User Behaviour and Task Characteristics: A Field Study of Daily Information Behaviour

Jiyin He
Centrum Wiskunde & Informatica
Science Park 123
1098XG Amsterdam
j.he@cwi.nl

Emine Yilmaz
University College London
66-72 Gower Street
London WC1E 6EA
emine.yilmaz@ucl.ac.uk

ABSTRACT

Previous studies investigating task based search often take the form of lab studies or large scale log analysis. In lab studies, users typically perform a designed task under a controlled environment, which may not reflect their natural behaviour. While log analysis allows the observation of users' natural search behaviour, often strong assumptions need to be made in order to associate the unobserved underlying user tasks with log signals.

We describe a field study during which we log participants' daily search and browsing activities for 5 days, and users are asked to self-annotate their search logs with the tasks they conducted as well as to describe the task characteristics according to a conceptual task classification scheme. This provides us with a more realistic and comprehensive view on how user tasks are associated with logged interactions than seen in previous log- or lab-based studies; and allows us to explore the complex interactions between task characteristics and their presence in naturalistic tasks which has not been studied previously.

We find a higher number of queries, longer timespan, as well as more task switches than reported in previous log based studies; and 41% of our tasks are zero-query tasks implying that large amounts of user task activities remain unobserved when only focused on query logs. Further, tasks sharing similar descriptions can vary greatly in their characteristics, suggesting that when supporting users with their tasks, it is important to know not only the task they are engaged with but also the context of the user in the task.

Keywords

Task-based search, field study, search log analysis

1. INTRODUCTION

Users of search systems often have a specific task in mind, ranging from simple fact look-up to complex quests, e.g., organising a wedding. User interactions with systems during tasks are recorded in logs, however, logs do not typically capture information about which interactions belong to which tasks; nor the specific characteristics, e.g., difficulty, of a task. Therefore little is known about when users start, stop, switch, or resume tasks, and in order for systems to better support users' information seeking processes in a task-aware manner, it is necessary to understand the relation between task characteristics, how they affect users' behaviour, and how this is reflected in logs.

Tasks in search have received increasing attention as a factor affecting user search behavior and satisfaction [15]. Early work in task-based search has focused on conceptualising tasks in the context of information seeking [5, 6, 15, 29] and characterising tasks along several dimensions [22]. The resulting conceptual models and classifications of tasks have provided a theoretical framework for the discussion of tasks and task characteristics.

To study specific characteristics (aspects) of tasks, lab user studies observe user behaviour during artificial tasks in a controlled setting. Analyses of interaction logs combined with interviews, questionnaires, or think aloud observations have provided rich contextual information about the relation between users' behavior and specific task characteristics. However, findings are limited to the tasks designed by the experimenters and the time spent in the lab, which may not necessarily reflect users' natural behaviour. Alternatively, field studies observe user behavior when engaging with tasks in a naturalistic setting [17]. However, these studies often focus on a single type of task, e.g. a course project [4].

Finally, large scale log analysis allows observing and identifying activity patterns of users naturally engaging in tasks at scale. However, the actual tasks users engage in are unobserved and strong assumptions need to be made by the analyst to associate tasks and log signals. Further, current log-based studies are typically focused on the level of "search tasks", e.g., by defining user tasks as topically coherent query sequences [13, 16, 21, 24]; few studies go beyond search tasks to consider tasks at a level of information seeking tasks (e.g., [2]) or work tasks [5].

In this study, we combine the log analysis and field study approaches to gain insights in the relation between users' searching and browsing behaviour and their tasks in a naturalistic setting. With user consent, we install a browser plug-in on participants' computer and record their daily searching and browsing activities for 5 days. Each day users are asked to annotate their own logs with the tasks they engaged in. This results in a data set that provides a more accurate view on how user tasks are associated with interactions recorded in search logs than seen in previous studies. By further asking participants to characterise their tasks according to the conceptual classification scheme of task facets and attributes [22], we obtain a holistic view of the relation between user tasks and task characteristics that was previously not possible due to the limitations of both in-lab user studies and log analysis.

We relate and contrast our results to some of the assumptions and observations made in previous log based studies on short term tasks, seeking answers to the research question (Section 4):

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

ISBN 978-1-4503-4677-1/17/03. . . $15.00

DOI: http://dx.doi.org/10.1145/3020165.3020188

RQ1 From a user annotated search log with records of rich types of interactions, do we observe similar statistics of task sessions as those observed from an expert annotated "query-only" log?

Further, by performing an exploratory analysis of the interactions between a wide range of task characteristics and their presence in users' naturalistic tasks, we discuss (in Section 5)

RQ2 How do task characteristics relate to each other? and how do these characteristics co-occur within actual Web user tasks?

Our contributions are as follows. In terms of methodology, we contribute an alternative approach to study the tasks underlying users' daily online activities, enabling insights that were not possible before. With this approach, (i) we add to the existing log-based studies with new insights obtained with logs that contain rich types of interactions and accurate task annotations; and (ii) we add to the existing results of in-lab/field studies by analysing a comprehensive set of task characteristics not limited to specific task designs and characteristics. Further, we discuss the implications of these results for future log based studies and in-lab experiment designs.

2. RELATED WORK

Recent studies on task-based information retrieval generally fall into three categories: lab user studies, large scale log analyses, and field studies. Below, we review and discuss findings from work within each category as well as how these relate to our work.

Lab studies. Lab user studies provide experimenters with control over the experimental setting and have often been employed to gain a better understanding of how different task characteristics affect users' information seeking behaviours. In these studies users typically perform tasks that are carefully designed by the experimenter so that the tasks carry the characteristics of interest. For example, Liu et al. [23] studied the associations between measures of user search behaviours with task characteristics along four dimensions: task complexity, product, goals, and levels. In a journalism scenario, four search tasks were designed to reflect varying conditions in these dimensions.

Task complexity is a frequently studied dimension. In a study of factual information finding tasks with varying complexity, Gwizdka and Spence [12] found that task difficulty is correlated with the number of unique pages visited, time spent on each page, deviation from the optimal path, and the linearity of the navigation path. Task complexity was found to affect the relative importance of these factors, and to affect user perceived task difficulty.

Based on education theory Kelly et al. [19] proposed a framework for the design of tasks with varying levels of cognitive complexity to support design and evaluation of IIR experiments. By examing behavioural and self-reported user data from a lab study, they found that cognitively more complex tasks require significantly more search activity from participants, but participants do not perceive cognitive complex tasks as more difficult, and were satisfied with their performance across tasks.

In a large scale user study of struggling behaviour during search, Aula et al. [1] found that users tend to formulate more diverse queries, use more advanced operators and spend longer time on the search result pages as compared to the successful tasks.

While carefully designed tasks performed in a lab setting allows testing specific hypotheses under controlled conditions, it also limits the generalizability of the findings towards users' search behaviour in a naturalistic setting. For example, it is difficult to observe interactions between tasks such as the multi-tasking behaviour of users, which are prevalent in online search [13, 20, 24, 28] or tasks that take longer than a typical lab study session. Moreover, while it is possible to control a particular aspect of the designed tasks, it is almost impossible to control multiple task characteristics simultaneously and design their ways of interactions that naturally occur in reality. In our study we monitor users remotely and we do not ask users to engage in specific tasks in order to observe users' natural search behavior.

Large scale log analysis. As an alternative to lab user studies, analyses of large scale search logs—typically containing user search queries and activities on the search result page—allow making observations of users' natural search behaviour.

One of the first challenges here is the identification of the user tasks from the search logs. Jones and Klinkner [16] argued that traditional time-out based search session segmentation is not sufficient as a criterion for identifying sessions of user tasks. Based on annotations from external assessors[1] classifiers were trained to identify task (mission) boundaries between consecutive query pairs.

A number of studies have focussed on improving the recall in task identification by tackling semantic relations between queries, task interleaving, and session boundary detection. Lucchese et al. [24] studied two unsupervised methods to identify task sessions from query logs: a time-based thresholding heuristic method, and a clustering based method. Similar to [24], Li et al. [21] proposed an algorithm that combines topic models with Hawkes processes to identify search tasks from query logs, based on the assumption that queries belonging to the same tasks are temporally close and topically coherent. Hagen et al. [13] proposed a 2-stage method that identifies search missions from query logs in 6 steps. Unlike [24] which focused on identifying task sessions, this study considers interleaved task sessions that belong to the same mission. To verify the algorithms, a sample of query logs from a commercial search system were annotated by external annotators [13, 24].

In a different setup, Kotov et al. [20] proposed a classification based method to (1) identify related queries from previous sessions for a given query, and (2) to predict whether a user will return to the same task in the future for a given multi-query task. User searching and browsing episodes were recorded with a browser plug-in. For ground-truth collection, automatic labelling of search tasks was followed by a manual correction process. In a related study, Awadallah et al. [2] proposed a method to mine search tasks from search logs. An association graph was constructed to connect multiple tasks and used to recommend a set of interesting and diverse tasks to support searchers during complex search tasks.

Although significant advances have been made in terms of task detection in logs, a limitation of this type of work is that external annotators need to create ground truth data to allow training and evaluation. Further, often strong assumptions are made in order to select a subset of data, e.g., sessions with a certain length [25] or users that visited a particular site [9], for annotators to process. However, Russell et al. [26] noted in a study where self-labelled task sessions were compared to labels provided by external annotators: "*the most accurate labelling of search task session data is done by the searchers themselves, and that it is very difficult for an external observer or automatic classifier to infer where the task boundaries are or what the actual user task goal is.*"

Mechanical Turk has been proposed as a middle ground between log and lab studies [32]. However, existing crowdsourcing platforms do not overcome the limitations of experimenters designing tasks for workers and imposing time (or monetary) constraints on performing tasks, making such a set up unsuitable for our purposes.

[1] We refer to annotators or assessors other than the user him/herself as "external" annotators.

In our study we ask participants to annotate their own search logs in order to obtain both naturalistic observations of search behaviour as well as accurate task labels associated with the behaviour.

Field studies. In addition to lab studies and log analysis, field or longitudinal studies are another approach often employed in investigating user information seeking behaviour. These studies typically focus on a specific task in a particular setting. For instance, the change of user search terms and tactics during the writing of a research proposal [30], or users' use and preferences of different search interfaces [4] and query reformulation patterns [14] during a study course.

In a longitudinal study of users' online information seeking behaviour, Kelly and Belkin [18] found that document display times differ significantly according to specific tasks and specific users. A following up study concludes that tailoring display time thresholds based on task information improves implicit relevance feedback [31]. Greifeneder [11] studied effect of distraction on users' task completion behaviours in a naturalistic environment and made observations contradictory to previous findings from in-lab studies. In the mobile context, Church et al. [7] conducted a large scale study over a three month period, aimed at a comprehensive understanding of people's daily information needs, and how these needs are addressed and influenced by contextual, techonological, and demograhpical factors.

In a study closely related to ours [17], the Web usage activities of 21 participants were logged, and participants were asked to categorise their Web usage into 5 categories: fact finding, information gathering, browsing, and transactions. Analyses show that user behaviour differs between task types in terms of measures such as dwell time, number of pages viewed, and user browser activities.

We take a similar field study and diary based approach. The differences being: (1) instead of categorising tasks into 5 general types, we study task characteristics in greater detail by following the faceted task categorisation scheme introduced by Li and Belkin [22]; (2) while Kellar et al. [17] has focused on users' Web usage activities, we focus on users' search and browsing behaviour during information seeking tasks and relate our findings to some of the recent lab and log studies discussed above.

Theoretical perspectives. Apart from empirical studies, a large body of work in information studies addresses task-based information seeking from a theoretical perspective. These studies provide various alternative frameworks for the characterisation of different types of search and work tasks.

Different perspectives of tasks and consideration of different levels of task granularity have led to different definitions of tasks [5, 15, 29]. Byström and Hansen [5] suggest that tasks can be studied at three levels: work tasks, information seeking tasks, and information search tasks, with the latter being sub-tasks of the former. In the context of this study, the user tasks we investigate can be seen as somewhere in between the level of work tasks and information seeking tasks, which may be further divided into smaller information search tasks, e.g., through searching in different types of information, subject topics, and sources.

Other theoretical frameworks categorise tasks based on characteristics. Li and Belkin [22] reviewed and compared several task categorisation schemes, and proposed a faceted task classification framework, which applies to work tasks as well as search tasks. In this study, we follow this classification scheme for the purpose of characterising the user tasks identified from the user logs.

In addition, log-based studies often have their own working definitions regarding task and task search sessions, making a direct comparison of their observations difficult. We therefore conduct a conceptual comparison of these different concepts from three representative studies [13, 16, 24] (Section 4), based on which we discuss and contrast our observations to those reported previously.

3. USER STUDY

3.1 Study procedure

The study consisted of three stages as detailed below.

Introduction session (30 min). In this session, participants were invited to the lab brining in their own laptops. We explained the procedure of the study and installed the logger on the participant's laptop. The participant was asked to fill in a pre-questionnaire to collect demographic information. A training session was then provided to familiarise participants with the logging software and the annotation tool used to annotate their logs with task labels.

Diary study (5 days). After the introduction session, participants were told to go home and use their computers as usual. Meanwhile, the logger automatically recorded their online searching and browsing activities. By the end of each day, the participants were asked to review their logs in order to: (1) remove any log entries that they did not wish to share, and (2) to associate each search/browsing event with the task they were engaged in. Allowing participants to review the information that is being shared improves a bond of trust between experimenter and participant and improves the chances of participants completing the study [27]. Participants could review and annotate their logs at any time during the experiment period. Kellar et al. [17] found in their field study that there is equal user preference between real time and delayed annotation.

De-brief session (30 min). After 5 days, participants were invited back to the lab for a debrief session. In this session, participants were asked to select 5 to 10 tasks of their choice, and to describe the characteristics of these tasks by filling in a post questionnaire.

3.2 The logger

We log both search and browsing activities of the participants. To this end, we developed a Chrome extension that records events triggered by browser operations, and information attached to these events such as user input. Four types of events are recorded (see below), each associated with a timestamp, a URL, and a tab ID. Participants can submit a blacklist of URLs or URL prefixes to prevent events with matching URLs to be recorded.

Search events. In terms of search events, we recorded participants' activities with three major search engines: Google, Bing, and Yahoo. Related information includes: the query issued, the type of vertical (e.g. image, news, etc.), and the engine used. We consider the events triggered by issuing a query or switching between verticals as an individual "search event". Of course, users can perform search in other sites such as in an organisation's intranet or dedicated domain like Amazon. However, it is not always possible to recognise search events from each of those sites due to their diverse interface designs. Therefore, in this study we focused on Web search activities, with the belief that the majority happen with the three major Web search engines.

Link click events. We identify two types of link clicks: clicks on a result in a SERP, and clicks on a regular page leading to an external or internal target page. Information related to click events includes the clicked anchor text and the linked target page URL.

Tab events. We also record user operations on tabs, including: open-a-new-tab, close-a-tab, switch-to-a-tab, open-a-link-in-new-tab, as well as the tab-loading status. This information helps us to determine for example whether a user has "viewed" a page or the dwell time a user has been on a page. For instance sometimes users

Task characteristics	Description	Values
Frequency (FQ)	How frequent would you say the following task have occurred? (Not limited to the experiment period, think about e.g. in the last year).	(1) One-time task—Routine tasks (5)
Length (TL)	How quickly do you think the following task can be finished?	(1) Very quick (< 1 day)—long term (≥ 1 month) (5)
Stage (STG)	To what extend did you manage to complete the task so far?	(1) Just started—(Almost) finished (5)
Cognitive level (CL)	Different tasks involve cognitive activities of different levels of complexity. At which level would you rate the activities involved to complete the following task?	(1) Remember; (2) Understand; (3) Apply; (4) Analyse; (5) Evaluate; (6) Create. (For each option, we provided explanations following [19]).
Collaboration (COL)	To what extend would you say you were responsible for the task?	(1) Solely responsible—Collaborates with many people (5)
Importance (IMP)	How would you rate the importance of the task?	(1) Unimportant—Extremely important (5)
Urgency (UR)	How would you rate the urgency of the task?	(1) Not urgent—Extremely urgent (5)
Difficulty (DIF)	How do you feel about the difficulty of the task? (e.g. difficult to find relevant information, or requires great effort in thinking/understanding).	(1) Easy—Extremely difficult (5)
Complexity (COM)	How do you feel about the complexity of the task? (e.g. it may involve many steps or subtasks in order to complete the task).	(1) Simple—Extremely complex (5)
Knowledge of topic (KT)	How would you rate your knowledge on the topic of the task?	(1) No knowledge—Highly knowledgeable (5)
Knowledge of procedure (KP)	How would you rate your knowledge on the procedure to complete the task?	(1) No knowledge—Highly knowledgeable (5)
Satisfaction (SAT)	Were you satisfied with the process of information seeking activities for completing the task?	(1) Unsatisfied, unable to find the information needed—Very satisfied, could find needed information easily (5)

Table 1: The characteristics of user tasks enquired in the post-questionnaire.

open links in new tabs, but stay on the original page and would not see the content of the linked page.

Navigation events. Navigation events contain information about how users arrive at a page, e.g. via link, via direct URL input, by form submission, or by forward/backward navigation. We keep this information to enrich the context of the rest of the logged events. For instance, it allows us to observe how much time the user has spent on a page based on the switches between different tabs.

3.3 Log review and task annotation

Annotation tool. To facilitate the diary study, we developed an online annotation tool. Figure 1 shows the annotation interface.

Participants review their log entries in the log review panel (1) and remove any items that they would not like to be recorded. In order to improve the readability of the log, only two types of events are shown to the participants: search queries and visited pages. Once a query or page visit is marked to be removed, related events with the same URL and tab ID are also removed from the log.

To annotate log entries with tasks, participants first need to enter a list of tasks in the task editing area (2), which works like a to-do list. The tasks entered by participants show up automatically as candidate labels under each log entry. Participants can then indicate the task that is associated with a given log entry by clicking the corresponding label. Participants can edit the task list before or during the annotation process.

To speed up the annotation process, participants can use the batch operation toolbar (3) to search and filter entries, and to batch remove and label selected entries. With batch operations it typically takes about 15 to 30 minutes for a participant to annotate a one-day log. Participants can review their annotation progress in the progress bar by selecting the corresponding date (4).

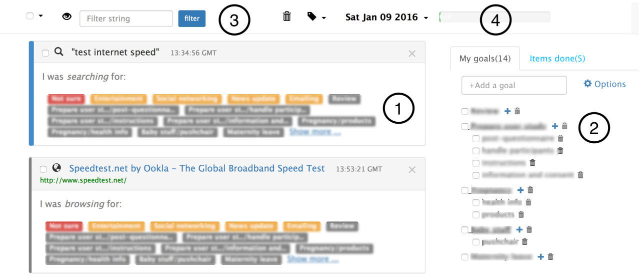

Filter string
filter
3
Sat Jan 09 2016
4
"test internet speed" 13:34:56 GMT
I was *searching* for:
1
Speedtest.net by Ookla – The Global Broadband Speed Test 13:53:21 GMT
http://www.speedtest.net/
I was *browsing* for:
My goals(14)
Items done(5)
+Add a goal
Options
2

Figure 1: Annotation tool for diary study: (1) log review; (2) task list; (3) batch annotation tool bar; (4) progress bar. Items in task list and candidate task labels are blurred out for privacy reasons.

Annotation instructions. To encourage participants to think of the notion of "tasks" at a level comparable to that are typically considered in the literature, in the introduction session we asked the participants to think of 3 to 5 tasks they plan to conduct, or have been recently engaged with and will continue to conduct in the next few days; we further gave example tasks such as "planning a vacation", "write a report".

On the other hand, not all search/browsing activities can be categorised to a specific named task. Apart from participant specified tasks, we also provide pre-defined task labels for typical online activities not always considered as a work task or search task, including: Emailing (001), Social networking (002), Entertainment (003), and News update (004), as well as a "Not sure" label (000) in case participants do not remember what they were doing.

3.4 Task characteristics

Questionnaire. The classification scheme by Li and Belkin [22] provides us with a rather comprehensive list of both objective and subjective properties that can be used to characterise an information task. Following this scheme, we translated the identified facets and attributes of tasks into a post-questionnaire.

In order to reduce participants' effort both in understanding and in making answers, we made the following adaptations to this classification scheme: (1) We excluded attributes that can be derived from the log directly, e.g. whether a task contains multiple subtasks/goals. (2) After pilot testing we excluded questions that participants feel difficult to interpret and answer. In particular, we replaced the "product" facet and the "objective complexity" attribute from Li and Belkin [22]'s scheme by the 6-level cognitive task complexity framework proposed by Kelly et al. [19], as this framework provides clear instructions for participants to categorise the cognitive activities involved for a task. Also, the outcome of these cognitive activities resembles the task outcomes described by the "product" facet, but with more consistent categories. In total we derived 12 questions (see Table 1).

Participant instructions. Participants were asked to rate their answers or perceptions of these characteristics on a 1-5 point Likert style item (1 - 6 levels in the case of rating the cognitive complexity levels). In all cases, the ratings can be treated as ordinal variables.

To limit participants' effort, they are asked to select 5 - 10 tasks of their choice and only answer questions for these selected tasks.

According to Li and Belkin [22], the first 5 characteristics (FQ—COL) are objective, while the latter 7 (IMP—SAT) are subjective. In order to help participants to answer these objective questions in a comparable manner, we provided detailed definitions for the options (see Table 1), and provided further explanation when needed.

3.5 Data obtained

We recruited participants by distributing flyers on the university campus. Participants volunteered to take part in the study and were told that they are free to quit at any moment. Each participant was paid an Amazon voucher worth 20 pounds for participating.

A total of 23 participants completed the study, including 13 males and 10 females; and 11 are between 18 - 24 years old, and 12 are between 25 - 34 years old. Participants also self-rated their experience of using search engines on a 1-5 scale (median = 5, IQR = 1.0). In addition, as people often use multiple devices, it is likely that what we observe during the study is only a sample of users' actual information seeking activities if participants spend most of their time searching on other devices such as mobiles. Among the 23 participants, 18 (78 %) indicated that they always or mostly search with their laptop computer, while 5 (22 %) indicated that about half of their search happens with mobiles/tablets.

There are 289 user defined tasks, i.e., which excludes tasks that were annotated with general labels such as "emailing" as defined in Section 3.3. Among those 17 have subtasks. Further, 135 tasks were selected by participants to provide characteristics using the post-questionnaire. In total, 2566 queries and 32,902 page visits were annotated, where 1768 queries and 17,313 page visits were annotated with user defined tasks.

4. USER TASK ACTIVITIES IN LOGS

A key differences between our study and previous studies on search log analysis of user tasks is that, we analyse a log that contains rich types of user activities annotated by task doers, as compared to a *query-only* log annotated by experimenters. In this section we investigate whether, and if so how, tasks annotated by users themselves leads to new observations in the scope of tasks and observed statistics in log analysis (**Q1**).

We chose three studies for this investigation, namely Jones and Klinkner [16], Lucchese et al. [24], and Hagen et al. [13], for the following reasons. (i) All studies describe tasks in terms of search sessions, and reported a number of common measures to characterise user task behaviours. (ii) They all used human external assessors to annotate tasks in search logs—in particular, [16] attempted to recreate the user experience for the external annotators. Both factors make these studies more comparable among each other, as compared to, e.g. studies using alternative annotation methods.

We start with a conceptual exploration of the notion "task" as defined in these studies, followed by an empirical comparison of the observations made from these studies. On the one hand, this helps position our work in the literature and provides context for our findings, as in, to which extend our "tasks" correspond to the tasks discussed in previous studies. On the other hand, this provides implications for reading the conclusions drawn from previous studies as well as the hypotheses considered and assumptions made when training and validating various task identification algorithms.

4.1 Task & sessions: a conceptual exploration

Each of the studies defined a set of concepts to operationalize user tasks in terms of various types of search session. While these concepts share similarity across studies, different terminology has been used. (Explained below.) In Table 2 we attempt to provide a loose mapping between the concepts defined in different studies.

Concepts	Physical session	Logical session	(Complex) task
Definition	All user queries or activities within a time window.	Consecutive queries belonging to the same task [13].	A set of related information needs span over one or more logical sessions.
Terminology			
Jones [16]	session	goal	mission
Lucchese [24]	time-gap session	task session	–
Hagen [13]	physical session	logical session	mission
Ours	physical session	logical/subtask	task

Table 2: A mapping between concepts as used in the literature.

Physical session. Physical sessions are typically defined by a timeout threshold on user inactivities. For instance [24] set a threshold of 26 minutes and [13] set a threshold of 90 minutes on time-gaps between queries, with the assumption that users are likely to switch tasks after a long pause. Like [13], we take a conservative threshold of 90 minutes as threshold to determine physical sessions, simply assuming that users have left the session (to do something else) after such a long inactivity.

Logical session. Although named differently, in all three studies, a logical session consists of queries related to a same information need. However, each study has a slightly different version of this concept. In [13] it consists of consecutive queries, and in [24] non-consecutive queries, within a physical session. The related concept "goal" in [16], however, consists of queries that are neither consecutive nor within the same physical session. In our study, we refer "logical session" to consecutive queries related to the same task within a physical session (cf. [13]). The notion of "goal" in [16] is similar to our notion of "subtask". However, our "subtask" does not necessarily represent an *atomic* information need [16], as it can still be a complex information need. For example, one of our participants wrote "plan vacation trip Christmas" as the main task, and "find out how to go from Denmark to Norway" as one of the subtasks, which is a rather open question.

(Complex) task. This concept corresponds to the notion of "mission" in both [16] and [13], and we simply refer to it as "task". It does not correspond to a concept in [24], as the "task-session" defined in [24] does not go beyond a physical session. In practice, we see that while physical sessions do not always represent units of user tasks, both logical sessions and subtasks (or goals as defined in [16]) are units that constitute a complex task. While logical sessions can be directly observed from a user annotated log, in our study very few participants actually decided to detail the subtasks (only 17 out of 289 tasks have subtasks specified).

4.2 Task and sessions: empirical analysis

When analysing user task behaviour in search logs, all three studies reported the following measures: the number of queries issued and the timespan over task sessions, as well as the relation between physical sessions and user tasks. In this section, we report analyses over the same set of variables and seek to answer RQ1.

4.2.1 Task session statistics

We measure the number of queries issued and the timespan at two levels, namely the logical session level and the task level. We filter out logical sessions that have a duration of 0 seconds (e.g. the user closed a tab related to a task when working on another task) as in such cases the user was not really working on the related task. We compute a task session as the collection of all logical sessions with the label of that task. Table 3 lists the result of these measures, as compared to those reported in the previous studies.

Dataset	per Logical sessions		per Tasks	
	# queries	timespan	# queries	timespan
Jones [16]	–	–	2 (md)	38 secs (md)
Lucchese [24]	2.57 (m)	–	–	–
Hagen [13]	–	–	6.42(m)	–
Ours (all)	0 (md)	12.9 secs (md)	1 (md)	10.5 mins (md)
	0.28 (m)	112 secs (m)	6.79 (m)	44.8 mins (m)
Ours (queries)	1 (md)	19.3 secs (md)	4 (md)	15.7 mins (md)
	1.66 (m)	134 secs (m)	11.5 (m)	57.9 mins(m)

Table 3: A comparative view of user task behaviour at the level of logical sessions and tasks. (md) refers to median, and (m) refers to mean. Ours (queries) refers to statistics computed over our data excluding zero-query sessions/tasks.

Number of queries. One immediate observation from Table 3 is that our median number of queries per logical session is 0. Indeed, a further examination shows that 82% of the logical sessions, and 41% of the user tasks, do not contain query requests. Unlike our annotation which includes user activities other than query requests, analysis of the three previous studies are based on *query logs*, hence the effect of zero-query sessions would not have been observed.

Looking at logical sessions and tasks that do contain query requests, we see that in terms of logical sessions, the average number of queries (1.66) is lower than that of the task sessions reported in [24] (2.57). In terms of tasks, the median (4) and mean (11.5) number of queries per task are relatively high compared to those reported previously (median 2 [16] and mean 6.42 [13]). This implies that users may do more task switching than one would think, i.e. a task may contain many logical sessions (meaning tasks are being interrupted a lot) while each session contains few queries.

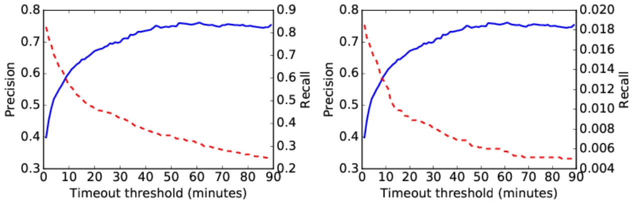
Timespan. We compute the timespan of a task as the sum of the duration of all logical sessions of this task. While there is no report on timespan of logical sessions for a direct comparison, we see that at task level, our observed task duration is much higher than what was reported previously (38 secs [16]). In fact, the task (mission) duration reported by Jones and Klinkner [16] is more comparable to the timespan of our logical sessions.

That our data show both more queries and longer timespan at the task level may be due to two reasons: (1) we have observations over a longer period (5 days), compared to 3 days reported in [16]; and (2) our data contains more than just querying activities, which contributes to the accumulated observed time spent on a task.

Task revisits. A common observation among the three studies is that, tasks (or goals and missions) are interleaved. Jones and Klinkner [16] reported 17% of the missions were revisited, and Hagen et al. [13] reported each task contains 2.09 logical sessions on average. In our data, we see that on average each task contains 23.9 logical sessions (median of 8.0), and among those 86% have more than one logical session (i.e. task was interrupted and revisited). Part of the difference in observation is due to the fact that our data contains task activities other than query events. If we only consider logical sessions with queries—as with a query log, we see that on average a task contains 6.9 logical sessions (median of 2.0), and about 68% of them were interrupted and revisited.

4.2.2 Physical sessions and task boundaries

When analysing physical sessions in relation to tasks, the main focus of the previous studies were whether a time-out threshold on gaps between user queries would signify switches of user tasks. Using the concepts discussed previously, it is equivalent to examine how well physical sessions (i.e. time-out based segmentation of logs) matches the logical sessions (i.e. task annotation based segmentation of logs). Although all studies have shown that applying a diverse set of features and learning algorithms can significantly improve the boundary detection accuracy, time-out remains an important feature, e.g. it alone achieves a F1-score of 0.65 [16].



(a) Query based task boundaries

(b) Event based task boundaries

Figure 2: Precision (blue solid line) and Recall (red dashed line) of time-out based task boundary identification.

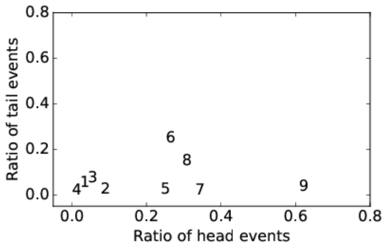
Following [16], we create physical sessions by applying varying time-out thresholds on time gaps between *user queries*, and inspect the precision and recall of each threshold in correctly identifying task boundaries. In terms of ground truth of task boundaries, previous studies were all focused on logical sessions that purely consist of queries. However, we have already seen that users switch tasks in between queries. We therefore also compare the segmentation results to *event-based* logical sessions that include queries as well as other user activities.

To compute precision and recall, we need to determine when a task boundary is correctly identified. When comparing against a query-based ground truth, we consider a boundary between two consecutive queries correct if the two queries belong to two different tasks. However, when computing recall using the event-based ground truth, we do not require an exact match of task boundaries. For example in an event sequence $q_1, e_1, ...e_i, q_2$, it may happen that the time-out threshold identifies that q_1 and q_2 are task boundaries, while the actual task switch happens at event $e \in e_1, ..., e_i$. We consider the task boundary between two consecutive queries correct if the two queries indeed belong to two different tasks and there is no more than 1 task switch on the other events between the two queries. This way we avoid over-counting failures to detect task boundaries.

Figure Fig. 2a and Fig. 2b show the effectiveness of thresholds on time-out between queries, compared against query-based and event-based ground truth, respectively. In terms of precision (obviously, the two setups should have the same precision), we see that our data generally agrees with what was reported in [16]. That is, as the threshold varies, the maximum precision can be achieved is in the range of 70% to 80%; and above certain threshold values e.g. 30 mins, there is no obvious change. While Jones and Klinkner [16] did not report recall, we see from Fig. 2a that reasonable recall values can be achieved; further, at a threshold of 5 minutes a maximum F1 score can be achieved at 0.59 (cf.0.65 [16]). The optimal threshold, as we have seen, differs from study to study (e.g. 26 minutes by [24] and 13 minutes by [16]). This can be both a result of differences in the logs (e.g. differences in how people use a particular search engine), as well as an artefact of how these studies pre-process the log—for instance in [24] annotators discarded queries that they considered as meaningless. Nevertheless, across all these studies, including ours, the achievable accuracy of time-out based task boundary identification is in a similar range.

On the other hand, from Fig. 2b we see that when comparing the query-based physical sessions to the event-based ground truth, recall significantly drops to the range below 0.02. That is, there is a majority of task switches happening in between queries that are missed out if we only look at queries to identify task switches.

Having observed that many task switches happen with user ac-



id	event type	% M
1	form submit	94.3
2	for/backward	91.9
3	link click	90.0
4	pagination	100
5	query	75.7
6	tab close	51.6
7	tab new	66.7
8	tab switch	57.3
9	go to url	37.3

Figure 3: Percentage of an event type being the head or tail of a logical session. Legend on the right side also lists the percentage of an event type being neither heads or tail (% M).

tivities other than querying, we further look into what other events may signify task switches. To this end, we compute the percentage that an event type is at the head (the first event), the tail (the last event), and in the middle of a logical session. Fig. 3 shows the result. Event types 1–4 allocated in the lower left corner of the plot are events that rarely occur as the boundary of a logical session (i.e. switch of task): >90% of the times they occur in the middle of a logical session. Indeed activities such as link clicks and paginations are most likely performed during a task. Event types 5–8 (issuing a query and various tab operations) are, however, more evenly distributed as being boundary- or non-boundary-events. In particular, events like issuing a new query (5) and opening a new tab (7) are more likely to signify a start of a new task than ending a task. In addition, when users directly go to a page by typing in the URL (9), it is quite likely to start a new task (about 60%).

5. TASK CHARACTERISTICS AND USER ACTIVITIES

While user activities can often be directly observed from logs, task characteristics are usually unavailable to search systems. Lab studies have therefore focussed on studying the relation between user activities and tasks designed to have specific characteristics. However, in labs studies typically a limited number of task characteristics is controlled for and a small set of tasks is employed. Therefore, we lack a holistic view of *how task characteristics relate to each other* and *how these characteristics co-occur within actual Web user tasks*. The naturalistic setup and the explicitly annotated task characteristics obtained in our study allow us to gain insights into the relations between task characteristics as well as to provide examples of tasks in which these characteristics naturally occur (**Q2**). We present our findings below.

5.1 Interactions between task characteristics

5.1.1 Clustered task characteristics

A correlation analysis provides insights in the degree of association between variables. Here we apply correlation analysis to the task characteristics of the 135 tasks that users annotated. We measure the correlation between two task characteristics using Kendall's τ.

Table 4 presents the correlations between the task characteristics. To facilitate the discussion, we cluster the characteristics. These clusters represent groups of task characteristics that often co-occur with a task. We employ Affinity Propagation [10] as the clustering method, given that it automatically searches for the appropriate number of clusters; and the absolute values of the correlation coefficients between two characteristics as the measure of similarity.

Table 5 shows the groups of task characteristics with strong mutual correlations. We observe that each group has a certain "theme".

Group	Member characteristics
1	CL, COM, DIF, TL, SAT
2	COL, KT, KP
3	IMP, STG, UG
4	FQ

Table 5: Clusters of task characteristics.

Group 1 consists of variables relating to task complexity (CL, COM), difficulty (DIF), expected length (TL), and user satisfaction (SAT). A set of variables that are often discussed together in studies (e.g., [19]). Group 2 concerns the knowledge of users (KT, KP), and collaborations on tasks (COL). Group 3 features the importance (IMP) and urgency (UR) aspects of the tasks, and their relation with task stages (STG). We discuss each group in more detail next.

G1: CL, COM, DIF, TL, SAT. Figure 4a shows the mutual correlation between the G1 variables. We observe that the level of cognitive complexity (CL), user perceived task complexity (COM), task difficulty (DIF), and expected task length (TL) have a significant positive correlation with each other. Further, user perceived satisfaction (SAT) has a significant negative correlation with the above four variables, indicating that the more difficult/complex/lengthy a task is, the less likely that the user is satisfied.

Interestingly, the observations here are somewhat contradictory to those reported by Kelly et al. [19], where users do not perceive a cognitively complex task more difficult or complex, and that users were equality happy with their performance across tasks of different CL levels. One possible explanation is that while tasks used in [19] were designed to control for levels of cognitive complexity, they were not necessarily controlled in terms of difficulty/complexity. Given the small sample of designed tasks it may not have been possible to detect variations across levels of difficulty.

G2: COL, KT, KP. We see (Figure 4b) that a user's knowledge of a task topic (KT) and task procedure (KP) has a significant positive correlation—which aligns with our intuition. Further, the level of task collaboration (COL) has a negative correlation with both KT and KP, i.e., the more collaboration is involved in a task, the less knowledge the user has about the topic or the procedure of the task.

G3: IMP, STG, UR. In Figure 4c we observe that user perceived task urgency (UR) has a significant positive correlation with user perceived task importance (IMP), as well as the task stage (STG). That is, the more important a task is, or the more advanced stage the task is at, the more urgent the user feels about the task.

5.1.2 Between group correlations

We now continue to examine the correlations between variables from different clusters, as illustrated in Figure 5.

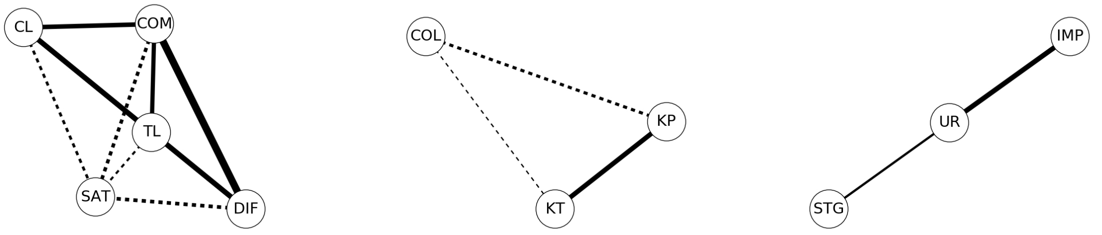
G1 - G2. From Figure 5a we see that both user knowledge of task topic (KT) and procedure (KP) have a significant negative correlation with perceived task complexity (COM), difficulty (DIF), expected task length (TL), and level of cognitive complexity (CL); but a significant positive correlation with their task satisfaction (SAT). These correlations feel intuitively right: the more users knows about the task (KT, KP), the less they perceive the task as difficulty/complex, and the more they are likely to be satisfied with the task.

Further, there is a significant positive correlation between task collaboration (COL) and perceived task difficult (DIF), complexity (COM), and length (TL), i.e., the more collaboration is involved in a task, the more likely that the task is perceived as complex/difficult.

G1 - G3. Between G1 and G3 (shown in Figure 5b), we observe that task stage (STG) has a significant negative correlation with all

	CL	COL	COM	DIF	FQ	IMP	KP	KT	SAT	STG	TL	UR
CL	--	0.01	**0.43****	**0.49****	-0.07	**0.30****	**-0.21****	**-0.18****	**-0.26****	**-0.19****	**0.47****	0.10
COL	0.01	--	**0.23****	**0.14***	-0.08	**-0.13***	**-0.29****	**-0.12***	-0.07	**-0.14***	**0.20****	-0.10
COM	**0.43****	**0.23****	--	**0.69****	-0.07	**0.37****	**-0.35****	**-0.25****	**-0.34****	**-0.15***	**0.40****	**0.24****
DIF	**0.49****	**0.14***	**0.69****	--	-0.10	**0.28****	**-0.37****	**-0.24****	**-0.35****	**-0.19****	**0.46****	**0.14***
FQ	-0.07	-0.08	-0.07	-0.10	--	**0.12***	**0.19****	**0.24****	**0.18****	0.05	-0.05	0.01
IMP	**0.30****	**-0.13***	**0.37****	**0.28****	**0.12***	--	-0.08	0.03	0.00	0.04	0.11	**0.47****
KP	**-0.21****	**-0.29****	**-0.35****	**-0.37****	**0.19****	-0.08	--	**0.45****	**0.33****	**0.13***	**-0.23****	-0.09
KT	**-0.18****	**-0.12***	**-0.25****	**-0.24****	**0.24****	0.03	**0.45****	--	**0.34****	0.09	**-0.22****	-0.00
SAT	**-0.26****	-0.07	**-0.34****	**-0.35****	**0.18****	0.00	**0.33****	**0.34****	--	**0.21****	**-0.20****	0.06
STG	**-0.19****	**-0.14***	**-0.15***	**-0.19****	0.05	0.04	**0.13***	0.09	**0.21****	--	**-0.44****	**0.21****
TL	**0.47****	**0.20****	**0.40****	**0.46****	-0.05	0.11	**-0.23****	**-0.22****	**-0.20****	**-0.44****	--	-0.00
UR	0.10	-0.10	**0.24****	**0.14***	0.01	**0.47****	-0.09	-0.00	0.06	**0.21****	-0.00	--

Table 4: Correlation (Kendall's τ) between task characteristics. *: p-value < 0.05; **: p-value < 0.01.

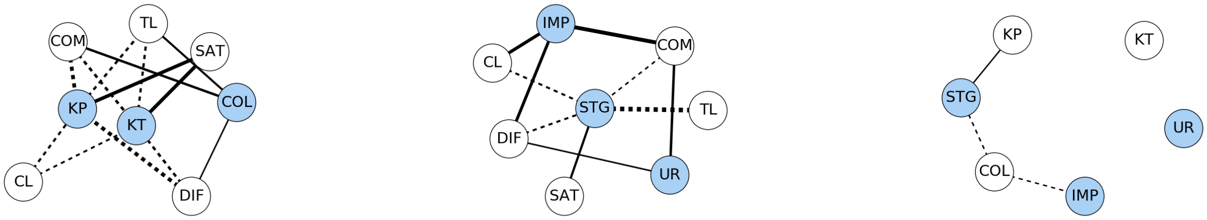


(a) G1: CL, COM, DIF, TL, SAT

(b) G2: COL, KT, KP

(c) G3: IMP, STG, UG

Figure 4: Illustration of the mutual correlation between the variables within a cluster. Solid edges indicate a significant positive correlation ($p < 0.05$); dashed edges indicate a significant negative correlation. The width of edges are proportional to the correlation coefficient τ.



(a) G1 - G2; dark nodes are from Group 2

(b) G1 - G3; dark nodes are from Group 3

(c) G2 - G3: dark nodes are from Group 3

Figure 5: Illustration of the mutual correlation between the variables from two cluster. Solid edges indicate a significant positive correlation ($p < 0.05$); dashed edges indicate a significant negative correlation. The width of edges are proportional to the correlation coefficient τ.

the difficulty/complex variables (CL, COM, DIF, TL), but a significant positive correlation with task satisfaction (SAT). That is, the more advanced stage the user is at a task, the more likely he/she is satisfied with the task, and the less he/she find the task complex/difficult.

Both task urgency (UR) and importance (IMP) have a significant positive correlation with the perceived task complexity (COM) and difficulty (DIF). In addition, IMP is also positively correlated with the level of cognitive complexity (CL) which means that the higher the level of cognitive complexity of a task is, the more likely users perceive it as important.

G2 - G3. Figure 5c shows the correlation between variables from group 2 and 3. We see that task stage (STG) has a significant positive correlation with users' knowledge of task procedure (KP)—the more advanced the stage the user is at in a task, the more likely that the user has knowledge of the task procedure.

Task collaboration (COL) has a negative correlation with both task importance (IMP) and task stage (STG). While one can imagine that the more collaboration is involved, the less importance the individual user may feel towards the task, the correlation between STG and COL is rather surprising. It suggests that less collaboration is involved at advanced stages of a task.

G4 - G1, G2, G3. Finally, we also observe (Table 4, column 5) that there is a positive correlation between task frequency (FQ) from group 4, and task importance (IMP), users' knowledge (KT and KP), and task satisfaction (SAT). That is, frequently performed tasks tend to be important, and users are more likely to have more knowledge about the task and to be satisfied.

5.2 Task characteristics in naturalistic user tasks: a case with cognitive complexity

To study user behaviour in context of certain types of tasks, in lab studies users are typically given one or more tasks which are designed with the desired characteristics. However, the design process is far from trivial. For instance Kelly et al. [19] proposed a framework to design search tasks of different levels of cognitive complexity and reported observations of user behaviours and perceptions in relation to the resulting tasks. We add to this framework with concrete examples of tasks that, from a user's personal perception, fall into one of the six categories of cognitive complexity.

Task anonymisation and abstraction. Since many of the user tasks share the same topic, e.g. look for jobs, we aggregated these tasks into a single topic. To avoid over-interpreting users' intent and activities, we perform the abstraction only for obvious cases, e.g. when the topic or its synonyms were mentioned in the task description. We then show tasks within popular topics distributed over different cognitive complexity levels, and the typical topics at each level. We anonymised the task examples by masking the identifiable information in the task description with "X".

Discussion. We list the top 8 most frequent topics (shared by at least 5 tasks) and their cognitive complexity levels, and one example task for each topic and cognitive complexity level (Table 6).

We see that each of the popular task topics spans over multiple cognitive complexity levels. In particular, tasks related to travel planning and job hunting almost cover all levels. This suggests that when people describe their tasks, although sometimes it seems that they are doing the same thing, the actual intention and activities involved can be very different. Indeed, a close check on the logged queries of "travel planning/booking" tasks reveals that in some cases users explored different touristic options, and in other cases users were just checking some facts such as currency rates.

The diversity of cognitive levels within a task topic certainly contributes to the difficulty of categorisation of tasks by external annotators, cf. [26]. We also considered to abstract the tasks using a taxonomy of Web activities, e.g., the taxonomy defined by Russell et al. [26] or the approach by Dumais [8] which uses a verb and topic to describe a task. The main problem we encountered is that it is almost unavoidable that we need to guess the intent of the user with these taxonomies, as shown in the above "travel planning" example. In fact, this observation is also reflected in Dumais [8]'s taxonomy, where tasks such as "plan travel" are mapped to both "Explore/Learn" and "Locate/Acquire" categories of Russell et al. [26]'s taxonomy and to "informational" and "transactional" categories in Broder [3]'s taxonomy. In addition, users may have mixed activities of multiple categories. For example a user was writing a report, and her log shows that she was both searching information related to the research topic (*Explore/Learn* [26]), and looking up latex commands (*find-simple* [26]).

Further, we see that different cognitive complexity levels are not evenly distributed across task topics. This suggests that some task topics are more likely to involve certain levels of cognitive complexity than others. For example, tasks related to writing, doing projects, programming and research tend to involve higher levels of cognitive complexity—which makes sense. Meanwhile tasks such as travel planning and shopping can be both low (as a simple booking/purchasing action) and high (by comparing, evaluating, and analysing different options). Sometimes a task and the associated cognitive complexity are surprising. For example, one user annotated watching online videos as "analyse" for the task "binge watching"—it seems that he/she was carefully researching a number of TV series before deciding to watch them.

6. DISCUSSION AND CONCLUSIONS

We conducted a diary style field study focusing on the tasks user engage in while searching and browsing online. We examined statistics of users' task activities in their search logs, correlations between task characteristics, and task characteristics of naturalistic user tasks. Here, we conclude our study by summarising and discussing the implications of our observations.

User tasks in search logs. We started this analysis with a comparison of the concepts used to describe task search sessions defined in three closely related studies that all aimed to identify task sessions from query logs (Section 4). We find that search sessions can be described at three levels: physical, logical, and task. However, each study we examined has a slightly different definition of these concepts, which makes a direct comparison of the observations made from these studies difficult. We hope that our review of these concepts will provide guidance for future work on task identification.

We continued with an empirical comparison of the statistics obtained from our self-annotated search log to that reported from these studies (RQ1). The main messages can be summarised as follows. (i) We have observed 41% zero-query tasks and 82% zero-query logical sessions, which implies that a large number of user task activities remain unobserved if we only focus on search queries. (ii) When restricting the analysis to only the queries in our log, we observed a higher number of queries and timespan per task compared to those reported in previous studies. This could be both an artefact of the differences in the log samples of different studies, as well as the way the logs were annotated (e.g. self vs external annotators). However, since the three studies we compared reported statistics in different measures (e.g. mean vs. median) and at different session levels, it is hard to verify if there were consistent results. However, our results suggest that tasks can require more queries and take longer than previously thought. (iii) Finally, we find that the canonical time-out for physical sessions is reasonably accurate in detecting task boundaries between queries within user defined tasks. However, in order to capture task switches in-between queries, activities other than queries need to be considered.

Task characteristics and user activities. Previous lab studies have investigated task characteristics with tasks specifically designed to address a single or a limited set of task dimensions. With designed tasks, however, it is difficult to investigate how different dimensions interact and how these occur in the daily information tasks. With the logs obtained in out naturalistic setup and explicitly annotated with task characteristics, we were able to perform an exploratory analysis and provide a comprehensive view of how task characteristics interact with each other and how they relate to user online activities (RQ2).

We identified groups of task characteristics that have strong mutual correlation. This has implications for task designs for lab studies. For instance, task collaboration is seen related to complex/difficult tasks, implying that studies of complex/difficult tasks may need to consider collaboration as an additional variable.

We further illustrated that tasks that share similar descriptions can vary greatly in their characteristics (cognitive complexity as an example). This supports the observation that it is very difficult for external annotators to classify user tasks or activities with a taxonomy [26]. This also has implications for studies that aimed at identifying user tasks from search logs in order to support their tasks. For example, we need to know not only what task the user is engaged with, but also *what status* the task is in, e.g., in terms of complexity and difficulty for the user, which would require different types of support.

Acknowledgements. This research was supported by the Dutch Technol-

Topics	Remember	Understand	Apply	Analyse	Evaluate	Create	Tot
Shopping	**10 (56%)** “Amazon-Heater”	– –	2 (11%) “sort out X lights”	3 (17%) “Baby products”	3 (17%) “buy contact lenses”	– –	18 (13%)
Writing	1 (9%) “compile X paper	– –	2 (18%) “Complete X tutorial	– –	**4 (36%)** “X Essay”	**4(36%)** “X paper”	11 (8%)
Travel	**3 (30%)** “weekend travel”	1 (10%) “X trip”	1 (10%) “Book trip to X”	1 (10%) “Flight home”	2 (20%) “book tickets for X ”	2 (20%) “Plan trip for X”	10 (7%)
Job	1 (14%) “Look for jobs”	– –	1 (14%) “Tutor jobs”	1 (14%) “Internship applications”	**3 (43%)** “job hunt”	1 (14%) “Finding job”	7 (5%)
Project	– –	1 (17%) “Project management	**2 (33%)** “ X project”	1 (17%) “X proj”	**2 (33%)** “research project-X	– –	6 (4%)
Research	– –	– –	**3 (50%)** “Research”	1 (17%) “...research for X”	1 (17%) “X research”	1 (17%) “X study”	6 (4%)
Programming	– –	1 (20%) “test X interface”	– –	**3 (60%)** “port X to java”	1 (20%) “...interface for X”	– –	5 (3%)
Watch X	2 (40%) “Youtube/streaming”	– –	– –	**3 (60%)** “Binge watch X”	– –	– –	5 (3%)
Other	21 “check location of X”	10 “stock knowledge	17 “Find solutions to X”	5 “learn about X”	9 “buy flat”	5 “study X”	67 (49%)
Total	38	13	28	18	25	13	135

Table 6: Example tasks at each cognitive complexity level. Each cell contains the counts of tasks within a topic at a CL level, and its percentage w.r.t the total number of tasks at that level. Boldface shows the most frequent CL level assigned to tasks within a topic.

ogy Foundation (STW) under project nr. 13675.

7. REFERENCES

[1] A. Aula, R. M. Khan, and Z. Guan. How does search behavior change as search becomes more difficult? In *SIGCHI'10*, pages 35–44. ACM, 2010.

[2] A. H. Awadallah, R. W. White, P. Pantel, S. T. Dumais, and Y.-M. Wang. Supporting complex search tasks. In *CIKM'14*, pages 1–10, 2014.

[3] A. Broder. A taxonomy of web search. In *ACM Sigir forum*, volume 36, pages 3–10. ACM, 2002.

[4] M. Bron, J. Van Gorp, F. Nack, L. B. Baltussen, and M. de Rijke. Aggregated search interface preferences in multi-session search tasks. In *SIGIR'13*, pages 123–132. ACM, 2013.

[5] K. Byström and P. Hansen. Conceptual framework for tasks in information studies. *JASIST*, 56(10):1050–1061, 2005.

[6] K. Byström and K. Järvelin. Task complexity affects information seeking and use. *IP&M*, 31(2):191–213, 1995.

[7] K. Church, M. Cherubini, and N. Oliver. A large-scale study of daily information needs captured in situ. *ACM TOCHI*, 21(2):10, 2014.

[8] S. Dumais. Task-based search: a search engine perspective. In *NSF Workshop on Task-Based Search*, page 1, 2013.

[9] C. Eickhoff, J. Teevan, R. White, and S. Dumais. Lessons from the journey: a query log analysis of within-session learning. In *WSDM'14*, pages 223–232. ACM, 2014.

[10] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[11] E. Greifeneder. The effects of distraction on task completion scores in a natural environment test setting. *JASIST*, 2015.

[12] J. Gwizdka and I. Spence. What can searching behavior tell us about the difficulty of information tasks. *A study of Web navigation. ASIST*, 6, 2006.

[13] M. Hagen, J. Gomoll, A. Beyer, and B. Stein. From search session detection to search mission detection. In *OAIR'13*, pages 85–92, 2013.

[14] J. He, M. Bron, and A. P. de Vries. Characterizing stages of a multi-session complex search task through direct and indirect query modifications. In *SIGIR'13*, pages 897–900. ACM, 2013.

[15] K. Järvelin and P. Ingwersen. Information seeking research needs extension towards tasks and technology. *Information Research*, 10 (1), 2004.

[16] R. Jones and K. L. Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *CIKM'08*, pages 699–708. ACM, 2008.

[17] M. Kellar, C. Watters, and M. Shepherd. A field study characterizing web-based information-seeking tasks. *JASIST*, 58(7):999–1018, 2007.

[18] D. Kelly and N. J. Belkin. Display time as implicit feedback: understanding task effects. In *SIGIR'04*, pages 377–384. ACM, 2004.

[19] 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. ACM, 2015.

[20] A. Kotov, P. N. Bennett, R. W. White, S. T. Dumais, and J. Teevan. Modeling and analysis of cross-session search tasks. In *SIGIR'11*, pages 5–14. ACM, 2011.

[21] L. Li, H. Deng, A. Dong, Y. Chang, and H. Zha. Identifying and labeling search tasks via query-based hawkes processes. In *SIGKDD'14*, pages 731–740. ACM, 2014.

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

[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. ACM, 2010.

[24] C. Lucchese, S. Orlando, R. Perego, F. Silvestri, and G. Tolomei. Identifying task-based sessions in search engine query logs. In *WSDM'11*, pages 277–286. ACM, 2011.

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

[26] D. M. Russell, D. Tang, M. Kellar, and R. Jeffries. Task behaviors during web search: The difficulty of assigning labels. In *System Sciences, 2009. HICSS'09.*, pages 1–5. IEEE, 2009.

[27] C. L. Smith. Conditions of trust for completely-remote methods: A proposal for collaboration. In *HCIR*, 2012.

[28] A. Spink, M. Park, B. J. Jansen, and J. Pedersen. Multitasking during web search sessions. *IP&M*, 42(1):264–275, 2006.

[29] P. Vakkari. Task-based information searching. *Annual review of information science and technology*, 37(1):413–464, 2003.

[30] P. Vakkari, M. Pennanen, and S. Serola. Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *IP&M*, 39(3):445–463, 2003.

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

[32] G. Zuccon, T. Leelanupab, S. Whiting, E. Yilmaz, J. M. Jose, and L. Azzopardi. Crowdsourcing interactions: using crowdsourcing for evaluating interactive information retrieval systems. *Information retrieval*, 16(2):267–305, 2013.