Department of Computer Science
Series of Publications A
Report A-2022-2

Behavioral Task Modeling for Entity Recommendation

Tung Vuong

Doctoral dissertation, to be presented for public examination with the permission of the Faculty of Science of the University of Helsinki in Hall PI, Porthania, on March 17th, 2022 at 14 o'clock.

University of Helsinki
Finland

Supervisor
Giulio Jacucci, University of Helsinki, Finland
Tuukka Ruotsalo, University of Helsinki, Finland

Pre-examiners
Katrien Verbert, Katholieke Universiteit, Leuven, Belgium
Shlomo Berkovsky, Macquarie University, Sydney, Australia

Opponent
Chirag Shah, University of Washington, United States

Custos
Giulio Jacucci, University of Helsinki, Finland

Contact information

Department of Computer Science
P.O. Box 68 (Pietari Kalmin katu 5)
FI-00014 University of Helsinki
Finland

Email address: info@cs.helsinki.fi
URL: http://cs.helsinki.fi/
Telephone: +358 2941 911

ISSN 1238-8645 (print)
ISSN 2814-4031 (online)
ISBN 978-951-51-7982-1 (paperback)
ISBN 978-951-51-7983-8 (PDF)
Helsinki 2022
Unigrafia

Behavioral Task Modeling for Entity Recommendation

Tung Vuong

Department of Computer Science
P.O. Box 68, FI-00014 University of Helsinki, Finland
vuong@cs.helsinki.fi
https://www.helsinki.fi/en/people/people-finder/thanh-tung-vuong-9135455

PhD Thesis, Series of Publications A, Report A-2022-2
Helsinki, March 2022, 85 + 171 pages
ISSN 1238-8645 (print)
ISSN 2814-4031 (online)
ISBN 978-951-51-7982-1 (paperback)
ISBN 978-951-51-7983-8 (PDF)

Abstract

Our everyday tasks involve interactions with a wide range of information. The information that we manage is often associated with a task context. However, current computer systems do not organize information in this way, do not help the user find information in task context, but require explicit user actions such as searching and information seeking. We explore the use of task context to guide the delivery of information to the user proactively, that is, to have the right information easily available at the right time. In this thesis, we used two types of novel contextual information: 24/7 behavioral recordings and spoken conversations for task modeling. The task context is created by monitoring the user's information behavior from temporal, social, and topical aspects; that can be contextualized by several entities such as applications, documents, people, time, and various keywords determining the task. By tracking the association amongst the entities, we can infer the user's task context, predict future information access, and proactively retrieve relevant information for the task at hand. The approach is validated with a series of field studies, in which altogether 47 participants voluntarily installed a screen monitoring system on their laptops 24/7 to collect available digital activities, and their spoken conversations were recorded. Different aspects of the data were considered to train the models. In the evaluation, we treated information sourced from several applications, spoken conversations, and various aspects of the data

as different kinds of influence on the prediction performance. The combined influences of multiple data sources and aspects were also considered in the models. Our findings revealed that task information could be found in a variety of applications and spoken conversations. In addition, we found that task context models that consider behavioral information captured from the computer screen and spoken conversations could yield a promising improvement in recommendation quality compared to the conventional modeling approach that considered only pre-determined interaction logs, such as query logs or Web browsing history. We also showed how a task context model could support the users' work performance, reducing their effort in searching by ranking and suggesting relevant information. Our results and findings have direct implications for information personalization and recommendation systems that leverage contextual information to predict and proactively present personalized information to the user to improve the interaction experience with the computer systems.

Computing Reviews (2012) Categories and Subject Descriptors:

Information systems → Information retrieval → Retrieval models and ranking

Human-centered computing → Human computer interaction (HCI) → Interactive systems and tools

General Terms:

information retrieval, user modeling

Additional Key Words and Phrases:

context information, task context, entity recommendation

Acknowledgements

This thesis describes my doctoral studies at the University of Helsinki. My adventures began in 2016, and I have the privilege of being part of the Ubiquitous Interaction group and the Department of Computer Science, where the research was carried out. My work has been financially supported by the MindSee and Coadapt projects.

This journey is not an easy one, however, I feel tremendous gratitude for the people whom I met and who have supported me along the way and also in life outside academia, making this experience fun, inspiring, and meaningful.

First of all, I would like to thank my supervisors, Prof. Giulio Jacucci and Associate Prof. Tuukka Ruotsalo, for believing in me and my research, for motivating me to follow many exciting ideas, for giving me their restricted time to answer all my questions, and for the opportunity to learn about the foundation of the academic world and how to do proper research. I am also grateful to Prof. Samuel Kaski and Prof. Antti Oulasvirta for providing me with invaluable comments on my works. Thank you for including me in many research projects.

I could not imagine a more inspiring research community than the University of Helsinki and Aalto University, which I have been privileged to work in. I am thankful for the support, friendship, and mentoring of many present and alumni colleagues who have influenced my work, including Salvatore Andolina, Mats Sjöberg, Markus Koskela, Baris Serim, Khalil Klouche, Antti Salovaara, Michiel Spape, Chen He, Oswald Barral, Imtiaj Ahmed, Patrik Floreen, Marie Al-Ghossein, Daee Pedram, and Reizaei Zeinab.

This journey will become harder without assistance from many faculty members and the administrative team in Kumpula. In particular, I want to thank Pirjo Moen for her guidance in many aspects of PhD studies and for many pizza evenings of PhD students. I also want to thank Marina Kurtén for her kind proofreading of the thesis.

Having freedom and fun in life also inspires my work. I am grateful to my family and friends for everything in my life. This work would not have been possible without the support from my wife. I want to thank my mom for the immense encouragement and support I have received throughout my life to follow my professional interest. I wish to thank my father for his love and support that has enabled a great deal of flexibility in work. Finally, I am grateful to my siblings for the endless fun on many ordinary days.

Helsinki, February 2022
Tung Vuong

List of Original Papers

This dissertation is based on the following peer-reviewed original publications. The publications are also referred to as Publication I-VI in the text. The publication list and the contributions of the author are described below. The publications are reproduced with permission from the copyright holders at the end of the dissertation.

Publication I Tung Vuong, Miamaria Saastamoinen, Giulio Jacucci, and Tuukka Ruotsalo. 2019. Understanding User Behavior in Naturalistic Information Search Tasks. Journal of the Association for Information Science and Technology, 70: 1248-1261.

Publication II Tung Vuong, Giulio Jacucci, and Tuukka Ruotsalo. 2017. Watching inside the Screen: Digital Activity Monitoring for Task Recognition and Proactive Information Retrieval. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. 1, 3, Article 109 (September 2017), 23 pages.

Publication III Reizaei Yousefi Zeinab, Tung Vuong, Marie Al-Ghossein, Tuukka Ruotsalo, Giulio Jacucci, and Samuel Kaski. 2022. Entity Footprinting: Modeling Contextual User States via Digital Activity Monitoring. In ACM Transactions on Interactive Intelligent Systems. (In revision)

Publication IV Tung Vuong, Salvatore Andolina, Giulio Jacucci, and Tuukka Ruotsalo. 2021. Does More Context Help? Effects of Context Window and Application Source on Retrieval Performance. ACM Transactions on Information Systems. 40, 2, Article 39 (April 2022), 40 pages.

Publication V Tung Vuong, Salvatore Andolina, Giulio Jacucci, and Tuukka Ruotsalo. 2021. Spoken Conversational Context Improves Query Auto-completion in Web Search. ACM Transactions on Information Systems. 39, 3, Article 31 (May 2021), 32 pages.

Publication VI Giulio Jacucci, Pedram Daee, Tung Vuong, Salvatore Andolina, Khalil Klouche, Mats Sjöberg, Tuukka Ruotsalo, and Samuel Kaski. 2021. Entity Recommendation for Everyday Digital Tasks. ACM Transactions on Computer-Human Interaction. 28, 5, Article 29 (Oct. 2021), 41 pages.

The author was among the main contributors to all of the articles. For Publications I, II, the author carried out the study design and execution in collaboration with the rest of the authors while contributing major parts in data analysis and reporting. For Publications III and IV, the author developed the initial idea and carried out the main parts of all study phases, from study design to reporting. For Publications V and VI, after other co-authors initiated the study, the author coordinated it and was the main contributor in the study analysis and reporting. None of the articles have been a part of previous dissertations.

Contents

Chapter 1

Introduction

Knowledge workers are required to process and produce more information than ever before; they work on multiple tasks, collaborate with colleagues, and use various applications to get their jobs done. The problem faced most often by the users in progressing the tasks is how to allocate their limited cognitive abilities to manage a wide range of information [6]. That is critical due to the increased range of data and information resources in digital systems every day. Information supply and searching is a key activity that supports the user's task performance [72]. However, many information-retrieval systems require user effort and cognitive attention in formulating queries [11, 39]. In addition, it is hard to recall relevant information for the task-at-hand in the first place, for example, what information is needed or names of the known documents [12]. Rather than having to recall information, people often use other retrieval methods, such as bookmarks and recent file lists [2, 10, 11, 81]. However, recent studies show that people often forget to use documents that can be helpful, even when they are stored in an appropriate location [29, 75].

In response to this limitation, the contextual recommendation has recently risen to the top of the research agenda, intending to help the user have access to task-relevant information easily without requiring the part of the user [75]. Recommending information helps to find useful information contextually while it also helps to reduce the number of computer interactions required for the task and improve the user's perceived usability with information retrieval systems [88]. As an example scenario on how contextual recommendation supports the user task, we consider a knowledge worker who is a person working mainly with information. She uses and produces information and works on different tasks in a day. A typical workday of such a knowledge worker can be described by a combination of activities. Some activities are organizational, such as handling e-mail

messages or attending meetings, making up a specific task. For the task, she needs to work on different documents and open different applications, search through the web with specific topics and keywords related to that task, talk with colleagues, and discuss the task. A recommendation system could be of help by having these documents opened for her, to spare her the time to navigate to them or look for them. Therefore, if a model can infer the user's task context given the history of user activities, it can predict the next activity and consequently recommend more relevant entities to the user.

Typically, the contextual recommendation system proactively suggests relevant information after considering the context of the user task, such as applications or documents being used [16, 69]. Many types of contextual signals have been considered, and numerous approaches have been proposed [52, 89]. In particular, the context of the user task has been mostly determined from the user's Web activity, such as recent Web queries that have been issued [31, 60] or the blogpost or Web document the user is composing [16, 36, 52]. Conventionally, most approaches leveraged pre-defined interaction logs or associated data acquisition has been confined to a certain application or a set of services, while there are many other sources of contextual information that can be useful in determining the task context and that have not been considered. In this thesis, we show the benefit of considering two types of novel contextual information: 24/7 behavioral recordings and spoken conversations. They are the extensive sources of context, and the contextual information determined from these sources are not restrained to a specific application range or a type of user input. We explore the use of these novel contextual signals to infer the user's task to facilitate information access through an entity prediction and recommendation system.

To collect the user's 24/7 behavioral recordings, we employed a screen monitoring approach that captured all user interaction data and generated visual content (e.g., visual content presented to the user on the screen) across application boundaries. To collect spoken conversational information, we utilized voice recording of spoken conversations the user engaged in with other individuals. To train the prediction models, we treated the data as multiple aspects (temporal, social, and topical aspects) that can be contextualized by several entities, such as applications, documents, people, time, and various keywords determining the task. By tracking the association amongst the entities, it is possible to model the task context, predict future information access, and consequently recommending more relevant information to use at the right time.

To understand which contextual signals and aspects of the data are useful in improving recommendation quality, we built several prediction models. Each model incorporated a different contextual signal. We also combined all of these signals and considered different aspects of the data; for example, we considered a combination of search history and interaction history on non-search applications; a combination of search history and spoken conversational input; and interaction history with or without temporal information.

1.1 Research Questions

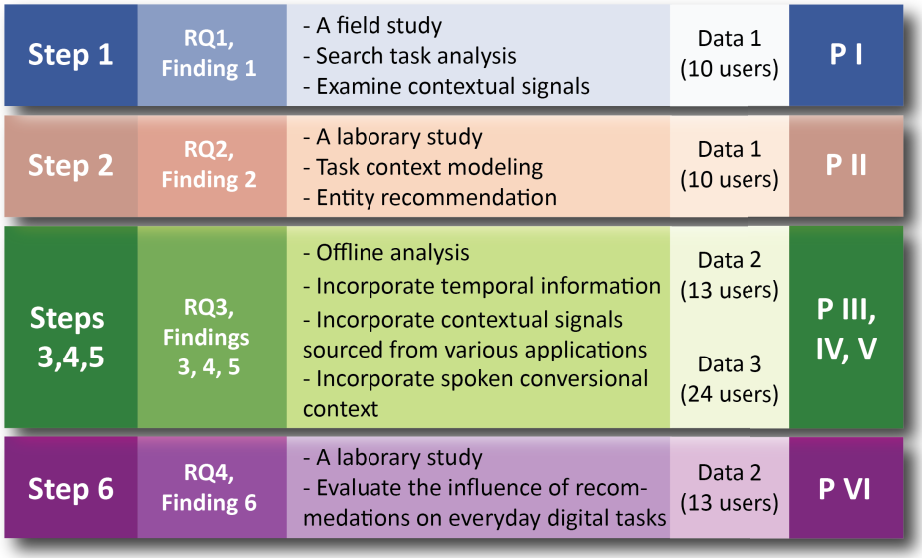
Taking into account more extensive sources of context and various aspects of the data for modeling, this dissertation seeks the answer to four research questions (RQs). Figure 1.1 presents how the research questions are covered in the publications. The first two deal with the modeling problem and the third one directly relates to observing the effects of considering task context for the recommendation. A fourth research question addresses how we study and examine the effect of recommendation on the user's task performance in realistic settings.

RQ1. Are there interdependencies between the tasks, searching behavior, and contextual entities? This RQ is closely related to search task analysis; we wanted to understand how users typically worked with their computers to find information in real-life settings and what contextual factors and what parts of the task mostly influence them to search.

RQ2. Can the association amongst the entities be used to model the user's task context for recommendation? In RQ2, we studied how the rich data gathered from extensive behavioral recordings can be used to model the user task and predict the future context that the user would probably be involved in.

RQ3. Does the use of more extensive sources of context improve recommendation quality? This RQ investigates the utility of more extensive signals from different context sources for entity recommendation. This RQ can be divided into the following sub-RQs:

- RQ3-1. Does the use of temporal information improve recommendation quality?
- RQ3-2. Does the use of 24/7 behavioral recordings improve recommendation quality?



RQ 1: Are there interdepencies between the tasks, search behavior, and contextual entities?

Finding 1: Searches are strongly influenced by the task context. Contextual entities such as, various types of application being used and specific types of content had appeared on the screen could trigger the user's information needs.

RQ 2: Can the association amongst the entities be used to model the user's task context for recommendation?

Finding 2: Topical relatedness amongst the entities are useful in inferring the user's task context.

RQ 3: Does the use of more extensive sources of context improve recommendation quality?

Finding 3: Temporal information is useful in improving recommendation quality

Finding 4: Contextual signals sourced from any types of application are useful in improving recommendation quality.

Finding 5: Contextual signal sourced from spoken conversation is useful in improving recommendation quality.

RQ 4: Does contextual recommendation improve users' task performance?

Finding 6: Recommendations positively influences the user's task performance. Task-based entity recommendation approach enable effortless information access. Recommendations contain insightful information that help the users to complete their tasks.

Figure 1.1: Research Questions, Methodology, Publications (P I-VI), and Findings.

- RQ3-3. Does the use of spoken conversational input improve recommendation quality?

RQ4. Does contextual recommendation improve users' task performance? This RQ investigates whether entity recommendation could provide information beyond what the user can find without it and whether entity recommendation could positively influence the user's information behavior and lead to improved task execution.

1.2 Methodology

To advance our understanding of whether the two novel sources of context: spoken conversations and 24/7 behavioral recordings, are useful in improving entity recommendation, we have conducted a series of experimental studies, including field studies, laboratory experiments, and offline analyses. When choosing the empirical methods, we considered the trade-off between the criteria [61]: 1) Generalizability which is the validity of the results across the population, 2) Precision of measurement of the behaviors that are being studied, and 3) Realism of the studied tasks relative to the context in which the evidence we gathered is applicable. For that, in each study, we set the focus in accordance with the research goal (RQs) at hand and selected the empirical method accordingly.

Field studies are rare in information retrieval research [48, 49, 72], where it allows researchers to study users' real-life tasks and topics, information searching context, and user behavior in naturalistic environments. Although field studies do not allow rigorous control as laboratory studies do, they can be more realistic, and the studied tasks are real-life; thus, the results and findings can be more congruent to the user's real-life situations. Understanding users' search behavior in real life provides valuable insights into the types of needs that occur from everyday digital activities, how those needs are addressed, and how contextual factors impact those needs. In field studies, the information search context could be characterized by the participant's self-identified tasks and topics and several factors of these. In our research, these factors were (a) application context or application being used; and (b) content observed previously by the users on the screen that triggered them to perform searches or how often the participants see a keyword on the screen and subsequently use it as a query in information-seeking activities that are related to a task. In Publication I (to address RQ1), we conducted the first data collection experiment wherein we studied the activities of 10 users during 14 days (Data 1). The observations of user

activities were collected by installing screen monitoring and digital activity monitoring systems on the participant's laptop. The participants were also asked to keep a diary reporting their everyday tasks that required computer support and entity usage (applications, documents, keywords, and persons involved in the task). The data include a wide range of search tasks occurring as part of the broader work task, for example, local file search activities using the OS-specific applications (e.g., Finder, Spotlight, and Explorer) as part of project work, search using map interfaces (e.g., Google Maps, with typed queries, drags, clicks, and searches in email clients) as part of planning a travel task, as well as custom searches on websites as part of the entertainment. This extensive data set and rare in information retrieval research allowed detailed statistical analysis, aiding in linking contextual factors to the users' tasks and their search performance.

Laboratory studies focus on precise measurements of recommendation system performance or prediction accuracy rather than studying realistic situations. For that, a context-aware recommendation system could be evaluated in an online interactive setting with the users. The goal in our studies was to evaluate task prediction models and recommendation systems, for example, how accurately the model can predict the task context and how accurately the model can predict the entities that the users would use next. In such a case, each user needs to work as she/he normally does while receiving recommendations from one of our systems that is being evaluated. During the experiment, we evaluated whether the contextual recommendation could lead to improved task execution in terms of usability of the system and quality of recommendation. To address RQ2, in Publication II we studied how the rich data gathered from 24/7 behavioral recordings (Data 1) could be used to model the user task. The prediction model was evaluated by asking the participants to specify what tasks they were doing and rated whether the system correctly predicted their current task. To study the actual effectiveness of proactive recommendation in real-world situations (to answer RQ4), in Publication VI, we conducted the second data collection experiment in which we collected 24/7 behavioral recordings from 13 participants (Data 2). After a 14-day monitoring period, we asked the participants to resume their real-world tasks for evaluation of a recommendation system. For example, the participants worked on the task, performing new unseen activities, while the system predicted the context and provided entity recommendations. The study aimed to understand how the recommendation system influenced the user task performance and how useful the participants perceived the recommendations. We quantified the influence and usefulness of recommendations by studying whether

the system allowed the participants to find more relevant entities and open more applications/documents than in a situation where no recommendations were offered.

Offline Analyses focuses on collecting real-world data and tasks for the evaluation of the prediction models. This methodology offers greater realism as the data represents the user's real-world computer usage. In Publications III, IV, V (to address RQ3), the aim was to collect the sets of data from which task context can be determined and to study the impact of different contextual signals on prediction accuracy. We employed two novel data collection approaches: 1) screen monitoring and 2) recording of spoken conversations. The data sets consist of 24/7 behavioral recordings (Data 2) and speech-to-text information of the conversations (Data 3). The resulting Data 2 collected from 13 participants in the second data collection experiment was considered. To collect spoken conversational information, we conducted the third data collection experiment in which 24 participants were recruited. The participants were formed into pairs, and they were asked to engage in an informal discussion with other peers about movies and travel destinations that they would want to watch or to go next while their spoken conversations were being recorded. The participants could use their laptops to search for information to support the conversations, and their search records were also collected for the analysis. The data set that resulted in the third data collection experiment contained spoken conversational information and all computer interaction data that were recorded during the conversation (Data 3). Both data sets (Data 2 and Data 3) were then used in the offline prediction studies. An offline evaluation method has the advantage that the impact of different data input in the model can be evaluated more easily. It provides the possibility to reproduce results with small changes in the model and provides control over the variables that we would study in the experiments. To understand whether two types of novel contextual information are useful for improving recommendation, a non-context-dependent recommendation system was used as a control condition. We manipulated the context source leveraged to construct the prediction model for experimental conditions: an experimental condition with the model utilizing search history, an experimental condition with the model utilizing 24/7 behavioral recordings, an experimental condition with the model utilizing spoken conversational information, and an experimental condition with the model considering temporal information.

1.3 The Structure of the Thesis

Continuing the groundwork and motivation in the introductory chapter, Chapter 2 provides a theoretical background for the present work. It gives an overview of the related research on the use of context to improve recommendation systems and the effects of different types of data input. The second chapter also reviews state-of-the-art task modeling approaches and evaluation practices for context-aware recommendation systems that have been considered in prior work. Chapter 3 describes our unique datasets, the setup of the data collection experiments, and the methodology used. In Chapter 4, we present an approach to study the user's information searching context and behavior in naturalistic environments. We identified which aspects of search context should be considered when implementing the recommendation system; what contextual factors (application context, keywords seen by the user prior to search) influence the user's search behavior, and how particular aspects of the task (individual intention, task goal, and substance of the tasks) are related to the contextual factors. In Chapter 5, we demonstrate our task modeling approach and the recommendation application. The heart of the thesis is Chapters 6-8, in which we investigate the effects of different kinds of contextual signals on recommendation quality. In Chapter 6, we investigate the effect of temporal information on recommendation quality. In Chapter 7, we investigate the effect of 24/7 behavioral recordings on recommendation quality. In Chapter 8, we investigate the effect of spoken conversational input on recommendation quality. In Chapter 9, we provide an evaluation for the context-aware recommendation. In Chapter 10, we then discuss the contributions of findings on the use of novel contextual information in recommendation systems from our empirical studies, provide a summary and discussion of the implications of our work, and highlight future directions for research into information personalization and recommendation.

Chapter 2

Background

In this chapter, we provide some background to position our work. In particular, we discuss earlier research that focused on using the tasks as search context and how contextual information was leveraged for document ranking and recommendation. Then, we review previous work on task modeling and the roles of different contextual sources in recommendation, highlighting our contributions with respect to prior work.

2.1 Task as Search Context

A user's task is often analyzed as an important context that invokes a user's information needs and influences searching behavior [42, 53]. For example, while the user is writing or reading a document, he/she may need to use related information to make sense of the information is being read or supporting the writing process [53, 65]. Because the available information or the user knowledge at that time is inadequate for her to complete a task [7], it prompts her to interrupt the task, look for additional information, or triggers information needs. Information retrieval is based on information needs and is operationalized by search task actions such as queries or following links [53]. It has been generally agreed that search tasks are part of broader user activity that can be referred to as a primary task or a broader work task [17]. Understanding the user's work task and capturing its context can help predict user needs; consequently, it can help improve system functionalities assisting information recommendation.

Kelly and Belkin [50] conducted a longitudinal study to elicit task context while the users were doing a web search. Participants were asked to think about their online information-seeking activities in terms of tasks and create personal labels for each task by classifying documents that they

viewed according to those tasks. Researchers then used this information to develop various implicit feedback models for each task grouping, and the models were used to predict the user's search intent and improve retrieval effectiveness by query expansion [90]. They found that the use of task context can help improve the ranking of Web documents. However, their study rather focused on investigating the effect of using task information in Web search ranking but did not apply the learning model to real-time online situations. In addition, their approach also required the labeled data for making query predictions, and this may be a burden on the user in practice. In contrast, we infer task context automatically and generate training data from observed user behavior for learning functions.

A more recent work did not rely on labeled data but utilized an unsupervised approach that could extract task context from search logs [31]. Their approach was first identifying multiple searches that shared the same information need. This sequence of queries was considered as a task context. Then, they used the task context to generate query recommendations. While we do not only consider search logs but also other interaction logs, our objective is to come up with a broader characterization of tasks that covers more extensive sources and aspects including task information that can be sourced from varying applications such as, local files, email, communication platforms, as well as spoken conversations. Additionally, we characterize the opportunities for recommendation systems to provide more comprehensive task support to their users, such as recommending user entities that are interesting and useful for the task the user is performing.

Research in characterizing tasks does not only focus on tasks carried out by an individual but also tasks performed by teams [92]. The study found that knowledge workers often engaged in activities with team members or groups of people. Consequently, the information they encountered, used, and later re-find and search can be intertwined with different parts of team activity. Recently, the influence of social context [68] has been used to analyze information behavior. They found that searches could be socially motivated and prompted by conversation. However, these studies do not focus on the support of recommendation systems that can help the workers complete their tasks.

2.2 Task-centric Information Management

Another relevant area within the context of this thesis is research in personal information management [35]. Researchers have been investigating how an explicit representation of information related to a task can help

improve a user's productivity at work. In an exploratory study, Dragunov et al. [26] demonstrated how the data transformed from user interaction with information objects (e.g., files, emails, documents, contacts, etc.) into a task template that can support the users doing their work. This template could be used to determine the task, aggregating information and associating relevant resources to the task. Here, information resources are usually documents and tools that support the users' daily tasks. Some information management systems allowed users to organize application windows in different folders associated with different tasks [70]. The users were able to bring up a task-specific folder showing relevant information while the task was being performed.

Related to our research is Brdiczka et al.'s [14] work, tasks were characterized by temporal relationships amongst the user actions (e.g., application switches, window switches). Users were first asked to label the tasks, associating each task with a set of documents and applications. The model was trained using the labeled data and outputted a task representation that is based on a distribution of temporal patterns of window switches characterizing the user's routine. The trained models and found patterns were then used to recognize the task from the unseen sequences. However, they did not focus on predicting the task context or considering more extensive sources of contextual information, but the model was trained on pre-defined interaction data such as interaction history on specific applications.

Bellotti et al. [8] focused on organizing task-centric workspaces in an email application. Filters were applied to threads of messages, files, links, and drafts that can represent the tasks. Users were able to view their tasks from the system, get notifications of upcoming deadlines regarding the task, or contact relevant people without sifting through thousands of emails. Their models considered emails as a context source to determine the user task; however, other important information that can be found from other non-email applications may not be recorded. For example, files, textual documents, and Web pages used for the task and related to the emails were not included in the model.

To be of assistance to a user, a personal information management system should understand what the user is currently doing by monitoring information behavior on a variety of applications. Kaptelinin [47] addressed this issue by collecting more information about users' tasks and activities by monitoring all running applications and the entire file system. The problem with this approach is huge amounts of data may create overhead and may not be needed and maybe even challenging to draw inferences from. Our research completes the picture by investigating the effect of varying contex-

tual signals sourced from different applications and spoken conversations. We also combined contextual signals from all the sources and revealed insights about how they, either separate or combined information, have an impact on the performance of the prediction model.

2.3 Context-Aware Recommendation

Research in using task context is not only limited to the support of information management, but it is also found useful in recommendation systems [41, 78, 83]. For instance, Rhodes and Maes [69] demonstrated the benefit of considering context for proactive information recommendation. In their work, the context was determined from the document that a person is writing, and this information was used as a query to the search system. The recommendation system is proactive in the sense that the querying took place in the background, and search results are automatically presented to the user. A limitation to this approach is that the contextual information derived from other application sources, such as news that is being read, an e-mail message being composed, a person, and spoken conversational information related to the writing document, are all ignored. In addition, the historical behavior of users such as long-term Web browsing history has proven to benefit recommendation were also not considered [89].

The use of contextual information has been extensively studied [31, 51, 83] and numerous approaches have been proposed. The main technique has been to construct the context models from observed past user behaviors that are often sourced from search history, e.g., query logs [28, 32, 37] or Web browsing logs [5, 38, 46]. The context model was then used to redefine the list of initial generated recommendations, e.g., reranked query suggestions or automatically generated search results. For instance, Eickhoff et al. [28] considered search engine result pages of the prior query as context; the signal value was the set of terms that the user paid attention to on the pages. Then, the candidate terms for query expansion were reranked according to the semantic correlation to those contextual terms. However, Web searches are often conducted as part of a more general task [58], and therefore considering search history as the only source of context may be a factor limiting the effectiveness of recommendation. Another approach is to take all the desktop data (documents stored on the computer) as context [20]. The authors first identified the set of terms that were closely related to the current query as candidates. Then, they restricted the query suggestions to only those semantically related to terms appearing in the desktop data. More recent research has also utilized data from other sources that

involve a richer context. Singh et al. [79] logged user behavioral signals, including clicks and page visits, on a real-world e-commerce site to predict user query intent. Li et al. [57] considered user context based on recently read emails. Tan et al. [80] collected recently opened documents as context for recommendations. However, prior work focused on access to partial data, which is only obtainable through predefined applications or services. Consequently, this would limit the advantages of using a recommendation system. While our work focuses on modeling the task context comprehensively by considering more extensive sources from various applications and spoken conversations, we also consider temporal associations of the user's past interactions and provide them with information relevant to the task at hand.

2.4 Social Context in Web Search

With the growing recognition that information interactions evolve from the users' social contexts, the information retrieval community has led to an increased awareness of how social elements influence the information search process [3, 22, 74]. However, research in capturing and using contextual information from the social aspect of the task (e.g., spoken conversational context) has received little attention. For example, Church et al. [22] conducted an exploratory study to understand how the presence of others influences user searches. They asked participants to carry out a task in collaborative settings (e.g., two participants engaged in a discussion and looked for a restaurant where they would like to eat lunch). They found that the users actually took into account others' opinions while searching, e.g., using named entity of restaurant or place to conduct the search or even sharing search results. However, the study did not focus on task context or recommendation of automatically generated search results but explored how people collaborate while conducting the mobile search.

Further, prior research found that searches may be socially motivated and prompted by conversations [23, 66]. McMillan et al. [62] suggested that a continuous speech stream could be considered as context and that could be used to identify users' following actions such as searches. Similar to prior work, we also consider the use of spoken conversational context; however, instead of using the context for Web searches, we use it to improve the ranking of entity recommendation. That is, we do not rely on explicit queries prompted by the user but use the conversational context to predict the entities (applications, documents, contacts, and keywords) that the users would use in the future.

Andolina et al. [4] conducted a study to investigate how information from spoken conversational context could be used for recommendation. Participants were asked to engage in the task of planning for a travel trip or the next movie to watch. The recommendation system proactively performed the search in the background given the user's spoken words; search results were automatically generated and presented to the users in real-time to support the conversations. They found that the contextual information such as location and person entities being referred to in the conversation was useful information and could be used as implicit queries for proactive information retrieval. However, their approach did not predict the user task context but only focused on studying how the recommendation system effectively supports the conversation.

2.5 Summary

This section summarizes the literature review with the concluding remarks. From the literature, it could be concluded that:

- The primary approach to context-aware recommendation has been to determine context from observed past user behaviors, which are often sourced from the search engine interaction logs [28, 32, 37] or Web browsing data [5, 38, 46]. However, search history considered in isolation often offers limited contextual information, while task context that considers user activities on various applications could provide richer information about the user's information need.

- There is research that utilizes data from other sources that involve richer context, such as behavioral user signals in the Web (clicks and page visits [79]) or email conversations [57]). However, there has been little related work in considering spoken conversational context for a recommendation.

- Other research works have attempted to exploit contextual information from more application sources [9, 89]. Different sources of contextual information have been used to support file navigation, such as file recommendation [15, 33] and general web browsing [64] such as Web page suggestion and document recommendation. However, little is known about the value of each contextual source for this purpose.

To fill this research gap, in this thesis, we (i) explore the use of task context information derived from two novel sources: spoken conversations

and 24/7 behavioral recordings (digital activities on a variety of applications); (ii) examine the effect of several aspects of the data (temporal, topical, social) on recommendation quality; and (iii) compare the utility of different sources of contextual information (queries, search history, Web browsing history, non-search interaction history, and spoken conversations) for building prediction models. Our goal is to facilitate information access through an entity prediction and recommendation system. We present a recommendation system that automatically offers entities as the user works on the task. In addition, we also (iv) investigate the effectiveness and usefulness of recommended entities in everyday digital tasks by a study using users' real-life data and tasks.

Chapter 3

Data Collection Experiments

We used three datasets collected from 47 knowledge workers who voluntarily took part in a series of user studies. Participants were recruited from university and industrial settings with varying professions: university students, computer scientists, engineers, entrepreneurs, and accountants. The participants were recruited via a posting that was distributed to mailing lists. A questionnaire was attached to the recruiting message to collect background information on potential candidates. Upon agreeing to participate in the experiments, the participants were informed of our privacy guidelines and data protection. Our studies and research plan were reviewed and accepted by the ethical review board of the University of Helsinki in Finland.

3.1 Data 1

The first dataset contained 14-day 24/7 behavioral recordings of 10 participants (5 females, 5 males). It consists of 1) screen captures, digital activity logs, and 2) task diaries. We used screen monitoring and digital activity monitoring systems to collect screen captures and Operating System (OS) logs from participants' laptops. The participants were asked to keep a diary reporting their tasks every day. For each task, the participants would need to write a short description, applications used for the task, several keywords describing the task, and people involved in the task.

3.1.1 Screen Monitoring and Digital Activity Monitoring Systems

The system automatically records user-computer system interaction. The system has two main modules: Screen Monitoring (SM) and Operating System (OS) Logging.

The *SM module* records the user screen and continuously takes screenshots of the active window at 2-second intervals or screen frames that indicate information changes on the screen. Screen frames were then converted into text using Optical Character Recognition (OCR). The SM module was developed in Mac OSX and MS Windows OS. We used Accessibility API, a native OS library, to implement both versions. The libraries in both versions performed identical functions, saving a screen frame as an image. To extract the textual content of the screen frames, we used Tesseract 4.0, which was a very accurate OCR.

The *OS Logging module* logged information that is associated with the screen frames, including the titles of active windows, the names of active applications, the Uniform Resource Locators (URLs) of Web pages if active applications are Web browsers or file path if active windows are local files and documents, and the timestamps indicating when the windows became active.

Given OS logs, we merged screen frames that belonged to a single *information object*. An information object describes the user's access to an information resource on the computer, such as a textual document, an email, a folder, a file, an instant message, a Web page, and an application window with a unique title. We focused on the content of the information object that the users read/produced by extracting only information change on the screen. For this process, we utilized a frame difference technique in which the two temporally adjacent screen frames (of a single information object) were compared, and the differences in pixel values were determined. That is, terms that appeared in the same pixels in the two adjacent screen frames were excluded from the information object.

3.1.2 Behavioral Recordings and Diaries

We installed screen monitoring and digital activity monitoring systems on participants' laptops and set them to run continuously for 14 days. After the installation, the participants were each asked to keep a diary of their daily tasks. For the convenience of writing a diary, we provided the participants with a diary template including three fields: a brief statement describing the task, specific keywords related to the task, and the names of the available people involved in the task. The participants used pen and paper to write in the diaries, and they could write the diaries whenever they felt comfortable throughout the day. We intentionally advised the participants to focus on writing a broader task consisting of several activities. We encouraged the participants to use their conceptual understanding of what activities could make for a meaningful broader task.

3.2 Data 2

The second dataset was collected from 13 participants (5 males, 8 females). The data collection experiment followed the same procedure as the first data collection experiment (Data 1). We installed a screen monitoring and digital activity monitoring system on the participants' laptops and set it to run continuously in the background for 14 days. The participants were also asked to keep a diary reporting their everyday tasks.

3.3 Data 3

This dataset was collected from 12 participant pairs (14 females, 10 males). A controlled task-based information-seeking experiment was conducted. Participants (in pair) had conversations about movies or travel lists and supported that conversation by performing Web searches. The data includes (1) Automatic speech-to-text transcripts and ideal speech-to-text transcripts produced by a professional text-to-speech transcribing service, (2) The queries that the participants inputted into the search interfaces, and (3) The Web pages were visited during their conversations.

3.3.1 Task

The participants were asked to converse with the other participant in the group on two topics: a list of movies that they planned to watch or a list of places that they wanted to visit. Each group was assigned a single task in a counterbalanced order. The designed work task was not meant to generate a specific outcome, but rather, it was intended to facilitate the discussion in the conversations. More specifically, we asked the participants to share their own experience regarding the movies they were impressed with, places they found attractive, and to get inspiration from the other participant.

3.3.2 Apparatus

In the experiment, participant pairs sit directly opposite each other across the table (Figure 3.1). Each participant could use a Macbook Pro 15" laptop connected to a Samson Meteor microphone to perform searches whenever they feel like doing so. The laptop screen was recorded using Screencast-O-Matic software, recording the participant's face with the webcam. Each experimental session was video-recorded using a Panasonic camcorder.

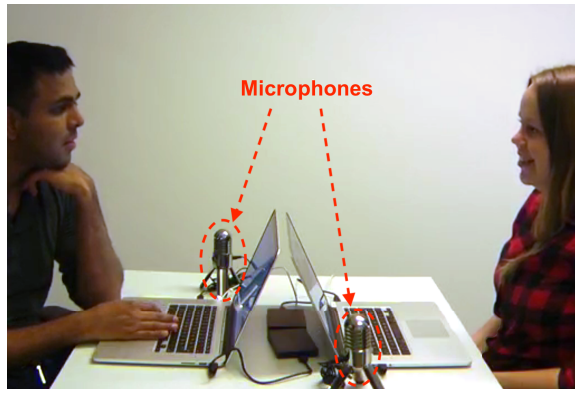


Figure 3.1: Experimental setup. The participant pairs sat opposite one another across a table. The participants could use laptops to perform searches and web browsing in front of them. High-quality microphones were also placed before the participants and continuously recorded the conversations.

3.3.3 Procedure

First, the researcher in charge began each session by welcoming the participants and introducing the overall procedure of the experiment. Participants signed informed consent forms upon joining the experiment. Then, the researcher described the tasks and left the room in order for the participants to talk about the topic freely. The researcher in charge followed the experiment through a video connection and could be reached if the participants needed help. The participants were not forced to perform a Web search, but they were allowed to do so if they needed additional information to support the conversation. The only service was available on the laptop was our Web search interface. This Web search interface was customized to record all queries and Web page visits. We used Google Custom Search to implement the Web search interface, and we also disabled the personalization of the Web search outcome.

3.3.4 Transcript and Web Search Logging

We used two transcription methods: automatic and ideal. The automatic transcription was conducted using an automatic speech recognition service. The ideal transcription was manually conducted by a professional transcription service. Figure 3.2 illustrates a snippet of a conversation in which the two transcription methods transcribed speakers' utterances. Web-search logs were also collected and temporally associated with the conversations.

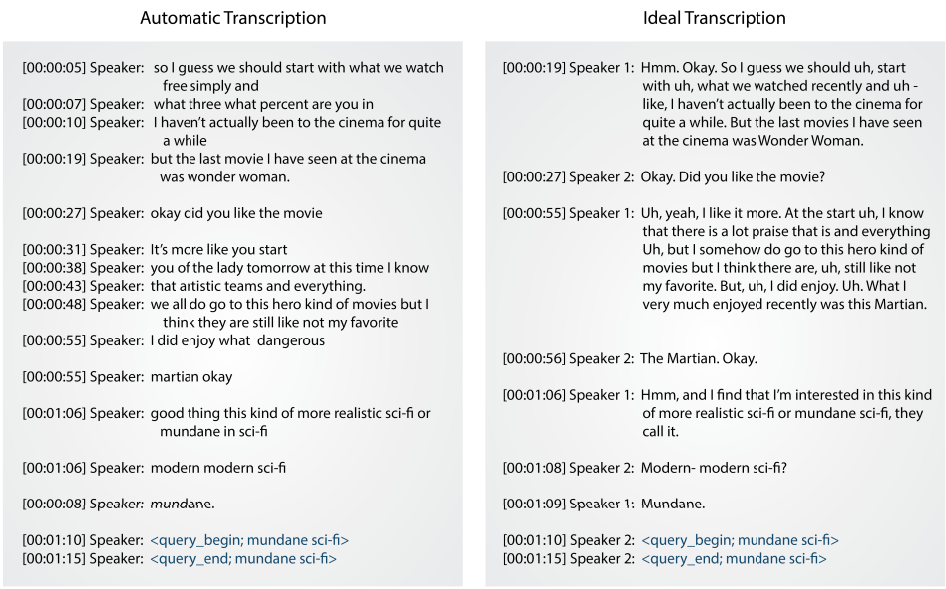


Figure 3.2: Examples of automatic and ideal transcriptions.

An automatic speech-to-text transcription system continuously transcribes the user's recorded speech. Speech recognition was performed using Google's Web Speech API [1]. The API took each audio recording as voice input and outputted an associated sentence transcript. Speech recognizer only transcribed speech whenever there was a voice activity. As soon as the voice transcript became available, it was saved as a text unit with a timestamp of when the speech was recorded, as illustrated in Figure 3.2. This procedure ensured that the speech recognizer had access to only the conversations that occurred before the search happened and could not use post-search conversation when creating the transcripts.

Ideal Transcription

Besides automatically processed transcripts, the output of the data-collection experiment also contained high-quality video recordings. We obtained ideal transcriptions through manual annotation of the video recordings. A professional transcription company was hired to transcribe the video recordings. Speakers' turns were identified, and each turn was associated with an end timestamp, as shown in Figure 3.2. The end timestamps were obtained whenever the speaker changed. Furthermore, we manually checked and

[1] https://www.google.com/intl/en/chrome/demos/speech.html

verified the correctness of the individual timestamps. Two coders manually transcribed the recordings and agreed on 100% of the transcribed texts except for the use of plurals and prepositions, which were challenging to identify. However, these did not affect the results because the text was also pre-processed by stemming and stop-words removal.

Web Search History

The effectiveness of using users' search history to contextualize recommendations was also investigated. The search history consists of queries submitted and Web pages browsed in the same session prior to searching. To extract the text from HTML responses, we used the content and comment extractors[2] of the Dragnet [63].

3.4 Ethical Considerations

We are fully aware of privacy implications when using speech data and 24/7 behavioral recording data in the studies. We have taken an active step toward data privacy and security. All the data were encrypted and stored in a secured server in a locked room with a key. The data was only used for research purposes and deleted after the research completes.

Participation in the studies has been voluntary, and the users were informed about the data collection and management procedures. The data collection was also subject to the IRB process of the University of Helsinki. Participants also gave their consent upon joining the studies.

[2]https://github.com/dragnet-org/dragnet

Chapter 4

Interdependencies between Tasks, Search Behavior, and Contextual Entities

The research described in this chapter aims at answering the first RQ: **Are there interdependencies between user tasks, search behavior, and contextual entities?**. To inform the implementation of the task context model and the design of the recommendation system, Publication I [87] reports the analysis on the user's information retrieval in real life in connection to the work tasks. We focused on knowledge workers since their activities frequently require digital support. In the study, work tasks are categorized by various factors, and they form the context in which information retrieval was performed. Task factors are individual intentions (for example, being creative or checking facts), task goals (for example, communicating with someone or as a part of an intellectual work task), and substances (for example, free-time or programming). The dependent variables representing user behavioral factors are 1) application context (what are the application types that form the cross-system interactions before searching); and 2) content-triggers (how often the searches are dependent on the content that users have already seen on their computer). These are also the contextual entities we studied in the analysis. The study is exploratory, and the main data (Data 1) consists of screen captures, digital activity logs, and diaries describing the tasks.

This study aims to capture the user's information retrieval in real-life work tasks using the screen monitoring and digital activity monitoring method and subjective report of the user tasks. This investigation is important to understand how information retrieval is performed and what contextual factors affect it. Otherwise, the development of the user model

and design of user tests would be based on assumptions that may or may not add to the actual usefulness of recommendation systems.

Publication I [87], therefore, presents an analysis of the key characteristics of the tasks, primarily focusing on the two aspects: 1) What are the application types that the users are using before searching? 2) How often does the content that users have seen before searching trigger their searches? Tasks described by the users are real-life; therefore, we applied thematic and quantitative analysis to provide insight and understanding into the everyday tasks that people performed. The comprehensive classification of tasks into factors is also reported in Publication I [87].

Results of the study showed that there were dependencies between the measured behavioral factors and the task factors. The applications used and the content seen before searches are important contextual signals that should be considered in generating recommendations. Therefore, the results of the study answer RQ1, and the takeaway message could be concluded as:

Finding 1 Searches are strongly influenced by the user's task context. Contextual entities such as various types of applications being used and specific types of content that had appeared on the screen could trigger the user's information needs.

4.1 Task Classification using a Thematic Analysis approach

We followed the thematic analysis approach [13] to analyze qualitative data (user-free texts describing the tasks in the diary), which was agreed upon by three researchers in our group. This approach has often been seen as a fundamental technique for analyzing qualitative data. We adopted a six-step process that was previously outlined in [24]: 1) familiarizing with the data, 2) generating initial codes, 3) searching for themes, 4) reviewing themes, 5) defining and naming themes, and 6) producing the report.

The first author completed the first pass of Steps 1-2. For Step 2, we carried out an extensive literature review of task classification and a common coding scheme. We found tasks have been categorized based on the following common factors: *task goals*, *individual intentions*, and *substance domain*.

Task Goals: Goal-driven task categorization has been extensively studied and used in many early works [19, 40, 58, 71, 72, 91]. Researchers con-

sidered the output target of the task in the categorization scheme. The previous work has also proposed data-driven categorizations that do not include any domain-specific task types and hence are broadly applicable to other domains as well [72]. The categorization is task goal-driven and particularly suited for studying real-life search tasks. It aims to derive categories by seeking an answer to the question: "What goals are the users trying to achieve in the task?". Examples of goals are whether the user is trying to communicate information or learning or achieving intellectual targets.

Individual Intentions: The user's intention behind the tasks that influence the search process has been considered as a factor and studied in many research works [55, 59, 76]. People searching for information related to their hobbies or work can be driven by different individual intentions even though they would aim for a similar goal. We followed the abstract concept of the everyday life information seeking model [76]. The individual intentions factor refers to preferences given to a task based on the individuals' choices in everyday life, thereby answering the following question: "What individual intentions are the tasks serving?". The individual intention classification divides things into diverse groups according to their value to the searcher.

Substances: is a third often-used source for categorization, which answers the following question: "What is the main domain that defines the task?". This factor has been particularly used in modeling information seeking for one specific professional group in one study, for example, nurses [45], vault inspectors [82], clergy [25], or researchers [56], and city administration [73]. For instance, all business-related tasks belong to the same substance domain of business regardless of their goal or intention. Task categories regarding the substance factor are mutually exclusive, which means that every task must belong to only one substance category. However, in actuality, a task may have the features of several substance categories. For example, a studying and researching task may be related to programming work. In these cases, the category was separated, and we selected the category that was more emphasized by the participant's task description. For example, in the case of programming tasks, all programming tasks were separated under a new category.

In Step 3, for each task factor (*task goals*, *individual intentions*, and *substance domain*), the first author formed an initial list of detailed low-level categories for the tasks and a set of candidate themes. Steps 4 and 5 were iterated among the three authors. Then two authors independently categorized the whole set of tasks. Cohen's Kappa test indicated high

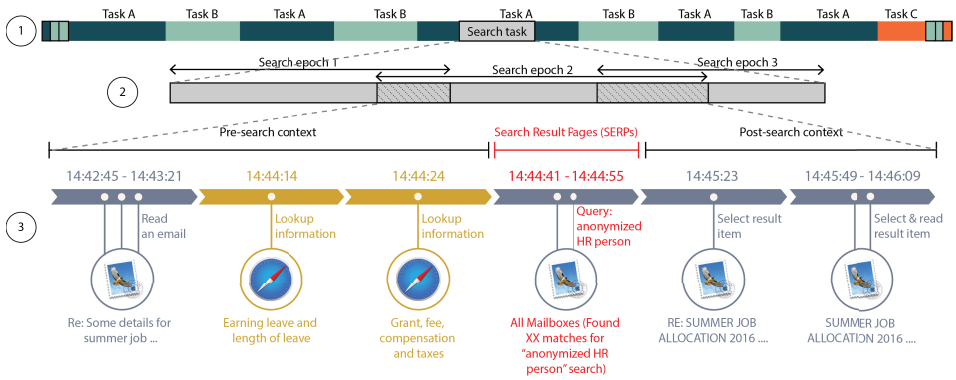


Figure 4.1: The granularity of a diary task that contains naturalistic search tasks and associated search epochs. 1) A user's digital activities include several diary tasks. 2) A search task is composed of three search epochs. 3) An overview of a search epoch. In *pre-search context*, a user engaged in a search task involving verification of factual information regarding human resource policies to reply to an email from a new employee. *A query frame* is the screen frame containing the user's issued query. In *post-search context*, a user used the retrieved information to respond to the email.

agreement between the coders. All categories were developed only from the task description participants wrote in the diary. The naming of the themes for the categories was done post-hoc.

4.2 Search Tasks Extraction

A search task includes a query or several queries. It has a uniform motivation or an information need that evolves seamlessly in the workflow of a diary task as a motivation for conducting immediate search activities. Figure 4.1 demonstrates how a search task was formed. To effectively identify search tasks, we decomposed a search task into one or several search epochs. Each search epoch contained a user-submitted query to the search engine and the associated pre-search and post-search context. To determine whether multiple continuous search epochs belonged to the same search task, we used the corresponding task in the diary as the context for understanding whether several search epochs shared the same search goal and belonged to the same broader diary task.

A broad spectrum of search tasks was extracted in the experiment. For instance, we extracted local file search activities using OS-specific applications, such as Finder, Spotlight, and Explorer. We also recorded searches

using map interfaces, such as Google Maps, with typed queries, drags, clicks, and searches in email clients and custom searches on websites.

4.2.1 A Search Epoch

The preliminary step of our analysis was to detect search epochs from the digital activity logs. Figure 4.1 (Part 3) illustrates a search epoch from a participant's digital activity logs. A search epoch comprises three parts: a query, pre-search context, and post-search context, which are described below:

- *A query* is a SERP that was logged in response to a query issued by a searcher. The regular expressions in Appendix[1] were applied to find all candidate queries in the participants' digital activity logs.

- *Pre-search context* is a temporal sequence of information objects recorded at two-minute intervals prior to the query frame. In the case of missing information objects due to the computer being idle within the pre-search context, we extracted one information object temporally preceding the query frame.

- *Post-search context* is a temporal sequence of information objects recorded at two-minute intervals subsequent to the query. Similarly, when there are no existing information objects in the post-search context, one temporally successive information object from the query is extracted.

4.2.2 A Search Task

Based on determining search epochs from the digital activity logs in the previous step, we formed a search task consisting of a set of search epochs. Figure 4.1 (Part 2) illustrates how a search task is formed. Search epochs can, but do not have to, follow each other temporally. In other words, a search task can be one isolated search epoch when the pre-search context and post-search context do not overlap with subsequent search epochs. In another case, several continuous search epochs sharing the same information need are combined as a search task.

[1] https://tinyurl.com/ybcyasd4

4.3 Task Categories

Given the diary task identifiers, each search task and its search epochs were mapped to the diary task; and diary tasks were classified to the task categories according to the methods described in Section 4.1 to obtain the following categories. The task factors and their categories are presented in Table 4.1.

Four categories were formed under the *Individual Intentions* factor: 1) Tasks with the intention of *Being Creative* shared the two dominant features, which were writing/composing documents. 2) Tasks with the intention to *Enjoy Oneself* shared two common features, which included social media activity and video streaming/music listening. 3) In *Gain Knowledge*, the tasks were described with the two features of learning and research-related activity. 4) The rest of the tasks fall into the *Daily Activity* category. These tasks represent a variety of routine activities, such as continuously making travel plans/accommodation arrangements, online shopping/daily e-commerce, following up-to-date news, and managing personal information.

The *Task Goal* categorization adopts the earlier categorization [72] and is based on the following generic features: 1) Tasks with a *Communication* goal have the main feature of communicating with other people as the precondition for success within the task. These can include going through email conversations, replying to the messages, or taking part in a live video call. 2) *Maintaining/advancing* category has the feature of whether the task is at the core of the Substance of the work or, instead, that supports the main function. These are typically information searches for administrative tasks or tasks where an expected larger output is approached gradually. They were easily recognizable from the task descriptions with "reviewing", "starting something new", or "continuously updating a document". 3) *Seeking or receiving information* are tasks that aim to acquire a specific piece of information by actively seeking it or passively receiving it. The diary entries corresponding to this goal often began with "finding something", "looking something up", or "watching something". 4) Tasks with *Intellectual* goal has the feature of demanding a degree of intellectual effort.

Categories in *Substance* task factor reflects the domain substance of tasks in the data. These include five general categories: 1) *Free-time*; 2) *Business* or industrial job-related tasks, these excluded tasks in the academy; 3) *Programming* tasks' scope can be the whole process of software development, not just coding or scripting; 4) *Social life* tasks mostly involved social media activity; and 5) *Studying and researching* tasks can be academic or industrial research and development tasks.

Task Factor	*Task Category*	*Examples*
Individual intentions	Be creative	Essay writing-marketing, promotion ...; Proactive search simulation analysis.
	Enjoy oneself	Watching movies and TV shows; Browsing an online forum for knitting and crochet.
	Gain knowledge	Reviewing discriminative representation learning; Foreign language studying with two teachers.
	Daily activities	Trying to find accommodation in a city; Following up latest news on BBC, CNN.
Task goals	Communication	Writing emails to a potential summer trainee; Arranging a job interview at a company.
	Maintaining and advancing	Updating a to-do list for a project; Starting ICT company papers.
	Seeking or receiving information	Finding scientific venues; Finding and listening to music.
	Intellectual	Studying C++; Preparing a pitch deck for a startup.
Substances	Free-time	Watching "Lost" seasons; Listening to music on a QuickTime player.
	Business	Generating business calculations; Writing an article about a startup with the entrepreneur.
	Programming	Modifying a user interface; Extracting keywords from software with KEA.
	Social life	Looking for friends on Facebook; Checking Slack.
	Studying and researching	Reviewing MTAP paper Face super resolution based on ... ; Reading LSTM RNN recurrent neural network.

Table 4.1: Task factors, the corresponding task categories, and examples are taken from the diaries.

Application Categories	*Description*
Social	Applications and websites where main function is to enable communication between people.
Search Engine	General Web search engines form a category of their own. Users re-visited the search engine application after some time to perform new searches.
Support Application	Applications or language support tools that support the search task.
Transactional Web	Websites that are typically used for manifold interactions and that support interaction and even enable transactions.
Static Web	Static websites that are typically used for browsing and that do not support or encourage much other interaction or transactions.
Local Application	Local applications that are installed on the participant's computer.
Other Web	Rare websites are placed in this category.

Table 4.2: Application categories are data-driven. They are categorized based on common function and type of use.

4.4 User Behavior Factors

We examine search behavior according to various aspects: 1) application context refers to an application used prior to a search, and 2) content-trigger refers to the content the user observed that triggers a search.

4.4.1 Application Context

We manually categorized applications into seven categories based on their common functions, types, and fields of use. The application categories are presented in Table 4.2. The *Social* category included applications where the main function was to enable communication with other people (e.g., Skype, Mail). General web search engines and information sources (e.g., Google, Wikipedia) were categorized into the *Search Engine* category. Participants used many dedicated tools to support their search tasks (e.g., various digital dictionaries), and these were categorized into the *Support* category. The *Transactional* category typically featured websites used to support interaction and even to enable transactions (e.g., online stores, journey planners). Meanwhile, the *Static* category included static websites that did not support interaction or transactions (e.g., personal weblogs or online tutorials). Locally installed applications were also grouped as the *Local* category, but

this category excluded applications that were categorized into the categories as mentioned earlier. For instance, the main function of instant messaging applications was to interact with others socially; thus, we classified it to the social category. Finally, any website that rarely occurred in the recorded data was placed into the *Other* category.

4.4.2 Content-Trigger

This factor refers to the textual content that triggers the search process based on the evidence obtained from the connection between the selection of query terms, and the content users have seen before the search. Content-triggers were determined by comparing the query's keywords with the content the users had seen on the screen in the pre-search context. A program was implemented to automatically extract whether any term in the set of keywords that existed in the content of the information objects in the pre-search context. Taking a search task in Figure 4.1 as an example, the query "anonymized HR person" was submitted to Mail's search interface. The phrase of this query originally appeared on the screen in the pre-search context that triggered user search activity. Content-triggers that were a combination of stopwords were discarded. We also noted that several keywords featured a set of stopwords during the process but were meaningful in the pre-search context. We further manually checked and verified the correctness of individual content-trigger.

4.5 Measures

The following set of measures was defined to operationalize user behavioral factors:

Application Context: The application context was measured as the share of application category appearances directly before searching across task categories. If no applications were used within 2 minutes before searching, the application context was assigned to the "search engine" application.

Content-Trigger: A binarized variable was used to characterize whether a search epoch contained content that triggers the users to search. If the query appeared in the pre-search context of the search epoch, we marked the search epoch having content-triggers. We computed the percentage of search epochs that have content-triggers across task categories.

Task Factor	*Task Category*	*Content-trigger*	*p-value*
Individual intention	Be creative	0.69	**0.010**
	Enjoy oneself	0.60	
	Gain knowledge	0.55	
	Daily activities	0.60	
Task goal	Communication	0.56	0.206
	Maintaining and advancing	0.67	
	Seek or receive information	0.59	
	Intellectual	0.61	
Substance	Free-time	0.63	**0.001**
	Business	0.58	
	Programming	0.82	
	Social life	0.59	
	Studying and researching	0.57	

Table 4.3: The results of content-trigger were measured as the percentage of search epochs that have content-trigger with respect to task categories. Bold values indicate significant difference among the tasks.

4.6 Results

A total of 688 naturalistic search tasks were identified in the screen monitoring and digital activity monitoring data. Participants reported 119 diary tasks, and 69 diary tasks containing search epochs which were analyzed in the study. In the following sections, we discuss the results for content-trigger and application context.

4.6.1 Content-Trigger

Table 4.3 presents the percentage of search epochs having conten-triggers in different task categories. The results indicated a statistically significant difference in how often the search was triggered by content observed on the screen when it came to individual intentions. While users were "being creative", information need was mostly triggered by the content. On another hand, "gain knowledge" tasks are less frequently influenced by the content-triggers.

The task goal was not dependent on the content-trigger, but the substance was dependent. The "programming" tasks showed a significantly high percentage of searches triggered by the content, while "social life" tasks had fewer searches triggered by the content seen on the screen.

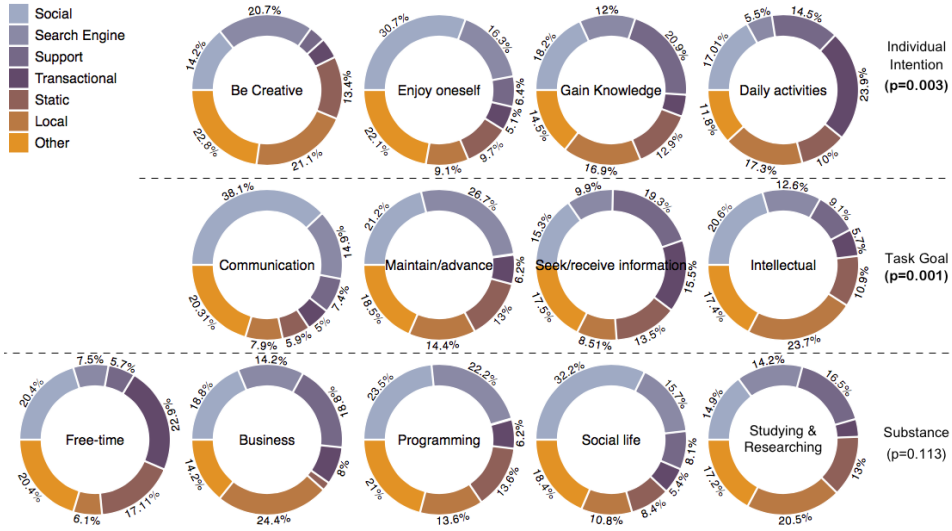


Figure 4.2: Results of application context with respect to task categories. The number of application categories occurrences prior to searches is measured as percentages.

4.6.2 Application Context

Figure 4.2 shows the general results of the percentages of the seven application categories. Overall, applications in social categories are by far the most common applications used in the pre-search context. Individual intentions and task goals were related to the application context, whereas substances of the tasks were not significantly related.

For the individual intention factor, the "be creative" category had the highest percentage of using "other" applications as a prior application context. Typically, while being creative, users frequently moved from rarely, or single-time used applications to the search engine. For the "enjoy oneself" category, most application context falls into the social application category. Users in "gain knowledge" tasks mostly used support applications before searching. Finally, transactional applications were mostly the application context before the search in "daily activities".

For task goals, interestingly, "communication" tasks had the highest percentage of the application context from social applications. For "maintaining and advancing" tasks, users often re-visited the SERP on search engines and carried out a new search. The "seeking and receiving information" category had the highest percentage of moving from support ap-

plications to search engines. "Intellectual" tasks mostly began with local applications (e.g., a PDF reader) and moved to a search.

4.7 Findings

Searching occurring in social application context was frequently induced by social applications when the task intent was to "enjoy oneself". This finding confirms that queries are often issued within social media services with the leisure intent, such as finding celebrities appearing in streaming movies or music videos, locating people with similar interests, and navigating to friends' pages to investigate social media activity. This result suggests that users' leisure information-seeking activity occurs inherently within social media services or raises from social communication platforms.

Search behavior in tasks with an intellectual goal, such as analyzing, researching, reviewing, and writing, was more often induced from utility applications, such as word processing applications, spreadsheet applications, or programming platforms. Although not surprising, this suggests that intellectual tasks are strongly associated with applications that support knowledge work. A large portion of tasks with the intent of being creative was found to have an intellectual goal. Consequently, search behavior in creative and intellectual tasks was induced by the artifacts that the users were producing and occurred in the context of utility applications. This suggests that search activity is integrally associated with the users' creative processes.

The analysis reveals that the applications are forming the context of users' digital activity before information search can inform us about the design of information access systems that could benefit from considering task context to identify and even predict specific kinds of search support that might benefit the user. Application sources can be linked to task categories constitute a useful step in adapting information retrieval environments. This can also be promising for customizing search by accounting for the importance of applications used and information seen prior to the search.

Chapter 5

Task Context Modeling

The research described in this chapter aims at answering RQ2: **Can the association amongst the entities be used to model the user's task context for the recommendation?** In this chapter, we introduce the implementation of the task context model. The 24/7 behavioral recordings consisting of screen monitoring and digital activity data were fed into an unsupervised machine-learning method to build a user model to detect the tasks that the user was engaged with. In general, the model could generate recommendations concerning the task context. In Publication II [86], we set out to study: 1) *How accurately can we detect user tasks from 24/7 behavioral recordings*, 2) *How accurately the resulting task model can be used to detect real-time task and proactively suggest relevant entities.*

The results show a task detection accuracy of more than 70% using the rich data and the corresponding model-based document recommendation with a Normalized Discounted Cumulative Gain of 95%.

Finding 2 Topical relatedness amongst the entities is useful in modeling and inferring the user's task context.

5.1 Data Description

Here we used 24/7 behavioral recordings (Data 1) to train the model. The data was collected from the continuous monitoring of the screen and digital activities of 10 participants for 14 days. The data consists of a collection of information objects (document, email, file, folder, etc.), each containing a merged OCR-processed document and application-specific operating system data. We treated the data as multi-aspects to train the model. These

aspects can be contextualized by several entities such as applications, documents, people, and various keywords determining the task. To recognize keywords and named entities in information objects, we utilized Watson Natural Language Understanding API [1]. To identify application names and document information, we considered the user's operating system data.

5.2 Modeling Technique

In this work, we use entities as the basic unit that represents the user's task context. Entities consist of applications, keywords, other named entities, and non-entities terms that provide a richer source of information to learn more expressive user models. Entities are stored in a vector representing the document and the document vector as matrix X. Here, we consider each information object as a document in the matrix. Therefore, each document is represented as a bag of individual entities in which non-zero elements are the entities present in the current information object.

Latent Semantic Analysis (LSA) was then run on the matrix X. It uses Singular Value Decomposition (SVD) to encode the original matrix to a low dimensional latent space:

$$\hat{X} = USV^{\mathrm{T}}, \tag{5.1}$$

where $\hat{X}$ is the low dimensional latent space of X. Then, we compute rankings for the entities and the documents from the decomposition for task context detection and entity recommendation, as explained in the following sub-section.

5.2.1 Task Detection and Labeling

The LSA resulted in a lower-dimensional representation, but as an unsupervised method, it is not directly interpretable. In order to allow interpretability of the lower-dimensional representation to the users, we developed an approach to label the tasks, i.e., find keywords and named entities that describe the dimensions in the lower-dimensional output space. Figure 5.1 presents an example of a detected task with visualized labels.

To select the keywords and named entities, we compute a ranking for the terms by using the matrix product US, which represents the relationship between terms and tasks. Five terms with the highest values in US were selected as seed terms. However, these seed terms were very general and not necessarily descriptive from the users' perspective. Therefore, we

[1] https://www.ibm.com/cloud/watson-natural-language-understanding

utilized Word2Vec to compute an embedding of keywords, named entities, and terms. We then selected keywords and named entities that frequently appeared in the task and were close to the highest-ranked seed terms in the Word2Vec embedding space. By doing so, we ensured that the entities were both related to the overall topic (close to the seed terms) and frequently appeared relevant to the task.

5.2.2 Entity Recommendation

Task labeling indicates how accurately the model can detect the tasks and make them interpretable. However, this does not indicate the model's usefulness for retrieving information. Therefore, we built an entity recommendation method that retrieves a ranked list of documents (or information objects: emails, instant messages, Web pages, textual documents, files, folders, etc.) and applications in response to a detected task. The rationale was to be able to study whether the task model can be used to proactively re-find documents and applications that could be valuable resources for the user in the task context.

The vector space model of information retrieval with cosine similarity ranking was used to retrieve and rank the documents represented in the lower-dimensional space. The input was the recent context vector (recent screen frames) and the lower-dimensional task vector. The model then ranked documents using the data from the original higher-dimensional space.

The documents were further grouped concerning the application from which they were captured. For example, in Figure 5.1, the documents that were opened using the same OBS video editing software were grouped under the application name of that OBS software on the user interface. Therefore, the document list on the user interface consists of two dimensions: documents relevant to the task and applications used for the task-related work.

5.3 Evaluation

The participants were called back to the laboratory to provide a ground-truth assessment on the quality of the task models. The participants were asked to compare the output of the methods to their diaries and assess the relevance of tasks, keywords, named entities, and documents.

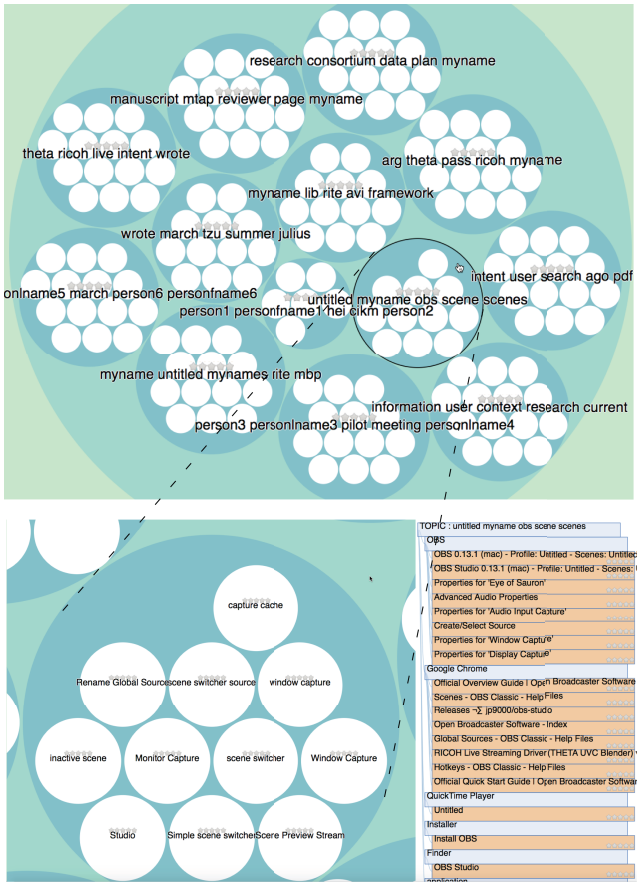


Figure 5.1: A screenshot of the system interface. *Top*: is the task view showing all detected tasks and the relevant terms (seed terms) for the tasks. Each blue circle visualizes a single independent task. *Bottom left*: a view of a detected task "Video capture with Theta" for which the associating circle was zoomed in. Inside the circle, a set of keywords and named entities extracted from OCR-processed documents were used to describe the task. Descriptive keywords and named entities could be video recording features, software, and tools that were used for the task, and the person entities who were involved in the task such as "capture cache", "window capture", "inactive scene", and "Rename Global Source". *Bottom right*: a recommendation view that presents a list of documents in response to the task. The recommended documents were from various applications such as text documents, files, folders, emails, etc. For instance, under "OBS" software, video projects or new scene files created and used for the video-making task were recommended; under "Chrome" browser, the relevant tutorials helpful information that the user looked up for the task were also suggested.

5.3.1 Task Detection Experiment

This experiment aims to answer Question 1) *How accurately can we detect user tasks from 24/7 behavioral recordings.*

Before starting the experiment, we trained the task model on the participant's screen monitoring data and digital activity data. The output of the task model was visualized and presented on a user interface (see Figure 5.1). The participants were asked to compare the elements (task description - five seed terms, keywords, named entities, and documents) on the interface to their diaries and rate each element accordingly. There were two levels of assessment:

- Accuracy: The participants were asked to explicitly indicate either (0) for a task in the diary if there were no matching tasks on the interface, and for the keywords, named entities, and documents, those that do not belong to the given task. Otherwise, above (0) indicates a task that was formulated correctly on the interface; and for keywords, named entities, and documents that belonged to a corresponding task. This provided the ground truth that was used to evaluate the accuracy of task detection and entity prediction.

- Relevance of the content of the task: We used the following scale (1) slightly relevant; (2) moderately relevant; (3) highly relevant; and (4) absolutely relevant. This was used to evaluate the relevance of task-related recommendations.

Measures

To measure the quality of the produced task models, we used mean score and detection accuracy. The score was computed simply as an average rating that the users gave for the task description. The accuracy was computed as the binarized output: scores greater than (0) were marked accurate, and scores of (0) were marked inaccurate.

To measure the quality of the recommendation, we used precision and normalized discounted cumulative gain (NDCG) at cut-off levels. Both are commonly used metrics in information retrieval [44].

Task Detection Results

Overall, the mean scores and detection accuracy for the tasks displayed to the participants were 2.754/4 (std=0.96) and 72.27%, respectively. For which, according to our scale, this indicates relevant to highly relevant task detection.

	Documents	*Keywords*
NDCG	0.92	0.72
Precision@1	0.93	0.80
Precision@5	0.93	0.70
Precision@10	0.92	0.70
Precision@20	0.89	

Table 5.1: The results of the document recommendation and keywords retrieval as precision at N and NDCG.

During the experiment, we noticed that some tasks were assigned a high score by the participants; however, they were not reported in the original diary. This result indicates that the system had detected meaningful tasks for the participants, although participants forgot to enter them in the diaries. This allowed us to compute the precision for task detection, which was higher than the accuracy, for there were additional tasks were found during the assessment and in the ground-truth pool. The precision was 76.85%.

Task Labeling Results

The "keywords" column in Table 5.1 shows the results of the task labeling. The NDCG and Precision at 10 were computed for the keywords and named entities that were retrieved for the task, as in general, more than ten keywords and named entities would not be useful for the participants to recognize a task, but rather cluttering. NDCG for the keywords was 0.72 and Precision for the first keyword was 0.8. This indicates that the quality of retrieved keywords was relatively high and that in over 80% of the cases, the participants could recognize the tasks with the first keyword.

Document Recommendation Results

The "document" column in Table 5.1 shows the results of the quality of recommended documents. The Precision and NDCG were computed for the top-20 documents that were recommended to participants in the experiment. NDCG for document recommendation was 0.92, and Precision was stable at 0.9 for the list of the top-20 documents. This indicates that most of the recommended documents were found to be highly relevant for the task.

5.3.2 Single-trial Task Detection and Recommendation Experiment

The purpose of the single-trial experiment was to study the usefulness of the resulting task model in real-time single-trial task detection and actual recommendation scenarios. It aims to answer Question 2) *How useful the resulting task model can be used to detect real-time tasks and proactively suggest relevant entities.*

We examined whether the model could correctly classify unseen input resulting from user interactions to a task in the task model and proactively recommend related information.

Participants and Apparatus

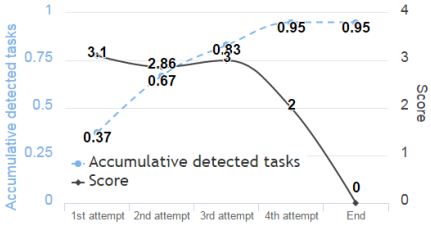
The same participants were invited back to the laboratory one week after the task detection experiment to provide assessments on the quality of recommendation in online interactive settings. Similar apparatus and relevance assessment were applied for the proactive recommendation of task-relevant information.

Procedure

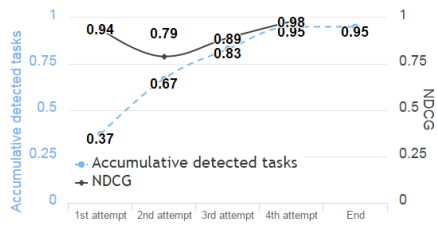
To begin the single-trial task detection and recommendation experiment, the participants were asked to select six tasks from their diaries. The participants used a computer running the screen monitoring and digital activity monitoring systems to perform activities related to the selected task one at a time while receiving recommendations from the system. Participants were explicitly advised to continue their tasks (i.e., to perform new activities dedicated to the chosen task). The new unseen screen monitoring data and digital activity data were fed into the LSA model, which resulted in the prediction of the participant's task.

Figure 5.1 shows the visualization of the predicted task for the user. When the system detected a task, it zoomed in to the circle representing that task and proactively recommended the documents that could be used for the task from the digital activity database.

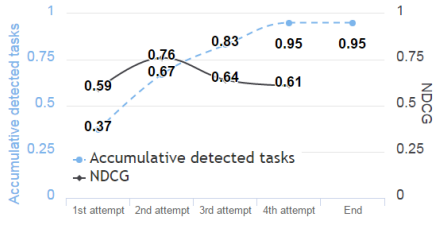
For each task, the participants were interrupted every 30 seconds to provide a relevance assessment on the task detection and recommendation. More precisely, every 30 seconds (up to 120 seconds), we asked the participant to look at the system interface and assess the relevance of the detected task on which the system zoomed in. We also asked the participants to assess the keywords and documents if the task was detected correctly. After



(a) Task detection accuracy over time



(b) Document retrieval NDCG over time



(c) Labeling NDCG over time

Figure 5.2: Results of the single-trial task detection and recommendation. The figures present measures: score and accuracy (y-axis) concerning the elapsed task time: the attempts (x-axis).

4 attempts (120 seconds), if the task was not detected correctly, the participant marked the task as failed, and the participant could continue with the next task.

Measures

The main evaluation criterion was the quality score that the participants provided. We also used NDCG and precision at N to measure the performance of the information retrieval performance of documents and keywords. All measures were computed at every interruption point (at 30, 60, 90, and 120 seconds).

Results

The overall result after all trials and attempts (mean over the tasks and participants at 120 seconds) shows a task detection accuracy of 95%. More precisely, 57 out of 60 tasks (6 tasks per participant, 10 participants) were correctly detected.

Temporal graphs in Figure 5.2a presents the results at different interruption points. After the first attempt, the system was able to detect the

Attempt	*1*	*2*	*3*	*4*
Score	3.31(0.61)	3.21 (0.67)	3.18 (0.79)	2.67 (0.98)
NDCG	0.94	0.79	0.89	0.98
P@1	1	0.78	0.9	1
P@10	0.95	0.82	0.87	0.92
P@20	0.93	0.89	0.87	0.83

Table 5.2: Document precision at 1, 10, and 20 in the single-trial task detection and document recommendation experiment. Results are reported with respect to attempts (task interruption at 30 seconds intervals).

user tasks with an accuracy of 37%, and after the second attempt (60 seconds), the detection accuracy was 67%. This indicates that the model could detect the majority of the tasks after one minute the user started to interact with the computer. After the third attempt, 83% of tasks could be detected correctly. The results indicate that the best task detection accuracy could be achieved only after only two minutes of digital activity monitoring with over 95%.

The temporal results also showed that the average scores that the users assigned were high after the third attempt (3.18-3.31/4), but then the score dropped after the fourth attempt due to the tasks becoming harder to detect. This indicates an expected trade-off of a majority of the tasks being easy to detect even in a single-trial setup, and a small portion of the tasks being very difficult to detect.

The results in Figure 5.2b present the NDCG for top-20 documents concerning the attempts. NDCG was found to be high throughout the attempts, varying between 0.79 to 0.98, with a lower value at the second attempt. It should be noted that both the quality score and the NDCG were already high starting from the first interruption, indicating that when the tasks were detected correctly, the document recommendation also worked well with real-time monitoring input.

Table 5.2 shows a summary of the results for the document recommendation. The results show that precision and NDCG for the top 20 recommended documents were consistently high throughout the attempts, with NDCG over 0.9 at the first attempt (after 30 seconds of digital activity monitoring). We observed a slight drop at the second attempt. This indicates that some documents were harder to retrieve and required more evidence for the task model to converge to the correct task and the improved document ranking. Similarly, Table 5.3 presents the results for task

Attempt	*1*	*2*	*3*	*4*
Score	1.20(0.67)	1.91 (0.67)	1.66 (1.01)	1.63 (0.73)
NDCG	0.59	0.76	0.64	0.61
P@1	0.59	0.89	0.80	0.86
P@5	0.60	0.72	0.64	0.60
P@10	0.52	0.62	0.63	0.60

Table 5.3: Labeling precision at 1, 5, and 10 in the single-trial task detection and document recommendation experiment. Results are reported with respect to attempts (task interruption at 30 seconds intervals).

labeling. The values are generally low for labeling, indicating that labeling tasks remain challenging compared to recommending task-relevant documents.

5.4 Findings

This research aimed to investigate the modeling of digital tasks from 24/7 behavioral recordings for task-aware information recommendation. We explored the two sub-questions defined earlier in each of the two experiments.

How accurately can we detect user tasks from 24/7 behavioral recordings using unsupervised learning? The task detection experiment showed that it is possible to detect participants' tasks by only monitoring their behaviors inside the screens. The unsupervised learning method was utilized, and the models were trained on 24/7 behavioral recording data with a task detection accuracy of 72.27%.

How accurately the resulting task model can be used to detect real-time tasks and proactively suggest relevant entities? The second experiment on single-trial task detection showed that participants' tasks could be detected in real-time at 95% of accuracy from their unseen interactions. Most of the tasks were detected after two attempts within one minute from the beginning of the screen and digital activity monitoring.

Overall, the document recommendation was successful in both experiments with a task detection accuracy of over 90%. In the first experiment, the NDCG for document recommendation was 0.92, and precision was stable around 0.9 for the list of the top-20 documents. In the second experiment, the NDCG of model-based document retrieval was also over 0.9 throughout the attempts. This indicates the high document recommendation effectiveness with our model. In both experiments, however, the

effectiveness of the task labeling was more challenging. The participants seemed to be more satisfied with the recommended documents than with the keywords and named entities.

To conclude, in this chapter, we presented the task modeling approach by exploiting 24/7 behavioral recordings. We showed that it was possible to detect tasks that the user performed by tracking the topical association amongst the entities in the data. We conducted a task detection experiment built an online system to proactively recommend real-time task-relevant information corresponding to screen monitoring and digital monitoring inputs. The experimental results show that our method can detect these tasks with high accuracy and recommend task-relevant information automatically using the task model.

Chapter 6

Effect of Temporal Information

In this chapter, we explore the use of temporal associations amongst the entities to model the context to answer RQ3-1: **Does the use of temporal information improve recommendation quality?** To achieve this goal, we tested the models in two conditions: a control condition with no temporal information considered (static model) and an experimental condition where the model incorporates temporal information. The static model was utilized with the assumption is that the next activity corresponds only to the current activity. On the other hand, the model with temporal information considers a sequence of a user's past activities to predict the next activity.

Publication III [67] reports a comprehensive description of our task context model that considers temporal information and evaluation of entity prediction. We evaluated our model by measuring the accuracy of predicting the user's task context and accuracy of predicting the entities that would occur next, e.g., predicting the subsequent applications and documents the user would open. This study is an offline analysis to evaluate our modeling approach; therefore, we only attempted to predict the entities rather than present the users' suggestions in real-time.

Results of the study showed that our model with temporal information could predict the user's task context with high accuracy. Temporal based task context model performs better than the static model in terms of prediction accuracy of document usage. Our proposed model, therefore, satisfies desirable properties in user behavior modeling. This performance improvement is enabled by modeling the time-varying properties of user activities and their dependencies. This answers RQ3-1, and the finding can be framed as:

Finding 3 Temporal information is useful in improving recommendation quality. This suggests that the user behaviors are often influenced by time and that tasks can be characterized by temporal regularity and repetition of the user's activities.

6.1 Modeling Technique

In this section, we describe a static model that does not take into account the temporal dynamics in digital activities and a temporal-based model that considers sequential patterns over a user's interaction data.

6.1.1 Static Model

We have entities extracted from information objects. We considered a user activity at each time step as a user state. At each time instant, the user is providing input as an information object based on her state describes the access of information resources on the computer (e.g., documents, files, e-mails, chat messages. Therefore, a user state is defined as a vector comprised of all entities that represent a user's context at that specific time. Each user is modeled as a sequence of states. We treat each information object, including extracted entities, as a document. Inspired by the bag of words (BoW) model, each document is represented as a bag of individual entities in which non-zero elements are the entities present in the current screen frame. The logged context is stored in the matrix $X \in R^{|E|\times|S|}$ where $|E|$ and $|S|$ denote the sets' size. We encode the user states into a low dimensional latent space such that co-occurring entities in an information object should have a similar representation.

Latent Dirichlet Allocation (LDA) was run on the matrix X and projects each information object onto latent principal factor space. Each information object is generated as a mixture of multiple distributions. The generative model can be described s follows:

1. Choose $\theta \sim \mathit{Dirichlet}(\alpha)$
2. For each entity e in information object d
   - Choose a topic vector $z \sim \mathit{Multinomial}(\theta)$
   - Choose an entity e from the multinomial $p(e|z, \beta)$

Where θ, α, and β are topic proportions, Dirichlet parameter, and topic-entity density, respectively.

6.1.2 Temporal based Model

The topic representation of each document is aimed to compress what is in the information object at each time frame, and we can also compress what happens over time. Topic representation at each time step is considered a state at that specific time. The second component of our model is to implement sequence learning on the reduced size vector of a state. This module is aimed to predict the next user state in the sequence, which is the future topic vector that is expected to be produced by the user. We use the Bidirectional Long-Short-Term-Memory (BiLSTM) [77] based sequence learning system that helps model the user state. BiLSTM based recurrent neural network has demonstrated state-of-the-art performance in various kinds of tasks with sequential data such as machine translation, speech recognition, time series prediction, etc. It is used to process the sequence of input and predict the most likely future continuation of the sequence. This capability of learning long-range dependencies makes BiLSTM networks particularly attractive for user modeling.

Formally, given a topic vector z_t representing the state at the time step t, the corresponding hidden state h_t can be derived using equations defining the various gates. We assume that topic distribution in each information object represents the state of the user. We build a sequence of states and utilize the BiLSTM to show the time-series relations amongst the topic vectors. Our intuition behind using the BiLSTM neural network is to use all available information and effectively model the local dependencies between certain states of the user temporally. We divide a sequence of states $z_1, z_2, ..., z_t$ into a fixed-sized sliding window of size W, and each sequence is formed as $\{z_{t-W+1}, ..., z_{t-1}, z_t\}$. Given the last W of user states in this window, the BiLSTM network learns to predict the next state of the user. The loss $MSE(\hat{z}_{t+1}, z_{t+1})$ is measured using mean-squared-error and is used to train the model using back-propagation. The trained network is then used to predict the future latent vector in the test data set. The output of the network depends not only on the latest latent vector but also on a window of previous latent vectors.

The LDA model provides the probability of entities at each topic. The BiLSTM model predicts the probability of each topic at the next time step. By knowing these probability values at time t, the probability of a given entity e_t assuming N topics is computed by:

$$p(e_t) = \sum_{n=1}^{N} p(e_t|z_t = n)p(z_t = n). \tag{6.1}$$

Top-k entities are generated by sorting entities in descending order. That

is, entities in each type (applications, documents, persons, and keywords) that are most consistent with the future state are retrieved.

6.2 Experimental Setup

In this section, we begin by describing the data used for modeling. This data contains users' real-world information behaviors useful in understanding and modeling task context. Then, we discuss the measures used to evaluate the predictive performance of the models.

6.2.1 Data Description

We conducted a data collection experiment in which we continuously monitored digital activities for 14 days (Data 2). The screen monitoring and digital activity monitoring systems were installed on 13 participants' laptops to collect the documents they opened, applications they used, and other people they talked to (on instant messaging applications and email).

The data was first pre-processed. All the application/document usage records were extracted, including application window, active application, email sender/recipient, the person in instant messaging applications (e.g., Skype and Whatsapp). Application window refers to the current opening file using a specific application (e.g., EXCEL.EXE) and the current window title (e.g., ConsoleWindow or MessageBox or InternetExplorerPanel1). Note that a user typically has several windows per application. Active information object refers to the document/file/email/message that the user is currently working on. Email senders and recipients are determined in an email. Similarly, persons in instant messaging applications are extracted from a chat window. Keywords are extracted from the textual content of the active window (OCR-processed document). The timestamp is the time when the application window becomes active. Finally, stopword removal and lemmatization were applied to the OCR-processed document.

6.2.2 Conditions

To understand whether leveraging temporal information improves recommendation quality, we tested our models in two conditions: a control condition with no temporal information and an experimental condition in which temporal information was considered.

- **Control condition**: In this condition, we included the static model (LDA), which considered only topical association amongst the entities.

- **Experimental condition**: In this condition, we included the temporal based model (LDA +BiLSTM), which considered both topical and temporal association amongst the entities.

6.2.3 Evaluation

The evaluation aimed to show that the temporal information incorporated in the task context model is beneficial for inferring context and predicting entities relevant to the context. Therefore, we compared the model's prediction performance in the experimental condition constructed for the collected data set to the model in the control condition, which does not account for temporal information. The analysis was conducted for each participant using the data fields of the logging data trace.

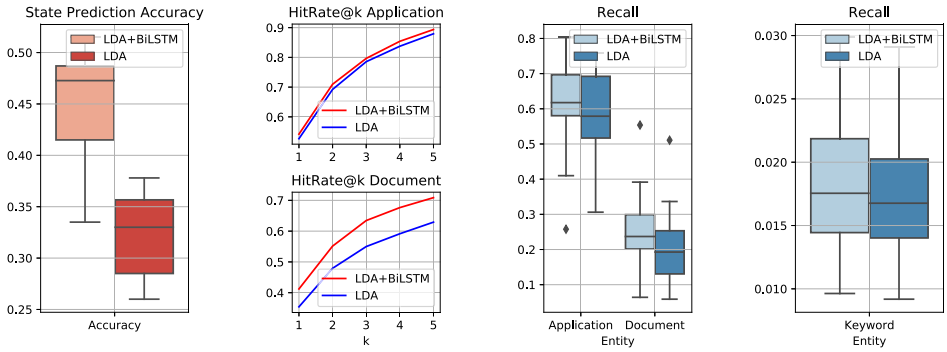
We conducted three-fold cross-validation. The prediction models were evaluated by splitting the participants' collected data into train and test sets (60% data for training = first 10 days, and predicting the remaining 40% data = 4 remaining days). That is, the testing set is independent of the training set. We measure the prediction accuracy by 1) prediction accuracy of the user's task context or attempt to predict the topic of the next user action e.g., the topic describing the active document; 2) hitrate@k and recall of predicting the documents and applications the participant will open in the next time step, the content (keywords) will occur in the information object e.g., document, email, etc.

6.3 Results and Findings

The obtained prediction accuracies are depicted in Figure 6.1. Task context model utilizing temporal information outperforms the static model. This temporal structure information significantly improves entity prediction. The context model performed well when predicting document usage.

Our model is constructed based on the assumption that routine tasks can be characterized by temporal regularity of user states that can be contextualized by co-occurring entities. Temporal information incorporated into the model utilizing the deep learning approach demonstrated the effectiveness in predicting the next state of the user, inferring task context, and outperforming non-temporal machine learning methods in the experiment.

The results suggest that considering the temporal aspect in modeling provides an efficient mean to recognize and infer the user's context. By recognizing which context a user probably engages in, a collection of related content (e.g, documents, emails) can be dynamically recommended.



(a) Accuracy of user state prediction.

(b) Hitrate for application and document.

(c) Recall for application and document.

(d) Recall for keyword.

Figure 6.1: User state and entity prediction performance

.

The ability to model and predict the user's task context indicates that tasks are often influenced by time. The temporal behavior of a user is reflected by the use of a similar set of entities (applications, documents, topics) for the tasks over time. We can, therefore, mine users' temporal behaviors by analyzing their historical interactions and making use of the mined temporal behaviors for task-aware recommendations.

Chapter 7

Effect of 24/7 Behavioral Recordings

The research described in this chapter aims at answering RQ3-2: **Does the use of 24/7 behavioral recordings improve recommendation quality**. In the previous chapter, we took the user data as a whole and learn the evolving context over time. In this chapter, we investigate the effect of different application sources of contextual information. For each user, the data will be divided into parts; each describes user activities on a specific type of application (e.g., only search activity logs or user interaction history on email clients). The same set of application categories reported in Chapter 4 was used in the study. We considered each application category as a single source of contextual information. Most of the application sources listed in this chapter include research issues that have been addressed before (e.g., search activity and browsing history) and many possible contexts of the task defined along the sources have not yet been explored.

To understand the effect of contextual information sourced from various applications, we built several prediction models for contextual query augmentation for Web search rankings (Publication IV [84]). Data 2 was used in this study. The data of thirteen participants include all Web search queries and the associated task context derived from various applications. The effects of various context sources were determined by training models with varying application sources.

The study results showed that the user's task context could be inferred from varying application sources. The model utilizing contextual signals sourced from many types of applications demonstrated its effectiveness in re-ranking the correct Web documents that the user clicked by expanding the Web search queries with additional contextual terms. This answers RQ3-2, and the finding of the study can be framed as:

Finding 4 Contextual signals sourced from any type of application are useful in improving recommendation quality.

7.1 Experimental Setup

In this section, we describe the data used for the experiments. In particular, we present our approach to annotate and classify the data, and how the formed data was used for modeling and query augmentation.

7.1.1 Data Annotation and Classification

Data annotation and classification were conducted for the collected data before data analysis experiments. Queries, clicked documents on SERPs, information objects (files, documents, emails, instant messages, etc.) were extracted and classified according to their application sources.

Query and Clicked Document Link Extraction

The preliminary step of data annotation and classification was to extract the participants' Web search queries from digital activity logs. We ran a script programmed to automatically identify all Web searches and queries from commercial search engines, including Google Search, Bing Search, DuckDuckGo, Yandex, and Yahoo Search. Search engine usage was identified in the Web URLs of the collected screen frames. The queries were then extracted directly from the URLs. The corresponding clicked document links from the SERPs following the queries were also extracted.

Application Classification

The application classification phase aimed to classify applications into a set of categories based on their common functions, types, and fields of use. The application names were extracted from the collected OS log information. The same application categories described in Chapter 4 were used in this study.

7.1.2 Contextual Query Augmentation

We leveraged the recent digital activity of the user to model context and augmented the current search query. The sources determined the information used to build the context models. As part of the analysis, we varied the sources used to construct the four models, which are described below.

- **Search history model**: The search history model was constructed based on the user's search activity followed by a subsequent search or the current query. We applied a constraint to the data, accepting only the content of SERPs of prior searches to train the model.
- **Application-specific model**: A model for each application type was created using the data assigned to the application category described earlier. We assumed that if a user opened a specific application, the application window contained useful content for modeling. All information objects captured on that application were used to train the model.

7.1.3 Modeling Technique

To build contextual query augmentation, we constructed a context model (search history model or application-specific model) and integrated the context model with a conventional query augmentation model. Our approach is based on the three following steps:

1. Use contextual information sourced from a specific application type to build a topic model of the task context before the search. We used Dirichlet Hawkes processes [27] for topic modeling of task context.
2. Use the content of Web search results in response to the original query to build a conventional query augmentation model.
3. Use the task context model to re-rank the conventional query augmentation model.

7.1.4 Conditions

To study whether contexts sourced from a specific type of application helps in query expansion, we tested the model in varying conditions: the control condition, the search context condition, and the application-specific context condition.

- **Control condition.** The initial ranking from the Bing search engine was used as a control condition. Rankings were obtained by sending a search request using the original query to Bing API to retrieve 1000 ranked Web documents.
- **Search context condition.** Search history was leveraged for contextual query augmentation. In this condition, a search history model was utilized.

- **Application-specific context condition.** Application-specific interaction history was leveraged for contextual query augmentation. In this condition, an application-specific model was utilized.

Context models were then evaluated by testing whether the conditions with the models using different application-specific contexts generated based on the six sources of contextual information can be useful in improving the quality of search results.

7.2 Results and Findings

In general, the results indicate that the contextual query augmentation in other application-specific conditions performed equally well with the model in the search context condition. All application-specific models consistently improved the performance over the control condition. In particular, the models in four application-specific context conditions (Social Application, Office, E-commerce, and Rare Web) performed better compared to the model in other conditions.

We found that the different application sources of contextual information are all important. Therefore, it seems that the user context should not be limited to the information available on the search systems themselves, but there are many equally good sources of contextual information that can be leveraged for query augmentation. Search history, in general, is an effective source of contextual information, but context from other sources can be used to complement or replace search history when extensive search history is not available. If many useful sources of context are available to the search system, it may be possible to address many cold-start problems [34].

Chapter 8

Effect of Spoken Conversational Input

The research described in this chapter also aims at answering RQ3-3: **Does the use of spoken conversational input improve recommendation quality**. Now, we focused our attention on spoken conversational information; more specifically, we considered this type of data input in improving query suggestions.

To explore the impact of spoken conversational information on recommendation quality, here we focused on query auto-completion suggestions. Our aim was to try to predict the queries from the voice input such that the user typed initial letters, and the spoken conversational context was used to predict the completion of the query. We conducted a study in which twelve pairs of participants engaged in spoken conversations about movies and travels (Data 3). Their tasks were to discuss what movies they intended to watch or where to travel next. Participants could perform a search during the discussion to support their conversations. The conversations were automatically transcribed, and all the search logs and Web browsing activities were collected for the study.

In Publication V [85], we conducted an offline analysis on the effect of the task context model by investigating whether spoken input from conversations can be used as context to improve query auto-completion (QAC). That means the participants did not see the suggestions, and they had to write the entire query without support from the recommender system. To evaluate our model, we compared the ranking of query suggestions with and without context to understand how spoken conversational context affects the quality of query suggestions.

Results of the study showed that it was possible to infer user context from spoken conversations, and consequently, the context can be used to

improve the query suggestions. This answers RQ3-3; that is, the answer can be framed as:

Finding 5 Contextual signals sourced from spoken conversations are useful in improving recommendation quality.

8.1 Experimental Setup

We model the spoken conversational context preceding queries and use these models to re-rank the query auto-completion suggestions. The following sections describe the data used and how the context models were constructed.

8.1.1 Data Description

Data 3 includes Web queries inputted to the search interface, Web browsing activity (Web pages visits), and transcripts of the conversations. We first segmented the data into search activities, each with a query with recent context: prior queries, Web browsing history, and the participants' utterances. The data was pre-processed with stopword removal and lemmatization.

8.1.2 Context Models

Two context sources were leveraged - spoken conversational input and search history (browsed Web pages and prior queries) - for re-ranking QACs. The sources determined the information used to build the context models. The sources used to construct the three models are described below.

- **Search Context Model.** The search context model was constructed based on a user's Web search activity followed by a subsequent search or the current query. The textual content of browsed Web pages and queries of prior searches were utilized for training the model. We assumed that if a user searched and opened a Web document, the content might influence the user's subsequent search and contain useful information for modeling. Text units of browsed Web pages and prior queries processed in the early step were used to train the model.

- **Spoken Context Model.** The spoken context model was constructed based on the spoken conversation between users that oc-

curred prior to the current search query - the information comprised text units produced from automatic or ideal transcription.

- **Combined Context Model (Spoken + Search Context).** The combined context model was created using a combination of spoken conversational inputs and a user's search history. Outputs from the two separate models were combined.

The collected data and the QAC model in varying conditions were evaluated in an offline experiment. Here, we explain the configuration for each condition and evaluation metrics used to measure the QAC performance in these conditions.

8.1.3 Modeling Technique

To build contextual query auto-completion models, we integrated the context models with non-contextual query auto-completions. Here, we opt to use query completions provided by Google service. Our approach is based on the three following steps:

1. Use contextual information sourced from spoken conversation or search history or both spoken conversation and search history to build a topic model of the task context before search. We used Dirichlet Hawkes processes [27] for topic modeling of task context.

2. Use prefix of a query to retrieve query auto-completion suggestions from Google service.

3. Use the task context model to re-rank the query auto-completion suggestions.

8.1.4 Conditions

To study the utility of spoken context in QAC, we tested the QAC model in four conditions: the control condition, the search context condition, the spoken context condition, and the combined context condition.

- **Control.** In the control condition, QAC was initially produced by the Google Query Suggestion Service[1] was used. However, the QAC did not account for any context information from the conversation. We turned off the personalization feature in Google Service

[1] http://clients1.google.com/complete/search?&q=¡prefix¿&client=chrome

to avoid any confounding factors that might affect the initial ranking of QAC. For instance, different users might have different tastes in movies and travel present in their long-term search history before the experiment, and Google Service would have this information and personalized QACs, which would have become a factor in the experiment.

- **Search Context.** In this condition, we included the search context model, which leveraged only a user's search context information, to re-rank Google QACs.
- **Spoken Context.** In this condition, we included the spoken context model, which leveraged only spoken context information, to re-rank Google QACs.
- **Combined Context.** In this condition, both spoken and search context information was leveraged to re-rank Google QACs.

8.2 Results and Findings

We compared the effects of spoken conversational input in four conditions: a control condition without contextualization; an experimental condition with the model using search query logs; an experimental condition with the model using spoken conversational input; and an experimental condition with the model using both search query logs and spoken conversational input.

In general, QAC in the spoken context condition performed better compared with the model in control and search context conditions. By considering spoken conversational information, QAC ranking performance improves, indicating that such contextual information was useful in improving query prediction. We also found the advantage of combining the spoken conversational information with the Web search context for improved retrieval performance. Our results suggest that spoken conversations provide a rich context for supporting information searches beyond current user-modeling approaches.

Chapter 9

Entity Recommendation in Everyday Task

The research described in this chapter aims at answering RQ4: **Does contextual recommendation improve users' task performance?** Our context-aware recommendation system was evaluated in light of the goal to support knowledge workers in everyday digital tasks. We focus on supporting the users through entity recommendations (documents, applications, people, and keywords describing the task). The context-aware entity recommendation system for finding information was evaluated with users in an online interactive setting. A user would work on the task as she/he normally does while receiving recommendations from the system.

In Publication VI [43], 13 participants who took part in the second data collection (Data 2) were invited back for the laboratory study. We evaluated whether the recommendations lead to improved task execution in terms of context relevance and usefulness. For context relevance, the participant was asked to rate the suggestions she/he receives at the end of the task. The usefulness of the recommendation was measured by how many entities were used (applications and documents were opened from the recommendation by the users). We investigate the effect of contextual recommendation by requesting the participants to resume and perform two of their tasks that were reported in the diaries. Tasks were randomly assigned into two conditions: a condition with recommendations visible for the user and a condition with recommendations not visible for the user.

Results of the study showed that the recommendation system has a direct influence on task performance. The participants used the recommended entities, e.g., opened the documents used the applications to complete the task. In some cases, the recommendation reminded the participants of information they had seen before and that information was relevant

at the time of the task. The participants also considered the recommendation system a companion: a personal assistant that helped carry out the task with new ideas and made it easier for the participants to stay on track. Such results help to answer RQ4 in that we could have a takeaway message as:

Finding 6 Recommendations positively influence the user's task performance. Our modeling and task-based entity recommendation approach enable effortless access to information for the users. Recommendations are the source of inspiration and contain insightful information that helps users complete their tasks.

9.1 Experimental Setup

In this section, we describe the data used for the study and how the data was collected. Then, we discuss the study conditions and how to evaluate the recommender system.

9.1.1 Data Description

Dataset (Data 3) of thirteen participants voluntarily who took part in the study was used in this study. Participants installed our digital activity monitoring for 14 days for data collection. The purpose was to collect a set of entities that the users had accessed before the lab study. The entities were extracted after visiting web pages in the browser or accessing applications and documents stored locally on personal computers. The collected data was used for recommendation. In addition, participants were asked to write diaries about their tasks every day.

9.1.2 Procedure

After the monitoring period, participants came back to our lab for the second part of the experiment. We asked participants to review their diaries and select two tasks they performed during the monitoring. In particular, we asked participants to pick two tasks that they felt were similar in category, with the same level of complexity, and comparable in duration. Participants were asked to write descriptions of the two selected tasks in a note. Then, the experimenter randomly assigned the two tasks to the two conditions:

- **Experimental condition**: with the user model and the recommendation was visible for the user. New unseen digital activities were considered as task context, and the user could also provide explicit interaction with the recommender system.

- **Control condition**: with the user model, but the recommendation was not visible for the user. Input to the model was also the recent context derived from the new unseen digital activities. In this condition, recommendations did not affect the information presented to the user on the screen or user information behavior.

To counteract fatigue and other carryover effects, we counterbalanced the order in which the participants were subjected to each condition. After selecting the tasks, the participants were briefed about the experiment's procedure: they were asked to resume the tasks on their laptops, engage in a short interview, and assess the relevance of the entities.

9.1.3 Evaluation

We address two possibilities to support knowledge workers in the case of finding information:

- **Entity Relevance**: A knowledge worker can be supported by having documents suggested to him that contain relevant information for the (writing) task he is working on. Where context relevance evaluates whether there is a general topical match with the current activities, entity relevance is aimed at a more detailed evaluation of how much a suggested document contributes to the task process.

- **Influence of Recommended Information**: Although a topical match of recommendation with the user's active context has been a common way to evaluate the effectiveness of recommendation, but it does not mean an entity that is suggested can be used by the users. Therefore, we additionally consider 1) *Attention on recommended information*, which evaluate how long the user looks at the recommended entities measured by the total duration of the gaze fixations of the users on the recommendation screen; and 2) *Utilization of recommended information*, which evaluate how useful a suggested entity is for a task that the user is working on. That is, how many documents and applications are actually opened after the user saw them.

9.2 Results and Findings

We found participants engaged in tasks ranging from thesis writing, data processing, and coding to travel planning and other social tasks. Within such tasks, participants were involved in various activities, and their interests frequently changed. Results on the relevance and influence of recommendations, together with qualitative findings, show that the system effectively captured participants' rapidly evolving interests and provided them with recommendations that positively influenced their tasks.

Participants were found to access recommendations that were subsequently used to perform the tasks. Participants looked at the recommendations and used the hyperlinks provided by the system to access documents useful for their tasks. In particular, we also observed that the utility of recommendation comes very fast upon the participant resuming the task. This indicates that our system provided good recommendations helping the participant to return to the task quickly.

In general, the task model could be used to effectively retrieve useful entities for the user in the task context. The results indicated that the list of recommendations related to the task containing the entities that the user wanted to refind. Another advantage of our task-centric recommendation approach could support the user's memory. The system provided forgotten or unknown resources, reminding the users of specific activities or a piece of information that they have seen before and that would be relevant now for the task.

Chapter 10

Conclusion

This thesis explored the use of task information derived from more extensive sources and social, temporal, and topical aspects to predict and recommend task-relevant information. We found that task context affected the user's information behavior. By analyzing varying types of contextual information about the task, we can learn user activity patterns and, to some extent, predict the entities (applications, documents, contact information, and keywords) that the users would use next. This work suggests new opportunities to further investigate whether more extensive contextual signals mined from even more aspects, e.g., mobile and physical context, could affect retrieval performance.

10.1 Summary of Main Findings

To summarize the main findings of the work, we revisit the research questions defined earlier and reflect on the answers below:

RQ1: Are there interdependencies between the tasks, searching behavior, and contextual entities? Yes, our work with task analysis has shown that contextual factors derived from varying aspects of the task have a crucial effect on searching behavior. We found that application contexts and topical interest before searching directly impact search performance. In particular, we found that the topic the users observed before searching often triggers a search in various tasks, such as a task that demands a degree of intellectual effort, such as writing or literature reviewing. Searching is more induced from a certain application context (such as utility applications: word processing, programming platform, or spreadsheet) in which the user intent was to process and produce information artifacts (such as programming, writing a document). Our findings are in line with previous literature in the case of a single contextual factor

(e.g., Web browsing activity) [18, 89]. However, contrary to their work, which used a lab-based analysis and focused on pre-determined tasks assigned by the experimenter for specific research design, our work provided new insights into real-life tasks. Our results and findings indicated the possibility of modeling the real-life task context from association amongst the entities (applications and topics) for many practical use cases such as recommendation or entity prediction.

RQ2: Can the association amongst the entities be used to model the user's task context for recommendation? Yes, we found it was possible to model the user task from the latent thematic structure of the data (the structured collection of entities). By observing the long-term history of user activities, the model could detect the user task predict the future context and relevant entities for the task. Our results also revealed that applications used for the tasks were very diverse, such as writing tasks; the users would also need to perform various online activities: Web browsing, searching, and emailing. The analysis allows us to understand how the users performed their tasks in work settings and the need for a recommender system that could provide access to information from multiple application sources and synchronize the content in one place, promoting ubiquitous access to information. Although in this thesis, we only focused on information entities that occurred from computer usage and routine tasks that have been executed on a computer, the model can be applied for a wide variety of tasks and events that could be observed in the physical environment, such as events created by location constraint or mobile devices.

RQ3: Does the use of more extensive sources of context improve recommendation quality? Yes, we found temporal information was an important factor in the task context modeling (RQ3-1). We have shown that contextual information sourced from any application could affect the performance of the prediction of the task and entity recommendation (RQ3-2). We also incorporated more signals into the prediction model, such as combing user information behavior from all application sources that could give even richer information about user context and intention and further improve the recommendation quality. Contextual signals sourced from spoken conversations could complement the conventional prediction model for generating recommendations using a search context when extensive search history is not available (RQ3-3). It would also be interesting to study the impact of other personalized cues such as an individual's past behavior in mobile devices or location history on the performance of the recommendation system.

RQ4: Does contextual recommendation improve users' task performance? Yes, we found that the task-related context structures extracted by the system contained insightful information that users had not previously noticed and that positively affected the task. Participants generally reported an improved user experience when performing tasks with the support of our recommendation system. Participants perceived the system as a companion that provided valuable insights into performing the task. The recommendations could function as a memory extension for the participant. By reminding the participants about the entities related to the task, the system permitted the reconstruction of the typical activities they performed during their task. Furthermore, the system stores explicit associations amongst the entities similar to how the user would associate the information to the task. This would allow the participants to have faster access to information. During the tasks, the participants needed to access various documents, such as PDFs, emails, code snippets, or websites, however, forgetting the name or location of a file, the applications used, or the subject of an important email make it difficult to retrieve the needed information. With our system, participants could easily retrieve important documents they needed for their tasks, either without any explicit input (e.g., by just opening a different relevant document) or by selecting key entities in the system. It allowed the participants to save time with a consequent improvement of the perceived experience. With our entity-centric approach, it would be one step towards the design principles for personal information management systems formulated by Elsweiler et al. [30]. Finally, the benefits provided by the system came with moderate costs in terms of division of attention, as reported by some participants. Overall, the participants reported a better user experience when the implemented system was available.

10.2 Implications of Research

The ability to model task context from more extensive sources and multiple aspects implies that the users completing the task need to perform a wide range of activities and access a variety of application sources such as Web browsing, search, e-mailing, or conversing with other individuals. In line with early research on context modeling [75], the findings showed that incorporating task information in recommendations has the benefit of minimizing the user's search effort and supporting the work. However, the findings can go further and complement previous research by considering the practical usefulness in real-world tasks rather than artificial tasks in a lab-based study. We foresee implications for using contextual information

to support researchers in designing user studies and experiments and for practitioners designing information access systems and privacy preservation strategies.

10.2.1 Designing user studies and experiments

There are practical implications of this work for designing user studies and experiments. Search history and Web browsing actions may not be the only sources that are representative of users' real underlying tasks. This may have set limits on the current experimental paradigm and the datasets used in information retrieval experimentation. Our findings showed that the user's task context could be predicted by speech data and various behavioral information sourced from local documents and applications, suggesting more task information that could appear in other non-search applications. Such knowledge can help researchers to design experiments with more face validity. Given the assigned work task and its potential queries, contextual evidence sourced from spoken conversations, different local applications, and Web services can be leveraged and an appropriate duration for the task can be established accordingly to improve the lab-based experiments.

10.2.2 Designing information access systems

Our results demonstrate that it is possible to model the user's task context from the association amongst the entities. The data was collected by simply logging the user interactions with the entities. Although this may sound like a limitation compared to the conventional approach that was based on more structured data such as clicks or other direct inputs in search applications, our approach could enable a general user modeling that can be used across applications. Utilizing task context information from one application in another may help to resolve an issue with cold-start recommendations [93]. Our findings also showed that there are distinct sets of queries for which context models utilizing specific sources perform most effectively, suggesting that query information is likely important in selecting sources and temporal contextual lengths. The richer models that we developed can be used to interpret a user's search intent for a wide variety of search applications, including proactively retrieving information of likely interest to the user, suggesting useful queries contextually, or document ranking and filtering. Search systems could also use the context model and assign a source and amount of contextual information based on the query to improve the quality of search rankings by promoting results that are consistent with the inferred user intent. The systems may need to vary the

sources depending on the modeling task, e.g., short-term models should use recent contexts, such as recently opened documents, conversations, emails, and instant messages, whereas long-term models should use information from longer-term learning behaviors and historical online transactions.

10.2.3 Designing privacy preservation strategies

More contextual signals from spoken conversations and various application sources provide benefits in generating recommendations or predicting when and where users can complete tasks. Our research has implications for implementing data minimization and retention requirements of some regulations for data protection and privacy. In particular, our study shows that, in most cases, only behavior from a limited set of applications is sufficient for effective search ranking. In some cases, using contextual information does not always yield an improvement. For instance, information behaviors from many applications may not be needed and do not affect the recommendation performance. As previous research has demonstrated [94], users are often reluctant to share information with a search system, and they do not understand why certain queries have been suggested or where they came from. Future work on search systems may consider providing users with various privacy thresholds and their explanations for suggestions and allow them to freely explore the relationship between privacy preservation, contextual sources, and search quality. Another important implication derived from our approach is that the ability to model the user's task context from the end-user device. That means this does not require access to the data provided by the service provider, such as large-scale search logs for modeling. The data could be owned and utilized by the users without sending private information to the service provider. This echoes the recent development that promotes user-centered information management and processing, putting the users at the center and in control of their data [21].

10.3 Limitations

This thesis presents some limitations which may have some impact on the generalizability of the results. This study was carried out as part of a field study and, because of this, followed an unusual experimental procedure. Two of the more important aspects of this procedure that might have impacted our results are that some of the user behaviors may have been restrained on purpose because users were fully aware that their digital activities were being tracked, and the two-week monitoring period may not

reflect the entire blueprint of their behaviors, including activity changes due to seasons/holidays or monthly routines. Furthermore, while the 14-day digital activity monitoring and conversation recording of 47 participants resulted in large data, it was a fairly small sample compared to large-scale data logs with a larger population, such as the context information that could be sourced from long-term interactions on a Web search engine. Although the potential impact of the experimental procedure and recruitment on the validity and reliability of our results cannot be ignored, we feel that our results are important and make an essential contribution to the research on recommendation and user modeling.

The observed impacts in this study may be related to the nature of the data that were collected, for example, data of a specific cohort (young individuals, mostly students and researchers). It may have been better also to include data of other groups of users. Other aspects of the task have also not been analyzed herein, such as user behaviors on smartphones or physical aspects (e.g., locations), and will be in future work. However, given that this study was computer-based and that we had extracted all available texts in all applications and even spoken conversations, the definitions of task context we adopted here seem reasonable.

Another limitation to the evaluation of the recommendation system is lacking the real-life performance of the user task because participants performed their tasks in the lab. This setting was planned on purpose as we would want to control the factors we would study. This aspect is important because the actual usability of the system can only be measured when it is put into practice, especially for context-aware systems. A system may not be likely to be useful to the knowledge workers when suggestions are not provided on time. Because our data set was not sufficiently large to pose modeling problems, we have not considered any potential errors or low system performance that may be made by data overhead.

Finally, the specific difficulty encountered with the introduction of the screen monitoring and digital activity monitoring method was privacy. Participants were fully aware of their activities being monitored, and thus some of their activities might be concealed on purpose. However, this limitation was predicted. We expected that the participants could share most of their activities that were considered to be less sensitive. Besides, this has no negative effect on the results of the study, as what we aimed for was to investigate the possibility of making inferences about the user's task context given the user behaviors collected from more extensive sources, and that was useful in improving the quality of recommendations. In addition, the research followed the ethical guidelines established by the University

of Helsinki and received ethical approval. However, we also see the opportunity raised by this work to support the recent MyData movement and related research to empower individuals having the right and practical means to manage their data and privacy [1, 54].

10.4 Future Work

This work provides a foundation for further development of context-aware recommendations. In future research, it is important to investigate what users value most in the context-aware recommendation. How often should the system recommend information, and how many entities should be in the suggestion list? Additionally, in our dataset, there were no actual relevance judgments by the users during work situations. Our evaluation approach was dependent on the prediction of the entities. That is, we did not know whether interesting entities are useful at the time they are recommended or whether they can be a potential source of distraction. However, asking the users to rate the suggestions the user receives at a certain moment is challenging; it might divert the user's attention from the task. A proper evaluation of the recommendation system in real-life settings requires further exploration.

Continued improvement of the comfort of using recommendation systems is an important task for the future. For instance, the scenario of the knowledge worker is different from typical lab-based context-aware recommendation scenarios, as the context is more dynamic, and there is a larger negative impact of irrelevant recommendations. In this thesis, we only presented and used the data set collected from the laptop to facilitate our research. We have no evidence of whether recommendations would lead to improved task execution in terms of time and profit. This requires further investigation and different evaluation approaches in realistic knowledge worker settings where the context is given by the interaction of users with their regular office computers.

Overall, we believe that our task-modeling approach is most promising for a context-aware recommendation. It performed well in terms of action prediction while providing relevant results. Moreover, our approach is not dependent on manually selecting the source for detection of the context. Nevertheless, there is room for improvement when it comes to document and context relevance. The flexibility of the system (e.g., number of entities are scalable and can be applied on any type of context data) provides ample opportunity to investigate these respects. Finally, we conclude that the evaluation approach to understanding how recommendations influence user task has an added value in real-world task-based evaluations.

References

[1] MyData Declaration 2019. https://mydata.org/declaration/. Accessed: 2020-03-10.

[2] David Abrams, Ron Baecker, and Mark Chignell. 1998. Information Archiving with Bookmarks: Personal Web Space Construction and Organization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Los Angeles, California, USA) *(CHI '98)*. ACM Press/Addison-Wesley Publishing Co., USA, 41–48. https://doi.org/10.1145/274644.274651

[3] Shahriyar Amini, Vidya Setlur, Zhengxin Xi, Eiji Hayashi, and Jason Hong. 2013. Investigating Collaborative Mobile Search Behaviors. In *Proceedings of the 15th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Munich, Germany) *(MobileHCI '13)*. Association for Computing Machinery, New York, NY, USA, 213–216. https://doi.org/10.1145/2493190.2493198

[4] Salvatore Andolina, Valeria Orso, Hendrik Schneider, Khalil Klouche, Tuukka Ruotsalo, Luciano Gamberini, and Giulio Jacucci. 2018. Investigating Proactive Search Support in Conversations. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 1295–1307. https://doi.org/10.1145/3196709.3196734

[5] Ioannis Antonellis, Hector Garcia-Molina, and Chi-Chao Chang. 2008. Simrank++: Query Rewriting through Link Analysis of the Clickgraph (Poster). In *Proceedings of the 17th International Conference on World Wide Web* (Beijing, China) *(WWW '08)*. Association for Computing Machinery, New York, NY, USA, 1177–1178. https://doi.org/10.1145/1367497.1367714

[6] David Bawden and Lyn Robinson. 2009. The dark side of information: overload, anxiety and other paradoxes and patholo-

gies. *Journal of Information Science* 35, 2 (2009), 180–191. https://doi.org/10.1177/0165551508095781

[7] Nicholas J Belkin, Robert N Oddy, and Helen M Brooks. 1982. ASK for information retrieval: Part I. Background and theory. *Journal of documentation* (1982).

[8] Victoria Bellotti, Nicolas Ducheneaut, Mark Howard, and Ian Smith. 2003. Taking Email to Task: The Design and Evaluation of a Task Management Centered Email Tool. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 345–352. https://doi.org/10.1145/642611.642672

[9] Paul N. Bennett, Ryen W. White, Wei Chu, Susan T. Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. 2012. Modeling the Impact of Short- and Long-Term Behavior on Search Personalization. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Portland, Oregon, USA) *(SIGIR '12)*. Association for Computing Machinery, New York, NY, USA, 185–194. https://doi.org/10.1145/2348283.2348312

[10] Ofer Bergman, Ruth Beyth-Marom, Rafi Nachmias, Noa Gradovitch, and Steve Whittaker. 2008. Improved Search Engines and Navigation Preference in Personal Information Management. *ACM Transactions on Information Systems* 26, 4, Article 20 (Oct 2008), 24 pages. https://doi.org/10.1145/1402256.1402259

[11] Ofer Bergman, Maskit Tene-Rubinstein, and Jonathan Shalom. 2013. The Use of Attention Resources in Navigation versus Search. *Personal Ubiquitous Computing* 17, 3 (March 2013), 583–590.

[12] Tristan Blanc-Brude and Dominique L. Scapin. 2007. What Do People Recall about Their Documents? Implications for Desktop Search Tools. In *Proceedings of the 12th International Conference on Intelligent User Interfaces* (Honolulu, Hawaii, USA) *(IUI '07)*. Association for Computing Machinery, New York, NY, USA, 102–111. https://doi.org/10.1145/1216295.1216319

[13] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[14] Oliver Brdiczka, Norman Makoto Su, and James Bo Begole. 2010. Temporal Task Footprinting: Identifying Routine Tasks by Their Temporal Patterns. In *Proceedings of the 15th International Conference on Intelligent User Interfaces* (Hong Kong, China) *(IUI '10)*. Association for Computing Machinery, New York, NY, USA, 281–284. https://doi.org/10.1145/1719970.1720011

[15] Jay Budzik and Kristian Hammond. 1999. Watson: Anticipating and contextualizing information needs. In *Proceedings of the Sixty-second Annual Meeting of the American Society for Information Science*. Citeseer.

[16] Jay Budzik and Kristian J. Hammond. 2000. User Interactions with Everyday Applications as Context for Just-in-Time Information Access. In *Proceedings of the 5th International Conference on Intelligent User Interfaces* (New Orleans, Louisiana, USA) *(IUI '00)*. Association for Computing Machinery, New York, NY, USA, 44–51. https://doi.org/10.1145/325737.325776

[17] Katriina Byström and Preben Hansen. 2005. Conceptual framework for tasks in information studies. *Journal of the American Society for Information Science and Technology* 56, 10 (2005), 1050–1061. https://doi.org/10.1002/asi.20197

[18] Katriina Byström and Preben Hansen. 2005. Conceptual framework for tasks in information studies. *Journal of the American Society for Information Science and Technology* 56, 10 (2005), 1050–1061.

[19] Katriina Byström and Kalervo Järvelin. 1995. Task complexity affects information seeking and use. *Information Processing & Management* 31, 2 (1995), 191 – 213.

[20] Paul Alexandru Chirita, Claudiu S. Firan, and Wolfgang Nejdl. 2007. Personalized Query Expansion for the Web. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Amsterdam, The Netherlands) *(SIGIR '07)*. ACM, New York, NY, USA, 7–14. https://doi.org/10.1145/1277741.1277746

[21] Eun Kyoung Choe, Nicole B. Lee, Bongshin Lee, Wanda Pratt, and Julie A. Kientz. 2014. Understanding Quantified-selfers' Practices in Collecting and Exploring Personal Data. In *Proceedings of the SIGCHI*

Conference on Human Factors in Computing Systems (Toronto, Ontario, Canada) *(CHI '14)*. ACM, New York, NY, USA, 1143–1152. https://doi.org/10.1145/2556288.2557372

[22] Karen Church, Antony Cousin, and Nuria Oliver. 2012. I Wanted to Settle a Bet! Understanding Why and How People Use Mobile Search in Social Settings. In *Proceedings of the 14th International Conference on Human-Computer Interaction with Mobile Devices and Services* (San Francisco, California, USA) *(MobileHCI '12)*. Association for Computing Machinery, New York, NY, USA, 393–402. https://doi.org/10.1145/2371574.2371635

[23] Karen Church and Barry Smyth. 2009. Understanding the Intent behind Mobile Information Needs. In *Proceedings of the 14th International Conference on Intelligent User Interfaces* (Sanibel Island, Florida, USA) *(IUI '09)*. Association for Computing Machinery, New York, NY, USA, 247–256. https://doi.org/10.1145/1502650.1502686

[24] Victoria Clarke and Virginia Braun. 2013. *Successful Qualitative Research: A Practical Guide for Beginners.*

[25] Jacob Dankasa. 2017. Seeking information in circles: The application of Chatman's life in the round theory to the information small world of Catholic clergy in northern Nigeria. *Journal of Information Science* 43, 2 (2017), 246–259.

[26] Anton N. Dragunov, Thomas G. Dietterich, Kevin Johnsrude, Matthew McLaughlin, Lida Li, and Jonathan L. Herlocker. 2005. TaskTracer: A Desktop Environment to Support Multi-Tasking Knowledge Workers. In *Proceedings of the 10th International Conference on Intelligent User Interfaces* (San Diego, California, USA) *(IUI '05)*. Association for Computing Machinery, New York, NY, USA, 75–82. https://doi.org/10.1145/1040830.1040855

[27] Nan Du, Mehrdad Farajtabar, Amr Ahmed, Alexander J. Smola, and Le Song. 2015. Dirichlet-Hawkes Processes with Applications to Clustering Continuous-Time Document Streams. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. ACM, New York, NY, USA, 219–228. https://doi.org/10.1145/2783258.2783411

[28] Carsten Eickhoff, Sebastian Dungs, and Vu Tran. 2015. An Eye-Tracking Study of Query Reformulation. In *Proceedings of the 38th*

International ACM SIGIR Conference on Research and Development in Information Retrieval (Santiago, Chile) *(SIGIR '15)*. ACM, New York, NY, USA, 13–22. https://doi.org/10.1145/2766462.2767703

[29] David Elsweiler, Ian Ruthven, and Christopher Jones. 2007. Towards memory supporting personal information management tools. *Journal of the American Society for Information Science and Technology* 58, 7 (2007), 924–946. https://doi.org/10.1002/asi.20570

[30] David Elsweiler, Ian Ruthven, and Christopher Jones. 2007. Towards memory supporting personal information management tools. *Journal of the American Society for Information Science and Technology* 58, 7 (2007), 924–946. https://doi.org/10.1002/asi.20570

[31] Henry Feild and James Allan. 2013. Task-Aware Query Recommendation. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland) *(SIGIR '13)*. Association for Computing Machinery, New York, NY, USA, 83–92. https://doi.org/10.1145/2484028.2484069

[32] Karim Filali, Anish Nair, and Chris Leggetter. 2010. Transitive History-Based Query Disambiguation for Query Reformulation. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Geneva, Switzerland) *(SIGIR '10)*. Association for Computing Machinery, New York, NY, USA, 849–850. https://doi.org/10.1145/1835449.1835647

[33] Stephen Fitchett and Andy Cockburn. 2012. AccessRank: Predicting What Users Will Do Next. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 2239–2242. https://doi.org/10.1145/2207676.2208380

[34] Vreixo Formoso, Diego Fernández, Fidel Cacheda, and Victor Carneiro. 2013. Using profile expansion techniques to alleviate the new user problem. *Information Processing & Management* 49, 3 (2013), 659–672. https://doi.org/10.1016/j.ipm.2012.07.005 Personalization and Recommendation in Information Access.

[35] Eric Freeman and David Gelernter. 1996. Lifestreams: A Storage Model for Personal Data. *SIGMOD Record* 25, 1 (March 1996), 80–86. https://doi.org/10.1145/381854.381893

[36] Ang Gao and Derek Bridge. 2010. Using Shallow Natural Language Processing in a Just-In-Time Information Retrieval Assistant for Bloggers. In *Artificial Intelligence and Cognitive Science*, Lorcan Coyle and Jill Freyne (Eds.). Springer, Berlin, Heidelberg, 103–113.

[37] Jianfeng Gao and Jian-Yun Nie. 2012. Towards Concept-Based Translation Models Using Search Logs for Query Expansion. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (Maui, Hawaii, USA) *(CIKM '12)*. Association for Computing Machinery, New York, NY, USA, Article 1, 10 pages. https://doi.org/10.1145/2396761.2530275

[38] Mihajlo Grbovic, Nemanja Djuric, Vladan Radosavljevic, Fabrizio Silvestri, and Narayan Bhamidipati. 2015. Context- and Content-Aware Embeddings for Query Rewriting in Sponsored Search. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 383–392. https://doi.org/10.1145/2766462.2767709

[39] Jacek Gwizdka. 2010. Distribution of cognitive load in Web search. *Journal of the American Society for Information Science and Technology* 61, 11 (2010), 2167–2187. https://doi.org/10.1002/asi.21385

[40] Preben Hansen. 2011. *Task-based information seeking and retrieval in the patent domain: processes and relationships*. Ph.D. Dissertation. University of Tampere. http://urn.fi/urn:isbn:978-951-44-8497-1

[41] Chen He, Denis Parra, and Katrien Verbert. 2016. Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications* 56 (2016), 9–27. https://doi.org/10.1016/j.eswa.2016.02.013

[42] Peter Ingwersen and Kalervo Järvelin. 2005. Information Retrieval in Context: IRiX. *SIGIR Forum* 39, 2 (Dec. 2005), 31–39. https://doi.org/10.1145/1113343.1113351

[43] Giulio Jacucci, Pedram Daee, Tung Vuong, Salvatore Andolina, Khalil Klouche, Mats Sjöberg, Tuukka Ruotsalo, and Samuel Kaski. 2021. Entity Recommendation for Everyday Digital Tasks. *ACM Transactions on Computer-Human Interaction* 28, 5, Article 29 (Oct. 2021), 41 pages. https://doi.org/10.1145/3458919

[44] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-based Evaluation of IR Techniques. *ACM Transactions on Information Systems.* 20, 4 (Oct. 2002), 422–446. https://doi.org/10.1145/582415.582418

[45] Jenny Johannisson and Olof Sundin. 2007. Putting Discourse to Work: Information Practices and the Professional Project of Nurses. *The Library Quarterly: Information, Community, Policy* 77, 2 (2007), 199–218.

[46] Rosie Jones, Benjamin Rey, Omid Madani, and Wiley Greiner. 2006. Generating Query Substitutions. In *Proceedings of the 15th International Conference on World Wide Web* (Edinburgh, Scotland) *(WWW '06)*. Association for Computing Machinery, New York, NY, USA, 387–396. https://doi.org/10.1145/1135777.1135835

[47] Victor Kaptelinin. 2003. UMEA: Translating Interaction Histories into Project Contexts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 353–360. https://doi.org/10.1145/642611.642673

[48] Diane Kelly. 2006. Measuring online information seeking context, Part 1: Background and method. *Journal of the American Society for Information Science and Technology* 57, 13 (2006), 1729–1739.

[49] Diane Kelly. 2006. Measuring online information seeking context, Part 2: Findings and discussion. *Journal of the American Society for Information Science and Technology* 57, 14 (2006), 1862–1874.

[50] Diane Kelly and Nicholas J. Belkin. 2004. Display Time as Implicit Feedback: Understanding Task Effects. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Sheffield, United Kingdom) *(SIGIR '04)*. Association for Computing Machinery, New York, NY, USA, 377–384. https://doi.org/10.1145/1008992.1009057

[51] Weize Kong, Rui Li, Jie Luo, Aston Zhang, Yi Chang, and James Allan. 2015. Predicting Search Intent Based on Pre-Search Context. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 503–512. https://doi.org/10.1145/2766462.2767757

[52] Markus Koskela, Petri Luukkonen, Tuukka Ruotsalo, Mats Sjöberg, and Patrik Floréen. 2018. Proactive Information Retrieval by Capturing Search Intent from Primary Task Context. *ACM Transactions on Interactive Intelligent Systems* 8, 3, Article 20 (July 2018), 25 pages. https://doi.org/10.1145/3150975

[53] Markus Koskela, Petri Luukkonen, Tuukka Ruotsalo, Mats Sjöberg, and Patrik Floréen. 2018. Proactive information retrieval by capturing search intent from primary task context. *ACM Transactions on Interactive Intelligent Systems* 8, 3 (2018), 1–25.

[54] Kai Kuikkaniemi, Antti Poikola, and Harri Honko. 2015. *MyData A Nordic Model for human-centered personal data management and processing.* WorkingPaper.

[55] Sanna Kumpulainen. 2014. Trails across the heterogeneous information environment: Manual integration patterns of search systems in molecular medicine. *Journal of Documentation* 70, 5 (2014), 856–877.

[56] Sanna Kumpulainen and Kalervo Järvelin. 2010. Information Interaction in Molecular Medicine: Integrated Use of Multiple Channels. In *Proceedings of the Third Symposium on Information Interaction in Context* (New Brunswick, New Jersey, USA) *(IIiX '10)*. Association for Computing Machinery, New York, NY, USA, 95–104. https://doi.org/10.1145/1840784.1840800

[57] Cheng Li, Mingyang Zhang, Michael Bendersky, Hongbo Deng, Donald Metzler, and Marc Najork. 2019. Multi-View Embedding-Based Synonyms for Email Search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 575–584. https://doi.org/10.1145/3331184.3331250

[58] Yuelin Li and Nicholas J. Belkin. 2008. A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management* 44, 6 (2008), 1822–1837. https://doi.org/10.1016/j.ipm.2008.07.005

[59] Jiqun Liu, Matthew Mitsui, Nicholas J. Belkin, and Chirag Shah. 2019. Task, Information Seeking Intentions, and User Behavior: Toward A Multi-Level Understanding of Web Search. In *Proceedings of the 2019 Conference on Human Information Interac-*

tion and Retrieval (Glasgow, Scotland UK) *(CHIIR '19)*. Association for Computing Machinery, New York, NY, USA, 123–132. https://doi.org/10.1145/3295750.3298922

[60] Jiqun Liu, Shawon Sarkar, and Chirag Shah. 2020. *Identifying and Predicting the States of Complex Search Tasks*. Association for Computing Machinery, New York, NY, USA, 193–202. https://doi.org/10.1145/3343413.3377976

[61] Joseph E. McGrath. 1995. Methodology Matters: Doing Research in the Social and Behavioral Sciences. In *Readings in Human-Computer Interaction*, Ronald M. Baecker, Jonathan Grudin, William A.S. Buxton, and Saul Greenberg (Eds.). Morgan Kaufmann, 152–169. https://doi.org/10.1016/B978-0-08-051574-8.50019-4

[62] Donald McMillan, Antoine Loriette, and Barry Brown. 2015. Repurposing Conversation: Experiments with the Continuous Speech Stream. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. ACM, New York, NY, USA, 3953–3962. https://doi.org/10.1145/2702123.2702532

[63] Matthew E. Peters and Dan Lecocq. 2013. Content Extraction Using Diverse Feature Sets. In *Proceedings of the 22nd International Conference on World Wide Web* (Rio de Janeiro, Brazil) *(WWW '13 Companion)*. Association for Computing Machinery, New York, NY, USA, 89–90. https://doi.org/10.1145/2487788.2487828

[64] Benjamin Piwowarski and Hugo Zaragoza. 2007. Predictive User Click Models Based on Click-through History. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management* (Lisbon, Portugal) *(CIKM '07)*. Association for Computing Machinery, New York, NY, USA, 175–182. https://doi.org/10.1145/1321440.1321467

[65] Mari Carmen Puerta Melguizo, Lou Boves, and Olga Munoz Ramos. 2009. A proactive recommendation system for writing: Helping without disrupting. *International Journal of Industrial Ergonomics* 39, 3 (2009), 516–523. https://doi.org/10.1016/j.ergon.2008.10.004 Selected papers from ECCE 2007, the 25th Anniversary Conference of the European Conference on Cognitive Ergonomics.

[66] Ahmad Rahmati and Lin Zhong. 2013. Studying Smartphone Usage: Lessons from a Four-Month Field Study. *IEEE*

Transactions on Mobile Computing 12, 7 (2013), 1417–1427. https://doi.org/10.1109/TMC.2012.127

[67] Zeinab Reizaei, Tung Vuong, Al-Ghossein Marie, Tuukka Ruotsalo, Giulio Jacucci, and Samuel Kaski. 2022. Entity Footprinting: Modeling Contextual User States via Digital Activity Monitoring. *ACM Transactions on Interactive Intelligent Systems* (July 2022).

[68] Yongli Ren, Martin Tomko, Flora Dilys Salim, Kevin Ong, and Mark Sanderson. 2017. Analyzing Web behavior in indoor retail spaces. *Journal of the Association for Information Science and Technology* 68, 1 (2017), 62–76. https://doi.org/10.1002/asi.23587

[69] B. J. Rhodes and P. Maes. 2000. Just-in-time information retrieval agents. *IBM Systems Journal* 39, 3.4 (2000), 685–704. https://doi.org/10.1147/sj.393.0685

[70] George Robertson, Maarten van Dantzich, Daniel Robbins, Mary Czerwinski, Ken Hinckley, Kirsten Risden, David Thiel, and Vadim Gorokhovsky. 2000. The Task Gallery: A 3D Window Manager. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (The Hague, The Netherlands) *(CHI '00)*. Association for Computing Machinery, New York, NY, USA, 494–501. https://doi.org/10.1145/332040.332482

[71] Daniel E. Rose and Danny Levinson. 2004. Understanding User Goals in Web Search. In *Proceedings of the 13th International Conference on World Wide Web* (New York, NY, USA) *(WWW '04)*. ACM, New York, NY, USA, 13–19.

[72] Miamaria Saastamoinen and Kalervo Järvelin. 2016. Queries in authentic work tasks: the effects of task type and complexity. *Journal of Documentation* 72 (10 2016), 1114–1133. https://doi.org/10.1108/JD-09-2015-0119

[73] Miamaria Saastamoinen, Sanna Kumpulainen, and Kalervo Järvelin. 2012. Task Complexity and Information Searching in Administrative Tasks Revisited. In *Proceedings of the 4th Information Interaction in Context Symposium* (Nijmegen, The Netherlands) *(IIIX '12)*. ACM, New York, NY, USA, 204–213.

[74] Alan Said, Shlomo Berkovsky, and Ernesto W. De Luca. 2011. Group Recommendation in Context. In *Proceedings of the 2nd Challenge on Context-Aware Movie Recommendation* (Chicago, Illinois, USA)

(CAMRa '11). Association for Computing Machinery, New York, NY, USA, 2–4. https://doi.org/10.1145/2096112.2096113

[75] Maya Sappelli, Suzan Verberne, and Wessel Kraaij. 2017. Evaluation of context-aware recommendation systems for information re-finding. *Journal of the Association for Information Science and Technology* 68, 4 (2017), 895–910. https://doi.org/10.1002/asi.23717

[76] Reijo Savolainen. 1995. Everyday life information seeking: Approaching information seeking in the context of way of life. *Library & Information Science Research* 17, 3 (1995), 259 – 294.

[77] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* 45, 11 (1997), 2673–2681.

[78] Chirag Shah. 2018. Information Fostering - Being Proactive with Information Seeking and Retrieval: Perspective Paper. In *Proceedings of the 2018 Conference on Human Information Interaction amp; Retrieval* (New Brunswick, NJ, USA) *(CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 62–71. https://doi.org/10.1145/3176349.3176389

[79] Gyanit Singh, Nish Parikh, and Neel Sundaresan. 2012. Rewriting Null E-Commerce Queries to Recommend Products. In *Proceedings of the 21st International Conference on World Wide Web* (Lyon, France) *(WWW '12 Companion)*. Association for Computing Machinery, New York, NY, USA, 73–82. https://doi.org/10.1145/2187980.2187989

[80] Jiwei Tan, Xiaojun Wan, Hui Liu, and Jianguo Xiao. 2018. QuoteRec: Toward Quote Recommendation for Writing. *ACM Transactions on Information Systems.* 36, 3, Article 34 (March 2018), 36 pages. https://doi.org/10.1145/3183370

[81] Jaime Teevan. 2008. How People Recall, Recognize, and Reuse Search Results. *ACM Transactions on Information Systems.* 26, 4, Article 19 (Oct. 2008), 27 pages. https://doi.org/10.1145/1402256.1402258

[82] Tiffany C Veinot. 2007. "The Eyes of the Power Company": Workplace Information Practices of a Vault Inspector. *The Library Quarterly: Information, Community, Policy* 77, 2 (2007), 157–179.

[83] Katrien Verbert, Nikos Manouselis, Xavier Ochoa, Martin Wolpers, Hendrik Drachsler, Ivana Bosnic, and Erik Duval. 2012. Context-Aware Recommender Systems for Learning: A Survey and Future

Challenges. *IEEE Transactions on Learning Technologies* 5, 4 (2012), 318–335. https://doi.org/10.1109/TLT.2012.11

[84] Tung Vuong, Salvatore Andolina, Giulio Jacucci, and Tuukka Ruotsalo. 2021. Does More Context Help? Effects of Context Window and Application Source on Retrieval Performance. *ACM Transactions on Information Systems.* 39, 3, Article 1 (Oct. 2021), 1 pages. https://doi.org/10.1145/3474055

[85] Tung Vuong, Salvatore Andolina, Giulio Jacucci, and Tuukka Ruotsalo. 2021. Spoken Conversational Context Improves Query Auto-Completion in Web Search. *ACM Transactions on Information Systems.* 39, 3, Article 31 (May 2021), 32 pages. https://doi.org/10.1145/3447875

[86] Tung Vuong, Giulio Jacucci, and Tuukka Ruotsalo. 2017. Watching inside the Screen: Digital Activity Monitoring for Task Recognition and Proactive Information Retrieval. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies.* 1, 3, Article 109 (Sept. 2017), 23 pages. https://doi.org/10.1145/3130974

[87] Tung Vuong, Miamaria Saastamoinen, Giulio Jacucci, and Tuukka Ruotsalo. 2019. Understanding user behavior in naturalistic information search tasks. *Journal of the Association for Information Science and Technology* 70, 11 (2019), 1248–1261. https://doi.org/10.1002/asi.24201

[88] Simon Wakeling, Paul Clough, and Barbara Sen. 2014. Investigating the Potential Impact of Non-Personalized Recommendations in the OPAC: Amazon vs. WorldCat.Org. In *Proceedings of the 5th Information Interaction in Context Symposium* (Regensburg, Germany) *(IIiX '14)*. Association for Computing Machinery, New York, NY, USA, 96–105. https://doi.org/10.1145/2637002.2637015

[89] Ryen W. White, Peter Bailey, and Liwei Chen. 2009. Predicting User Interests from Contextual Information. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Boston, MA, USA) *(SIGIR '09)*. Association for Computing Machinery, New York, NY, USA, 363–370. https://doi.org/10.1145/1571941.1572005

[90] Ryen W. White and Diane Kelly. 2006. A Study on the Effects of Personalization and Task Information on Implicit Feedback Performance. In *Proceedings of the 15th ACM International Conference on*

Information and Knowledge Management (Arlington, Virginia, USA) *(CIKM '06)*. Association for Computing Machinery, New York, NY, USA, 297–306. https://doi.org/10.1145/1183614.1183659

[91] Ricbard Whitley and Penelope Frost. 1973. Task Type and Information Transfer in a Government Research Laboratory. *Human Relations* 26, 4 (1973), 537–550.

[92] Jessica L. Wildman, Amanda L. Thayer, Michael A. Rosen, Eduardo Salas, John E. Mathieu, and Sara R. Rayne. 2012. Task Types and Team-Level Attributes: Synthesis of Team Classification Literature. *Human Resource Development Review* 11, 1 (2012), 97–129. https://doi.org/10.1177/1534484311417561 arXiv:https://doi.org/10.1177/1534484311417561

[93] Chirayu Wongchokprasitti, Jaakko Peltonen, Tuukka Ruotsalo, Payel Bandyopadhyay, Giulio Jacucci, and Peter Brusilovsky. 2015. User Model in a Box: Cross-System User Model Transfer for Resolving Cold Start Problems. In *User Modeling, Adaptation and Personalization (Lecture Notes in Computer Science)*, Francesco Ricci, Kalina Bontcheva, Owen Conlan, and Séamus Lawless (Eds.). Springer International Publishing AG, Switzerland, 289–301. https://doi.org/10.1007/978-3-319-20267-9$_2$4

[94] Yuxi Wu, Panya Gupta, Miranda Wei, Yasemin Acar, Sascha Fahl, and Blase Ur. 2018. Your Secrets Are Safe: How Browsers' Explanations Impact Misconceptions About Private Browsing Mode. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 217–226. https://doi.org/10.1145/3178876.3186088