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].

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 | Usefulness: 3 (Fairly) |
| Click #2 | Air Canada - Baggage Information |
| Relevance: 4 | Usefulness: 2 (Somewhat) |
| Query #2 | carry-on baggage liquids |
| Click #3 | The Best Way to Pack a Suitcase |
| Relevance: 2 | 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 assessor 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 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.
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).

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.

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.

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

<table>
<thead>
<tr>
<th>Concepts</th>
<th>Measures</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevance</td>
<td>Relevance annotation (R)</td>
<td>3-level graded relevance annotations made by external assessors in Stage II.1</td>
</tr>
<tr>
<td>Usefulness</td>
<td>Usefulness feedbacks (U_f)</td>
<td>3-level graded usefulness feedbacks collected in Stage I.4 (see Figure 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-level graded usefulness feedbacks made by external assessor in Stage II.2</td>
</tr>
<tr>
<td>Satisfaction</td>
<td>Usefulness satisfaction annotation</td>
<td>3-level graded usefulness annotations made by external assessors reviewing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>usefulness</td>
</tr>
<tr>
<td></td>
<td>Usefulness</td>
<td></td>
</tr>
<tr>
<td></td>
<td>prediction or predicted usefulness</td>
<td></td>
</tr>
<tr>
<td>Query-level</td>
<td>Satisfaction</td>
<td></td>
</tr>
<tr>
<td></td>
<td>satisfaction feedbacks (QSATS)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(see Figure 1 and Section 3.2)</td>
<td></td>
</tr>
<tr>
<td>Task-level</td>
<td>Satisfaction</td>
<td></td>
</tr>
<tr>
<td></td>
<td>satisfaction feedbacks (TSATS)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(see Figure 1 and Section 3.2)</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Task</th>
<th>Query</th>
<th>Description</th>
<th>Description</th>
</tr>
</thead>
</table>
|      |       | 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) | travel restrictions...
| 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) | long-term care...
| Quit smoking | Your friend would like to quit smoking. You would like to provide him with relevant information about: the different ways to stop smoking, benefits of quitting smoking, second effects of quitting smoking. (2013-12) | quitting smoking effects...
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 ($U_i$: 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 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 ($T_{SAT_i}$). 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 (with full task description). As shown in Figure 1(II.2), for each a search session in which a participant completed a single search task (e.g. “Please provide three suggestions for quitting smoking.”), the participant was 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 assessor was instructed to

**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 ($U_i$) for all clicked documents; (3) query-level satisfaction annotations ($QSAT_i$) for all issued queries; and (4) task-level satisfactions ($T_{SAT_i}$) for all search sessions.

**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.
we chose to use the same 4-level graded scale for usefulness, and 5-level graded scale for satisfaction. We also showed behavioral information 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 see that the collected judgments are not reliable enough to serve as ground truth, so they are discarded. We also examined the 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_e and R_o. R_e are the relevance annotations for clicked documents, which will be compared with usefulness measures. R_o 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_o and QSAT_u, and moderate agreements are reached for R_e, U_o, and QSAT_u, which indicates the annotation data are of reasonable quality.

4. USEFULNESS VS. 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_o)

³Landis et al. [30] characterize κ values as < 0 as no agreement, 0 – 0.20 as slight, 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_o) 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_o); (b) joint distribution of usefulness annotations (U_a) and usefulness feedbacks (U_o).

The marginal distribution of R and U_o are shown in Figure 3. Note that the distributions are computed per click, so only the relevance annotations of clicked documents (R_e in Section 3.3) are used. We can see an obvious difference between these two distributions (Chi-Square test, χ²(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 engine topically related to the short query. Meanwhile, for usefulness feedbacks U_o, we spot a nearly uniform distribution with a little more clicks with U_o = 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_o = 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_o 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) = 0.332⁴, p < 0.001, two tails. The computed Weighted κ is 0.209 (κ = 0.017), just reaching a fair agreement level [30]. We plot the heat map for the joint distribution of U_o and R in Figure 4 (a) and find that R and U_o are not aligned well. Except for the document with perfect relevance (R = 4), other documents are likely to be not useful at all (U_o = 1). Even for the clicks on documents with fair relevance (R = 3), 29.3% are not useful from the users’ perspective. However, only a few clicks on documents have low relevance (R ≤ 2) and high usefulness (U_o ≥ 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],

⁴The 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 satisfaction. 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: \( CG, DCG, MAX, \) and \( CG/\#clicks \). Click cumulated gain (\( CG \)) 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 \( U_u \). \( CG/\#clicks \) is the average gain per click. Click discounted cumulative gain (\( cDCG \)) is defined as:

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

\( cMAX \) 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 \( r \), 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: \( CG, CG/\#queries, CG/\#clicks \), and \( DCG \). \( CG \) is defined as the sum of each query’s gain [22].

**Table 6: Correlations with query-level satisfaction feedback \( QSAT_u \)**

<table>
<thead>
<tr>
<th>Correlation</th>
<th>All Queries ((df = 933))</th>
<th>Queries with top 5 clicks ((df = 635))</th>
</tr>
</thead>
<tbody>
<tr>
<td>( dCG )</td>
<td>0.724( ^* )</td>
<td>0.746( ^* )</td>
</tr>
<tr>
<td>( dCG )</td>
<td>0.575( ^* )</td>
<td>0.637( ^* )</td>
</tr>
<tr>
<td>( cMAX )</td>
<td>0.751( ^* )</td>
<td>0.759( ^* )</td>
</tr>
<tr>
<td>( cCG/#clicks )</td>
<td>0.733( ^* )</td>
<td>0.587</td>
</tr>
<tr>
<td>MAP@5</td>
<td>-0.192</td>
<td>0.255</td>
</tr>
<tr>
<td>DCG@5</td>
<td>-0.293</td>
<td>0.363</td>
</tr>
<tr>
<td>ERR@5</td>
<td>-0.258</td>
<td>0.332</td>
</tr>
<tr>
<td>Weighted Rel. [20]</td>
<td>-0.229</td>
<td>0.273</td>
</tr>
</tbody>
</table>
We use $cCG$ to measure a query $q_j$’s gain. So $sCG$ is computed by:

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

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:

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

The correlations between these click-sequence-based metrics and the task-level satisfaction feedbacks $TSAu_r$ are shown in Table 7. Except for $sCG$, the other metrics significantly correlate with $TSAu_r$. The click-sequence-based metrics based on $Ua$ are significantly more correlated with $TSAu_r$ than their counterparts based on $R$ (with $p < 0.01$, two-tailed). $sCG(Ua)/#clicks$ is moderately correlated with task-level satisfaction $TSAu_r$ 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 $Ua$ or assessor’s relevance annotation $R$ with query-level satisfaction feedbacks $QSATu_r$. We also show $cCG$ and $cDCG$ are significantly more correlated with $TSAu_r$ than their counterparts based on $R$ (with $p < 0.01$, two-tailed). $cCG(Ua)/#clicks$ is correlated with task-level satisfaction $TSAu_r$ with $r(223) = 0.525$.

### 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 these two approaches can reflect the user-perceived usefulness of documents.

From usefulness annotation approach, we assess its reliability by calculating the inter-assessor agreement, and its validity by comparing usefulness annotations $Ua$ with usefulness feedbacks from users $Ua$ and correlating them with query-level satisfaction feedbacks $QSATu_r$. For usefulness prediction approach, we also assess its validity by comparing the predicted usefulness scores with $Ua$ and $QSATu_r$, 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 $Ua$.

To measure the reliability of usefulness annotation, we use Cohen’s Weighted $\kappa$ to assess the agreement between different assessors. As shown in Table 5, the $\kappa$ for $Ua$ ($\kappa_{Ua} = 0.530$, $\sigma_{Ua} = 0.008$) is larger than those for $R$ ($\kappa_{R} = 0.413$, $\sigma_{R} = 0.010$) and $Rnc$ ($\kappa_{Rnc} = 0.344$, $\sigma_{Rnc} = 0.008$). The standard error of weighted $\kappa$ is computed by the method introduced by Cohen [11]. The difference between $\kappa_{Ua}$ and $\kappa_{R}$ and the difference between $\kappa_{Ua}$ and $\kappa_{Rnc}$ 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 can help them do more judgements. This is corroborated by the large marginal distribution of $Ua$ shown in Figure 3: unlike the relevance distribution concentrated on $R = 3$, the distribution of $Ua$ is a more similar to $Ua$ 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 $Ua$ with the usefulness feedbacks $Ua$, 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 $\kappa$. 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 $Ua$ and $Ua$ are detected. The correlation between $Ua$ and $Ua$ is significantly stronger than that between $R$ and $Ua$. We also show the joint distribution of $Ua$ and $Ua$ in Figure 4(b). The diagonal blocks are the darkest block in almost every rows and columns, showing a fair agreement between $Ua$ and $Ua$. From the correlation metrics and the joint distribution we can see that although $Ua$ and $Ua$ 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 $QSA_{T_u}$. We use usefulness annotations $U_u$ to compute four click-sequence-based metrics defined in Section 5.1: cCG, dCCG, cMAX, and cCG/$\#clicks$, and correlate them with $QSA_{T_u}$. 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 $QSA_{T_u}(q_i) > QSA_{T_u}(q_j)$. For each query pair, if an evaluation metric also indicates the same relative preference, then we say the evaluation metric agrees with $QSA_{T_u}$ 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 $QSA_{T_u}$, measured in Pearson’s $r$ and the agreement ratios in the preference test, in Table 9. We compare the correlations related to $U_u$ 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 ($QSA_{T_u}$) as another baseline. The results show that, although usefulness annotations $U_u$ do not correlate with query-level satisfaction feedbacks $QSA_{T_u}$ 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 $U_u$ outperform their counterparts based on $R$, in terms of correlation with $QSA_{T_u}$. It is also interesting to observe that query-level satisfaction annotations from external assessors ($QSA_{T_u}$) are quite different from query-level satisfaction feedbacks from users ($QSA_{T_u}$), which is also observed by Liu et al. [32]. Some of click-sequence-based metrics based on $U_u$ are significantly better than satisfaction annotations ($QSA_{T_u}$), 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 $U_u$) 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).

#### 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 ($QSA_{T_u}$).

#### 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 example, $U_Q$ 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 $U_u$ 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 $U_{All}$ 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 $U_u$ 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 $U_u$, we can automatically generate document-level usefulness labels $U_{All}$ of at least equal validity to $U_u$, 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 $U_r$. "(or **) indicates the difference between $U_r$ and $R$ is significant at $p < 0.05$ ($p < 0.01$). The darker/lighter shadings indicates the difference between $U_{11}$ and $U_r$ is significant at $p < 0.05$/$0.01$.

<table>
<thead>
<tr>
<th></th>
<th>Pearson s $r$</th>
<th>MSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U_{11}$</td>
<td>0.398</td>
<td>1.198</td>
<td>0.894</td>
</tr>
<tr>
<td>$U_{Q}$</td>
<td>0.410</td>
<td>1.186</td>
<td>0.889</td>
</tr>
<tr>
<td>$U_{A}$</td>
<td>0.461</td>
<td>1.103</td>
<td>0.851</td>
</tr>
<tr>
<td>$U_{A+B}$</td>
<td>0.467</td>
<td>1.105</td>
<td>0.845</td>
</tr>
<tr>
<td>$U_{A+R}$</td>
<td>0.519</td>
<td>1.021</td>
<td>0.815</td>
</tr>
<tr>
<td>$U_{A+B+R}$</td>
<td>0.524</td>
<td>1.023</td>
<td>0.803</td>
</tr>
<tr>
<td>$U_{r}$</td>
<td>0.415</td>
<td>1.172</td>
<td>0.852</td>
</tr>
<tr>
<td>$R$</td>
<td>0.352</td>
<td>1.186</td>
<td>1.020</td>
</tr>
</tbody>
</table>

Table 12: Correlations with query-level satisfactions $QSAT_r$.
Measured in Pearson’s r ($df = 933$). "(or **) indicates the difference between $U_{11}$ and $U_r$ is significant at $p < 0.05$ ($p < 0.01$). The darker/lighter shadings indicate the difference between $U_{11}$ and $U_r$ is significant at $p < 0.05$ ($p < 0.01$). The darker/lighter shadings indicate the difference between $U_{11}$ and $U_r$ is significant at $p < 0.01$ and $0.05$.

<table>
<thead>
<tr>
<th></th>
<th>$QSAT_{11}$</th>
<th>$QSAT_{A+B}$</th>
<th>$QSAT_{A}$</th>
<th>$QSAT_{r}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$cCG$</td>
<td>0.459</td>
<td>0.490</td>
<td>0.466</td>
<td>0.572</td>
</tr>
<tr>
<td>$cDCG$</td>
<td>0.580</td>
<td>0.612</td>
<td>0.518</td>
<td>0.724</td>
</tr>
<tr>
<td>$cMAX$</td>
<td>0.601</td>
<td>0.635</td>
<td>0.580</td>
<td>0.751</td>
</tr>
<tr>
<td>$cCG$/#clicks</td>
<td>0.571</td>
<td>0.608</td>
<td>0.548</td>
<td>0.733</td>
</tr>
<tr>
<td>$QSAT_{r}$</td>
<td>0.508</td>
<td>0.529</td>
<td>0.539</td>
<td>0.563</td>
</tr>
</tbody>
</table>

Summary

In this section, we proposed two usefulness labeling methods: usefulness annotation and automatic usefulness prediction, and conducted analyses to demonstrate their reliability and 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 naturalistic 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, we can still use them to complement the former in different situations.
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