Understanding Ephemeral State of Relevance

Jiepu Jiang[†], Daqing He[‡], Diane Kelly[§], James Allan[†]
[†] Center for Intelligent Information Retrieval (CIIR), University of Massachusetts Amherst
[‡] School of Information Sciences, University of Pittsburgh
[§] School of Information Sciences, University of Tennessee, Knoxville
jpjiang@cs.umass.edu, dah44@pitt.edu, dianek@utk.edu, allan@cs.umass.edu

ABSTRACT

Despite its dynamic nature, relevance is often measured in a context-independent manner in information retrieval practice. We look into this discrepancy. We propose a contextual relevance/usefulness measurement called *ephemeral state of relevance* (ESR), which is defined as the amount of useful information a user acquired from a clicked result as assessed just after examining the result during an interactive search session. We collect ESR and context-independent usefulness judgments through a laboratory user study and compare the two. We examine factors related to both judgments and examine their differences.

Our study demonstrates a few advantages of ESR: it captures users' real-time state of mind and perceptions; it measures how much useful information the user is able to acquire from a result rather than how much there is in the result; it better reflects users' needs and criteria of useful results during a session, highlighting novelty as a salient factor. However, we also find that users may not be able to correctly assess the credibility of information during a session, which may reduce the reliability of the collected ESR judgments.

We evaluate ESR, context-independent usefulness judgments, and TREC-style topical relevance judgments by correlating with user experience in a session. The results demonstrate that switching the judgment criterion from topical relevance to usefulness is fruitful, but moving from context-independent judgments to contextual ones has only limited advantages with respect to its cost and complexity. Our study enriches current understanding on the dynamics of relevance in a search session and identifies both opportunities and challenges for collecting contextual relevance judgments.

Keywords

Relevance judgment; usefulness; user experience; web search.

1. INTRODUCTION

Relevance is a key notion in information retrieval. Most search systems are designed and trained to rank results by relevance, and almost all offline methods for evaluating search systems are based on relevance judgments. Previous studies had numerous discussions on the notion of relevance [5, 26, 27, 28] and its measurement [35, 36, 37]. Most of these studies acknowledged that relevance depends on not only topic aboutness but also many other factors such as novelty, understandability, reliability, search task, search context, users' interaction with the search results, and so on.

However, in practice, relevance is usually assessed by external annotators, without genuine search context, and primarily focusing on topical relevance. A typical example is the TREC ad hoc tracks [34] and web tracks [9], where external assessors judged a preassigned set of documents one after another, using criteria focusing on topical relevance.

In this paper, we propose a relevance judgment measurement called *ephemeral state of relevance* (ESR), which is defined as the perceived amount of useful information a user acquired from a clicked result just after the user finished examining the result during a search session. Particularly, ESR has the following characteristics:

- ESR is assessed by real searchers in real time under a real search context.
- ESR is assessed by the criterion of being useful to the problem at hand.
- ESR distinguishes between the relevance contained in a result and the relevance acquired by a user.

In order to obtain data for studying ESR, we conducted a laboratory user study including 28 participants' ESR judgments on 736 clicked search results in 112 search sessions. We answer the following questions in the rest of the article:

- *What contributes to ESR judgments?* We examine the influence of topicality, novelty, understandability, reliability, and effort factors on ESR judgments in Section 5, along with a wide range of other factors as controls, including user background, search task attributes, and user behavior signals.
- *How do ESR and context-independent usefulness judgments differ from each other?* To examine the influence of context on relevance/usefulness judgments, we also collected users' usefulness judgments in a context-independent setting after a search session. Section 5 also analyzes the influencing factors for the context-independent usefulness judgments and compares with those for ESR.

CHIIR '17, March 07-11, 2017, Oslo, Norway

DOI: http://dx.doi.org/10.1145/3020165.3020176

- *How well do ESR judgments correlate with users' search experience?* Section 7 evaluates ESR, static usefulness judgments, and TREC-style topical relevance judgments by correlating with user experience in a session.

2. RELATED WORK

Many articles reviewed the concept and dynamics of relevance [5, 27, 28]. Due to the limited space, we only summarize a few most important understandings of relevance that are related to our study: relevance is not only topic aboutness or relatedness; relevance can be subjective and needs to be assessed by users themselves; relevance and relevance judgments are influenced by context; relevance is time-dependent during a search process; "relevance is derived" and users need to interact with a search result to acquire relevant information. The de facto standard of relevance judgments in information retrieval practice, however, still largely focuses on topical relevance. A typical example is the criteria of the TREC web tracks 2009–2014 [9].

Recently, Belkin et al. [3, 4, 8] proposed to adopt usefulness as the criterion for evaluating interactive information retrieval systems. They [4, 8] proposed an evaluation model for interactive search systems based on three levels of usefulness regarding "the entire information seeking episode", "each interaction", and "system support", respectively. However, studies comparing usefulness and other search result judgment measures (such as topical relevance) are very limited. In a recent study, Mao et al. [24] measured searchers' judgments on the usefulness of search results and compared with topical relevance judgments by external annotators. However, the measured usefulness does not take into account real search context, and they had not excluded the influence of the difference between searchers and external annotators. The contextual judgment we measured in this paper (ESR) can be considered as a specific case of the usefulness of "each interaction" in Belkin et al.'s model, where the interaction is a visit to a search result. We compare the ESR judgments with a static usefulness judgments collected in a setting similar to Mao et al.'s study [24] to examine the influence of context on search result usefulness judgments.

We examine factors related to ESR judgments. The examined factors are based on previous studies of factors for relevance judgments. Xu et al. [35] examined five factors related to relevance judgments, including topicality, novelty, understandability, reliability, and scope. Except for scope, Xu et al. [35] found that the other four factors significantly contribute to relevance judgments. We suspect that usefulness judgments may also relate to the four factors. Many other factors [30, 31] may also relate to relevance judgments, but we do not examine them, such as to reduce the total number of judgments collected after a click activity.

We also examined the difference of users' perceptions on understandability and reliability during and after a search session as factors for the difference between contextual and contextual-independent usefulness judgments. A few previous studies also reported similar findings [29, 32], but they all examined very long search process (may last several months). In contrast, we examine the factors within a relatively short session (about 10 minutes).

3. EPHEMERAL STATE OF RELEVANCE

This paper studies a contextual relevance measurement called ephemeral state of relevance (ESR). We introduce the notion in this section and propose a few related hypotheses.

3.1 Definition

Without loss of generality, we define the *ephemeral state of relevance* (ESR) of an information object (such as a search result) as the amount of useful information a user would acquire from the object by interacting with it under a natural condition at a particular moment of a search process. More specifically, its working definition in this paper, for the purpose of measurement, is the perceived amount of useful information a user acquired from a clicked result right after the user finished examining the result during a search session.

We consider ESR as a "snapshot" of the *situational relevance* by Saracevic:

> "*Situational relevance or utility: Relation between the situation, task, or problem at hand, and information objects (retrieved or in the systems file, or even in existence). Usefulness in decision making, appropriateness of information in resolution of a problem, reduction of uncertainty, and the like are criteria by which situational relevance is inferred. This may be extended to involve general social and cultural factors as well.*" [28]

We also consider ESR as a particular implementation of Belkin et al.'s evaluation model [4, 8], where the second level measures the usefulness regarding "each interaction". In the case of ESR, the interaction being evaluated is a click on a search result.

ESR has the following characteristics:

- The criterion for assessing ESR is to be useful regarding the problem at hand. *Relevance* and *usefulness* are interchangeable in the context of ESR.

- ESR depends on the result, the user, the search task, and the time of a search process (a search session). Not only ESR but also the state of the user and the search task are time-dependent. Users may have both cognitive and affective changes during a search process [18], and the search task at hand may also evolve during a search session. Thus, ESR captures not only the dynamics of search result relevance, but also the state of the user and the search task in a search session.

- ESR measures the effectiveness of the interaction for acquiring relevant/useful information from a result. It is not an attribute of the result. For example, users do not necessarily read the whole result document in order to obtain all the information. They may skip reading a result if it costs too much effort to locate or understand the information they need. Thus, ESR measures not only how much useful information the result contains, but also how much the users are willing to acquire from the result in the particular search context.

- As its name denotes, ESR slips away soon and cannot be restored, because both the state of the user and the problem at hand are changing. ESR needs to be assessed by the user just-in-time in a search process.

We call ESR a *contextual* relevance/usefulness judgment measure because users assess ESR in real search context. ESR takes into account factors such as the status and background of the user, the search task, the time of a session, previously viewed results, the easiness to understand the content of the result, the effort spent on the result, and so on. In contrast, we call relevance/usefulness judgments that do not involve a search context *static* or *context-independent* relevance judgments. A typical example is the TREC approach for relevance judgments [9], where annotators are requested to judge a preassigned set of results one after another, without a real search context. ESR has many theoretical advantages, which may make it a more accurate measure than static relevance judgments in IR system design and evaluation. However, we also note that it requires a more complex setting (and possibly a higher cost) to collect ESR judgments. Thus, it requires a comprehensive analysis regarding both the pros and cons of ESR judgments. The purposes of this study are:

- to understand factors contributing to ESR and its differences to context-independent judgments;
- to evaluate ESR as a relevance judgment measure and to examine its advantages and limitations.

3.2 Hypotheses

After reviewing previous studies, we come to a few initial hypotheses regarding ESR:

- **H1** – ESR relates to users' perceptions on the topicality, novelty, understandability, and reliability of the search result at the time of assessing ESR. H1 is based on Xu et al.'s [35] and Zhang et al.'s studies [37].
- **H2** – ESR relates to the effort spent on the result, since it measures the amount of acquired useful information.
- **H3** – ESR relates to the user and the search task. H3 is based on previous studies that show user background and and search task influence user behavior patterns in a session [7, 15, 20, 22, 23]. We suspect ESR is also related to user background and search task.
- **H4** – ESR differs from context-independent usefulness judgments. We believe context influences the state of the user and the search task in a session and consequently makes contextual and context-independent usefulness judgments different.

We designed and conducted a user study (§ 4) to collect data to examine these hypotheses.

4. USER STUDY

We conducted a laboratory user study to collect the data. The participants worked on preassigned tasks in an experimental search system. The tasks range from locating facts to exploratory ones. For each task, the participant needed to perform a 10-minute interactive search session to fulfill the goal of the task. We recorded participants' search behaviors and collected their judgments on the clicked results and their search experience in a session.

4.1 Experiment Design

Each participant completed four tasks of different types. The tasks were developed by the TREC session tracks [6]. They were categorized into four types by the targeted task product and goal based on Li and Belkin's faceted classification scheme [19]. The targeted product of a task is either *factual* (to locate facts) or *intellectual* (to enhance the user's understanding of a problem or topic). The goal of a task is either *specific* (well-defined and fully developed) or *amorphous* (an ill-defined or unclear goal that may evolve along with the user's exploration).

We divided the participants into groups of four. Participants in the same group finished the same four tasks (one task for each type), but with a different sequence (rotated using a Latin square). We assigned different tasks to each group, which was to increase task diversity and reduce the concern regarding task parity in experiment design [17].

For each task, the participants went through two stages:

- **Search stage (10 minutes)**. The participants needed to perform an interactive search session for 10 minutes to fulfill the task goal. They could issue and reformulate any queries and click on any results. After clicking on a result's link, the participants switched to the result webpage in a new browser tab. When the participants had finished examining the result and turned back to the SERP, they needed to complete some judgments on the clicked results (called **post-click judgments**) before they could resume the search session.
- **Judgment stage (about 10 minutes)**. The participants rated their search experience in the session and finished additional judgments on each clicked result (called **post-session judgments**). We will introduce details of the judgments in Section 4.2.

The interface of the experimental search system is similar to popular search engines. It redirected users' queries to Google and returned filtered Google results. The system only showed ordinary "10-blue links", vertical results (except image verticals), and related queries. We removed other SERP elements such as ads, direct answers, and entity information to simplify the user study. The system displayed search results in the same way they would appear on Google. The main difference between the experimental search system and Google in SERP design was that our system showed task description on the top of a SERP. We made this change to help participants recall the task requirements.

In total, the experiment took a participant about 100 minutes to complete. First, we required the participants to work on a training task for 10 minutes. Then, the participants worked on four formal tasks (about 20 minutes for each task). We also required the participants to take a 5-minute break after they finished two formal tasks.

4.2 User Judgments

We collected users' judgments on the clicked results and the search sessions to verify the hypotheses in Section 3.2. Table 1 shows the questions for collecting users' judgments.

We collected three relevance/usefulness judgments:

- **ESR** is a contextual usefulness judgment assessed during a search session just after a user finished examining

Table 1: Questions for collecting post-click and post-session search result judgments and users' experience in a session.

Post-click Judgments

Ephemeral State of Relevance (**ESR**)	How much useful information did you get from this webpage? From 1 (none) to 7 (a lot of).
Novelty (**Nov.**)	How much new information did you get from this webpage? From 1 (none) to 7 (a lot of)
Effort (**Effort**)	How much effort did you spend on this webpage? From 1 (none) to 7 (a lot of).
Understandability (**Under.**)	How difficult is it for you to follow the content of this webpage? From 1 (very difficult) to 7 (very easy).
Reliability (**Relia.**)	How trustworthy is the information in this webpage? From 1 (not at all trustworthy) to 7 (very trustworthy).

Post-session Judgments

Topical Relevance (**TRel.**)	How relevant is this webpage? • *Key* (3): this page or site is dedicated to the topic; authoritative and comprehensive; it is worthy of being a top result in a web search engine. • *Highly Relevant* (2): the content of this page provides substantial information on the topic. • *Relevant* (1): the content of this page provides some information on the topic, which may be minimal. • *Not Relevant* or *Spam* (0).
Usefulness (**Usef.**)	How much useful information does this webpage provide for the task? From 1 (none) to 7 (a lot of).
Understandability (**Under.ps**)	How difficult is it for you to follow the content of this webpage? From 1 (very difficult) to 7 (very easy).
Reliability (**Relia.ps**)	How trustworthy is the information in this webpage? From 1 (not at all trustworthy) to 7 (very trustworthy).

Search Experience Measures

Satisfaction	How satisfied was your search experience? From 1 (very unsatisfied) to 7 (very satisfied).
Frustration	How frustrated were you with this task? From 1 (not frustrated) to 7 (very frustrated).
System Helpfulness	How well did the system help you in this task? From 1 (very badly) to 7 (very well).
Goal Success	How well did you fulfill the goal of this task? From 1 (very badly) to 7 (very well).
Session Effort	How much effort did this task take? From 1 (minimum) to 7 (a lot of).
Task Difficulty	How diffcult was this task? From 1 (very easy) to 7 (very difficult).

a clicked result. The wording of the question emphasizes the *acquired* amount of useful information (... *did you get* ...).

- Usefulness (**Usef.**) is a context-independent usefulness judgment collected after a search session terminated. The wording of the question is similar to that for ESR, but puts an emphasis on the amount of useful information the result contains (... *does this webpage provide for the task* ...).

- Topical relevance (**TRel.**) is a context-independent relevance judgment using the same criteria of the TREC web tracks [9], which uses topic aboutness as the main criterion. Table 1 shows the question and options for TRel. judgments.

In the search stage, we instructed the participants to examine results as they would normally do when using a search engine in their daily lives. For example, they did not need to fully read the content and they could abandon examining a result. Particularly, they were instructed that during post-click judgments, they should not revisit the results for the purpose of answering the judgment questions. This is to make sure that the post-click judgments only measure the utility of the latest search result examination activity.

In the judgment stage, we asked the participants to read the results in a better detail to finish the post-session judgments. The system also required the participants to revisit each clicked result and spend at least 30 seconds before they could submit their post-session judgments.

In addition to ESR, the post-click judgments also include users' perceptions on the novelty (**Nov.**), understandability (**Under.**), and reliability (**Relia.**) of the results, and their **effort** spent on the clicked results. In the post-session judgments, we collected understandability (**Under.ps**) and reliability (**Relia.ps**) judgments again to examine changes in users' perceptions. We did not collect novelty judgments again because participants of a pilot study reported confusions on the criteria of assessing novelty in the post-session judgments. Except TRel., the participants answered questions using a 7-point Likert scale.

In the judgment stage, participants also rated their search experience in the session using a 7-point Likert scale. We collected six search experience measures, including satisfaction [14], frustration [10], system helpfulness, goal success [13], session effort, and task difficulty [21].

4.3 Collected Data

We recruited 28 participants through flyers posted on the campuses of two universities in the United States. We required the participants to be English native speakers to exclude the influence of language fluency on relevance judgments [12]. All participants were college or graduate students studying different majors. 16 of them are female. They were reimbursed $15 per hour.

We collected 112 sessions by 28 participants on 28 tasks. Each participant worked on four tasks and each task was performed by four participants. The collected dataset includes judgments on 736 clicked results (6.6 results per session).

5. WHAT AFFECTS ESR JUDGMENTS?

This section studies factors related to the ESR judgments using multilevel regression analysis (§ 5.1). We also examine factors related to the static usefulness judgments (Usef.) and compare with those for ESR.

5.1 Multilevel Regression

We examine two regression models M1 and M2, where the dependent variables (DVs) are ESR and Usef. judgments, respectively. M1 and M2 include the same set of independent variables (IVs):

- Our main purpose is to examine the impacts of other judgments in Table 1 on ESR and Usef. judgments.

- User background information (collected from an entry survey): including gender (*Male* or *Female*), age (four levels; 0 for *18–24*, 1 for *25–30*, 2 for *31–40*, and 3 for

Table 2: Pearson's correlation matrix of variables.

	ESR	Novelty	Effort	Under.	Relia.	TRel.	Usef.	Under.ps
Novelty	0.70							
Effort	0.25	0.27						
Understandability	0.26	0.20	−0.36					
Reliability	0.47	0.43	0.11	0.28				
Topical Relevance	0.65	0.49	0.18	0.19	0.45			
Usefulness (post-session)	0.75	0.56	0.18	0.24	0.47	0.83		
Understandability (post-session)	0.27	0.23	−0.30	0.72	0.28	0.25	0.31	
Reliability (post-session)	0.45	0.42	0.09	0.23	0.82	0.51	0.54	0.30

Light , dark , and darker shadings indicate the correlation is significant at 0.05, 0.01, and 0.001 levels.

Table 3: Multilevel (hierarchical) regression analysis – ESR judgments as dependent variable (M1) and post-session usefulness judgments (Usef.) as dependent variable (M2). Independent variables without † are controls.

	Independent Variable	Standardized Coefficients Beta M1 (ESR)	M2 (Usef.)
	(Constant)	-	-
	Gender: *Male*	−0.05	−0.05
	Age	−0.02	0.01
	Highest degree: *Graduate*	0.03	0.03
	Search engine expertise	0.05	0.03
	Product: *Factual*	0.09	0.02
	Goal: *Specific*	0.06	0.03
	Topic familiarity	0.06	−0.00
	Time spent in the session	0.06	0.01
	Number of past queries	−0.03	−0.05
	Number of past clicks	0.01	0.03
	Dwell time (log)	0.08	0.07
†	Topical Relevance	0.33	0.67
†	Novelty	0.48	0.16
†	Understandability	0.10	0.02
†	Reliability	0.09	−0.04
†	Effort	0.07	0.02
†	Understandability (post-session)	0.01	0.06
†	Reliability (post-session)	−0.04	0.11
	Adjusted R^2	**0.639**	**0.741**

Light , dark , and darker shadings indicate the coefficient is significant at 0.05, 0.01, and 0.001 levels, respectively.

Over 40), highest degree obtained or expected (*Undergraduate* or *Graduate*), and the expertise of using web search engines rated using a Likert scale from 1 (*very badly*) to 5 (*very well*).

- Task attributes (assigned or collected before each session): product (*Factual* or *Intellectual*), goal (*Specific* or *Amorphous*), and user's familiarity with the topic of the task rated using a Likert scale from 1 (*very unfamiliar*) to 7 (*very familiar*).
- Search behavior related to a click: the total time spent in the current session while clicking, the number of past queries and clicks in the session, and the dwell time on the clicked result (in logarithm).

We conduct multilevel regression analysis [11, 16] because the collected observations are nested—each session can have many clicks, and each user performed four sessions. In such a case, some observations share the same contexts at the session or user levels, violating the independence assumption for regular regression analysis. Multilevel models address these issues. We construct models with three levels: Level 1 includes other judgments and search behavior, Level 2 includes task attributes, and Level 3 includes user background information. We perform analysis using SPSS 24.

We examine multicollinearity between variables using variance inflation factor (VIF). The IVs of both models satisfy VIF < 4. The VIF values are below the commonly suggested threshold (4–10) for concerns on multicollinearity issues [25].

Table 3 reports the standardized coefficients (β) of IVs in M1 and M2. Positive and negative β values indicate positive and negative relationships, respectively, between the IV and the DV. The absolute value of the standardized coefficient β is often interpreted as the impact of the independent variable on the variance of the dependent variable (normalized by the standard deviation of variables)—the magnitude of change in the DV (relative to its standard deviation) caused by one-unit change in the IV (relative to the IV's standard deviation) while other variables being equal. We also report the bivariate correlation (Pearson's r) of the collected search results judgments in Table 2 for reference.

5.2 Topical Relevance and Novelty

Topical relevance and novelty are the two most important factors for both ESR and context-independent usefulness judgments (Usef.). However, novelty shows a much greater impact on ESR judgments compared with its effect on the collected Usef. judgments.

Novelty shows a significant positive effect on ESR judgments in model M1 ($\beta = 0.48$, $p < 0.001$). According to the standardized coefficients, novelty has the most salient impact on ESR judgments among the examined independent variables. Topical relevance also has a significant positive effect on ESR judgments ($\beta = 0.33$, $p < 0.001$), which is the second strongest in model M1. In contrast, model M2 shows that topical relevance is the most significant factor affecting context-independent usefulness judgments (Usef.) ($\beta = 0.67$, $p < 0.001$). Although the second most important variable in M2, novelty exhibits a much weaker impact on the Usef. judgments ($\beta = 0.16$, $p < 0.001$) compared with its influence on ESR judgments.

While confirming the central place of topical relevance and novelty in both contextual and context-independent usefulness judgments (ESR and Usef.), results also disclose the different contribution of the two factors in the two usefulness judgment settings. Compared with its impact on context-independent usefulness judgments (Usef.), novelty plays a more important role in determining the actual perceived amount of useful information acquired from a clicked result assessed in a contextual setting (ESR). This suggests

that conventional context-independent relevance/usefulness judgments [9, 24, 34] may have the risk to overlook the novelty of search results.

However, we note the reported results are based on judgments collected for the clicked results from the top-ranked entries returned by a search engine (Google). Therefore, we are cautious regarding the seemingly most important role of novelty in determining ESR judgments. The top-ranked results from a search engine are mostly topically relevant, which may reduce the importance of topical relevance among the retrieved entries. In addition, users' click decisions and bias may also influence the representativeness of the clicked results, making it difficult to assess the importance of the factors sorely based on data collected from our study.

5.3 Understandability

We collected users' understandability judgments twice in the experiment (post-click and post-session). The two judgments have a strong correlation ($r = 0.72$, $p < 0.001$), but they also have differences in 38% of the results. The mean absolute difference of the two ratings is 0.63 (in a 7-point scale). This indicates that users' perceptions on the understandability of a result indeed undergo some changes during a 10-minute search session.

Model M1 shows that ESR only relates to the contextual (post-click) understandability judgments but not the context-independent (post-session) ones. Although ESR has a weak positive linear correlation with both understandability judgments ($r = 0.26$ and 0.27, respectively), only the post-click judgments show a significant positive effect in M1 ($\beta = 0.10$, $p < 0.01$). The post-session judgments do not show any significant effect in M1 at 0.05 level, suggesting that it provides little value in addition to the post-click judgments for explaining the variance of the ESR judgments. Similarly, model M2 shows that Usef. judgments only relates to the context-independent understandability judgments but not the contextual ones.

The relationship between ESR and the two understandability judgments discloses a potential advantage of ESR—it takes into account a user's ability to understand at a particular time of a search session. As a user's understanding varies over time, the user may prefer results with different understandability levels at different stages of a session, e.g., a user may expect to read easy-to-understand introductory texts such as a Wikipedia entry at the beginning of a session. Collecting ESR judgments potentially makes it possible to account for such issues in system design and evaluation.

5.4 Reliability

We also collected users' reliability judgments twice. The two reliability judgments have a strong correlation ($r = 0.82$, $p < 0.001$), but they are also different in 43% of the clicked results. The mean absolute difference of the two reliability judgments is 0.60 (in a 7-point scale). This suggests that users' perceptions on the reliability of a search result also undergo some changes in a 10-minute search session.

Similar to the findings regarding understandability judgments, ESR also only relates to post-click reliability judgments but not the post-session ones. Although ESR shows a moderate correlation with both reliability judgments ($r =$ 0.47 and 0.45, respectively), only the post-click reliability judgments have a significant positive effect on users' ESR judgments ($\beta = 0.09$, $p < 0.05$).

ESR seems to account for users' perceptions on the reliability of search results at the time they examined the results. However, we believe this brings in a risk of performing ESR judgments. Unlike the subjective nature of understandability, the reliability of a result is a rather objective existence. It is reasonable to believe that after a search session's exploration, searchers may have acquired more knowledge to assess the reliability of results with better accuracy in post-session judgments. The post-click reliability judgments may be less accurate than the post-session ones, since a user may fail to accurately assess the reliability of a result during a search session due to the limited knowledge on the task. As a significant factor for ESR judgments, the possibly defective contextual (post-click) reliability judgments may consequently reduce the quality of the ESR judgments as well.

5.5 Effort

ESR judgments also relate to the effort spent on the results in a positive way, partly confirming that ESR captures users' interaction for acquiring useful information from the results. Effort exhibits a significant positive effect on ESR ($\beta = 0.07$, $p < 0.05$). Also, there is a weak positive linear correlation between ESR and effort ($r = 0.25$, $p < 0.001$).

However, despite its statistically significant effect in M1, effort only seems to have a small practical contribution for explaining the variance of the ESR judgments compared with other variables. We suspect a possible reason is that the connection between ESR and effort is more complex than simply a linear relationship. Section 6 analyzes this issue in detail.

It is also worth mention that the significant effect of effort is observed with many other variables as controls. Previous studies often connect effort with dwell time and understandability [2, 33]. We indeed observed some correlation of effort with both dwell time ($r = 0.33$, $p < 0.001$) and understandability judgments ($r = -0.36$, $p < 0.001$). Despite these connections, both effort judgments and the other two variables show significant effects on ESR judgments, suggesting a non-replacable value of effort judgments in addition to time and understandability for explaining ESR judgments. This also suggests that the effort spent on a clicked result relates to factors other than dwell time and understandability.

5.6 Control Variables

We included a wide variety of variables as controls in the regression models. Results suggest that in addition to the collected search result judgments, ESR also significantly relates to gender, the expertise of using search engines, task attributes, and the dwell time on the clicked result.

Among the examined user background variables, gender and the expertise of using web search engines show significant effects on the ESR judgments, while age and the highest obtained or expected degree do not. Results suggest that male participants rated the ESR of the clicked results lower than female participants ($\beta = -0.05$, $p < 0.05$). The expertise of using search engines has a positive effect on the ESR of the clicked results ($\beta = 0.05$, $p < 0.05$), probably because it is easier for experienced searchers to find useful results.

All the three examined task attributes show significant effects on the ESR judgments. Searching in a session targeting a *factual* product (compared with one for an *intellectual* product) has a significant positive effect on ESR judgments

($\beta = 0.09$, $p < 0.001$). Searching in a session with a *specific* goal (compared with an *amorphous* one) also shows a significant positive effect on ESR ($\beta = 0.06$, $p < 0.01$). In addition, users' familiarity with the task topic also has a significant positive effect on ESR judgments ($\beta = 0.06$, $p < 0.01$). However, it is difficult to interpret why task attributes influence ESR judgments sorely based on regression analysis. A possible explanation is that task attributes are linked to search task difficulty and performance, such that users can more effectively find useful results in certain sessions (such as sessions dealing with simple tasks). However, it may also come from factors such as users' different criteria of assessing usefulness in different types of sessions and so on. We believe it requires further investigation to fully understand these issues.

5.7 ESR vs. Usef. Judgments

Model M2 examines factors for the context-independent usefulness (Usef.) judgments. In contrast to ESR, Usef. relates to a substantially different set of variables and by different strengths. A comparison between M1 and M2 discloses many differences between the contextual and context-independent usefulness judgments (ESR and Usef.):

- Usef. judgments relate to topical relevance by a greater extent than ESR judgments do (as discussed in § 5.2).
- ESR captures users' real time perceptions on the understandability and reliability of the results when they examined the results in a search session, while Usef. only significantly relates to those after a search session while users performed the Usef. judgments.
- ESR is significantly influenced by the attributes of the search task while Usef. judgments are not.
- Usef. judgments do not account for the actual effort spent in a session for acquiring relevant/useful information from the result but ESR judgments do.
- ESR is significantly related to the expertise of using search engine while Usef. judgments are not.

To sum up, the differences of M1 and M2 confirm that the ESR judgments collected in a contextual setting capture more contextual factors than the Usef. judgments collected in a context-independent setting. This suggests that by collecting context-independent relevance/usefulness judgments, we may fail to capture factors such as search task, users' actual interaction with the result, users' cognitive state in a session, and so on.

5.8 Summary

To summarize, the regression analysis performed in this section confirms our hypotheses H1–H4.

- **H1** – Similar to previous studies on factors for relevance judgments [35, 37], we found that ESR judgments also significantly relate to topicality, novelty, understandability, and reliability. Particularly, ESR depends on users' real time perceptions on the understandability and reliability of the results at the time of examining the results, confirming that ESR indeed captures users' changing state of mind in a search session.
- **H2** – ESR judgments are significantly affected by the effort spent on the results, indicating that ESR depends on users' actual interaction with the results for acquiring relevant/useful information.
- **H3** – ESR judgments significantly relate to two user background variables and three search task attributes, suggesting that we may expect certain variation of ESR judgments from different users and in different types of tasks.
- **H4** – Results confirm that ESR and Usef. judgments are different in many aspects. Particularly, Usef. only significantly relates to topicality, novelty, users' perceptions on understandability and reliability in post-session judgments, and gender.

6. ESR, EFFORT, AND INTERACTION

Section 5 examined the relationship between ESR and other variables using linear models. However, a deeper analysis shows that the regression models concealed complex and non-linear relationships of the variables.

This section examines the ESR of and the effort spent on the results with different understandability and reliability levels (based on post-click judgments). We group results into five levels to make the sample size of each group as close as possible (although group size still varies a lot due to the skewed distribution of users' judgments). The five understandability levels are 1–2 ($N = 45$), 3–4 ($N = 69$), 5 ($N = 84$), 6 ($N = 137$), and 7 ($N = 401$). The five reliability levels are 1–3 ($N = 117$), 4 ($N = 112$), 5 ($N = 152$), 6 ($N = 154$), and 7 ($N = 201$). Figure 1 plots the results.

ESR, Effort, and Understandability – Users acquired a lot of useful information (mean ESR = 4.87) with only a small amount of effort (mean effort = 1.93) from the results with the highest level of understandability (7). While encountering results that are more difficult to understand (Under. = 6 and 5), users spent significantly greater effort (mean effort = 3.01 and 3.49), and they were still able to acquire a similar amount of useful information (mean ESR = 5.04 and 4.92). When the results are even more difficult to understand (Under. =3–4), the trend of spending more effort stopped (mean effort = 3.54), and the acquired amount of useful information also declined significantly (mean ESR = 3.91). When the results are extremely difficult to understand (Under =1–2), users started to abandon examining results, spending fewer effort (mean effort = 3.16) and acquiring very limited amount of useful information (mean ESR = 2.67).

ESR, Effort, and Reliability – We also observed a similar pattern on results with different reliability levels. Users acquired a lot of useful information (mean ESR = 5.36 and 5.45) with a small amount of effort (mean effort = 2.57 and 2.42) from the results with the two highest reliability levels (Relia. =7 and 6). When the results provide less reliable information (Relia. = 5 and 4), users spent significantly greater effort (mean effort = 2.86 and 2.86), but started to acquire a significantly fewer amount of useful information (mean ESR = 5.01 and 4.04). They abandoned examining results when the reliability level is very low (1–3), spending the least amount of effort (mean effort = 1.90) and acquiring very limited useful information (mean ESR = 2.69).

The empirical findings in Figure 1 discloses that the process of examining search results and acquiring useful infor-

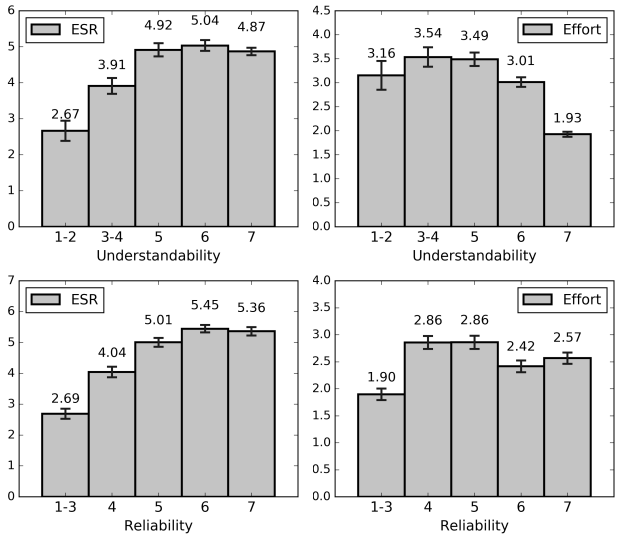

Figure 1: Mean ESR and effort (the error bars are standard error) for results with different novelty, understandability, and reliability levels (in post-examination judgments).

mation involves complex interaction and decisions. We propose a hypothesis for this process as in Figure 2:

- The acquired amount of useful information and the effort spent on a result depend on not only how much useful information the result contains but also the efficiency of acquiring useful information from the result, which further depends on both the result itself and the user.
- When the efficiency of acquiring useful information (x-axis) declines, users will first try to spend more effort (red lines) to compensate the limited efficiency, such that the amount of useful information acquired from the result (the blue line) still maintains at a relatively high level. However, when the efficiency is very low, users will abandon examining the result to avoid wasting effort or due to too much effort spent.
- The efficiency of acquiring useful information from a result correlates with factors such as the understandability and reliability of the results. For example, users need to spend more effort if the result is difficult to understand. They also need to spend more effort on the less reliable results such as to assess the credibility of information.

The empirical observations in Figure 1 fit well with the hypothesis. Here we do not have the resource to fully verify this hypothesis, but we believe it offers a new understanding to users' interaction with the search results. Nevertheless, results in Figure 1 demonstrate that the process of acquiring useful information from the search results involve complex interactions, which is also an expected advantage of ESR over static relevance/usefulness judgments.

7. ESR AND USER EXPERIENCE

The main purpose of performing relevance judgment is to collect ground truth data to optimize and evaluate search

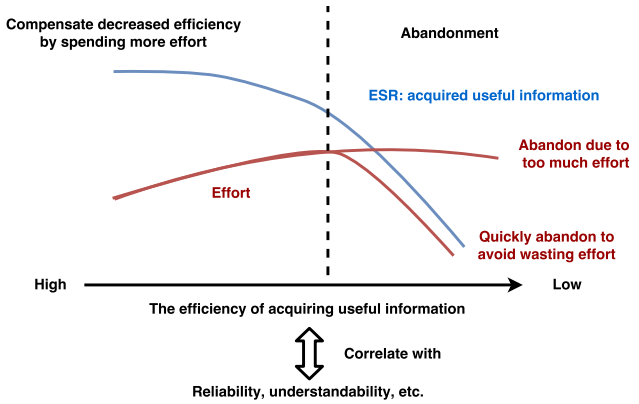

Figure 2: A hypothetical relationship between the amount of acquired useful information (ESR), the effort spent on the result, and the efficiency of acquring useful information.

Table 4: Pearson's correlation between user experience measures and the mean value of the clicked results' judgments.

User Experience	mean ESR	mean Usef.	mean TRel.
Satisfaction	**0.50**	0.48	0.44
Frustration	−0.37	**−0.41**	−0.30
System Helpfulness	**0.40**	0.38	0.31
Goal Success	**0.51**	0.49	0.37
Session Effort	**−0.42**	**−0.42**	−0.33
Task Difficulty	**−0.46**	−0.43	−0.36

All correlations are significant at least at 0.01 level.

systems. A good measure for relevance judgment should be able to identify high-quality results, such that presenting the results to users leads to a satisfactory search experience.

This section compares ESR, Usef., and TRel. judgments for their abilities to correlate with users' search experience. We assume the quality of the clicked results in a session is a factor for the user's experience in that session. For each session, we use the mean ESR, Usef., and TRel. of the clicked results as indicators for that session's search experience. We correlate the mean values of the judgments with users' perceptions on six search experience measures in the collected 112 search sessions. Table 4 reports the results.

Although ESR has many theoretical advantages, Table 4 shows that the collected ESR and Usef. judgments have only slight differences in terms of correlating with the six user experience measures. Mean ESR of results has slightly stronger correlations with satisfaction, system helpfulness, goal success, and task difficulty, while mean Usef. has a slightly stronger correlation with frustration. The differences in correlation values do not exceed 0.04, suggesting that whether to offer high ESR results or high Usef. ones may not differ much in terms of correlating with user experience in a session. However, we also note that our results are based on short sessions (10 minutes). We expect that in longer search processes, ESR and Usef. judgments may have greater differences.

The limited practical advantage of ESR compared with Usef. judgments in terms of correlating with user experience is not unexplainable. First, the collected ESR and Usef. judgments do not vary greatly ($r = 0.75$). Second, as we discussed in Section 5.4, users may not have enough knowledge to correctly assess the credibility of information

during a search session, which may consequently reduce the quality of the collected ESR judgments.

Considering that it requires a more complex setting (and probably a higher cost) to collect ESR judgments, collecting context-independent usefulness judgments seems a more practical choice. Table 4 shows a clear difference between TRel. and Usef. in terms of correlating with the six user experience measures (0.04–0.14). This suggests that, even in a context-independent setting, using usefulness as the criterion for assessing search results better correlate with users' search experience in a session.

8. DISCUSSION AND CONCLUSIONS

A key challenge of information retrieval is to determine which result is relevant and what accounts for a relevant result. We examined the long-lasting discrepancy between theoretical discussions of relevance and the actual measurement of relevance in IR practice. We proposed a contextual relevance measurement called *ephemeral state of relevance* (ESR), which is tightly connected to Saracevic's situational relevance [27, 28] and Belkin et al.'s evaluation model [3, 8]. We designed an experiment to collect ESR judgments. We compared this contextual judgments with context-independent ones, examined factors related to both, and looked into their differences. Our work makes the following contribution:

First, we successfully designed and collected contextual relevance judgments, confirming that it is possible to measure search result relevance/usefulness in a contextual and interactive manner. Numerous results verified that the collected ESR judgments indeed have the characteristics we expected, suggesting that our measurement of ESR is successful for its purpose. The experiment design also sheds lights on further studies with a similar purpose.

Second, through a thorough analysis of the related factors, our study offers new understandings on both contextual (ESR) and context-independent relevance/usefulness judgments. We note a few findings of particular importance:

- Similar to static relevance judgments, ESR judgments are also influenced by the four relevance judgment factors identified in previous studies (topicality, novelty, understandability, and reliability).
- ESR and static usefulness judgments weigh the four factors differently—ESR puts a higher weight on novelty, but Usef. puts a higher one on topicality. This shows different needs of searchers on search result at different periods and for different types of judgments.
- The state of mind of a user is changing over time, and such changes affect relevance judgments over time as well. Users' perceptions on understandability and reliability during and after a session are different, but both ESR and static usefulness judgments are only influenced by users' perceptions at the moment they performed the judgments.
- The acquired amount of useful information is affected by the effort spent on the results—it depends on not only how much the result contains, but also how much cost the user paid.

Third, this is the first study exploring the relationship between the acquired useful information, the effort spent, and the efficiency of acquiring useful information from the results. While encountering results from which it is difficult to acquire useful information (such as those with limited understandability and reliability), users would first spend more effort to compensate for the limited efficiency of acquiring useful information; but the users would also abandon examining the results if the efficiency declines below a threshold, acquiring limited useful information but costing only a small amount of effort as well. These observations complement understandings on the dynamics of user interaction at a search result level, while many previous studies stay at the session and SERP levels [1].

Last, we offer practical suggestions on the choice of relevance judgment measures by correlating different relevance judgments with user experience in a session—we show that switching the judgment criterion from topical relevance to usefulness is fruitful, while moving from context-independent judgments to contextual ones seems to have only limited improvements (in terms of correlating with user experience). Given what we now know, it seems more practical to simply collect usefulness judgments in a context-independent manner due to its reasonable correlation with user experience and low complexity in collection.

To conclude, our study uncovers the advantages and limitations for both contextual (ESR) and context-independent (static) relevance/usefulness judgments. The results demonstrate several advantages of ESR judgments compared with context-independent relevance/usefulness judgments:

- ESR judgments are closer to users' needs on the search results in a search session. As our results show, users' criteria for ESR judgments are different from those for static judgments in that ESR puts a higher weight on novelty. Using ESR judgments for system design and evaluation can better capture such needs of users.
- ESR captures the user's real-time state of mind when they examine the search results, which are subject to change during a search session. Using ESR judgments for IR system development can better capture such changes.
- ESR measures not only how much useful information a result contains but also how much the user is able or willing to acquire from the result. As we showed, users do not acquire all information from results. IR systems using ESR judgments can more accurately assess the influence of the results to the searchers.

Despite ESR's having several interesting characteristics, the results in this paper suggest that a few practical issues need to be solved to make ESR a practical and useful measurement. Results also shed light on a few important areas of applications in the future:

- Collecting ESR judgments requires a more complex experimental setting (and very likely also a higher cost) than that for collecting static judgments. Thus, an important area of application in the future is to develop prediction models for ESR judgments based on implicit feedback signals (such as dwell time) and static relevance/usefulness judgments. Such techniques may reduce the high complexity of collecting ESR judgments.
- The procedure of collecting ESR judgments brings in a selection bias—we can only collect judgments for

the clicked results, because ESR needs to be judged in a natural setting to preserve a genuine context. In contrast, static judgments fit well with the standard test collection development procedures such as pooling. The ESR prediction models discussed above may also solve the selection bias issues—as long as we can predict ESR reasonably well without click-related features (such as dwell time), we can apply the model to predict ESR judgments for the unclicked results.

- Capturing users' real-time states of mind in judgments is not always ideal, because searchers may not have the ability to correctly assess the results. As the study shows, users may overestimate or underestimate the reliability of a search result. Thus, we believe it will also be helpful to develop techniques to rectify ESR judgments by setting off users' inaccurate perceptions of results during a session.

Of course, our study also has a few limitations. First, our analysis is only based on data collected from a laboratory user study, which may not entirely represent real search scenarios. It is also worth noting that the adopted tasks in our study are more complex than regular web search information needs (such as navigational search). Second, we restrict our user study to relatively short search sessions (about 10 minutes), while real-world search process can be much longer and more complex (such as involving multiple sessions). We expect a greater difference between ESR and static judgments in such cases. We leave these issues for future work.

9. REFERENCES

[1] L. Azzopardi. Modelling interaction with economic models of search. In *SIGIR '14*, pages 3–12, 2014.

[2] L. Azzopardi and G. Zuccon. An analysis of the cost and benefit of search interactions. In *ICTIR '16*, pages 59–68, 2016.

[3] N. J. Belkin. Salton award lecture: People, interacting with information. In *SIGIR '15*, pages 1–2, 2015.

[4] N. J. Belkin, M. J. Cole, and J. Liu. A model for evaluation of interactive information retrieval. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, 2009.

[5] P. Borlund. The concept of relevance in IR. *J. Am. Soc. Inf. Sci. Technol.*, 54(10):913–925, 2003.

[6] B. Carterette, P. Clough, M. Hall, E. Kanoulas, and M. Sanderson. Evaluating retrieval over sessions: The TREC session track 2011-2014. In *SIGIR '16*, pages 685–688, 2016.

[7] M. J. Cole, J. Gwizdka, C. Liu, R. Bierig, N. J. Belkin, and X. Zhang. Task and user effects on reading patterns in information search. *Interact. Comput.*, 23(4):346–362, 2011.

[8] M. J. Cole, J. Liu, N. J. Belkin, R. Bierig, J. Gwizdka, C. Liu, J. Zhang, and X. Zhang. Usefulness as the criterion for evaluation of interactive information retrieval. In *HCIR '09*, 2009.

[9] K. Collins-Thompson, C. Macdonald, P. Bennett, F. Diaz, and E. Voorhees. TREC 2014 web track overview. In *TREC 2014*.

[10] H. A. Feild, J. Allan, and R. Jones. Predicting searcher frustration. In *SIGIR '10*, pages 34–41, 2010.

[11] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University, 2006.

[12] P. Hansen and J. Karlgren. Effects of foreign language and task scenario on relevance assessment. *J. Doc.*, 61(5):623–639, 2005.

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

[14] J. Jiang, A. Hassan Awadallah, X. Shi, and R. W. White. Understanding and predicting graded search satisfaction. In *WSDM '15*, pages 57–66, 2015.

[15] J. Jiang, D. He, and J. Allan. Searching, browsing, and clicking in a search session: Changes in user behavior by task and over time. In *SIGIR '14*, pages 607–616, 2014.

[16] J. Jiang and C. Ni. What affects word changes in query reformulation during a task-based search session? In *CHIIR '16*, pages 111–120, 2016.

[17] D. Kelly, J. Arguello, A. Edwards, and W.-c. Wu. Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework. In *ICTIR '15*, pages 101–110, 2015.

[18] C. C. Kuhlthau. Inside the search process: Information seeking from the user's perspective. *J. Am. Soc. Inf. Sci.*, 42(5):361–371, 1991.

[19] Y. Li and N. J. Belkin. A faceted approach to conceptualizing tasks in information seeking. *Inf. Process. Manage.*, 44(6):1822–1837, 2008.

[20] C. Liu, N. J. Belkin, and M. J. Cole. Personalization of search results using interaction behaviors in search sessions. In *SIGIR '12*, pages 205–214, 2012.

[21] C. Liu, J. Liu, and N. J. Belkin. Predicting search task difficulty at different search stages. In *CIKM '14*, pages 569–578, 2014.

[22] J. Liu and N. J. Belkin. Personalizing information retrieval for multi-session tasks: The roles of task stage and task type. In *SIGIR '10*, pages 26–33, 2010.

[23] J. Liu, M. J. Cole, C. Liu, R. Bierig, J. Gwizdka, N. J. Belkin, J. Zhang, and X. Zhang. Search behaviors in different task types. In *JCDL '10*, pages 69–78, 2010.

[24] J. Mao, Y. Liu, K. Zhou, J.-Y. Nie, J. Song, M. Zhang, S. Ma, J. Sun, and H. Luo. When does relevance mean usefulness and user satisfaction in web search? In *SIGIR '16*, pages 463–472, 2016.

[25] S. Menard. *Applied Logistic Regression Analysis*. Sage, 1997.

[26] S. Mizzaro. Relevance: The whole history. *J. Am. Soc. Inf. Sci.*, 48(9):810–832, 1997.

[27] T. Saracevic. Relevance reconsidered. In *CoLIS 2*, pages 201–218, 1996.

[28] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. part II: Nature and manifestations of relevance. *J. Am. Soc. Inf. Sci. Technol.*, 58(13):1915–1933, 2007.

[29] R. Tang and P. Solomon. Use of relevance criteria across stages of document evaluation: On the complementarity of experimental and naturalistic studies. *J. Am. Soc. Inf. Sci. Technol.*, 52(8):676–685, 2001.

[30] A. Tombros, I. Ruthven, and J. M. Jose. How users assess web pages for information seeking. *J. Am. Soc. Inf. Sci. Technol.*, 56(4):327–344, 2005.

[31] E. G. Toms, H. L. O'Brien, R. Kopak, and L. Freund. Searching for relevance in the relevance of search. In *CoLIS '05*, pages 59–78, 2005.

[32] P. Vakkari and N. Hakala. Changes in relevance criteria and problem stages in task performance. *J. Doc.*, 56(5):540–562, 2000.

[33] M. Verma, E. Yilmaz, and N. Craswell. On obtaining effort based judgements for information retrieval. In *WSDM '16*, pages 277–286, 2016.

[34] E. Voorhees and D. Harman. Overview of the fifth Text REtrieval Conference (TREC-5). In *TREC-5*.

[35] Y. Xu and Z. Chen. Relevance judgment: What do information users consider beyond topicality? *J. Am. Soc. Inf. Sci. Technol.*, 57(7):961–973, 2006.

[36] Y. Xu and D. Wang. Order effect in relevance judgment. *J. Am. Soc. Inf. Sci. Technol.*, 59(8):1264–1275, 2008.

[37] Y. Zhang, J. Zhang, M. Lease, and J. Gwizdka. Multidimensional relevance modeling via psychometrics and crowdsourcing. In *SIGIR '14*, pages 435–444, 2014.