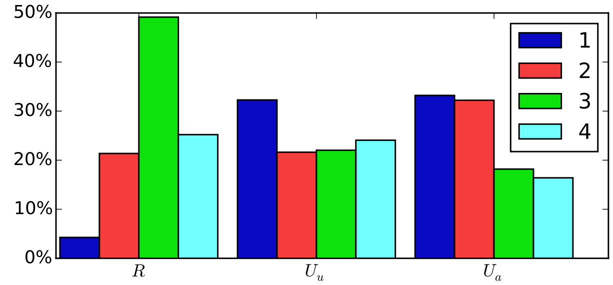

Figure 3: Marginal distributions of the relevance annotations (R), usefulness feedbacks (U_u) and usefulness annotations (U_a) for the clicked documents. For relevance, $R=1$: irrelevant; 2: somewhat relevant; 3: fairly relevant; 4: highly relevant. For usefulness, $U=1$: not useful at all; 2: somewhat useful; 3: fairly useful; 4: very useful.

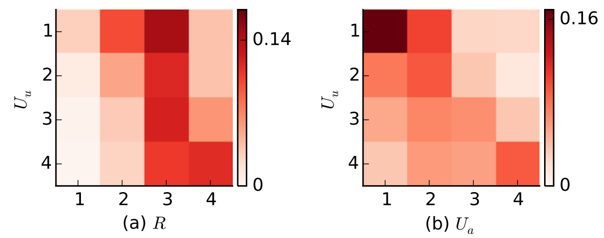

Figure 4: Joint distributions of document-level measures for clicked documents. Darker color indicates a higher frequency. (a) joint distribution of relevance annotations (R) and usefulness feedbacks (U_u); (b) joint distribution of usefulness annotations (U_a) and usefulness feedbacks (U_u).

as the ground truth labels for usefulness, to which the relevance annotation (R) for each clicked document, will be compared.

The marginal distribution of R and U_u are shown in Figure 3. Note that the distributions are computed per click, so only the relevance annotations of clicked documents (R_c in Section 3.3) are used. We can see an obvious difference between these two distributions (Chi-Square test, $\chi^2(3, N = 1,512) = 874$, $p < 0.001$). For relevance R, nearly 50% of clicked documents are annotated as fairly relevant ($R=3$). This is not surprising because all these documents ranked in high positions by a commercial search engineare topically related to the short query. Meanwhile, for usefulness feedbacks U_u, we spot a nearly uniform distribution with a little more clicks with $U_u = 1$, which implies that the user knows clearly whether an examined document is useful or not. As there are only a few clicks on the document with $R=1$ (4.3%), and a considerable number of clicks (32.3%) are reported as $U_u = 1$, we can conclude that a large proportion of the documents considered relevant by the assessors may not be useful to users.

To study the correlation between U_u and R, we compute Pearson's correlation coefficient r and Cohen's Weighted κ between these two document-level measures. A moderate positive correlation is detected, $r(1,510) = 0.332^4$, $p < 0.001$, two tails. The computed Weighted κ is 0.209 ($\sigma_\kappa = 0.017$), just reaching a fair agreement level [30]. We plot the heat map for the joint distribution of U_u and R in Figure 4 (a) and find that R and U_u are not aligned well. Except for the document with perfect relevance ($R=4$), other documents are likely to be not useful at all ($U_u = 1$). Even for the clicks on documents with fair relevancy ($R=3$), 29.3% are not useful from the users' perspective. However, only a few clicked documents have low relevance ($R \le 2$) and high usefulness ($U_u \ge 3$), suggesting that high relevance is a necessary condition for high usefulness. This finding may explain why some implicit signals for high usefulness (e.g. long dwell time [4, 46],

[4]The degree of freedom is given by $\#clicks - 2$

and last click [8, 24]) could be used as positive implicit relevance feedbacks in previous studies.

We are now interested in understanding why U_u and R are not aligned. We manually inspected the clicks with low relevance ($R \leq 2$) and high usefulness ($U_u \geq 3$), and the clicks with high relevance ($R \geq 3$) and low usefulness ($U_u \leq 2$). We find that the major reason for the users reporting that a document with low relevance is actually very useful, is that the document is useful for the overall search task but not so relevant to the current issued query (e.g. Click 3 that we showed in Table 1). On the other hand, the users will report low usefulness for some relevant documents because (1) the document is redundant in content with previously seen documents in the search session [5]; (2) the dwell time on the document is short, the user might not read it as carefully as the assessors did in relevance annotation process [47]. These observations confirm once again that the query-level relevance judgments are unable to fully capture user's perceived usefulness.

Summary

To summarize, regarding **RQ1**, we find that although there is a moderate positive correlation between assessor's relevance annotation R and user's usefulness feedback U_u, there is a significant gap between these two document-level measures in our dataset. High relevance seems to be a necessary but not sufficient condition for high usefulness, which This explains the success of the previous approaches using positive usefulness feedback as positive relevance feedback. The differences observed between assessor's relevance annotations and user's usefulness judgments also suggest that a system evaluation directly based on usefulness may be more appropriate.

5. RELEVANCE, USEFULNESS AND USER SATISFACTION

As stated by Kelly [27], "satisfaction can be understood as the fulfillment of a specified desire or goal". Satisfaction attempts to gauge users' actual feelings about the system. It is becoming an important criterion in the user-centric evaluation for Web search engines [1, 20]. As we observed in Section 4 that at document-level, user reported usefulness U_u is not well aligned with annotator's relevance annotation R, we further investigate their correlations with query-level and task-level user satisfaction ($QSAT_u$ and $TSAT_u$) to answer **RQ2**.

To do this, first we need to introduce some evaluation metrics to link document-level measures with query-level and task-level user satisfaction. In traditional batch evaluation paradigm, evaluation metrics, such as NDCG, MAP, and ERR, are used to summarize document-level relevance annotations to estimate query-level satisfactions. We refer to these classic metrics as *rank-based* metrics. On the other hand, the *click-sequence-based* metrics are computed based on the click sequences and document-level measures (i.e. usefulness or relevance) of clicked documents. We believe that this latter type of measure can better capture user satisfaction.

5.1 Correlation with Query-level Satisfaction

For query-level satisfaction, we use four click-sequence-based metrics: *cCG*, *cDCG*, *cMAX*, and *cCG/#clicks*. Click cumulated gain (*cCG*) for a query measures the total information gain, or utility, after submitting the query and viewing all the clicked documents in sequence. It is computed by summing up the document-level measures for all clicks under that query [23, 32]:

$$cCG(CS,M) = \sum_{i=1}^{|CS|} M(d_i)$$

Here, $CS = (d_1, d_2, \ldots, d_{|CS|})$ is the click sequence in which each element d_i is a clicked document. $M(d_i)$ is the document-level measure for document d_i. In this section, M can be either relevance annotation R or usefulness feedback U_u. *cCG/#clicks* is the average gain per click. Click discounted cumulative gain (*cDCG*) is defined as:

Table 6: Correlations with query-level satisfaction feedback $QSAT_u$.
All correlations (measured in Pearson's r) are significant at $p < 0.001$. *(or **) indicates the difference is significant at $p < 0.05 (p < 0.01)$, comparing to the same metric based on relevance annotation R.

	All Queries ($df = 933$)		Queries with only top 5 clicks ($df = 635$)	
	U_u	R	U_u	R
cCG	0.572^{**}	0.425	0.647^{**}	0.499
cDCG	0.724^{**}	0.498	0.747^{**}	0.535
cMAX	0.751^{**}	0.563	0.759^{**}	0.599
cCG/#clicks	0.733^{**}	0.551	0.751^{**}	0.587
MAP@5	-	0.192	-	0.255
DCG@5	-	0.295	-	0.363
ERR@5	-	0.258	-	0.332
Weighted Rel. [20]	-	0.229	-	0.273

$$cDCG(CS,M) = \sum_{i=1}^{|CS|} \frac{M(d_i)}{log_2(i+1)}$$

cMAX assumes that the user's satisfaction is largely dependent on the most relevant or useful document s/he finds. It is given by:

$$cMAX(CS,M) = \max(M(d_1), M(d_2), \ldots, M(d_{|CS|}))$$

We also use four rank-based metrics: *MAP@5*, *DCG@5*, *ERR@5* and Weighted Relevance introduced by Huffman et al. [20]. All these metrics use cut-off at rank 5, because we only collected relevance annotations for top 5 documents in the relevance annotation stage (see Section 3.2). We do not use *nDCG* [21] here, because the computation of ideal *DCG* is biased when we do not have an exhaustive list of relevant documents.

As we only have usefulness measures for clicked document, we compare the click-sequence-based metrics based on usefulness feedback U_u with those based on relevance annotation R. We compute their correlations with query-level satisfaction $QSAT_u$, and use the rank-based metrics based on relevance annotation R as references. In order to compare two correlation coefficients (rs), we construct a t-statistic to test the significance of the difference between dependent r's [12]. As the cut-off of 5 for rank-based metrics may affect their correlations with satisfaction, especially when the user goes deeper than rank 5, we further compute and report the correlations for 637 queries that only has clicks among top 5 results.

The correlations are shown in Table 6. First, we can see that the correlations between $QSAT_u$ and the click-sequence-based metrics (shown in upper part of Table 6) are stronger than those between $QSAT_u$ and rank-based metrics (shown in the lower part of Table 6). The best rank-based metric is *DCG@5* with $r(933) = 0.295$ (the degrees of freedom is given by $\#queries - 2$). However, all of the click-sequence-based metrics are more positively correlated with $QSAT_u$ than rank-based metrics, with all differences being significant at $p < 0.001$, two-tailed. Second, the click-sequence-based metrics based on U_u are more correlated with $QSAT_u$ than those based on R, with all the differences between two counterparts being significant at $p < 0.01$, two-tailed. $cCG(U_u)$, $cDCG(U_u)$ and $cCG(U_u)$ are strongly correlated with $QSAT_u$, with $r(933) > 0.7$. This result shows that user usefulness feedback is a much better indicator of user satisfaction than assessor's relevance annotations. Third, for the queries with only top 5 results clicked, the correlations between $QSAT_u$ and the rank-based metrics are slightly stronger than those for all queries; but they are still much weaker than those between click-sequence-based metrics and $QSAT_u$, with all differences being significant at $p < 0.01$, two-tailed. This suggests that click-based metrics can better capture user perceived satisfaction.

5.2 Correlation with Task-level Satisfaction

For task-level satisfaction, we only use four click-sequence-based metrics: *sCG*, *sCG/#queries*, *sCG/#clicks*, and *sDCG*. *sCG* is defined as the sum of each query's gain [22].

Table 7: Correlations with task-level satisfaction feedback $TSAT_u$.

Measured in Pearson's $r(df = 223)$. The darker and lighter shadings indicate the correlation is significant at $p < 0.01$ and 0.05. *(or **) indicates the difference is significant at $p < 0.05 (p < 0.01)$, comparing to the same metric based on relevance annotation R.

	U_u	R
sCG	0.110^{**}	-0.046
$sCG/\#queries$	0.437^{**}	0.330
$sCG/\#clicks$	0.525^{**}	0.320
$sDCG$	0.317^{**}	0.142

We use cCG to measure a query q_j's gain. So sCG is computed by:

$$sCG(M) = \sum_{j=1}^{n} gain(q_j) = \sum_{j=1}^{n} cCG(CS_j, M)$$

Here n is the number of queries in the session. CS_j is the click sequence for q_j. $sCG/\#queries$ and $sCG/\#clicks$ measure average gain per query and per click. $sDCG$ [22] discounts the gains for later queries in a search session:

$$sDCG(M) = \sum_{j=1}^{n} \frac{gain(q_j)}{1 + log(j)} = \sum_{j=1}^{n} \frac{cCG(CS_j, M)}{1 + log(j)}$$

The correlations between these click-sequence-based metrics and the task-level satisfaction feedbacks $TSAT_u$ are shown in Table 7. Except for sCG, the other metrics significantly correlate with $TSAT_u$. The click-sequence-based metrics based on U_u are significantly more correlated with $TSAT_u$ than their counterparts based on R (with $p < 0.01$, two-tailed). $sCG(U_u)/\#clicks$ is moderately correlated with task-level satisfaction $TSAT_u$, with $r(223) = 0.525$.

Summary

In this section, regarding **RQ2**, we compare a variety of evaluation metrics based on either user's usefulness feedbacks U_u or assessor's relevance annotation R with query-level satisfaction feedbacks $QSAT_u$ and task-level satisfaction feedbacks $TSAT_u$. Comparing to the rank-based metrics, the click-sequence-based metrics are more related to users' query-level satisfaction feedbacks. Comparing to relevance, usefulness has a stronger correlation with user satisfaction in all metrics. These empirical results further suggest that: (1) when the click sequence is known, we can exploit click-sequence-based metrics to make a better user-oriented evaluation; (2) usefulness can better reflect user's real feelings in Web search than assessor's relevance.

6. COLLECTING USEFULNESS LABELS

In Section 4 and 5, we showed that there is a significant difference between assessor's relevance and user's usefulness. Although usefulness may be more suited for evaluating the Web search engine, it is unrealistic to collect explicit usefulness feedback from users. We have to come up with alternative approaches to assess and acquire document-level usefulness labels. In this section, with regard to **RQ3** and **RQ4**, we test two such approaches. The first one is to rely on external assessors to review augmented search logs and make document-level usefulness annotations. The second one is to use a machine learning method and features extracted from behavior logs to estimate usefulness.

We evaluate these two usefulness estimation approaches in terms of their *reliability* and *validity*. As stated by Kelly [27] (p. 176), reliability is "the extent to which the method and measures yield consistent findings", and validity is "the extent to which methods and measures allow a researcher to get at the essence of whatever it is that is being studied". Reliability is a necessary condition for validity, and when combined together, these two criteria measure the extent to which the usefulness labels produced by theses two approaches can reflect the user-perceived usefulness of documents.

Table 8: Correlations with usefulness feedbacks U_u.

*(or **) indicates difference is significant at $p < 0.05 (p < 0.01)$, comparing to the same metric related to R

	Pearson's r	MSE	MAE	Weighted κ
U_a	0.413^{**}	1.51^{**}	0.852^{**}	0.321^{**}
R	0.332	1.79	1.020	0.209

Table 9: Correlations with query-level satisfactions $QSAT_u$.

* (or **) indicates the difference between U_a and R is significant at $p < 0.05 (p < 0.01)$. $^{\triangledown}$ (or $^{\blacktriangledown}$) indicates the difference between U_a and U_u is significant at $p < 0.05 (p < 0.01)$. The darker and lighter shadings indicate the difference between U_a and $QSAT_a$ is significant at $p < 0.01$ and 0.05.

	Pearson's $r(df = 933)$			Pref. agreement ratio		
	U_a	U_u	R	U_a	U_u	R
cCG	$.466^{\blacktriangledown/*}$	.572	.425	$.701^{\blacktriangledown/**}$	.751	.669
$cDCG$	$.518^{\blacktriangledown/*}$	.724	.498	$.742^{\blacktriangledown/**}$	.826	.698
$cMAX$	$.580^{\blacktriangledown/*}$	.751	.563	$.681^{\blacktriangledown/**}$	.779	.632
$cCG/\#clicks$	$.548^{\blacktriangledown}$	.733	.551	$.716^{\blacktriangledown/*}$	.807	.689
$QSAT_a$	.508			.584		

For usefulness annotation approach, we assess its reliability by calculating the inter-assessor agreement, and its validity by comparing usefulness annotations U_a with usefulness feedbacks from users U_u and correlating them with query-level satisfaction feedbacks $QSAT_u$. For usefulness prediction approach, we also assess its validity by comparing the predicted usefulness scores with U_u and $QSAT_u$, and we use cross-validations and significance tests to ensure the results are reliable.

6.1 Usefulness Annotation

The detailed procedure of usefulness annotation is described in Section 3.2. So here we only describe and discuss the reliability and validity of collected usefulness annotations U_a.

To measure the reliability of usefulness annotation, we use Cohen's Weighted κ to assess the agreement between different assessors. As shown in Table 5, the κ for U_a ($\kappa_{U_a} = 0.530, \sigma_{\kappa_{U_a}} = 0.008$) is larger than those for R_c ($\kappa_{R_c} = 0.413, \sigma_{\kappa_{R_c}} = 0.010$) and R_{nc} ($\kappa_{R_{nc}} = 0.344, \sigma_{\kappa_{R_{nc}}} = 0.008$). The standard error of weighted κs are computed by the method introduced by Cohen [11]. The difference between κ_{U_a} and κ_{R_c} and the difference between κ_{U_a} and $\kappa_{R_{nc}}$ are both significant at $p < 0.001$ (two-tailed independent t-tests). These results suggest that, measuring at the inter-assessor agreement level, the usefulness annotations are more *reliable* than the conventional relevance annotations. The possible reason is that providing search context and behavioral information (e.g. the full search task, search session and dwell times) to assessors may help them make judgments. This is corroborated to some extent by the marginal distribution of U_a shown in Figure 3: unlike the relevance distribution concentrated on $R = 3$, the distribution of U_a is a more similar to U_u than R, which indicates that, with the help of search context and user behavior information, the assessors can detect low usefulness clicks and make more discriminative judgements.

To assess the validity of usefulness annotation, we first compare U_a with the usefulness feedbacks U_u, which are used as the ground truth labels for usefulness. The correlations are measured in Pearson's r, Mean Squared Error (MSE), Mean Absolute Error (MAE), and Cohen's Weighted κ. The results are shown in Table 8. A moderate positive correlation ($r(1,510) = 0.412$, $p < 0.001$, two tailed) and a fair agreement ($\kappa = 0.321$, $\sigma_\kappa = 0.017$) between U_u and U_a are detected. The correlation between U_a and U_u is significantly stronger than that between R and U_u. We also show the joint distribution of U_u and U_a in Figure 4(b). The diagonal blocks are the darkest block in almost every rows and columns, showing a fair agreement between U_u and U_a. From the correlation metrics and the joint distribution we can see that although U_u and U_a are not perfectly aligned, comparing to relevance annotation, usefulness annotation can better reflect the user-perceived usefulness.

As shown in Section 5.1, a strong correlation exists between usefulness feedbacks and query-level satisfaction. Therefore, a valid assessment of usefulness should also correlate well with query-level satisfaction feedbacks $QSAT_u$. We use usefulness annotations U_a to compute four click-sequence-based metrics defined in Section 5.1: cCG, $cDCG$, $cMAX$, and $cCG/\#clicks$, and correlate them with $QSAT_u$. Beside computing the Pearson's r for these correlations, we also conduct a naturalistic pairwise preference test. In the preference test, we extract 1,455 query pairs (q_i, q_j), where q_i and q_j belong to the same search session, and $QSAT_u(q_i) > QSAT_u(q_j)$. For each query pair, if an evaluation metric also indicates the same relative preference, then we say the evaluation metric agrees with $QSAT_u$ on that query pair. A similar method is used by Sanderson et al. [36]. As we only extract query pairs from the same search sessions, the preference test can effectively reduce the variabilities introduced by different users and different search tasks.

We report the correlations with $QSAT_u$, measured in Pearson's r and the agreement ratios in the preference test, in Table 9. We compare the correlations related to U_a to those related to relevance R (baseline) and usefulness feedbacks U_u (oracle performance). We also use the query-level satisfaction annotation from external assessors ($QSAT_a$) as another baseline. The results show that, although usefulness annotations U_a do not correlate with query-level satisfaction feedbacks $QSAT_u$ as well as usefulness feedbacks U_u from users (all the differences are significant at $p < 0.01$), most click-sequence-based metrics based on U_a outperform their counterparts based on R, in terms of correlation with $QSAT_u$. It is also interesting to observe that query-level satisfaction annotations from external assessors ($QSAT_a$) are quite different from query-level satisfaction feedbacks from users ($QSAT_u$), which is also observed by Liu et al. [32]. Some of click-sequence-based metrics based on U_a are significantly better than satisfaction annotations ($QSAT_a$), which suggests that document-level usefulness annotation may be more valid than query-level satisfaction annotation.

6.2 Usefulness Prediction

As previous studies show that there are substantial correlations between the user behavior signals (e.g. long dwell time [4, 31, 46], last click in a query [8, 24], and query position and reformulation types [34]) and evaluation-related measures like document relevance, search success, and user satisfaction, we attempt to use a regression model based on user behavior features and search context features to (1) automatically generate document-level usefulness labels, and (2) improve and enhance the document-level annotations (both R and U_a) so as to make them more aligned to users' usefulness feedbacks U_u.

Features

We list the features extracted from behavior logs in Table 10. We categorize these features into three groups: Query features (Q), Session features (S) and User features (U). Query features are the features related to a single query. With user behavior features, such as click numbers and dwell time included, they mainly describe how the user interacted with the search engine. Session features depend on the whole search session, and include short-term search context features like query position and query reformulation types. To compute User features, we need the long-term search history of that user. For (1) automatic usefulness label generation, only query features, session features and user features are involved (Q+S+U or referred to as All for simplicity). For (2) annotation enhancement, we extract relevance annotation features (R) and usefulness annotation features (A) from the annotation data. In particular, we use the document-level annotation itself, and the four interactive evaluation metrics computed by the document-level annotations, as relevance annotation features (R) and usefulness annotation features (A).

Table 10: Features to predict usefulness, extracted from the behavior logs.

Query features(Q)	
rank	The rank of clicked document in result list
#clicks	The number of clicks in the query
query length	The length of the query, in words and in characters
click position	Whether the click is the first/last/intermediate click in a query with more than one click, and whether the query has only one click
dwell time	click dwell time and query dwell time
Session features(S)	
#queries	The number of queries in the search session
#queries w/o click	The number of queries without click in session
query position	Whether the query is the first/last/intermediate query in a session with more than one query, and whether the session has only one query
time to completion	The total time spent on this search session
query reformulation	Whether the query is generated from a specification/ generalization/ parallel reformulation, and whether the query leads to a specification/ generalization/ parallel reformulation
User features(U)	
user #clicks	The average/max/min/standard deviation of #clicks per query of the user
user #queries	The average/max/min/standard deviation of #queries per session of the user
user #dwell time	The average/max/min/standard deviation of query/click dwell time of the user

Prediction Models

We frame the usefulness prediction as a supervised regression problem, and use usefulness feedbacks (U_u) for clicked documents as the target value of the regression model. We perform five-fold cross-validation over search sessions to ensure the results are reliable. All the user features are computed on the training set. Since the cross-validation are performed over sessions, each session belongs to either the training set or the test set, the query and session features for a test document will not be present in the training set. We use a Gradient Boosting Regression Tree (GBRT) [17] as our regression model, because it can naturally handles mixed types of features, has a good predictive power, and is robust to outliers. A variety of feature combinations are tested. Similar to usefulness annotation studied in Section 6.1, we evaluate the validity of usefulness predictions in terms of their correlations with usefulness feedbacks(U_u), and their correlations with query-level satisfaction feedbacks($QSAT_u$).

Prediction Results

We measure the correlations between predicted usefulness scores and usefulness feedbacks from users (U_u) in Pearson's r, MSE and MAE. The results are shown in Table 11. We use subscripts to indicate the feature groups used in usefulness prediction, for examples, U_{Q} refers to the predicted usefulness based on the query features and U_{All} refers to the predictions based on all the features extracted from the behavior logs (i.e. Q+S+U). Both relevance annotation R and usefulness annotation U_a are used as baselines.

The results show that, as we add more behavior features, the performance of usefulness prediction increases, which proves that search context features (Q) and user-specific features (U) are useful in usefulness prediction. Comparing to R, all the predicted usefulness scores $U_{(\cdot)}$ are significantly more correlated with users' usefulness feedbacks, which once again demonstrates the gap between relevance and user-perceived usefulness. When we combine all the features extracted from the behavior logs, the resulting U_{All} establishes a comparable or stronger correlation with U_u, than usefulness annotation U_a does. This result suggests that when some usefulness feedbacks from users U_u are available for training, instead of relying on external assessors to generate usefulness annotation U_a, we can automatically generate document-level usefulness labels U_{All} of at least equal validity to U_a, based on the features that can be implicitly collected from behavior logs (1).

On the other hand, for (2) annotation enhancement, when we

Table 11: Results for usefulness prediction.

Measured in the correlations with usefulness feedback U_u. *(or **) indicates the difference between $U_{(\cdot)}$ and R is significant at $p < 0.05 (p < 0.01)$. The darker / lighter shadings indicates the difference between $U_{(\cdot)}$ and U_a is significant at $p < 0.05/0.01$.

	Pearson's r	*MSE*	*MAE*
U_{Q}	0.398^*	1.198^{**}	0.894^{**}
$U_{\text{Q+S}}$	0.410^{**}	1.186^{**}	0.889^{**}
U_{All}	0.461^{**}	1.103^{**}	0.851^{**}
$U_{\text{All+A}}$	0.467^{**}	1.105^{**}	0.845^{**}
$U_{\text{All+R}}$	0.519^{**}	1.021^{**}	0.815^{**}
$U_{\text{All+A+R}}$	0.521^{**}	1.023^{**}	0.803^{**}
U_a	0.413	1.512	0.852
R	0.332	1.786	1.020

Table 12: Correlations with query-level satisfactions $QSAT_u$.

Measured in Pearson's $r(df = 933)$. *(or **) indicates the difference between $U_{(\cdot)}$ and U_a is significant at $p < 0.05 (p < 0.01)$. $^{\triangledown}$(or $^{\blacktriangledown}$) indicates the difference between $U_{(\cdot)}$ and U_u is significant at $p < 0.05 (p < 0.01)$. The darker and lighter shadings indicate the difference between $U_{(\cdot)}$ and Jiang et al. [23] is significant at $p < 0.01$ and 0.05.

	U_{All}	$U_{\text{All+A+R}}$	U_a	U_u
cCG	$0.459^{\blacktriangledown}$	$0.490^{**/\blacktriangledown}$	0.466	0.572
cDCG	$0.580^{**/\blacktriangledown}$	$0.612^{**/\blacktriangledown}$	0.518	0.724
cMAX	$0.601^{\blacktriangledown}$	$0.635^{**/\blacktriangledown}$	0.580	0.751
cCG/#clicks	$0.571^{\blacktriangledown}$	$0.608^{**/\blacktriangledown}$	0.548	0.733
$QSAT_a$	0.508			
Jiang et al. [23]	0.539			

combined behavior features (All) with document-level annotations (A or R), significant improvements over the annotation-based baselines (U_a and R) are found. While it is not surprising to see $U_{\text{All+A+R}}$ achieving the best performance, it is interesting to find that $U_{\text{All+R}}$ is better than $U_{\text{All+A}}$. A possible reason for this is that while usefulness annotations inevitably depend on some behavior features like dwell time, the relevance annotations of documents are in some sense more complementary to the behavior features than usefulness annotations, thus U_{All+R} has a broader coverage of useful features than U_{All+A}.

Correlations with Query-level Satisfaction

We further demonstrate the validity of usefulness prediction approach by showing the correlations between predicted usefulness labels and query-level satisfaction feedbacks $QSAT_u$. Due to the lack of space, we only show the correlations in Pearson's r, since the preference test gives similar results. Usefulness annotation U_a and usefulness feedback U_u are used as document-level baselines; query-level satisfaction annotation $QSAT_a$ and a graded satisfaction prediction method based on user behavior features developed by Jiang et al. [23] are used as query-level baselines. Although our goal is not to predict satisfaction, we use this method as a baseline because we share similar behavior features, and the performance of a recently proposed satisfaction prediction model sets up a relatively high standard about how well one can predict satisfaction based on these features.

The results (Table 12) show the following facts: Firstly, because the correlations related to U_{All} are comparable or stronger when compared to those related to U_a, the usefulness predictions based on behavior features are at least as valid as usefulness annotations. Secondly, the usefulness predictions based on both behavior features and annotation features $U_{\text{All+A+R}}$ are significantly better than U_a (therefore better than R). We can thus use behavior features to enhance usefulness and relevance annotation. Finally, although there is still a significant gap between usefulness predictions (U_{All} and $U_{\text{All+A+R}}$) and usefulness feedbacks (U_u), the click-sequence-based metrics based on document-level usefulness predictions outperforms the query-level satisfaction prediction baselines in terms of correlations with $QSAT_u$, which indirectly proves that these usefulness predictions indeed reflect users' opinions and perception to some extent.

Summary

In this section, we proposed two usefulness labeling methods: usefulness annotation and automatic usefulness prediction, and conducted analyses to demonstrate their reliability ad validity. With regards to **RQ3**, we find that usefulness annotations are more reliable than conventional relevance annotations. The assessors in usefulness annotation process can detect low usefulness clicks effectively. The usefulness annotations collected in this process are shown to be valid due to their consistence with usefulness feedbacks and query-level satisfaction feedbacks from users. With regards to **RQ4**, we show that using behavior features, we can automatically generate valid usefulness labels, and improve existing document-level annotations so as to make them more aligned to usefulness feedbacks.

To summarize, we can collect reliable and valid usefulness labels by different approaches. When there is no usefulness feedback from any users at all, we can hire external assessors to generate usefulness annotation when provided with sufficient search context information. When there are some usefulness feedbacks from real users, we can use machine learning techniques and features extracted from behavior logs, to generate usefulness labels for other search sessions. We can also combine manual annotations from assessors and features from behavior logs to better estimate usefulness. In this case, if the cost of the additional annotations is taken into account, it is better to ask the assessors to give relevance judgments instead of usefulness annotations.

7. CONCLUSIONS AND DISCUSSIONS

In this work, through a carefully designed user study and dedicated annotation processes, we collected a comprehensive dataset that consists of behavior logs, user feedback data, and corresponding annotation data. Based on this dataset, we first investigated the difference and relationship between two document-level measures: the system-centric, highly-independent, objective relevance, and the user-centric, situational, and sometimes subjective usefulness. The results suggest that high relevance by assessors is a necessary but not sufficient condition for high usefulness for users, thus, in general, these two document-level measures are not aligned well. We further studied the correlations between relevance, usefulness, and user satisfaction, and found that usefulness is potentially of a great value in the evaluation of Web search engines since it is highly correlated with query-level satisfaction feedbacks. These findings partially explain why traditional system-centric evaluation metrics are not well aligned with user satisfaction. Finally, we proposed two approaches to collect usefulness labels in practical Web search settings, and evaluate them in terms of their reliability and validity.

Our findings and conclusions are based on a laboratory user study in which a set of predefined tasks are used and 29 participants are treated as real search users. Compared to a naturalist study based on real search logs, a laboratory user study has its limitations in its scale and the ecological validity of the collected data. However, the laboratory study has the advantage to be able to control the variabilities that lie in the different information needs from different users. To enhance the ecological validity and ensure our findings can generalize, we carefully chose the search tasks and designed the experimental search system to simulate practical Web search scenarios.

Although the main theme of this paper is contrasting usefulness perceived by users with relevance annotations by assessors, we do not hope to fully replace the latter with the former in all situations. Traditional relevance annotations have the advantage to be reusable, thus can be used to evaluate the system *in prior to* its deployment; while usefulness is suited in a more user-centric *post hoc* evaluation. Although the latter evaluations are of great importance for commercial Web search engines, the former is still indispensable.

Our study makes a first step towards a new user-centric evaluation framework. A variety of click-sequence-based evaluation metrics (e.g. *cCG* and *cDCG*) are shown to be better suited for user-centric evaluations in this work. Their properties, and the assumptions and user models behind them are worth being investigated in the future. To fully establish a new evaluation framework based on usefulness and these metrics, more user studies that involve multiple search systems and more users are required in the future.

8. ACKNOWLEDGMENTS

This work was supported by Tsinghua University Initiative Scientific Research Program(2014Z21032), National Key Basic Research Program (2015CB358700) and Natural Science Foundation (61532011, 61472206) of China.

9. REFERENCES

[1] A. Al-Maskari, M. Sanderson, and P. Clough. The relationship between ir effectiveness measures and user satisfaction. In *Proc. SIGIR '07*, pages 773–774, New York, NY, USA, 2007. ACM.

[2] A. Al-Maskari, M. Sanderson, and P. Clough. Relevance judgments between trec and non-trec assessors. In *Proc. SIGIR '08*, pages 683–684, New York, NY, USA, 2008. ACM.

[3] N. J. Belkin, M. Cole, and R. Bierig. Is relevance the right criterion for evaluating interactive information retrieval. In *Proc. SIGIR '08 Workshop on Beyond Binary Relevance: Preferences, Diversity, and Set-Level Judgments.*, 2008.

[4] G. Buscher, L. van Elst, and A. Dengel. Segment-level display time as implicit feedback: A comparison to eye tracking. In *Proc. SIGIR '09*, pages 67–74, New York, NY, USA, 2009. ACM.

[5] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proc. SIGIR '98*, pages 335–336, New York, NY, USA, 1998. ACM.

[6] B. Carterette, E. Kanoulas, M. Hall, and P. Clough. Overview of the trec 2014 session track. 2013.

[7] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proc. CIKM '09*, pages 621–630, New York, NY, USA, 2009. ACM.

[8] O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking. In *Proc. WWW '09*, pages 1–10, 2009.

[9] 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 *Proc. SIGIR '08*, pages 659–666, 2008.

[10] C. Cleverdon. The cranfield tests on index language devices. In *Aslib proceedings*, volume 19, pages 173–194. MCB UP Ltd, 1967.

[11] J. Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[12] J. Cohen and P. Cohen. *Applied multiple regression/correlation analysis for the behavioral sciences*, chapter 2, pages 53–54. Lawrence Erlbaum Associates, 1975.

[13] M. Cole, J. Liu, N. Belkin, R. Bierig, J. Gwizdka, C. Liu, J. Zhang, and X. Zhang. Usefulness as the criterion for evaluation of interactive information retrieval. *Proc. HCIR*, pages 1–4, 2009.

[14] W. S. Cooper. On selecting a measure of retrieval effectiveness. *JASIS*, 24(2):87–100, 1973.

[15] Z. Dou, R. Song, and J.-R. Wen. A large-scale evaluation and analysis of personalized search strategies. In *Proc. WWW '07*, pages 581–590, New York, NY, USA, 2007. ACM.

[16] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM TOIS*, 23(2):147–168, 2005.

[17] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[18] A. Hassan, R. Jones, and K. Klinkner. Beyond dcg: User behavior as a predictor of a successful search. In *Proc. WSDM '10*, pages 221–230, 2010.

[19] A. Hassan, R. W. White, S. T. Dumais, and Y.-M. Wang. Struggling or exploring?: Disambiguating long search sessions. In *Proc. WSDM '14*, pages 53–62, New York, NY, USA, 2014. ACM.

[20] S. Huffman and M. Hochster. How well does result relevance predict session satisfaction? In *Proc. SIGIR '07*, pages 567–574, 2007.

[21] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM TOIS*, 20(4):422–446, Oct. 2002.

[22] K. Järvelin, S. L. Price, L. M. Delcambre, and M. L. Nielsen. Discounted cumulated gain based evaluation of multiple-query ir sessions. In *Advances in Information Retrieval*, pages 4–15. 2008.

[23] J. Jiang, A. Hassan Awadallah, X. Shi, and R. W. White. Understanding and predicting graded search satisfaction. In *Proc. WSDM '15*, pages 57–66, New York, NY, USA, 2015. ACM.

[24] S. Jung, J. L. Herlocker, and J. Webster. Click data as implicit relevance feedback in web search. *Information Processing & Management*, 43(3):791–807, 2007.

[25] E. Kanoulas, B. Carterette, P. Clough, and M. Sanderson. Evaluating multi-query sessions. In *Proc. SIGIR '11*, pages 1053–1062, 2011.

[26] J. Kekäläinen and K. Järvelin. Using graded relevance assessments in ir evaluation. *JASIST*, 53(13):1120–1129, 2002.

[27] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1—2):1–224, 2009.

[28] D. Kelly, X. Fu, and C. Shah. Effects of rank and precision of search results on users' evaluations of system performance. *University of North Carolina*, 2007.

[29] Y. Kim, A. Hassan, R. W. White, and I. Zitouni. Modeling dwell time to predict click-level satisfaction. In *Proc. WSDM '14*, pages 193–202, New York, NY, USA, 2014. ACM.

[30] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[31] C. Liu, J. Liu, N. Belkin, M. Cole, and J. Gwizdka. Using dwell time as an implicit measure of usefulness in different task types. *Proc. ASIST*, 48(1):1–4, 2011.

[32] Y. Liu, Y. Chen, and et. al. Different users, different opinions: Predicting search satisfaction with mouse movement information. In *Proc. SIGIR '15*, pages 493–502, 2015.

[33] A. Moffat, P. Thomas, and F. Scholer. Users versus models: What observation tells us about effectiveness metrics. In *Proc. CIKM '13*, pages 659–668, New York, NY, USA, 2013. ACM.

[34] D. Odijk, R. W. White, A. Hassan Awadallah, and S. T. Dumais. Struggling and success in web search. In *Proc. CIKM '15*, pages 1551–1560, New York, NY, USA, 2015. ACM.

[35] T. Sakai and R. Song. Evaluating diversified search results using per-intent graded relevance. In *Proc. SIGIR '11*, pages 1043–1052.

[36] M. Sanderson, M. L. Paramita, P. Clough, and E. Kanoulas. Do user preferences and evaluation measures line up? In *Proc. SIGIR '10*, pages 555–562, New York, NY, USA, 2010. ACM.

[37] T. Saracevic. Relevance reconsidered. In *the Second Conference on Conceptions of Library and Information Science*, volume 1, pages 201–218, 1996.

[38] M. Shokouhi, R. W. White, P. Bennett, and F. Radlinski. Fighting search engine amnesia: Reranking repeated results. In *Proc. SIGIR '13*, pages 273–282, New York, NY, USA, 2013. ACM.

[39] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *Proc. SIGIR '05*, pages 449–456, New York, NY, USA, 2005. ACM.

[40] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *Proc. SIGIR '06*, pages 11–18, 2006.

[41] P. Vakkari and E. Sormunen. The influence of relevance levels on the effectiveness of interactive information retrieval. *JASIST*, 55(11):963–969, 2004.

[42] S. Verberne, M. Heijden, M. Hinne, M. Sappelli, S. Koldijk, E. Hoenkamp, and W. Kraaij. Reliability and validity of query intent assessments. *JASIST*, 64(11):2224–2237, 2013.

[43] E. M. Voorhees. The philosophy of information retrieval evaluation. In *the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, CLEF '01, pages 355–370, 2002.

[44] E. M. Voorhees and D. Harman. Overview of trec 2001. In *Trec*, 2001.

[45] R. W. White and S. M. Drucker. Investigating behavioral variability in web search. In *Proc. WWW '07*, pages 21–30, 2007.

[46] R. W. White and D. Kelly. A study on the effects of personalization and task information on implicit feedback performance. In *Proc. CIKM '06*, pages 297–306, New York, NY, USA, 2006. ACM.

[47] E. Yilmaz, M. Verma, N. Craswell, F. Radlinski, and P. Bailey. Relevance and effort: An analysis of document utility. In *Proc. CIKM '14*, pages 91–100, New York, NY, USA, 2014. ACM.