Investigating Users' Time Perception during Web Search

Cheng Luo[†],Xue Li[†], Yiqun Liu[†], Tetsuya Sakai[◇], Fan Zhang[†], Min Zhang[†], Shaoping Ma[†]
[†]Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[◇]Waseda University
yiqunliu@tsinghua.edu.cn

ABSTRACT

Due to the tremendous economic value, search engine companies have invested a lot to improve the quality of their search results. Recently much effort has been made to directly model key aspects of users' interactions with search system, for example, *Benefit* and *Cost*. Time has been widely adopted in both of the two aspects since benefit and cost must be expressed in meaningful units in practical application. Psychological studies have demonstrated that the subjectively perceived time might be different from the objective time measured by timing device and the time perception process of human beings is affected by some psychological factors, such as motivation and interest, which are closely related to the search process. Considering that time is usually used to describe the subject experience of search users, it is necessary to investigate the difference between perceived time and objective time in search process. In psychology, there is a temporal illusion effect named *Vierordt's law*, i.e. shorter intervals tend to be overestimated while longer intervals tend to be underestimated. In this work, we carefully designed a lab-study to examine the impacts of duration length on user's time perception in the context of search. Experimental results show that the Vierordt's law is consistently observed in Web search environment. This work could help us to correct the estimation of users' perceived time and provide insights about the mechanism of satisfaction.

Keywords

Interactive Information Retrieval; User Behavior; Time Perception

1. INTRODUCTION

Search engines have became one of the most important tools to access Web resources for many years. Due to the tremendous economic value, search engine companies put great effort to improve their search results. Thus search effectiveness evaluation has attracted a lot of attention from both industry and academia.

Beyond traditional Cranfield evaluation paradigm based on test collections, query set, relevance judgments and metrics [14], recently much effort has been made to directly model key aspects of users' interaction with the search system [12]. Several theories have been proposed to glean insights and generate hypothesis about users' behavior [5], for example, Information Foraging Theory [52], Interactive Probability Ranking Principle [28] and Search Economic Theory [4]. The concepts of *Benefit* and *Cost* sit at the central place of all these theories. Time is widely adopted in both of the two aspects, because benefits and costs must be expressed in meaningful units in practical calculation. Search cost is often measured by the time of a series of actions, such as formulating queries, examining snippets, clicks on results and etc. Search benefit might be measured in Time Well Spent (TWS), i.e. the time spent viewing relevant material [13]. Temporal information is also incorporated into traditional evaluation metrics. Time is either explicitly used as a parameter of decay function (for example, in Time Biased Gain [54], and Expected Latency-discounted Gain [3]) or implicitly encoded in other measures such as the length of text read by user in U-measure [53].

To summarize, time plays an important role in varying perspectives of search evaluation. We find that the time in most existing works is the *objective* time measured by clock instead of the *subjective* time perceived by human beings. However, we argue that it is more intuitive to adopt perceived time while time is serving as an estimation of effort, because effort itself represents the exertion of mental power and is highly likely to be subjective. Time perception is the subjective experience of time, which is measured by someone's own perception of durations of the indefinite and unfolding of events [30]. A number of researches have established that the perception of temporal information is influenced by both psychological factors (for example, attention [55], complexity [35] and emotion [22]) and physical factors (for example, body temperature [58] and age [21]). Although many of these factors are also regarded as important research issues in IR community, their effects on perceived time have not received enough attention.

To the best of our knowledge, few works investigate the difference between objective time and perceived time in Web search scenario. Luo et al. showed that high level of Temporal Relevance [47] and irrelevant document [46] would make the users overestimate the durations than usual. Czerwinski et al. [18] found that the perceived duration

CHIIR '17, March 07-11, 2017, Oslo, Norway
© 2017 ACM. ISBN 978-1-4503-4677-1/17/03…$15.00
DOI: http://dx.doi.org/10.1145/3020165.3020184

of an uncompleted task would be overestimated, while the durations on tasks completed successfully would be underestimated in various application scenarios (also known as Zeigarnik effect [59]). Although these studies shed some light on time perception in Web search environment, we wonder whether some other fundamental factors would affect the perceived durations.

In this paper, we investigate users' time perception in the context of search. In psychology, there is a long standing theory named *Vierordt's law* proposed by Karl von Vierordt [25], which states that "short" intervals of time tend to be overestimated and "long" intervals of time tend to be underestimated. We wonder whether the Vierordt's law exists in Web search, i.e. whether and how the absolute dwell time length would influence the users' time perception. Based on existing research [33], relative long dwell time are usually assumed to be more likely to accompany with relevant materials and users' satisfaction than short ones. Considering that satisfaction would affect user's attention and interest, which have potential effects on time perception, we further explore whether the impact of object interval length exists under different satisfaction conditions.

To answer these questions, we conducted a controlled user study with 50 participants. The participants were instructed to complete several search tasks and report their perceived durations afterwards. Experimental analysis indicates that in Web search, the perceived time was affected by the absolute duration length as Vierordt's law states and this effect exists consistently across different users' satisfactions. The findings of this work suggests that when using time as an estimation of users' effort, we should adopt the user perceived time by making an adjustment based on duration length. What's more important, the users' perceived time would provide insights about the mechanism of information need fulfillment in the search process. To summarize, the contributions of this paper are stated as follows: (1) We carefully designed a experimental framework to investigate the users' time perception in Web search environment. (2) We analyzed the impact of different factors on time perception and found that the perceived time was affected by absolute duration length as in Vierordt's law: long durations tend to be underestimated and short durations tend to be overestimated. (3) The difference in dwell time usually accompanies with the differences in users' satisfaction. We examined the influence of absolute duration length under different satisfaction conditions. Experimental results show that the influence exists consistently.

2. RELATED WORK

2.1 Time in Web Search

Time and temporal information is widely used in multiple perspectives of IR researches.

In **Search Intent Understanding**, the queries which have temporally dependent intents are usually recognized as *time-sensitive* queries [20], whose best search results change with time, for example "Presidential elections" or "CHIIR conference". The temporal aspects are identified [6, 51, 32] and integrated into the overall ranking mechanism to improve the freshness and relevance of search results [49, 24, 19, 40, 11]. From the aspect of time urgency, some queries are *time-critical*, where users have urgent information needs in the context of an acute problem, for example, "stroke in woman" [50]. Crescenzi et al. showed that time pressure would lead to changes in user behavior [17, 16]. Mishra et al. proposed a model to predict urgent information needs with features including user behavior [50]. On the opposite side of time-critical query, Teevan et al. explored "slow search", a class of search where traditional speed requirements are relaxed in favor of a high quality search experience [56].

In **User Behavior Analysis**, different temporal measurements have been proposed as users' implicit feedbacks [9]. For example, *time-between-clicks* is an estimation of users' dwell time on landing page. It is widely used in multiple applications: satisfaction prediction [42], search success evaluation [33], result usefulness [44] and task difficulty prediction [45].

In **Search Evaluation**, time is taken into consideration in both offline and online evaluation methods. In offline evaluation, time is either explicitly used as the parameter in decay function (for example, Time Biased Gain [54] and Expected Latency-discounted Gain [3]) or implicitly encoded in other measures, such as examination depth in Precision and Recall, and the length of trailtext in U-measure [53]. Recently, researchers focus on directly modelling essential aspects of users' interactions, for example, benefits and costs. Time is often used as an estimation of users' search cost in practical computation of corresponding theories, for example, Information Foraging Theory [52], Interactive Probability Ranking Principle [28] and Search Economic Theory [4]. It is intuitive that the more time user spent on a specific action (examine a snippet, read a document and etc.) usually indicates the more cognitive resources he/she has invested. Similar to system-centric evaluation methods, time is also encoded in other measures like the number of queries in a session [38]. For benefits, Time Well Spent (TWS) [13], expressed as the total time spent on relevant material, measures the utility users have gained in search.

2.2 Time Perception in Psychology

Time perception, referred to as the subjective experience of the objective time has been carefully studied for decades in the fields of psychology and neuroscience [8, 27]. While time itself is objective, the perceived succession and duration of time is subjective.

Based on phenomenological and experimental data, huge amount of efforts has been invested to explore what the human being is able to know about time through perception and estimation of durations. The experiments are usually adopted in two paradigms: *prospective* timing and *retrospective* timing [7]. It is called prospective when a participant is aware of the necessity to judge the experience of time before the duration. Otherwise, it is called retrospective. Dan Zakay [60] summarizes several models of time perception theory and proposed that time perception is manipulated by the following factors: non-temporal information processing load (simple or complex stimuli), type of judgment (absolute or relative), and experiment paradigm (prospective or retrospective).

Time perception would be affected by both psychological (interestingness, attention, cognitive load etc.) and physical factors (body temperature, drug usage, etc.) [55]. Some of these concepts have also attracted the IR community. For example, the cognitive complexity of search tasks for interactive IR experiments has been explored by Kelly et al. [41]. User would be more engaged in interesting tasks and they spent longer completing these tasks [23]. However, in Web search environment, how these factors would affect the users' perception of time has not received enough attention. A laboratory study conducted by Luo et al. [47] shows that in high temporal relevance situations, users would tend to overestimate the duration length than usual. They also found that on document level, search users tend to shorten their perceived time on relevant documents [46].

Although these work opened the door to time perception in Web search environment, we wonder whether and how some more central factors, for example, the user satisfaction and dwell time length would affect the perceived time of users.

Time perception is a subjective feeling about the duration of the indefinite and continuous unfolding of events [30]. That is to say, there is not a straightforward way to measure the perceived time. In experimental psychology, several methods have been developed to assess estimations of the perceived time. The first method named *Verbal Estimation* require the participants to provide a verbal estimation of a duration using temporal units, such as minute and second. In the second method *Reproduction*, the participant was first shown a duration and then asked to reproduce the interval by some operations, for example, push and hold a button for some time. Similar to Reproduction, in the third method *Production*, the participant needs to produce an interval according to a duration in temporal units. The last method *Comparison* presents two durations and require the participants to make a judgment about which one is longer.

3. METHOD

To investigate the perceived time of users in search tasks, we designed and conducted a laboratory user study with several tasks. Although the controlled experiment had a smaller scale comparing to search log analysis, it enabled us to control variabilities and to collect users' time perception with effective psychological methods.

3.1 Scenario, Tasks and System

In our user study, the participants need to perform 9 ad-hoc search tasks in the experimental search system. The first task is for instruction and training, while the remaining ones are for formal experiments. For each task, the participants need to read the *task description* and then search for relevant information with a predefined *query* in our system. Then they were instructed to estimate how long they have spent in the searching process. Throughout the experiment, a participant completed several questionnaires including a demographic questionnaire, pre/post-search questionnaires and an exit questionnaire.

We created 9 informational search tasks which are similar with the topics of TREC Session Track[1]. Several criteria were taken into consideration when organizing these tasks: Firstly, the tasks should be of varying levels of cognitive complexity. Cognitive complexity refers to the amount of learning and cognitive effort required to complete the search. Based on psychological theory [35], cognitive complexity might have an effect on time perception. Following previous Interactive IR researches [10, 36, 41], our tasks vary across four domains (psychology, culture, healthcare and science) and across four levels of cognitive complexity according to the taxonomy of learning proposed by Anderson and Krathwohl [1]: (1) *Remember*: recalling relevant knowledge from long-term memory, (2) *Understand*: constructing meaning through summarizing and explaining, (3) *Analyze*: breaking material into constituent parts and determining how the parts relate to each other, and (4) *Evaluate*: making judgements through checking and critiquing. In the eight tasks for formal experiment, for each level of cognitive complexity, we have two tasks.

Table 1: An example of the search tasks

#	Topic	Cognitive Complexity	Goal	Product
1	Science	Remember	Clear	Factual
The Halo Effect is a very interesting psychological effect, please search for some information about this effect, and find an appropriate example from our daily lives.				
Query	Halo Effect			

Secondly, we consider the goal of the tasks, which is associated with interactive search behavior [37]. Li and Belkin [43] identified a variety of generic facets of tasks in information seeking, such as Source of task, Action, Goal, Product and etc. We follow Jiang and Ni's experimental design and consider two characteristics of search tasks [39]: the goal of a search task is either clear or amorphous (`goal`); the product of a search task is either factual information, or enhanced intellectual understanding of the user (`product`). In our experiment, all of the tasks have a clear goal, because we have to make sure that the tasks are clearly stated and all the participants can interpreted the task descriptions in the same way. For task product, we have 4 intellectual tasks and 4 factual tasks. In Jiang and Ni's work, they also take the user's self-rated familiarity (`familiarity`) into consideration, we investigate this in our pre-task questionnaire.

For each search task, we provided a predefined query to perform search in our system. Although the fixed query might threaten the ecological validity of our experiment, it would make sure that all the participants were presented with identical Search Result Page (SERP) for each task. An example of the tasks is shown in Table 1.

To simulate a real Web search environment, we developed an experimental search system, which is very similar with general Web search engines. When the user begins search, our system will provide ten search results, which are manipulated (as described in Section 3.4) based on SERPs crawled from a commercial search engine. We removed all the query suggestions, sponsored results, ads and vertical results to prevent potential distraction. For the purpose of variabilities controlling, query reformulation and pagination are not supported in this experiment. All the users' interactions including clicks, scrolling, tab switching and mouse movements are logged in the back-end database using an injected Javascript script.

After the setup of the experimental system, the authors of this paper double-checked the tasks carefully to make sure: (1) The tasks and topics are geared towards our participants, i.e. university students. (2) The manipulated search results could provide enough information to fulfill the information needs proposed in task descriptions.

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

3.2 Study Participants

In our experiments, we recruited 50 undergraduate students from a university located in China, which represent a typical type of subject in IR researches. The average age of the participants was 20.02 (SD=2.00) and 32 of them were female. A variety of majors were represented across the natural sciences (N=10), social sciences (N=8), arts (N=4), and engineering (N=28). All of the participants searched at least 1 to 3 times per day. 37 of the 50 participants reported at least 4 to 6 searches per day. The participants reported a high level of familiarity with search engine (4.88/7, SD=1.17). They were informed in advance that they would be paid $15 for the participation. The experiments actually lasted about 60 minutes and the participants all signed a post facto participation form revealing the content of the experiment.

3.3 Experiment Procedure

The experiments were performed in a quiet room to avoid external disturbance. The participants were asked to take off their watches and turn off any device which would provide temporal information. They were also asked to remove all jewelries or anything which might be a distraction during the experiment.

The study used the following protocol as shown in Figure 1. (I) Firstly, the participants need to complete a demographic questionnaire, which is about their age, gender, and familiarity with search engine. (II) The participants then received instruction via a video on screen. In the video, we introduced the procedure of the experiment and completed the training task as an example. More specifically, the participants were instructed as follows: *"First, please read the task description and make sure you have understood the information needs, then you will see a search result page. Please find relevant information as you are using a search engine in a natural manner."*. The participants were not informed about the purpose of our experiment. They were instructed to estimate the durations spent on searching after each task. After instruction, the participants would go through the training task to get familiar with the procedure and the experimental system. (III) For each formal task, the participants were first shown the task description, then a pre-task questionnaire (III-a) was used to investigate their familiarity, interest, expected difficulty and understanding about the topic. Once they finished the questionnaire, they would enter the SERP (III-b), which had ten search results presented as a commercial search engine. While no task time limits were imposed, the search process ended when the participants felt satisfied or hopeless, or just finished examining all the ten results. Then the participants were redirected to another page in which they reported their minimal/maximal estimations (III-c) to the nearest 10 seconds about the dwell time on the SERP. This estimation method is following Grondin et al.'s approach [31] about multi-minute intervals estimations. The participants were then asked to report their perceived quality of search results, perceived difficulty, consciousness of time elapsing, and perceived urgency (III-d). As suggested by Mao et al. [48], the participants were asked to summarize their outcome after each task. This would encourage the participants to concentrate on the search process to find useful information. The answer would be recorded by voice instead of keyboard, which would reduce their efforts as much as possible.

Finally, after the participants completed the search tasks, they were directed to an exit-questionnaire (IV), which investigates their overall experience during all the formal tasks. The exit-questionnaire is about their average interests, consciousness of quality manipulation, fatigue and confidence of time estimations.

For all the questions in pre/post-task questionnaires and exit-questionnaires, participants were asked to respond their agreement on 7-point Likert scale (from strong disagreement to strong agreement) to the predefined statements such as: "You felt tired after all the tasks were completed" as in Arapakis et al.'s experiment [2]. Moreover, we conducted a pilot experiment to confirm that the tasks and procedures are appropriate and the manipulation of search results is effective.

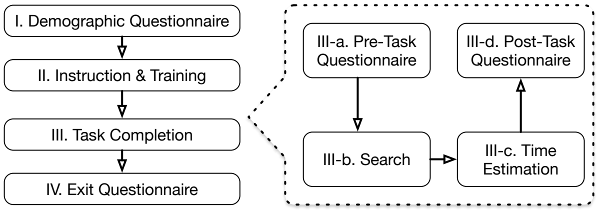
I. Demographic Questionnaire
II. Instruction & Training
III. Task Completion
IV. Exit Questionnaire
III-a. Pre-Task Questionnaire
III-b. Search
III-c. Time Estimation
III-d. Post-Task Questionnaire

Figure 1: Procedure of the experiment

3.4 Experiment Design

In this work, we focus on the impact of objective duration length on users' perceived time. However, as we know, the dwell time of a search session is a user's reaction towards many factors, such as system effectiveness and result presentation. It is also affected by some subjective factors, e.g. reading speed and cognitive ability.

Another factor accompanying with dwell time is user's satisfaction. There are a number of studies using dwell time as a feature to predict users' satisfaction [34, 38]. According to psychological studies [22, 30], satisfaction would have an impact on users' motivation and emotional states, which further lead to illusions of perceived time. However, there is not a certain relationship between dwell time and satisfaction. For example in Table 2, a search session with long dwell time could be a complex information needs solved by several relevant documents or a struggling process with several irrelevant documents. On the other hand, a session with short dwell time may be an information need resolved by a high quality document ranked at top position or an early abandonment on a disappointed SERP.

In our experiment, the two factors which may influence users' time perception, absolute dwell time length and satisfaction are nested together to some degree. Therefore, to tackle this problem, we attempted to present SERPs of varying qualities in different tasks. Following Smucker et al. [54], we manipulated the precision of the SERPs. More specifically, for each task, there are two different SERPs as follows:

Poor Quality: three of the ten search results are relevant (precision=0.3) while none of the top five ones is relevant.

High Quality: seven of the ten search results are relevant (precision=0.7) while none of the top five ones is irrelevant.

The SERPs are manipulated from two aspects: the quantity and position of relevant results. Comparing to High

Table 2: Example of sessions with long/short dwell time under different satisfaction (SAT/DISSAT) conditions

	Long Dwell Time	Short Dwell Time
SAT	The user viewed a few relevant results to resolve a complex problem. For example, "history of WW2".	The user was satisfied by a top relevant result. For example, "NYC Weather".
DISSAT	The user examined several results however few of them were relevant. For example, "inline bug in C program".	The user felt disappointed by the SERP and gave up this query.

Quality SERPs, the Poor Quality SERPs have fewer relevant results, which are presented at lower positions.

For a specific search task, we first crawled the top search results of the predefined query from a commercial search engine. Then the binary relevance score (relevant, irrelevant) of each search result is judged by three professional assessors. We only kept the results for which the assessors reached an agreement, i.e. all of the assessors labelled the search results as relevant or irrelevant. At last, we sequentially selected 10 search results to fill the slots on Poor Quality SERP and High Quality SERP and made sure that the quantity and position of relevant results meet our definitions.

In the user study, participants search for relevant information for 9 tasks. The first one is for training purpose and we presented High Quality to make sure all the participants received identical instruction. For the remaining 8 ones, we randomly assigned 4 of them to High Quality group and the other 4 tasks to Low Quality group. We rotated the sequence of 8 formal tasks using a Latin square to avoid presentation order bias. For each task, the participants in the two groups are balanced, i.e. half of the participants were presented with High Quality SERPs while the other half were presented with Low Quality SERPs.

4. DATA

4.1 Users' Perceived Satisfaction v.s. Manipulated SERP Quality

The participants were not informed the purpose and the manipulation of our experiments, we begin the analysis by examining the perceived satisfaction under SERPs of varying qualities.

Recall that in post-task questionnaire, participants were asked to report their agreement about "You are satisfied with the search results about this topic" on a 7-point scale (1 to 7, from strong disagreement to strong agreement). To fold the users' satisfaction ratings into binary categories (SAT, DISSAT), we arbitrarily used 4 as the threshold of satisfaction, i.e. for a specific topic and SERP, if a user gives a rating which is equal or greater than 4, we think that it is a satisfied search.

We assume that users would be satisfied with High Quality SERPs and dissatisfied with Low Quality SERPs. Then we can measure the consistency of users' perceived satisfaction and SERP quality by calculating the accuracy. More specifically,

$$accuracy = \frac{\#SAT\_on\_HQ + \#DISSAT\_on\_LQ}{\#Tasks} \quad (1)$$

The results are shown in Figure 2. We can see that for most of the users (46/50), the accuracy is above 60%, except for participant #4 (50.0%), #23 (50.0%), #24 (50.0%) and #27 (25.0%). There might be various reasons which lead these divergences. First, the participant may misunderstand the feedback mechanism. For example, #27 gave relatively high ratings (7, 5, 5 and 3) on Low Quality SERPs and low ratings (3, 2, 2 and 5) on High Quality SERPs. Second, the users may have various levels of satisfaction even on SERPs with identical content, because their expectations about the performance of the search engine are also varying with individuals [57]. In our experiment, participant #4, #23 and #24 they give ratings which are greater than 4 for all of the search sessions. Third, it is possible that the participants would make mistake due to fatigue or distraction during the experiment. We remove the data of these 4 participants in the further analysis to reduce potential noise.

4.2 Average Dwell Time on Different SERPs

The average dwell time on different SERPs of each task is presented in Figure 3 and the error bars indicate the corresponding standard variations.

Figure 3 illustrates that for most of the tasks, the average dwell time on the High Quality and Low Quality SERPs are very close. This observation is possible against the intuition that users should spent longer time on High Quality SERPs, because there are more relevant results and the durations on relevant results are usually longer [26]. A potential reason is that dwell time on SERP level is affected by multiple factors. Though previous studies found that users spend longer on relevant documents, in practical search scenario, on a High Quality SERP, a user may leave earlier since he/she may be satisfied by a relevant result ranked at top positions while on a Low Quality SERP, users may have to examine more results, which also leads to long durations. We conducted a two-sided t-test and found only the difference on Task #6 is significant (t-stat=-2.130, p-value=0.038, df=43.90, ES=0.642). The Task #6 requires the participants to investigate the ranking of movie box office value in 2015. We found that on the High Quality SERP, most participants were satisfied by the first relevant result.

4.3 Feedbacks

We grouped the response to the items on the pre/post-task questionnaire as follows. We put the search sessions in which the High Quality SERPs were presented into the "HQ Group", and then placed the remaining ones into "LQ Group". For the questions which were reported on a 7-point Likert scale, we converted the responses to numeric values from 1 to 7. We also calculated the average dwell time length and perceived time length of different groups. A two-sided t-test was conducted to examine whether the difference between two groups is significant. The results are presented in Table 3.

We can see that for pre-task questionnaire, the responses from both groups are quite similar. The participants reported moderate level of *familiarity with the topic* and relatively high level of *interests in the topics*. They are very confident that they had fully understood the

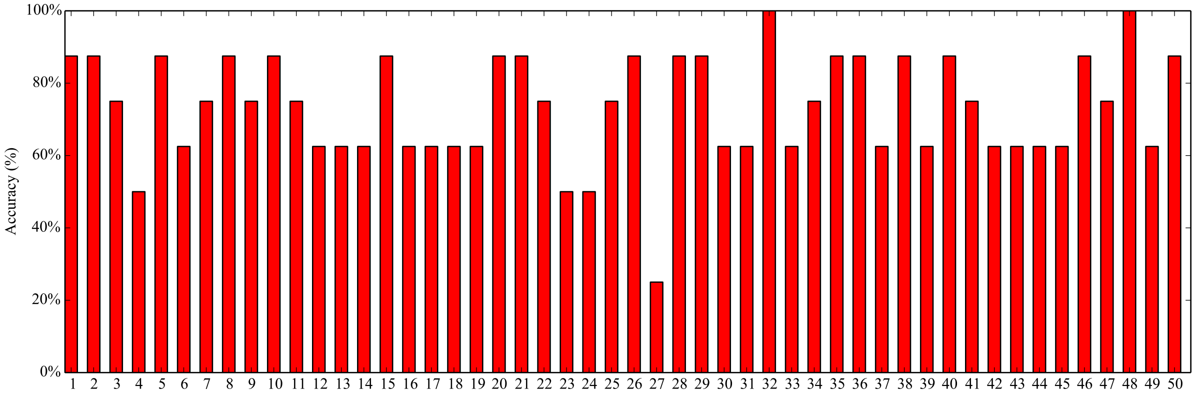

Figure 2: Consistency of Users' Satisfaction v.s. SERP Quality (High Quality/Low Quality)

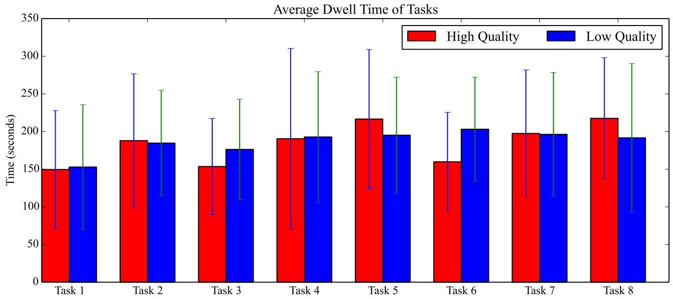

Figure 3: Average Dwell time on High Quality/Low Quality SERPs of Different Tasks

Table 3: Statistics of Participants' feedback in Questionnaire and Behavior (* indicates significant at p<0.01;** indicates significant at p<0.05)

	HQ Group	**LQ Group**
	Pre-task Questionnaire	
Familiarity with the topic	3.920(1.703)	3.829(1.804)
Interests in the topic	5.150(1.614)	5.166(1.625)
Understanding about info. need	6.305(0.879)	6.351(0.685)
Expected search difficulty	3.705(1.599)	3.577(1.690)
	Post-task Questionnaire	
Perceived Satisfaction**	5.325(1.410)	3.397(1.721)
Perceived search difficulty**	2.635(1.422)	3.748(1.700)
Awareness of time*	6.095(0.752)	5.970(0.789)
Perceived time pressure	4.910(1.030)	4.930(0.948)
	Behavior	
Dwell time length (seconds)	182.985(87.372)	184.804(80.214)
Perceived time length (seconds)	163.925(75.334)	165.954(74.576)

information needs of the tasks (*understanding about info. need*). The expected search difficulty in HQ Group (Mean=3.705, SD=1.599) is slightly higher than that in LQ Group (Mean=3.577, SD=1.690) but the difference is not significant. It is not surprised that for none of the items in pre-task questionnaire, the differences between two groups is significant since the participants were not aware of the purpose and the manipulation of the experiment.

In the post-task questionnaire, the participants perceived significantly higher satisfaction in HQ Group (Mean=5.325, SD=1.410) than in LQ Group (Mean=3.397, SD=1.721). This proves that our SERP manipulation effectively affects users' satisfaction. They also felt that it is much easier to find relevant information in HQ Group (Mean=2.635, SD=1.422) than in LQ Group (Mean=3.748, SD=1.700). The participants from both HQ Group and LQ Group reported that they are aware of the elapsing of time when conducting the search tasks. They also perceived a relatively high level of time pressure. Although we did not give any time limit in the experiment, the participants might feel some pressure because that they were required to report their perceived time after each task.

The dwell/perceived time length in HQ Group and LQ Group is quite close. This is consistent with our observations on each task (Figure 3). It should be noted that the variations of dwell time is quite large, which allows us to observe the influence of dwell time length on time perception.

5. EXPERIMENT RESULTS

This section discusses the following research questions:

RQ1: How can we measure the relationship between users' dwell time and perceived time?

RQ2: Does absolute dwell time length have an effect on users' time perception? Does Vierordt's law exist in Web search?

RQ3: Considering satisfaction is another factor which would influence users' time perception, is the effect of absolute duration length consistent across different satisfaction conditions?

5.1 Measuring the Relationship between Dwell Time and Perceived Time

In our experiment, the participants provided minimal/maximal estimations about how long they had spent on search in temporal units. Following Grondin et al.'s approach [31], we use the mean of minimal and maximal duration length as their estimations. To measure the relationship between perceived time and actual time (**RQ1**), we first define a measure named *perceived rate*. For a specific user u in task t, the *perceived rate* denoted by $P\text{-}rate\,(u,t)$ is defined as:

$$P\text{-}rate(u,t) = \log \frac{perceived\ time}{dwell\ time} \quad (2)$$

where *perceived time* and *dwell time* denotes the perceived duration length and actual duration length respectively. *P-rate* is defined as a logarithm function to ensure that the absolute value represents the difference between perceived time and dwell time and the signal(+/-) reflects whether the perceived time is shorter or longer.

P-rate reflects the ratio between the perceived time and actual dwell time. The perceived time is normalized by the actual dwell time since we are focusing on whether the perceived duration is overestimated or underestimated comparing to the actual duration length. For a specific search session, $P\text{-}rate < 0$ means that the perceived time is shorter than the dwell time, i.e. the duration length is underestimated. On the opposite side, $P\text{-}rate > 0$ indicates that the perceived time is longer than dwell time and the duration length is overestimated.

The distribution of *P-rate* on dwell time is shown in Figure 4. Each point in the figure denotes a certain search session. An intuitive observation is that the *P-rate* correlates negatively with dwell time length. Most of the points whose *P-rate* is positive appear at the positions where dwell time is shorter than 200 seconds, which means that the overestimated sessions usually have a relatively short dwell time. For almost all the points whose dwell time is longer than 300 seconds, the *P-rate* is negative, i.e. the long durations tend to be underestimated. We would to examine the effect of dwell time on *P-rate* with regression analysis in Section 5.2.

5.2 Impacts of Different Factors on Time Perception

To investigate the impact of absolute dwell time length on *P-rate* (**RQ2**), we ran a regression analysis to analyze the effect of duration length by controlling different factors following Crescenzi et al. [15] . We developed several linear regression models with different factors. More specifically,

- **Model 1** dependent variable: *P-rate*; independent variable: absolute dwell time length.
- **Model 2** dependent variable: *P-rate*; independent variables: absolute dwell time and the factors reported in pre-task questionnaire.
- **Model 3** dependent variable: *P-rate*; independent variables: absolute dwell time, the factors reported in pre-task questionnaire and the information of tasks and participants.

In Model 1, we focus on the effect of absolute dwell time length while in Model 2 more factors reported in pre-task questionnaire are taken into consideration. We did not involve factors which were reported in the post-task questionnaire since these factors are probably dependent on the dwell time length. In Model 3, we further took several the task and participant specific factors into consideration to capture the potential variability: participants (one per participant), task cognitive complexity (4 levels, remember, understand, analyze and evaluate respectively), task goal (2 levels, clear or amorphous) and task product (2 levels, factual or intellectual).

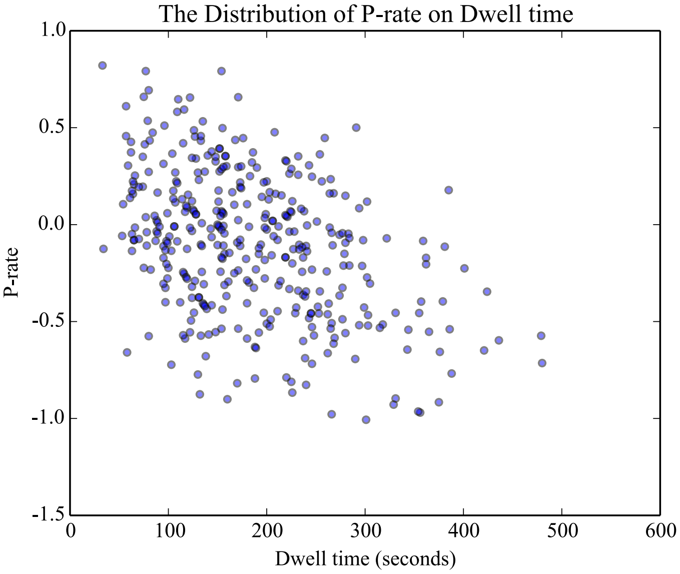

Figure 4: The Distribution of *P-rate* on Dwell Time

Table 4 displays the regression coefficients (β) and intercepts (constant). In Model 1, we can see that the dwell time length is a significant predictor of *P-rate*. With each additional unit (seconds) of dwell time, the *P-rate* would decrease 0.0018. Although the effect looks slightly weak, based on the fact that the dwell time ranges from tens to hundreds of seconds and the *P-rate* is defined as a logarithm function, the effect of absolute dwell time is significant ($p < 0.001$) and could not be ignored.

In Model 2, we added several factors which were reported by the participants in pre-task questionnaire: familiarity with the topic, interests in the topic, understanding about the information need and the expected difficulty. The dwell time length remained significant ($p < 0.001$). The other factors presents insignificant correlations with *P-rate* and no interaction effect between dwell time length and these factors was found.

In Model 3, the dwell time shows similar effect with that in Model 1 & 2. The participant presents a coefficient varying from -0.4436 to 0.5467. Among all the 46 participants, 18 bring significant ($p < 0.05$) impact on *P-rate*. This finding is in line with previous psychological studies [30], time perception is a subject cognitive process of human beings and highly affected by individual factors. For the task based factors, only cognitive complexity has a significant ($p < 0.05$) influence on *P-rate*. With each additional unit of cognitive complexity, the model predicts that *P-rate* will increase by 0.0223. When the task becomes more complex, the user tends to overestimate perceived time. This phenomenon may also be supported by some theories in psychology. When the task become

Table 4: Regression Models for *P-rate* with Different Strategies (*** indicates $p < 0.001$ and * indicates $p < 0.05$)

Variables	Model 1	Model 2	Model 3
Absolute dwell time length	-0.0018***	-0.0018***	-0.0015***
Familiarity with the topic		-0.0119	-0.0098
Interests in the topic		0.0010	0.0015
Understanding about info. need		-0.0130	-0.0116
Expected search difficulty		-0.0049	0.0058
Participant			-0.4436 to 0.5467
Task Cognitive Complexity			0.0223*
Task Goal			-0.0224
Task Product			-0.0451
Constant	0.2359 ***	0.3785*	0.1717*
R-squared	0.3776	0.3828	0.6857

more complex, users would have lower motivation since he/she has to invest more cognitive resource and mental power. Psychological studies find that motivation causes the time to be underestimated during pleasant experiences and overestimated during struggling experiences [29].

We can see that in Model 1, dwell time length explained a sizeable proportion of the variance in *P-rate*, in Model 1, $R^2 = 0.3776$, F(1,365)=80.01, $p < 0.001$.

5.3 Vierordt's law in Time Perception

In psychological studies about time perception, there is a long standing temporal illusion named *Vierordt's law* [25], which states that human beings tend to overestimate short durations and underestimate long durations. We want to investigate whether the Vierordt's law works in the Web search environment.

In our experiments, we can see that when the dwell time is relative short, the *P-rate* tends to be positive. According to the definition of *P-rate*, it indicates that the perceived time is longer than the dwell time, i.e. the dwell time is overestimated. Based on the regression analysis, *P-rate* correlates negatively with dwell time length. As dwell time grows, *P-rate* decreases and becomes negative. That is to say, when the duration is relative long, $\frac{perceived\ time}{dwell\ time}$ decreases and goes below 1, i.e comparing to the absolute dwell time, the perceived time is underestimated.

To examine whether this effect is consistent across search sessions with different length, we split the search sessions into several parts according to the dwell time length and ran regression analysis on each part.

The search sessions were divided into the following groups:

- **Short Group (N=60):** dwell time is shorter than 100 seconds.
- **Middle Group (N= 272):** dwell time is longer than 100 seconds and shorter than 300 seconds.
- **Long Group (N=36):** dwell time is longer than 300 seconds.

Table 5: Regression Models for *P-rate* on Search Sessions with Different Lengths (*** indicates $p < 0.001$, ** indicates $p < 0.01$ and * indicates $p < 0.05$)

Variables	Short	Middle	Long
Absolute dwell time length	-0.0058***	-0.0013**	-0.0014
Constant	0.5382 **	0.1632*	-0.3280
R-squared	0.4776	0.2828	0.0480

Table 6: Regression Models for *P-rate* on Search Sessions Under Different Satisfaction Conditions (*** indicates $p < 0.001$ and * indicates $p < 0.05$)

Variables	Satisfied	Dissatisfied
Absolute dwell time length	-0.0022***	-0.0012***
Constant	0.3079 ***	0.1127*
R-squared	0.3601	0.3828

We conducted regression analysis on all the three groups and the results are presented in Table 5. We can see that the absolute dwell time length presents a negative effect across all the three groups. On the Short Group, the effect is much more stronger than that on the Middle and Long Groups. Comparing to the *Underestimation of Long Dwell Time*, the *Overestimation of Short Dwell Time* is more obvious and significant. The effect on the Long Group is not significant, a potential implication is that the data in Long Group is too sparse.

In summary, we find that the absolute dwell time has a significant effect on the *P-rate*. When the dwell time is relatively short, *P-rate* is positive, i.e. user tends to overestimate the duration. As time grows, *P-rate* decreases and becomes negative, which indicates user underestimates the duration.

The results in our experiments correlates with the Vierordt's law. This would help us to better understand the cost model of users in Web search, e.g we can estimate the perceived time based on the duration length to better users' effort during Web search.

5.4 Impact of Dwell Time Length Under Different Satisfaction Conditions

As we mentioned before, satisfaction is another factor which may influence the perceived time. We are wondering whether the effect of dwell time length is consistent across different satisfaction conditions. We split the search sessions into two groups according to the self-reported satisfaction of users: **Satisfied Group** (perceived satisfaction$\geq$4) and **Dissatisfied Group** (perceived satisfaction<4). Since satisfaction may have an impact on perceived time by affecting motivation and emotional states, the user perceived satisfaction may be more appropriate than the predefined SERP quality.

We ran a regression analysis on each group and the coefficients (β) and intercepts (constant) were shown in Table 6. The dwell time length presents a negative and significant correlation on both the two groups. As discussed in Section 5.3, we can conclude that the Vierordt's law was consistently observed under different satisfaction conditions.

On Satisfied Group, the effect of dwell time length was slightly stronger than on the Dissatisfied Group. This result suggests that when users' satisfaction might be another factor which will influence the time perception in Web search.

6. CONCLUSION AND FUTURE WORK

In this work, we investigated the subjective perception of time in the context of Web search. Based on a controlled user experiment, this paper provides insights on the impact of dwell time length on perceived time. We found that Vierordt's law had a consistent and significant effect on users' perceived time under varying satisfaction conditions: participants tend to underestimation relatively long durations and overestimate short durations. The findings in this paper may help us better understand the time perception mechanism and cost model of Web search users. In the future work, we would like to explore adopting the perceived time in existing evaluation framework.

Our study has a few limitations. Our experiment was conducted in a lab environment in which participants have to feedback their perceived time explicitly. This might be slightly different from the natural environment. Time perception is a subjective mental activity. When a person is using a search engine, many factors (presentation, response latency etc.) would influence the time perception process. In this experiment, we tried our best to control these factors and findings in this work may provide guidance for further experimental design in a more practical environment.

7. ACKNOWLEDGMENTS

This work was supported by Natural Science Foundation (61622208, 61532011, 61472206) of China and National Key Basic Research Program (2015CB358700).

8. REFERENCES

[1] L. Anderson, D. Krathwohl, and B. Bloom. *A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives*. Longman, 2001.

[2] I. Arapakis, X. Bai, and B. B. Cambazoglu. Impact of response latency on user behavior in web search. In *SIGIR '14*.

[3] J. A. Aslam, M. Ekstrand-Abueg, V. Pavlu, F. Diaz, and T. Sakai. Trec 2013 temporal summarization. In *TREC '13*.

[4] L. Azzopardi. The economics in interactive information retrieval. In *SIGIR '11*.

[5] L. Azzopardi and G. Zuccon. An analysis of theories of search and search behavior. In *ICTIR '15*, pages 81–90. ACM, 2015.

[6] K. Berberich, S. Bedathur, O. Alonso, and G. Weikum. A language modeling approach for temporal information needs. In *ECIR '10*.

[7] R. A. Block and D. Zakay. Prospective and retrospective duration judgments: A meta-analytic review. *Psychonomic bulletin & review*, 4(2):184–197, 1997.

[8] R. A. Block and D. Zakay. Psychological time at the millennium: Some past, present, future, and interdisciplinary issues. *Time: Perspectives at the millennium (The study of time X)*, 2001.

[9] A. Borisov, I. Markov, M. de Rijke, and P. Serdyukov. A context-aware time model for web search.

[10] R. Capra, J. Arguello, A. Crescenzi, and E. Vardell. Differences in the use of search assistance for tasks of varying complexity. In *SIGIR '15*.

[11] S. Cheng, A. Arvanitis, and V. Hristidis. How fresh do you want your search results? In *CIKM'2013*.

[12] C. L. Clarke, M. D. Smucker, and E. Yilmaz. Ir evaluation: Modeling user behavior for measuring effectiveness. In *SIGIR '15*.

[13] C. L. A. Clarke and M. D. Smucker. Time well spent. In *IIiX '14*.

[14] C. W. Cleverdon and M. Keen. Aslib cranfield research project-factors determining the performance of indexing systems; volume 2, test results. 1966.

[15] A. Crescenzi, R. Capra, and J. Arguello. Time pressure, user satisfaction and task difficulty. In *Proceedings of the 76th ASIS&T Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries*, ASIST '13, pages 122:1–122:4, Silver Springs, MD, USA, 2013. American Society for Information Science.

[16] A. Crescenzi, D. Kelly, and L. Azzopardi. Impacts of time constraints and system delays on user experience. In *CHIIR'2016*.

[17] A. Crescenzi, D. Kelly, and L. Azzopardi. Time pressure and system delays in information search. In *SIGIR'2015*.

[18] M. Czerwinski, E. Horvitz, and E. Cutrell. Subjective duration assessment: An implicit probe for software usability. In *IHM-HCI'2001*.

[19] N. Dai, M. Shokouhi, and B. D. Davison. Learning to rank for freshness and relevance. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 95–104. ACM, 2011.

[20] W. Dakka, L. Gravano, and P. G. Ipeirotis. Answering general time sensitive queries. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 1437–1438, New York, NY, USA, 2008. ACM.

[21] J.-C. Dreher, A. Meyer-Lindenberg, P. Kohn, and K. F. Berman. Age-related changes in midbrain dopaminergic regulation of the human reward system. *Proceedings of the National Academy of Sciences*, 105(39):15106–15111, 2008.

[22] S. Droit-Volet, S. L. Fayolle, and S. Gil. Emotion and time perception: effects of film-induced mood. *Frontiers in integrative neuroscience*, 5, 2011.

[23] A. Edwards and D. Kelly. How does interest in a work task impact search behavior and engagement? In *CHIIR '16*, pages 249–252. ACM, 2016.

[24] M. Efron and G. Golovchinsky. Estimation methods for ranking recent information. In *SIGIR '11*.

[25] C. Fortin and R. Rousseau. Interference from short-term memory processing on encoding and reproducing brief durations. *Psychological Research*, 61(4):269–276, 1998.

[26] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Transactions on Information*

Systems (TOIS), 23(2):147–168, 2005.

[27] P. Fraisse. Perception and estimation of time. *Annual review of psychology*, 35(1):1–37, 1984.

[28] N. Fuhr. A probability ranking principle for interactive information retrieval. *Information Retrieval*, 11(3):251–265, 2008.

[29] P. A. Gable and B. D. Poole. Time flies when you're having approach-motivated fun effects of motivational intensity on time perception. *Psychological science*, page 0956797611435817, 2012.

[30] S. Grondin. *Psychology of time*. Emerald Group Publishing, 2008.

[31] S. Grondin and M. Plourde. Judging multi-minute intervals retrospectively. *The Quarterly Journal of Experimental Psychology*, 60(9):1303–1312, 2007.

[32] D. Gupta and K. Berberich. Identifying time intervals of interest to queries. In *CIKM '14*.

[33] A. Hassan. A semi-supervised approach to modeling web search satisfaction. In *SIGIR '12*.

[34] A. Hassan and R. W. White. Personalized models of search satisfaction. In *CIKM'13*.

[35] R. E. Hicks, G. W. Miller, G. Gaes, and K. Bierman. Concurrent processing demands and the experience of time-in-passing. *The American Journal of Psychology*, pages 431–446, 1977.

[36] B. J. Jansen, D. Booth, and B. Smith. Using the taxonomy of cognitive learning to model online searching. *Information Processing & Management*, 45(6):643–663, 2009.

[37] K. Järvelin, P. Vakkari, P. Arvola, F. Baskaya, A. Järvelin, J. Kekäläinen, H. Keskustalo, S. Kumpulainen, M. Saastamoinen, R. Savolainen, et al. Task-based information interaction evaluation: The viewpoint of program theory. *ACM Trans. Inf. Syst.*, 33(1):3:1–3:30, Mar. 2015.

[38] J. Jiang, A. Hassan Awadallah, X. Shi, and R. W. White. Understanding and predicting graded search satisfaction. In *WSDM '15*.

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

[40] N. Kanhabua and K. Nørvåg. Learning to rank search results for time-sensitive queries. In *CIKM'2012*.

[41] 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*.

[42] Y. Kim, A. Hassan, R. W. White, and I. Zitouni. Modeling dwell time to predict click-level satisfaction. In *WSDM '14*.

[43] Y. Li and N. J. Belkin. A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management*, 44(6):1822 – 1837, 2008.

[44] C. Liu, J. Liu, N. Belkin, M. Cole, and J. Gwizdka. Using dwell time as an implicit measure of usefulness in different task types. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–4, 2011.

[45] J. Liu, C. Liu, M. Cole, N. J. Belkin, and X. Zhang. Exploring and predicting search task difficulty. In *CIKM '12*.

[46] C. Luo, Y. Liu, T. Sakai, K. Zhou, F. Zhang, X. Li, and S. Ma. Does document relevance affect the searcher's perception of time? In *WSDM'17*.

[47] C. Luo, F. Zhang, X. Li, Y. Liu, M. Zhang, S. Ma, and D. Yang. Manipulating time perception of web search users. In *CHIIR '16*.

[48] 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?

[49] D. Metzler, R. Jones, F. Peng, and R. Zhang. Improving search relevance for implicitly temporal queries. In *SIGIR'09*.

[50] N. Mishra, R. W. White, S. Ieong, and E. Horvitz. Time-critical search. In *SIGIR'2014*.

[51] M.-H. Peetz, E. Meij, and M. de Rijke. Using temporal bursts for query modeling. *Information Retrieval*, 17(1):74–108, 2014.

[52] P. Pirolli and S. Card. Information foraging. *Psychological review*, 106(4):643, 1999.

[53] T. Sakai and Z. Dou. Summaries, ranked retrieval and sessions: A unified framework for information access evaluation. In *SIGIR '13*.

[54] M. D. Smucker and C. L. Clarke. Time-based calibration of effectiveness measures. In *SIGIR '12*.

[55] M. Sucala, B. Scheckner, and D. David. Psychological time: interval length judgments and subjective passage of time judgments. *Current psychology letters. Behaviour, brain & cognition*, 26(2, 2010), 2011.

[56] J. Teevan, K. Collins-Thompson, R. W. White, S. T. Dumais, and Y. Kim. Slow search: Information retrieval without time constraints. In *HCIR'2013*.

[57] D. K. Tse and P. C. Wilton. Models of consumer satisfaction formation: An extension. *Journal of marketing research*, pages 204–212, 1988.

[58] J. H. Wearden and I. S. Penton-Voak. Feeling the heat: Body temperature and the rate of subjective time, revisited. *The Quarterly Journal of Experimental Psychology*, 48(2):129–141, 1995.

[59] B. B. Weybrew. The zeigarnik phenomenon revisited: Implications for enhancement of morale. *Perceptual and Motor Skills*, 58(1):223–226, 1984.

[60] D. Zakay. Relative and absolute duration judgments under prospective and retrospective paradigms. *Perception & psychophysics*, 54(5):656–664, 1993.