Information Processing and Management 58 (2021) 102522

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Deconstructing search tasks in interactive information retrieval: A systematic review of task dimensions and predictors

Jiqun Liu

School of Library and Information Studies, College of Arts and Sciences, University of Oklahoma, Bizzell Library, Room 120, 401 West Brooks, Norman, OK 73019, USA

A B S T R A C T

Increasingly, users are interacting with information retrieval (IR) systems with the goal of addressing complex, ill-defined search tasks. As a result, the interest in deconstructing and predicting search tasks has grown among IR researchers. Given the multidimensionality of search tasks, researchers usually focus on one or multiple dimensions and study the associations between implicit task dimensions and observable predictors. Synthesizing the knowledge of tasks and predictors learned from individual user studies can clarify the progresses we have made as a research community and provide an intellectual benchmark for further explorations of task-based search interactions. This article presents an overview of 76 task-based interactive IR (IIR) studies published between 2000 and 2020 and systematically coded the papers using features such as task dimensions (definitions and operationalizations), task-predictor associations, as well as task prediction models. Results include 1) data illustrating the growth and interdisciplinarity of IIR studies; 2) a comprehensive typology of task dimensions along with the associated measures; 3) a summary of the statistically significant correlations between task dimensions and predictors; 4) a list of the task dimensions being predicted, ground truth labels, and the feature sets employed. In addition, our bibliography of IIR works can be used by students and junior researchers who are new to the area. The results of our review can facilitate the growth of knowledge in IIR community and serve as the basis for future research on new modalities of user-task interactions.

1. Introduction

People's interactions with search systems are often motivated by tasks that emerge from evolving, continuous problematic situations (Belkin, 2015). According to a 2018 Pew Research Center survey[1], a big majority of Americans (81%) rely on their online searches for gathering useful information and making decisions of varying types. Many Interactive Information Retrieval (IIR) researchers have reacted to this phenomenon by focusing studies on characterizing and predicting search task properties in users' search interactions (e.g., Capra et al., 2018; Li & Belkin, 2008; Liu, Liu, Cole, Belkin & Zhang, 2012; Liu et al., 2020) and developing task-aware recommendations for various modalities of information search (e.g., Ahn et al., 2008; Moshfeghi et al., 2017; Song & Guo, 2016; Yuan & Belkin, 2010; Zhang et al., 2018). With the support of existing IR models, search systems and technologies have achieved phenomenal success in addressing simple fact-finding and navigational search tasks (White, 2018). However, we still face plenty of challenges and obstacles when seeking to support users engaging in *complex tasks* that involve multi-round, multidimensional search iterations (e.g., writing a research proposal, planning a conference trip, preparing useful information for a job interview) (Awadallah et al., 2019). To address this problem, researchers have explored a variety of *dimensions of search tasks* (e.g., task complexity, task difficulty, task product, task state) as well as the *predictors* (e.g., search behavioral measures, quality of retrieved results) that may help infer and predict the nature of these dimensions. Investigating search task dimensions and the associated predictors can 1) help us

E-mail address: jiqunliu@ou.edu.

[1] https://www.pewresearch.org/fact-tank/2020/03/05/most-americans-rely-on-their-own-research-to-make-big-decisions-and-that-often-means-online-searches/

https://doi.org/10.1016/j.ipm.2021.102522
Received 29 June 2020; Received in revised form 13 November 2020; Accepted 17 January 2021
Available online 26 January 2021

empirically test the hypotheses derived from theories on task-based IR and 2) pave ways to simulating, developing, and evaluating task-aware search supports. A systematic review of the studies on search task dimensions and predictors can facilitate the accumulation of our knowledge on this area by 1) clarifying our current position within the research map of task-based IR, and more broadly, IIR in general, 2) revealing the implicit associations among individual unit theories, hypotheses, and predictors, and 3) pointing out the limitations, unexplored issues, and future directions within this research stream.

In much of previous IIR research which involves the concept of task, a search task was conceptualized either as an abstract representation of a set of desired search goals (static perspective) or as a sequence of search actions (dynamic perspective) (Byström & Hansen, 2005). Information search task is also defined as a task that a user seeks to accomplish through interactions with information systems (Li, 2009). According to Li (2009), a search task is usually situated within a broader work task, which refers to a separable activity people perform to fulfill their responsibility for their work (i.e. work-related tasks). In the rest of this paper, when discussing the concept of task, we are referring to search task, unless specified otherwise.

Given the *multidimensionality* of search tasks (Li & Belkin, 2008), researchers usually focus on one or multiple dimensions of a search task and study the association between task dimensions and users' search behaviors. The dimensions examined in previous IIR studies include both predefined, objective task facets (e.g., task goal, product, complexity) and user-task-combined dimensions (perceived task difficulty, task-related topical knowledge, users' interests on task). In addition to these session-level static factors, some researchers have also explored the dynamic aspect of search tasks and examined the predictability of evolving task states in search sessions (Liu et al., 2020). The fundamental assumption behind this dynamic perspective is that we cannot fully reveal the nature of a search task without understanding the process of performing the task. Adopting both static and dynamic perspectives will enable us to understand the connections between tasks and search interactions in a comprehensive manner.

To inform the design of intelligent search and task supporting systems, researchers have also examined the extent to which the nature of task dimensions can be inferred and predicted from the *predictors* extracted from different aspects of search interactions. A growing body of empirical research focusing on search task prediction and modeling has generated a variety of measures and techniques that have been applied and evaluated in different task contexts and modalities of search (Awadallah et al., 2019; Koolen et al., 2017). With the increasing complexity of study designs and interdisciplinary approaches, the variations and divergencies in task conceptualization, operationalization, and prediction often make it difficult to extract and synthesize common themes, standard practices, and collective insights from individual unique user studies. These divergencies call for an overview that systematically summarizes the studies and findings on the connections between search process and search task properties. This overview could facilitate meta-evaluations of predictors and reflections on the validity, generalizability and transferrability of the findings obtained from diverse study environments.

The aim of this study is to analyze how the dimensions of search tasks have been conceptualized and operationalized, and how different task dimensions are associated with predictors of different types. Specifically, we seek to answer following research questions:

- How do IIR studies conceptualize and operationalize search task dimensions?
- How are search task dimensions associated with predictors of varying types?
- Which search task dimensions have been inferred and predicted from existing predictors?

The article is structured as follows. We start by introducing the background and further explaining the rationale of our study. Then, this article describes the methods used in the study. After that, we analyze the definitions and operationalizations of search task dimensions investigated in task-based IR studies, with the goal of clarifying the advances and gaps in this area. Next, we summarize the predictors extracted from search interactions and present findings about the connections between search task dimensions and the associated predictors. We conclude by discussing the contributions and limitations of our study and suggesting possible directions for future research.

2. Background

This section focuses on the theoretical background of our work and discusses three interrelated topics: 1) dimensions of search tasks; 2) search task predicators, and 3) systematic reviews of IIR studies. To further clarify our rationale and motivation in this study, this section puts its emphasis on the existing gaps and limitations in previous studies with respect to conceptualizing and operationalizing task dimensions and extracting prediction measures.

2.1. Dimensions of search tasks

In IIR community, search tasks are often conceptualized as complex entities consisting of multiple dimensions (e.g., Li & Belkin, 2008; Liu, Cole et al., 2010). To gain a comprehensive understanding of the impacts of tasks on users' search activities, information seeking and IIR researchers have examined task dimensions or facets of different types (e.g., cognitive task complexity, task difficulty, task stage) and developed a variety of task typologies and operational measures. Given the importance of task characteristics, Kelly et al., and Capra (2013) organized an NSF workshop on task-based search at the University of North Carolina at Chapel Hill, where a group of researchers suggested a standardized task framework that would facilitate the discovery and documentation of universal task dependencies. Investigating different task dimensions allows researchers to deconstruct complex task-based IR problems into smaller, actionable questions and approach the nature of search tasks from different perspectives.

The task dimensions examined in previous studies can be roughly grouped into two categories: 1) *Objective search task dimensions*

(OTDs), which refer to the dimensions that can be objectively defined and measured without involving user traits or the processes of search interactions; 2) *Task-user combined dimensions* (TUDs), which refer to the dimensions that are jointly defined by task traits and user perceptions and involve the interaction between users and task features during search processes. Note that the major distinction between OTDs and TUDs is the way in which these task dimensions are operationalized and manipulated in empirical studies. OTDs are often determined by predefined task requirements and the characteristics of the environment in which the associated search task is performed. For instance, Kelly et al., and Wu (2015) defined cognitive task complexity based on the nature of cognitive process involved and the complexity of the learning goal(s) associated with a given task. At the operationalization level, Kelly and her colleagues manipulated cognitive task complexity by designing tasks associated with learning goals of varying complexity. In this sense, this task dimension was objectively designed and measured (without directly measuring the cognitive variation across different search scenarios). Urgo et al., and Capra (2019) further enriched this complexity-based task taxonomy by incorporating knowledge dimension (possible values: factual, conceptual, procedural, or metacognitive) into it. Similarly, Capra et al. (2018) created tasks of varying determinability by manipulating the items and dimensions involved. Differing from OTDs, the operationalizations of TUDs are usually associated with the interaction between users and their search tasks. Many widely discussed TUDs in IIR community involve using scales and self-reported ratings to measure users' perceptions and experiences regarding different aspects of search tasks or the process of doing tasks, such as topic knowledge (e.g. Liu et al., 2016), perceived task difficulty (e.g. Aula et al., 2010), level of interest in a task (e.g. Edwards & Kelly, 2017), and users' expectations or estimations of task difficulty (Liu & Shah, 2019a). In contrast to OTDs, TUDs emerge and vary during search interactions and thus cannot be manipulated beforehand.

Although there is a growing body of research on theorizing, measuring, and predicting search task dimensions, we still lack a systematic review that 1) explicates how different task dimensions are conceptualized and operationalized and 2) clarifies where we stand as a research community in terms of understanding and modeling search tasks. In addition, it would also be useful to know the overall distribution of research efforts on different topics (e.g. which dimensions are the center of previous research and which dimensions are on the periphery or scarcely studied) and the reason(s) behind the distribution pattern. A deep reflection on the connections between task dimensions and the associated predictors can 1) facilitate the accumulation and meta-evaluation of our existing knowledge on task-based IR and 2) help researchers utilize the knowledge about search task dimensions in supporting people engaging in complex search tasks via search recommendation and proactive intervention.

2.2. Search task predictors

A significant amount of prior research has been devoted to inferring and predicting search task characteristics from observable signals, aiming to inform the design of task-aware search and recommender systems. Identifying search task predictors based on observable signals enables researchers to 1) understand complex inputs (e.g. querying, eye movements, browsing patterns, other implicit usefulness feedbacks) from users at multiple levels and 2) develop systems that can detect, differentiate and support on-going search tasks in various scenarios.

According to the differences in the focus and method of data collection, search task predictors can be roughly classified into three groups: 1) *Online behavioral measures*, which covers the behavioral measures generated by users in on-going search interactions (e.g., query formulation and reformulation, clicking on search result surrogates, dwell time features, cursor movements) and can be directly extracted from search logs; 2) *Neuro-physiological measures* that capture signals about users' cognitive loads and activities, such as eye movement features (e.g. eye fixation, saccade), heart rates, electroencephalogram (EEG) signals, and Functional magnetic resonance imaging (fMRI) measures. These measures are argued to be effective in providing more direct evidence on users' cognitive processes compared to the behavioral features extracted from search logs (Mostafa & Gwizdka, 2016); 3) *Non-search situational measures* that go beyond search sessions and characterize the problematic situation that contextualizes a user's search activities, such as physical locations visited, time of the day, and on-going work tasks (e.g. Aliannejadi et al., 2019). Among the three types of measures employed in existing studies, online behavioral measures have attracted a majority of the research efforts as they can be easily extracted from search logs and do not depend on any external assessments or extra research devices (e.g., eye tracker, EEG machine, mobile and wearable devices).

Empirical evidences from existing studies demonstrated that there is no "one-size-fits-all" model or feature set for predicting all task dimensions and types (e.g. Jiang et al., 2014; Liu, Cole et al., 2010; Liu et al., 2020). A deep look into the observable features or task predictors proposed in prior IIR research allows researchers to 1) embrace a more comprehensive understanding of the connections between predictors of varying types and search task dimensions, and 2) identify useful measures as well as understudied measures, which can be of help for future research on task modeling and search personalization.

2.3. Systematic reviews of IIR studies

Systematic reviews seek to make literature review process more transparent, rigorous and replicable to some extents and are among the most highly cited documents in social sciences (Cooper et al., 2019). In IIR community, systematic reviews have contributed to the accumulation of knowledge learned from empirical studies and the connections between unit theories, concepts, and techniques. Existing systematic reviews in IIR have covered three main types of focuses: 1) *Empirical studies on core research problems of IIR* (e.g., IR system evaluation, Kelly & Sugimoto, 2013; personalization in text retrieval, Liu et al., 2019; usefulness of search results, Vakkari, 2020), 2) *IIR study methods and techniques* (e.g., Kelly, 2009; Liu & Shah, 2019b), and 3) *Concepts and theories of varying scopes* (e.g., Byström & Hansen, 2005; Ruthven, 2008).

According to Cooper et al. (2019), a complete systematic review consists of following steps: 1) State research questions; 2) Develop

guidelines for collecting literature (e.g. inclusion and exclusion criteria); 3) Develop a comprehensive search plan for finding relevant literature; 4) Develop a codebook and code form for grouping and characterizing literature; 5) Code and analyze the literature; 6) Synthesize the collected literature.

Built upon decades of IR research, Kelly and Sugimoto (2013) summarized the components of IIR evaluation studies and clarified the focuses and trends in IR evaluations. Similarly, Vakkari (2020) focused on one aspect of search interaction, *usefulness of search results*, and explained how result usefulness is conceptualized and measured in a variety of empirical studies. These two systematic reviews present a comprehensive picture of the research progresses in one or multiple sub-areas and clarify where we are as a community in research explorations. In addition to summarizing empirical evidences, IIR researchers also seek to sharpen the research methods of IIR by summarizing, comparing, and meta-evaluating different components of IIR study designs (e.g., Kelly, 2009; Liu & Shah, 2019b). For instance, following a systematic review of IIR user studies, Liu and Shah (2019b) proposed a faceted framework in which a user study can be deconstructed, characterized, and evaluated in terms of various dimensions (e.g. tasks, tools and devices, experimental design, user characteristics). With respect to theoretical progresses, Byström and Hansen (2005) focused on one of the core concepts of IIR, task, and discussed the differences and connections between multiple levels of tasks (i.e. work task, information seeking task, search task) based on empirical findings obtained both within and beyond library and information studies. Differing from this one-focus approach, Ruthven (2008) provided an overview of IIR area and summarized a series of classical theories and models of search interactions. These reviews on theories and concepts facilitate the accumulation of theorized knowledge and offer us a better sense of the advances and limitations in theory development.

Given the contributions and advantages of system reviews, we seek to utilize the method in synthesizing the knowledge of search task dimensions and associated predictors reported in existing studies and addressing the research gaps identified in the Introduction section.

3. Methods

Methodologically, inspired by Cooper et al. (2019) and Kelly and Sugimoto (2013), our systematic review included following steps:

1) Establish and clarify the inclusion and exclusion criteria for research paper selection.
2) Based on the criteria from step 1), implement the method (e.g. sources and databases involved, keyword combinations) of retrieving and selecting studies on task-based IR.
3) Propose a coding scheme for annotating research papers.
4) Synthesize the findings from selected papers and answer research questions.

These steps are explained in more detail.

3.1. Step 1: Inclusion and exclusion criteria

As the initial step of the review, this section clarifies the inclusion and exclusion criteria for paper selection, which guides the second step (retrieving and selecting papers) and set boundaries for our work.

We first covered empirical studies that seek to predict *search task type* defined by one or multiple task dimensions. We also included the research that investigated the correlations between task facets and a variety of predictors (e.g. behavioral signals, external assessments), with the ultimate goal of predicting task dimensions and informing the personalization of IR (e.g. Jiang et al., 2014; Liu, Cole et al., 2010). This review excluded the studies where researchers used automatically extracted keywords and topics as task representations (e.g., Mehrotra & Yilmaz, 2017), instead of designing simulated search tasks beforehand (e.g., Jiang et al., 2014) or eliciting authentic task information from participants in naturalistic settings (He & Yilmaz, 2017). We employed this exclusion criteria for two reasons: 1) the focuses of these studies are not on the multidimensionality of tasks, and thus are not relevant to the theme of our study; 2) regarding the techniques involved, the main contributions of these studies are on topic mining and factual task extraction, rather than characterizing and predicting complex tasks in interactions. Moreover, including the studies on automatic topic and task extractions would have complicated the analysis by undermining the comparability of literature and results.

Besides, we excluded the IR evaluation studies where researchers focused on ad hoc retrievals or assigned predefined search engine result pages (SERPs) to participants, without eliciting any authentic search sessions under tasks of varying types. This is because these studies cannot tell us how features of online search interactions help infer and predict search task characteristics. We also excluded the empirical studies that did not involve any variations in task characteristics or only manipulated task topics. In these studies, researchers utilized predefined tasks as simulated contexts for eliciting interactive search activities, without measuring the potential impact of any task trait. In addition, we excluded the user studies where the necessary details of search tasks and predictors are not fully reported (which would have created extra difficulties for paper coding and synthesis). For instance, in some conference short papers, the authors omitted some details of task design and predicting measures due to the page limits. In these cases, we traced back to the corresponding long papers (if applicable) where the full study procedure, metrics, and findings are explained in detail. The inclusion and exclusion criteria above offered us a reasonable scope of paper retrieval and allowed us to be more accurate on explaining the contributions and limitations of our work.

In this study, we focused on the task-based IIR studies published between 2000 and 2020, mainly for two reasons: 1) most of the IR studies published before 2000 were conducted with the ultimate goal of improving ad hoc retrieval performance and are thoroughly reviewed in Kelly and Sugimoto (2013); 2) this time range (2000-2020) covers a variety of IIR research published on diverse venues

associated with multiple domains and disciplines (e.g. Library and Information Science, IR and Computer Science, Human-Computer Interaction, Psychology, Economics, Ergonomics). This diversity allowed us to trace back to various theoretical and methodological origins of task-based IIR studies and enhanced our understanding of the value of interdisciplinarity in this research area.

3.2. Step 2: Paper retrieval

With the inclusion and exclusion criteria introduced above, we collected publications on task-based IR through searches in a variety of research publication search engines and databases. Specifically, we followed and expanded Vakkari (2020)'s paper retrieval approach and collected a variety of studies by searches in Google Scholar, Science Direct, Web of Science, the Association for Computing Machinery (ACM) Digital Library, Academic Search Premier, as well as Microsoft Academic. These search engines and databases cover all major venues where task-based IR research are communicated, such as ACM Special Interest Group on Information Retrieval (SIGIR) conferences (e.g., SIGIR, CIKM, CHIIR, JCDL, ICTIR, WSDM), ACM Transactions on Information Systems (TOIS), Information Processing and Management, Journal of the Association on Information Science and Technology (JASIST), and ASIS&T Annual Meetings. We formulated queries with following search terms: *search task, task facet, task dimension, task type, prediction, search task classification, task characteristics, user study, information search, Web search, search interaction, information seeking, search strategy, search action, search tactic, search behavior* and *knowledge*. We tried different combinations of the keywords, aiming to cover most, if not all, of the relevant research for further analysis. Whenever a formulated query resulted in only a few hits, we modified the structure and made the query more general and broader (e.g. deleting keywords, adjusting the logic of query).

Since relying on queries alone is not sufficient for fully collecting relevant studies due to the variety of labels that researchers assign to IIR works (Kelly & Sugimoto, 2013), we decided to identify missing studies and validate our search processes using manual method. Specifically, we complemented our subject searches by citation chaining (i.e. tracing relevant research and authors through citations in reference lists) and author searches of scholars who had published relevant research articles. After the paper retrieval was completed, we manually removed repeated papers indexed in multiple databases, topically irrelevant papers (e.g., psychology research paper on visual search), and other papers that do not fit with our inclusion criteria.

After the paper retrieval and manual selection process, a total of 76 papers were identified for inclusion in the dataset. Each paper functions as a basic unit or instance of our coding and comparative analysis, which is introduced in the following section.

3.3. Step 3 & 4: Paper coding and synthesis

To analyze the ways in which search tasks of varying types are conceptualized, deconstructed, and predicted in existing IIR studies, we defined several *coding features* and applied them in paper coding, including: 1) *Publication features*: paper title, venue and year; 2) *Task dimensions*: definitions and operationalizations in empirical studies; 3) *Predictors*: types and specific measures; 4) *Statistically significant associations* between task dimensions and predictors; 5) Performances of *machine learning predictions*. These coding features were not randomly selected based on convenience. Instead, they were proposed and defined according to the research questions (RQs) we sought to answer: The publication features give us an overview of the distribution of task-based IR research across different topics, subareas, and venues. The task dimension feature corresponds to the RQ1 (i.e. how task dimensions are conceptualized and operationalized). The predictors feature and task-predictor associations speak to the RQ2 (i.e. how task dimensions are associated with predictors of different types). The last coding feature, performances of Machine Learning (ML) predictions, addresses the RQ3 and sheds light on the progresses we have made so far as a research community on predicting task characteristics.

We employed the *coding scheme* consisting of the aforementioned coding features as the basis for extracting information from collected papers. To analyze the information associated with each coding feature, when reviewing papers, we mainly focused on the methods, results and discussion sections. Based on the typologies of task dimensions and predictors developed through literature review, we deconstructed and categorized the instances (i.e. papers) and contrasted them with respect to different features and factors (e.g. task dimensions involved, operationalizations of task dimensions, prediction measures). By comparing and contrasting different categories of instances, we gained a clearer understanding of the approaches through which researchers have conceptualized and operationalized various types of search tasks as well as the associated predictors. After the categories of task dimensions and predictors were finalized and enriched with specific instances, to answer the RQ2 and RQ3, we took a step forward by reviewing the task-predictor associations identified in empirical studies and summarizing the performances of ML models in predicting various task-related features.

The coding process started with two coders, the author and a Ph.D. student majoring in IR, coding the first twenty retrieved papers separately retrieved papers according to the predefined coding scheme. After this initial coding process was finished, we compared our coding results and found only two minor differences (the student annotator accidentally missed out two task dimensions covered in the result section of a paper). Note that we obtained this high agreement rate because the coding features we defined were straightforward and clearly reported in most of the reviewed papers. Thus, differing from regular coding analysis in qualitative research, there was no much room (if any) for different interpretations. After the initial coding for twenty papers, the author solely completed the coding work for the remaining 56 papers. It is worth noting that we did encounter obstacles when annotating some of the papers due to issues in result reporting (e.g., behavioral features used for predicting analysis were not fully reported; some details regarding task design were missing). These problems occurred because our work was restricted to the contents that are explicitly reported in the reviewed papers. Thus, we could not address these issues by adding more coders or altering coding schemes. This paper further explains the issues as well as possible solutions in the discussion section.

Synthesizing the knowledge learned from existing studies can inform us of the progresses and contributions that IIR researchers

have made in understanding the multidimensionality of search tasks and predicting task characteristics from various signals. Also, through the systematic review, we sought to identify and discuss the understudied task dimensions, missing predictors, as well as other broader contextual factors (e.g. task relevance to a specific time point and situation), which can be of help for further expanding the scope of future IIR research.

4. Result

The result section reports the findings from our synthesis and comparative analysis and describes the papers/studies, organized by the coding scheme and the three research questions.

4.1. Overall characteristics of publications

Task-based IIR is an interdisciplinary area in nature and attracts attentions of researchers coming from different backgrounds and disciplines. As a result, during paper retrieval and coding, we found that the collected papers were published in multiple and inter-disciplinary venues.

Given the divergent topical focuses and main disciplines involved, we roughly classified the venues into four categories: 1) *Library and Information Science (LIS) venues*, such as JASIST, Proceedings of ASIS&T Annual Meeting, Library and Information Science Research, Journal of Documentation and Journal of Academic Librarianship. 2) *Computer Science and Information Retrieval (CS&IR) venues*, such as most of the SIGIR sponsored conferences (SIGIR, ICTIR, CIKM, JCDL, WSDM), European Conference on Information Retrieval (ECIR), ACM Transactions on Information Systems (TOIS), Information Processing and Management. 3) *Human-Computer Interaction and Information Retrieval (HCI&IR) venues*, such as CHIIR and ACM SIGCHI, and 4) *Other venues*, such as journals from psychology and cognitive ergonomic areas. We studied the distribution of papers/studies across different types of venues and analyzed how the distribution change over time. This analysis gave us a better sense of the composition of the interdisciplinary research efforts on this topic and the trend and temporal variations in this composition. The result could also illustrate the diversity in specific research focuses (i.e. task dimensions and operationalizations), conceptual frameworks and methodologies, which are explained in detail in the following sections.

Fig. 1 shows how the distribution of studies across publication venues changes over time. Overall, the number of studies has kept growing from 2001-2015, with the LIS and CS&IR being the two major types of venues for publication. During this time period, a variety of user studies emerged in Information Seeking and IR communities, aiming to go beyond the Cranfield paradigm and explore the role of search tasks in people's information seeking and search episodes (e.g. Cole et al., 2014; Li & Belkin, 2008; Li et al., 2011; Liu, Cole et al., 2010).

Meanwhile, we observed the increasing research efforts from IIR and HCI communities, with more papers published in a relatively new venue, CHIIR, which is co-sponsored by SIGCHI and SIGIR. In recent years (2016-2020), we saw almost equal contributions from CS&IR and HCI&IR groups, with decreasing coverage from traditional LIS venues. During this time period, many researchers sought to 1) further explore the roles of task-user combined factors in various search scenarios (e.g., Edwards & Kelly, 2017; He & Yilmaz, 2017; Liu et al., 2020) and 2) support users' search episodes by designing new recommendations and interfaces (Choi et al., 2019; Kulahcioglu et al., 2017). As more studies focusing on search tasks were published in computer science and HCI venues, the proportion of conference papers has been rising over time, which provides timely exposures of new breakthroughs (see Fig. 2).

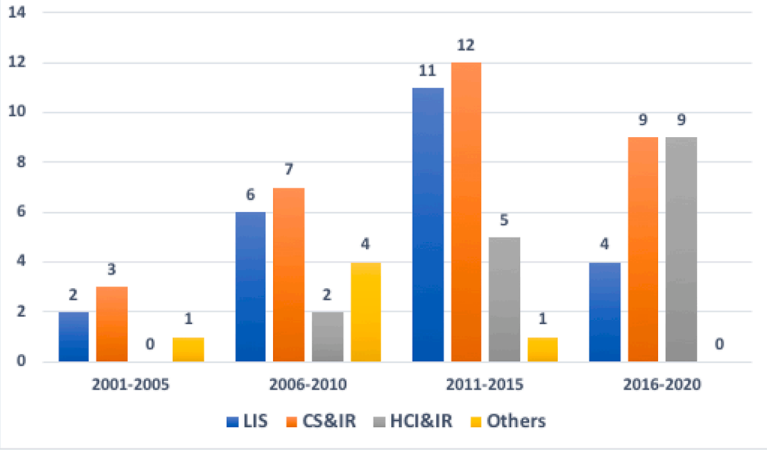

Fig. 1. Distribution of papers across disciplines.

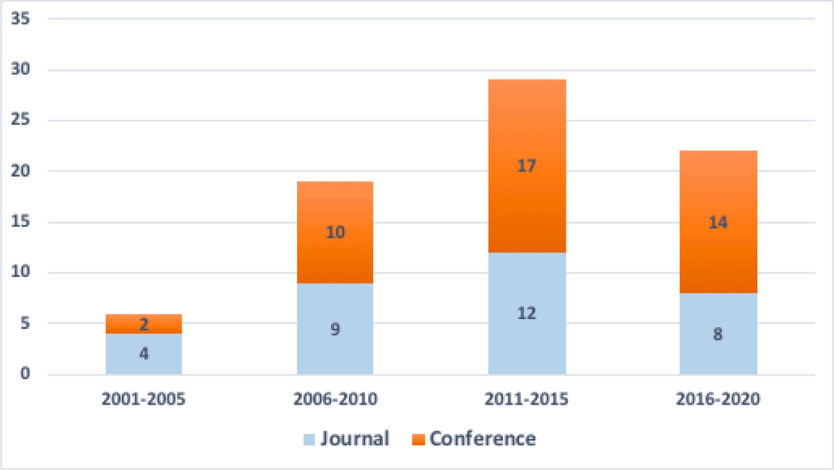

Fig. 2. Distribution of papers across venue types: Journal versus Conference.

4.2. RQ1: search task dimensions

IIR researchers have explored various dimensions of search tasks, which allows them to conceptualize tasks from different perspectives. However, the research efforts devoted to this area have not been equally distributed across different dimensions. Although we have a few task dimensions being thoroughly examined and measured in a variety of contexts (e.g., task difficulty, task complexity), many other task dimensions that may significantly affect search interaction and evaluation still remain insufficiently studied or even unexplored.

Table 1
Objective task dimensions (OTDs).

Task dimension	Conceptualization and Operationalization
Task complexity (n=13)	Levels of cognitive complexity/types of cognition required (remember, understand, analyze, create) (e.g., Arguello et al., 2012; Brennan et al., 2014; Choi et al., 2019; Ghosh et al., 2018; Kelly et al., 2015; Walhout et al., 2017) Two-dimensional taxonomy: cognitive process dimension and knowledge dimension (factual, conceptual, procedural, metacognitive) (Urgo et al., 2019) Task determinability (manipulated by specifying items, objective dimensions and subjective dimensions) (Capra et al., 2017; Capra et al., 2018) Search-strategy-based classification: e.g., low complexity task (a task that provided subjects with more information on what needs to be found; users evaluate important attributes separately) and high complexity task (a vaguely formulated task requiring information from multiple sources; a holistic strategy or multiple information sources is needed; task goals and schedules need to be continuously adjusted) (Bin, 2009; Lopatovska, 2010; Yoo, 2009) Search efforts needed: simple fact-finding problems (answer directly accessible on the SERP), difficult fact-finding problems (keywords have to be inferred), open-ended search problems (multiple answers possible and navigation necessary) (Sanchiz et al., 2017)
Task type (n=13)	Researchers employed tasks of varying types (within same or different topics), without defining or specifying any task dimension, for instance: Subject-oriented search, specific information search, known-item search (Kim, 2001; Xie & Joo, 2012) General purpose, information gathering and understanding task and specific fact-finding task (Liu et al., 2019; Pharo & Krahn, 2011; Thatcher, 2006) Factual task, exploratory task, and personal experience task ((Zhang et al., 2017)) Navigational goal, informational goal, and resource-seeking goal (Broder, 2002)
Task type – dimensions (n=11)	Researchers defined task types using combinations of multiple task dimensions, for instance: Combinations of (externally annotated) task product (factual versus intellectual), task goal (specific versus amorphous), task level (document versus segment), and task complexity (low objective complexity or high objective complexity) (e.g., Cole et al., 2011, 2014, 2015; Hienert et al., 2018; Li, 2009; Liu, Cole et al., 2010; Mitsui et al., 2018) Vertical and horizontal information needs; levels of information needs (professional practice, work task, information seeking task, information search task) (Byström & Kumpulainen, 2020)
Other dimensions (n=5)	Task scenario: government, library, commercial (Kessler et al., 2014) Task stage defined by subtasks and query modifications (He et al., 2013; Liu & Belkin, 2015)
Objective or predefined task difficulty (n=3)	Search efforts needed: e.g., ready to use (easy-to-find information with simple keywords), easy-to-interpret (requires organizing keywords in a meaningful manner), hard to interpret (needs to interpret the relations between the found information) (Aula et al., 2010; Şendurur, E. & Yildirim, 2015) Specificity of search goal: general task (ask participants to locate any tobacco cessation strategy), specific task (ask participants to locate a specific, well-defined tobacco cessation method) (Hong, 2006)
Task structure (n=3)	Hierarchical versus parallel (; Toms et al., 2007) Well-structured tasks versus poorly structured and indecomposable tasks (Dedema & Liu, 2019)

To answer the RQ1, we investigated how task dimensions (i.e. OTDs and TUDs) were defined and operationalized in previous research and summarized our findings in Tables 1 and 2. Among all OTDs, the most popular and widely studied task dimension is *task complexity*. As it is shown in Table 1, there are roughly two types of task complexity definitions: 1) defining task complexity based on the task characteristics and processes at the cognitive level, such as task cognitive complexity (e.g., Kelly et al., 2015), task determinability (Capra et al., 2018), and types of knowledge required (e.g., Urgo et al., 2019); 2) defining task complexity based on the efforts and tactics required at the behavioral level (e.g., Bin, 2009; Lopatovska, 2010; Sanchiz et al., 2017).

These two groups of definitions conceptualize task complexity at different levels and enrich our understanding of the reasons of a task being complex. At the operationalization level, however, both of them cause challenges to task modeling and the measurement of complexity. Specifically, for cognition-based definitions, it is often difficult, if not entirely impossible, to classify predefined tasks into clear, discrete categories. In many cases, the boundaries between different groups are not always clear or quantitatively measurable. For example, two different search tasks may involve both factual and conceptual knowledge and require users to both remember and understand new information for task performance. The lack of validity and reliability metrics for evaluating task typology also increases the ambiguity in complexity-based task classification. Regarding behavior-based definitions, inferring search tactics and efforts needed for pre-defined search tasks is challenging, especially in the studies where critical user characteristics (e.g., topic knowledge, search skill) cannot be controlled or manipulated.

Similarly, researchers have also explored other dimensions of tasks, such as objective task difficulty, task scenario, and task stage, which allows us to conceptualize tasks and develop task taxonomies from varying perspectives. Meanwhile, these studies also face similar challenges and obstacles associated with the research on task complexity. Additionally, in naturalistic studies, the qualitative nature of the task taxonomies discussed above (along with many other dimensions defined in Li & Belkin, 2008, such as task product, task goal, task level) often makes it difficult to quantitatively extract and model authentic search tasks in an online fashion. Transforming discrete task categories into specific thresholds and quantitatively separable groups, if possible, would be of help for bridging the gap between task-based IIR studies and computational research on Web-scale log-based task mining.

In addition to the above *one-dimension-based* task studies, researchers have also explored the nature of search tasks either 1) with no task dimension specified or 2) with task types defined by combinations of task dimensions. For instance, (Zhang et al., 2017) assigned three types of tasks (factual task, exploratory task, and personal experience task) to participants and examined the difference in search interactions across tasks of varying types. Similarly, Broder (2002) defined three types of goals (navigational, informational, and resource-seeking) and studied the associations between goal type and search tactics. These task types were developed using a bottom-up approach and were extracted from empirical evidences in qualitative analyses. As a result, researchers could easily map these task types to a majority of authentic tasks from naturalistic settings. However, it is intellectually challenging to 1) figure out the conceptual connections between these task types and the tasks defined in other studies and 2) evaluate the theoretical contributions from the studies using these no-dimension-specified tasks. Also, these task types are broad in scope and thus provide limited practical contributions to accurate task modeling and task-based search personalization.

To improve the flexibility and generalizability of task taxonomies, some researchers attempted to define task types using faceted or multidimensional approaches. Many of these faceted task designs were originated from Li and Belkin (2008)'s task classification framework, such as Cole et al. (2011, 2014, 2015), He and Yilmaz (2017), Hienert et al., and Belkin (2018), Liu et al. (2010), and Mitsui et al., and Shah (2018). These studies jointly illustrated a relatively flexible approach to deconstructing complex search tasks into separate dimensions and encouraged researchers to investigate the main and interaction effects of multiple "hidden" task characteristics on search activities and experiences via multivariate analyses. For instance, Mitsui et al. (2018) developed ML models to predict task type defined by four combinations of task product (factual versus intellectual) and task goal (specific versus amorphous). Regarding task product, intellectual task refers to a task which produces new findings or ideas (e.g., searching for useful information for preparing a research proposal), whereas a factual task refers to a task locating facts, data or existing other information objects. With respect to task goal, task with specific goal refers to a task with a measurable, explicit goal (e.g., information about today's weather), whereas a goal-amorphous task is defined as a task with a goal that has no explicitly defined outcome and is difficult to be measured quantitatively (e.g., finding useful information for learning python programming). Similar to He and Yilmaz (2017), Mitsui et al. (2018) also sought to represent authentic search tasks using the correlations among multiple facets.

Table 2
Task-user-combined dimensions (TUDs).

Task dimension	Conceptualization and Operationalization
Topic familiarity/ knowledge (n=5)	Self-assessed familiarity with search task topics on a 5-point or 7-point scale (e.g., Hu et al., 2013; Kelly & Cool, 2002; Liu et al., 2016)
Subjective task difficulty (n=4)	Users' perceived difficulty of fulfilling task goal(s) and requirement(s); self-rated on a 5-point or 7-point scale (e.g., Arguello, 2014; Liu, Gwizdka, Liu & Belkin, 2010; Liu et al., 2014)
Domain knowledge (n=3)	Users' knowledge about a domain or discipline associated with search task topics. Multi-stage assessment (before instruction/treatment, just after instruction, and a long period of time after the instruction) (Wildemuth, 2004) Self-rated knowledge on certain terms or terminology (Liu, Liu et al., 2012)
Other dimensions (n=4)	Self-rated interest in the task topic(s): users were asked to rank task topics based on their interest and preference (e.g. Edwards & Kelly, 2017) Task states: task states identified based on users' annotations of their intentions and in-situ search problems in task-based search sessions (Liu et al., 2020)

Besides the OTDs, IIR researchers have also explored the interactions between users and task characteristics in search sessions and investigated the roles and impacts of task-user-combined dimensions (TUDs). The TUDs introduced in Table 2 speak to different aspects of users' perceptions of tasks during search sessions, and are often measured through pre- or post-search self-assessments and annotations. For instance, topic familiarity is one of the widely discussed TUDs and significantly affects multiple aspects of search interactions (Kelly & Cool, 2002; Liu et al., 2016). At the operationalization level, researchers used *self-assessed familiarity* with search task topics (on a 5-point or 7-point scale) as a proxy of participants' topic familiarity. Likert scales as a low-cost, convenient tool can help researchers quantify topic familiarity and support statistical analyses of task effects (e.g. search behavioral variations across different levels of familiarity). However, the limitations of using discrete values in measurements and the potential biases in users' perceptions and memories may increase the difficulty in capturing the nuances and within-session variations of participants' topic familiarity. In addition, many of the self-reported measures as pseudo-independent variables cannot be controlled beforehand and thus may lead to skewed data distributions as well as the violations of the assumptions of many parametric tests. These limitations also exist in other scenarios where researchers use self-reported indicators to measure TUDs, such as perceived task difficulty, domain knowledge and self-rated task interest.

Besides the studies focusing on OTDs or TUDs, in our sample, there are also some research (n=12) that covered both OTDs and TUDs and sought to examine the behavioral impacts of multiple search task dimensions. For example, Liu and Belkin (2015) looked at several task dimensions, including task stage, task type as well as users' topic knowledge and examined how these contextual factors could help infer document usefulness from dwell time measures. Wildemuth et al., and Toms (2014) compared and contrasted two closely related concepts, task complexity and (perceived) task difficulty, and explained how these two concepts could be further deconstructed into separate items and smaller units of analyses. He and Yilmaz (2017) investigated the connections among three groups of task characteristics, which involve both objective task characteristics (e.g., task complexity, task stage, task urgency) and task-related perceptions and cognitive features (topic knowledge, task difficulty, user satisfaction).

In sum, different types of task dimensions were conceptualized and operationalized differently. Overall, OTDs were usually defined according to one or multiple dimensions. Based on the associated typologies, search tasks could be classified into discrete categories (e.g., factual task versus intellectual task, simple task versus complex task, easy task versus difficult task) and be represented using categorical variables. Regarding TUDs, previous research mainly relied on Likert-scale-based self-assessments and annotations. A variety of scales and metrics have been developed within this stream of research (e.g., Capra et al., 2018; O'Brien & Toms, 2008) and facilitated a wide range of IIR studies on users' perceptions and cognitive activities. To overcome the limitations associated with the subjectivity and biases in self assessments, researchers have also proposed various objective measures, such as search behavioral metrics, eye movement metrics, as well as other physiological features (e.g. fMRI), and sought to infer the nature of task dimensions objectively. These measures and their correlations with different task characteristics will be discussed in the following section.

4.3. RQ2: associations between predictors and search task dimensions

To bridge the typologies of multifaceted search tasks with the design of task-aware search systems, many existing studies sought to extract observable predictors from search interactions and use them to predict the nature of implicit search tasks defined by one or

Table 3
Types of measures.

Category	Predictors/Measures
Online behavioral measures (n=71)	Querying: query length, number of MeSH terms used, query term diversity, number of advanced operators used, number of unique keywords, query reformulation/modification type, number of query reformulations, time spent on each type of query reformulation, number of queries, number of unique queries, number of queries without a bookmark (e.g., Capra et al., 2018; He et al., 2013; Hu et al., 2013; Toms et al., 2007; Walhout et al., 2017; Wildemuth, 2004; Zhang et al., 2015) Browsing: number of clicks, number of clicks at different ranks, time to first result click, fraction of top ten results hovered, trail speed, mouse movement distance, number of scrolls, cursor idle time, number of clicks in search box, non-hyperlink result click count, number of clicks on vertical results, number of satisfied (SAT) clicks or dissatisfied (DSAT) clicks, number of clicks without a bookmark, total number of actions (e.g., Agichtein et al., 2012; Arguello et al., 2012; Buscher et al., 2012; Capra et al., 2018; Jiang et al., 2014) Dwell Time: dwell time on content page/document reading time, dwell time on SERP, dwell time on first SERP (within a session), dwell time on the first page/document, hovering time on search results surrogates, dwell time before saving first useful document/page, query dwell time, task completion time (e.g., Agichtein et al., 2012; Buscher et al., 2012; Liu & Belkin, 2015; Liu, Liu et al., 2012; Liu et al., 2014; Liu et al., 2020; Zhang et al., 2015) Usefulness Judgment: number of pages bookmarked/saved, ratio of pages saved to viewed/clicked (e.g., Liu, Liu et al., 2012; Zhang et al., 2015)
Neuro physiological measures (n=7)	Eye Movement: number of eye fixations, fixation duration, number of fixations in the reading sequence, pixels covered by reading sequences, number of long reading sequences (number of fixations>4), saccade length, number of single fixation ("scanning") sequences (e.g., Cole et al., 2011, 2014, 2015; Jiang et al., 2014; Mostafa & Gwizdka, 2016) Emotional and Physiological Measures: window heart rate, window skin conductance, mood measured by Positive Affect and Negative Affect Scale (PANAS) (e.g., Edwards & Kelly, 2017; Lopatovska, 2010)
Non-search situational measures (n=7)	User Experience and Engagement Measures: e.g., perceived search effort, focused attention, felt involvement, novelty, total engagement (all self-rated on scales) (e.g., Edwards & Kelly, 2016; Gwizdka & Spence, 2006; O'Brien et al., 2020) Situation of Search: device of search, task relevance (e.g., temporal relevance, personal relevance), action to task recommendation (actioned or dismissed) (e.g., Aliannejadi et al., 2019)

Note: the total number reported here is greater than 76 as some task-based IIR studies involved two or three types of predictors/measures.

multiple dimensions. Based on the systematic review, we identified three types of measures (i.e. online behavioral measures, neuro-physiological measures, non-search situational measures) which have been implemented in describing and predicting task characteristics in interactive search sessions.

We summarized the specific measures under each category in Table 3. Online behavioral measure as the main type of observable indicators covers four sub categories of search behaviors, including querying, browsing, dwell time features and usefulness judgment. Unlike other features that heavily rely on extra data collection devices and sensors (e.g., eye movement measures, heart rate and skin conductance), online behavioral measures can be directly extracted from search logs, sometimes with the help of extra tools (e.g. client-side JavaScript for capturing cursor movements). The independence from external devices and techniques create the potential for supporting Web-scale implementation and thus encourages researchers and engineers to keep exploring and extracting available online behavioral features.

Compared to online behavioral measures, neuro-physiological measures are expensive in the sense that they require extra devices, software for data analysis, as well as additional time for data cleaning and processing. However, neuro-physiological measures as more accurate, fine-grained measures could better capture the small but critical variations in users' search interactions and detect the nuances between similar cognitive and emotional states. Thus, to reduce the cost of building fine-grained task models, some researchers tried to find appropriate, low-cost behavioral measures as proxies of some neuro-physiological features. For instance, Huang et al., and Dumais (2011) presented a scalable approach to capture cursor movements and demonstrated that cursor position can serve as a good proxy of eye gaze, especially on SERPs. Using cursor movement features, one could approximate the movements of eye gaze, without implementing an eye tracker and cleaning eye movement data. Despite these efforts, some other neuro-physiological features still cannot be widely applied or approximated in search sessions due to various restrictions. For example, fMRI signals can be used to detect information needs and the anomalous state of knowledge in single-query/question contexts (Moshfeghi et al., 2016). However, it would be very difficult, if not entirely infeasible, to collect fMRI signals continuously in task-based search. The challenge of further bridging neuro-physiological signals with scalable metrics requires more research efforts and experiments.

Besides the features discussed above, researchers have also explored features outside search activities and investigated different aspects of users' overall search experience and the relevance of search tasks in broader environments (e.g., Aliannejadi et al., 2019; O'Brien et al., 2020). Despite the limitations in scalability and accuracy, these measures offer researchers a direct access to users' perceptions and memories of search experiences and thus can inform the research on broader impacts of task characteristics.

Based upon the summary of measures in Table 3, to answer the RQ2, we reviewed and summarized the statistically significant task-predictor associations demonstrated in previous empirical studies (see Tables 4 and 5). For each dimension, we extracted the predictors and measures that are either positively (+) or negatively (-) associated with the dimension. The "Correlated" list includes features for which the directions of correlations were not reported.

Overall, we found that complex search tasks that involve amorphous task goals, unclear structures, high-level cognitive activities (e.g., synthesizing, evaluation, and creating) and unfamiliar topics are often associated with deeper, more active search activities in

Table 4
Statistically significant associations between predictors and OTDs.

Task dimension	Statistically significant associations
Task complexity	+ number of queries, + number of keywords, + number of unique query terms, + query formulation/reformulation time, + number of abandoned queries (queries that do not lead to any click on SERPs), + SERP view/dwell time, + clicks on search results, + number of Web pages visited, + task completion time, + queries without a bookmark, + clicks without a bookmark - dwell time on SERPs in controversial topic task, - the frequency of selecting the first (top-ranked) search result
Task type	*Subject-oriented search, specific information search, & known-item search*: + use of embedded links by novice participants in known-item task than subject-oriented search; + use of jump tools by experienced participants in subject-oriented search. *General purpose, information gathering and understanding task and specific fact-finding task*: + task completion time in information gathering task, + click on embedded links in specific fact-finding task, + query dwell time in information gathering task - query length in information gathering task
Task type – dimensions	*Combinations of task complexity, level, product and goal* Correlated: task completion time, number of pages visited, number of sources/unique URLs clicked, number of queries, number of search sources, dwell time on content pages, decision time on usefulness judgment, read/scan ratio, saccade distance, number of eye fixations in reading sequences, number of transitions from scanning to reading *Combinations of task product and task goal* Correlated: number of pages visited, number of queries, task/session completion time, dwell time on content pages per SERP, % time on SERPs, average dwell time on each SERP
Other dimensions	*Task stage* + number of topics explored in initial stage, + focused search (longer queries and more overlaps between adjacent queries) in later stage
Objective or predefined task difficulty	+ query diversity, + number of advanced operators in query formulation, + dwell time on SERPs, + task completion time
Task structure	*Hierarchical versus parallel task* + number of queries in parallel task, + query dwell time in parallel task *Well-structured task versus poorly structured and indecomposable task* + frequency of query stopping/abandonment in poorly structured task

Note: +: positive correlation. -: negative correlation. Non-significant correlations are not included here.

Table 5
Statistically significant associations between predictors and TUDs.

Task dimension	Statistically significant associations
Topic familiarity/ knowledge	+ dwell time on SERPs, + dwell time on the first SERP - time spent on query reformulations, - ratio of dwell time on content pages, - number of query reformulations, - document reading time/content page dwell time, - query diversity
Subjective task difficulty	+ query dwell time, + dwell time on content pages, + number of pages visited, + dwell time on first SERP, + number of documents/ pages visited in the first query segment, + number of queries not leading to saving pages - number of saved/bookmarked document in the first query segment, - total number of saved/bookmarked documents, - number of queries leading to saved/bookmarked pages
Domain knowledge	+ average query length, + number of documents saved/bookmarked, + number of documents viewed, + average number of actions per task, + number of SERPs accessed, + average ranking position of the documents opened in SERPs
Other dimensions	*Self-rated interest in the task topic(s)* + number of scrolls on SERP, + query dwell time, + focused attention, + felt involvement, + novelty, + total engagement, + total task completion time – window heart rate

Note: +: positive correlation. -: negative correlation. Non-significant correlations are not included here.

almost all aspects. For instance, these complex tasks usually lead to more query reformulations, more diverse query terms, more result clicks, lower ranks of clicking, more page visits, more query abandonments, as well as longer dwell time on content pages. Users engaging in these tasks often face difficulties in identifying and evaluating useful documents (increasing number of clicks and queries without a bookmark). In particular, when task topic is unfamiliar, a user tends to spend less time on SERPs (especially the first SERP in a session). Meanwhile, users are less likely to make quick decisions, scan search results or only pick top-ranked results. These active search interactions in complex, unfamiliar and difficult search tasks with ill-defined goals are usually correlated with higher cognitive loads, which can be partially revealed by eye movement metrics (e.g., Cole et al., 2014, 2015), especially the relatively high cognitive loads in frequent query formulation and modification attempts (Gwizdka, 2010). In contrast, when search tasks are simple, straightforward and have well-defined goals, we are likely to observe significantly lower number of queries and visited pages, lower query diversity, less dwell time on content pages, and longer, more specific queries. It is worth noting that when users are familiar with the domain(s), they tend to issue longer, more specific queries, bookmark more useful documents, open more SERPs, and click lower ranked search results. This may be because a higher level of domain knowledge makes it easier for user to evaluate the relevance and usefulness of results and thereby encourages them to identify and bookmark more documents. The findings discussed above jointly illustrate the ways in which the complexity and difficulty of tasks and users' unfamiliarity with task topics are manifested through search interactions at multiple levels and thus could inform the design and application of task prediction models.

Additionally, we also found significant correlations between task interests with several groups of user experience measures.

Table 6
Search task prediction.

Task dimension	**Ground truth label**	**Best feature combination**	**Major predictors**
Task product (Mitsui et al., 2018)	binary: intellectual versus factual; annotated by researchers or external assessors	online behavioral features from the first query segment (query, click and page visits, dwell time features)	
Task goal (Mitsui et al., 2018)	binary: specific versus amorphous; annotated by researchers or external assessors	online behavioral features from the first query segment (query, click and page visits, dwell time features)	
Task type (product + goal) (Cole et al., 2014; Mitsui et al., 2018)	four type: combinations of product and goal	online behavioral features from the first query segment (query, click and page visits, dwell time features); eye movement features	
Subjective task difficulty (Arguello, 2014; Liu, Gwizdka et al., 2010; Liu, Liu, Cole, Belkin & Zhang, 2012;Liu et al., 2014)	binary: easy or difficult; split based on mean or kurtosis of Likert scale values (self-assessed by participants)	all online behavioral features combined (query, click, bookmark, mouse movement, scrolling, dwell time features)	dwell time features (e.g., total dwell time on lading pages, average time spent on a landing page), bookmarking features, and mouse movement features
Topic knowledge (Liu et al., 2016)	binary: split based by the mid-point of Likert scales: 1-3: low knowledge, 4-7 high knowledge	all online behavioral features from the first query segment (querying, dwell time on documents and SERPs)	first dwell time on the first SERP, first dwell time on first viewed document, first query length, number of SERP paginations in the first query segment
Domain knowledge (Zhang et al., 2015)	continuous measure based on users self-assessed familiarity with every search task topic	final model including number of documents saved, the average query length, the average ranking position of the documents opened	number of documents saved, the average query length, the average ranking position of the documents opened
Task state (Liu et al., 2020)	clusters (n>2) extracted based on users' annotations of search intentions and encountered search problems, and validated by external assessments	all online behavioral features combined (query, click and page visits, dwell time features, bookmarking) from both current query segment and previous sessions	

Specifically, some studies reported that users' self-rated level of interest in task topic were significantly associated with window heart rate and several user engagement dimensions (e.g., focused attention, felt involvement, novelty, total engagement), each of which was measured by a series of specific questions or items on Likert scales. These findings enhance our understanding of user characteristics as well as the broader contexts of search. However, they offer limited implications for task prediction and Web-scale search personalization due to the scalability issue and subjective biases in self assessments.

4.4. RQ3: search task prediction

To answer the RQ3, we summarized findings from the reviewed IIR studies where researchers build prediction models to infer and predict the nature of one or multiple task dimensions (see Table 6). *Ground truth label* defines the way in which the target factors (i.e. task dimension) were operationalized and thereby determines the type of prediction model (classification or regression). For each task dimension, we also reported the *best feature combinations* that produced best performing prediction models. We also reported the *major predictors* identified through feature ablation analyses and stepwise estimations in some of the studies. Note that a majority of IIR studies have focused on the correlations between various predictors and task dimensions and drew conclusions mainly from statistical testing (e.g., Capra et al., 2017; Jiang et al., 2014; Liu, Cole et al., 2010). The contributions of these studies were summarized in Table 4 and Table 5. However, statistically significant correlations between a task facet and a set of search behavioral measures do not necessarily mean that this task facet can be accurately predicted from the same set of features (Liu et al., 2020).

As it is shown in Table 6, in most cases, researchers turned task prediction problem into ML classification problem based on either researchers' annotations of facet values (e.g., task product, task goal) or participants' post-search self-ratings. For TUDs (e.g., topic knowledge, domain knowledge, subjective task difficulty), all-feature models usually produce the best performances in prediction/classification. This may be because users' perceptions and understanding of task characteristics are closely associated with their search behaviors. When a user encounters a difficult task or unfamiliar topic, he or she may adjust their search tactics as a response. Thus, incorporating online behavioral features in prediction models can help capture the behavior-task connection and thereby improve the prediction performance. Moreover, we found that different TUDs are associated with different set of major predictors. Specifically, dwell time features are usually effective in predicting task difficulty and topic knowledge, whereas the major predictors in predicting level of domain knowledge are query length and document clicking measures.

For the OTDs reported in Table 6, previous research emphasized the possibility of making reliable early predictions. Specifically, Mitsui et al. (2018) ran experiments based on two separate datasets and demonstrated that first query measures can be at least as good as, and sometimes even better than, whole-session measures in predicting task product, goal and type. Although this conclusion may require additional experiments and broader validations (based on other OTDs, user populations, and datasets), it indicates that a user's initial search interactions of a search session could offer valuable insights about the related search task and may serve as the basis for task-based, proactive search recommendations.

5. Discussion

Increasingly, users are interacting with search systems with the goal of addressing complex, ill-defined search tasks. In the work reported in this paper, we systematically reviewed seventy-six recently published task-based IIR studies, aiming to understand 1) how different facets or dimensions of search tasks have been conceptualized and operationalized in existing studies, and 2) how these dimensions are associated with various predicting measures. Our systematic review synthesizes the knowledge of the connections between predictors of varying types and search task dimensions. It also helps clarify the progresses we have made as well as the challenges we are still facing as a community on understanding the nature of complex search tasks in IIR.

5.1. Answers to the three research questions

To answer the RQ1, we explored how different task dimensions have been defined, operationalized, and measured in empirical research. After synthesizing the knowledge learned from different studies, we found that 1) overall, OTDs attracted a majority of the research attention, with task complexity being the most popular task characteristic in IIR research; 2) regarding OTD-based tasks, researchers emphasized the cognitive aspects (e.g., cognitive complexity, knowledge type required, task determinability) and behavioral aspects (e.g., search strategies and tactics, search efforts) of search interactions and designed experimental/simulated tasks based on predefined typologies as well as researchers' own annotations; 3) with respect to TUDs, researchers focused on users' perceptions of search task features (e.g., difficulty, topic and domain knowledge) and classify different tasks based upon users' self-reported measures; 4) based on the existing faceted typologies (e.g., Li & Belkin, 2008; Urgo et al., 2019), there are still many task dimensions that have been conceptually explored but scarcely studied in empirical research, such as task source (self-motivated or externally assigned), task urgency, learning stage during search, and the knowledge of task procedure.

As the response to RQ2, we summarized the statistically significant predictor-task correlations reported in the collected papers and synthesized the findings for every dimension type. We found that when users are engaging in complex and difficult search tasks with ill-defined goals and unfamiliar topics, they tend to be more active in exploring different search paths and information sources (e.g., more diverse queries and page visits) and paying more attention to search result evaluation (e.g., longer document reading time). Meanwhile, they are less likely to do quick judgments on SERPs or merely look at the top-ranked search results. The behavioral differences between easy and difficult tasks echo the variations at cognitive level (e.g., cognitive loads), which are revealed by eye movement measures. In addition, researchers have also employed other neuro-physiological features and self-rated measures in order

to capture other aspects of users' search contexts and experiences. These measures have not been widely applied in inferring and predicting search task features due to the scalability issues and raters' subjective biases. In new and emerging modalities of search interactions, such as conversational search (e.g., Qu et al., 2019; Vtyurina & Fourney, 2018) and reality-based information retrieval (e.g. Büschel et al., 2018), we can identify new sets of features (e.g., conversational cues, sentiment scores, number of turns in an information seeking dialog, gestures in interaction, facial expression) and leverage the knowledge of the associations between these features and task characteristics in personalizing novel forms of information seeking and retrieval.

With respect to the RQ3, we found that 1) some task dimensions that have been discussed in conceptual and descriptive analyses, such as task complexity, task structure and task interest, are not covered in existing task prediction studies; 2) different groups of task dimensions (i.e., OTDs and TUDs) correspond to different types of best performing feature sets and major predictors. Specifically, our review indicates that it is possible to predict some of the OTDs in initial stages of search, whereas perception-based TUDs usually require whole-session, full-feature models for achieving best prediction performances. To better support users engaging in complex search tasks, it is necessary to 1) implement, test and evaluate more fine-grained features extracted from search interactions and the associated contexts (e.g., time of the day, locations visited, parallel work tasks), 2) include more task dimensions for building a comprehensive task model, and 3) empirically evaluate the value of knowing certain task characteristics (e.g., task topic, types of product, task difficulty) in making recommendations and satisficing users' needs. In other words, to determine the priorities in characterizing and parameterizing different task dimensions, we need to compute and compare the costs of not knowing or not being able to predict certain task dimensions from different perspectives (e.g., decreases in a user's in-situ and session-wise search satisfaction, search productivity, knowledge learning, search fairness and transparency).

5.2. Implications and future directions for task-based IIR research

Our systematic review synthesized the knowledge learned from a variety of empirical studies and answers the proposed three RQs. Meanwhile, it also uncovered a series of open questions and offers implications for multiple aspects of future task-based IIR research.

5.2.1. Task dimensions and search scenarios

The existing task typologies (both one-dimensional and faceted) have contributed to the growth of knowledge in this area through 1) producing unit theories and hypotheses for empirical testing and evaluation, 2) presenting integrated theoretical research programs (e.g., faceted approach to characterizing search tasks) and describing the relations between different concepts, hypotheses, as well as unit theories. Under these typologies and frameworks, IIR research community have accumulated more empirical evidences to support and revise their theories, which can depict the research objects in a more accurate manner.

With respect to the limitations, there is a gap between conceptualizing search tasks and applying task knowledge in facilitating search interactions. Despite of the research efforts on characterizing search tasks, the value of understanding and predicting task dimensions is not quantitatively clear on the application side. This ambiguity of practical value also hinders the evaluation of the goodness of different task typologies. Addressing this meta-evaluation issue (which involves both the conceptual framework and the associated categories or metrics) would require reliable ground truth measure(s) and generalizable approach for scalable experiments. Moreover, even within the same typology, different task dimensions may need to be assigned with different weights in personalization algorithms as they are very unlikely to be equally important for supporting users. For instance, according to Mitsui et al. (2018), OTDs may play a major role in initial stages as the task type prediction model performs better at the end of first query segment, compared to that of the whole-session completion point. However, as the search process proceeds, TUDs may generate larger effects on a user's needs for system supports. Specifically, Liu et al. (2019) found that the impact of topic familiarity and perceived task difficulty on certain local information needs or intentions (e.g. evaluating usefulness, finding items with common features) tends to increase over time during a search session.

Beyond search task dimensions, users' interactions with information are also affected by factors from broader search scenarios, such as work task traits, devices and tools, accessible supports from the workplace, time of the day, physical locations, as well as other parallel work tasks (e.g., driving, cooking). These factors jointly shape the environment in which users' search for, evaluate and use information. In existing information seeking and IR studies, researchers have 1) described the connections between work task features and search task facets (Li, 2009) and 2) incorporated work task introduction into search task descriptions as "cover stories", with the purpose of simulating realistic information acquiring situations (Borlund & Schneider, 2010). However, these attempts alone are not enough for fully understanding the role and impact of task characteristics within a broader picture. To integrate task features with search scenarios, White and Awadallah (2019) employed features extracted from calendar appointment data to estimate work task duration. Similarly, Aliannejadi et al. (2019) went beyond the traditional features of search interactions and examined the overall task relevance to the evolving information search scenarios and problematic situations. These approaches can be applied to expand existing task frameworks and characterize search scenarios that contextualize users' tasks and information interactions. Note that when investigating the impacts of search scenarios and incorporating more features into models, researchers also have to deal with the challenge of balancing the controllability of the variables involved and the authenticity of the simulated scenarios.

5.2.2. Statistical analyses

According to our systematic review, a large proportion of the conclusions in task-based IR studies were drew from the results of various statistical testing and modeling. In IR research community, some researchers have reviewed a large body of IR literature, aiming to evaluate the reliability and validity of the statistical results and the associated conclusions. For instance, Sakai (2016) identified extremely overpowered and underpowered IR experiments after reviewing more than 1000 IR full papers and reported the

appropriate sample sizes for these experiments. Similarly, Fuhr (2018) also reviewed a series of common statistical issues in IR evaluation, such as ignoring multiple comparisons problem in significance tests, not reporting effect sizes, and the reproducibility issue of the experiments. In addition, the violations of model assumptions (e.g., normality, homogeneity of variances) may also affect the validity of statistical results. Although the issues discussed above are mostly identified in system-oriented IR evaluations, it is possible that the statistical analysis on task dimensions and predictors are also subject to these issues. Moreover, compared to well-controlled evaluation experiments, IR user studies often involve more implicit, uncontrolled factors associated with both dependent and independent variables, which may lead to endogeneity problem. Considering and addressing these critical issues can help researchers examine predictor-task correlations in a more accurate manner.

In addition, given the small number of task prediction research identified in this review, our understanding of task effects may be affected by the potential *survivorship bias* in publication: Some IIR studies that obtain non-significant, conflicting, or unexpected results may end up being rejected by researchers themselves (not reporting them) or publication venues. In our systematic review, we identified a variety of unique study design and distinctive experimental setups in different user studies. This lack of standard requirements or settings for session-based IR study design, evaluation, and reporting also exacerbates the problem associated with survivorship bias (Liu & Shah, 2019b). Consequently, we may consider some empirically supported hypotheses reported in a small amount of published papers as universally applicable findings, without being aware of the possible restrictions revealed in unpublished experiments. To mitigate the potential negative effect of this survivorship bias, researchers need to pay more attention to the reproducibility of their works and perhaps start a new platform (e.g., workshops, paper track) dedicated to communicating conflicting results and replicating existing analyses on task-based IR (e.g., task-predictor correlations, ML models for predicting task dimensions). Recent efforts on improving reproducibility and replicability of IR experiments (e.g., Ferro & Kelly, 2018; Wilson et al., 2014) could be a good point of departure for this line of research and practice.

5.2.3. Integrated prediction models

In addition to task characteristics, researchers have also investigated other features of search interactions, aiming to leverage the learned knowledge in improving the quality of ranking and search recommendations. For instance, Feild et al., and Jones (2010) built models based on query log features and signals from physical sensors to predict the current state of user frustration. Diaz et al., and White (2016) investigated users' search result examination behavior and sought to dynamically estimate the result a searcher will request next based on their prior cursor movements. Mitsui et al., and Shah (2017) demonstrated that users' information seeking intentions can be predicted with a classification model built with search behavioral features. Liu and Shah (2019c) took a step forward and integrated users' behavioral features with the knowledge of their intentions to proactively identify their potential failures in search. Apart from extracting and developing additional features from new modalities of search, combing multiple prediction models (e.g., task dimensions, information seeking intentions, in-situ search satisfaction and frustration) together may also enable IR systems to better capture users' needs in real time and thereby provide dynamic, proactive informational supports.

5.3. Limitations of the systematic review

This review was conducted with a narrowly defined set of papers and it is important to be mindful of this when making arguments and drawing conclusions. Since the exact amount of papers relevant to our research topic is unknown, it is difficult to assess the extent to which our sample cover the whole population. Nevertheless, in the last few rounds of paper retrieval, we did notice several indications of a good coverage: 1) citation chaining gradually led to repeated papers as found in earlier citation tracing and retrieval rounds; 2) using different keyword combinations in different search engines and databases gradually produced repeated papers as found in previous searches; 3) author/scholar searches kept producing same relevant papers found in previous rounds of paper searches. These results suggest that our sample at least included the most influential research papers on task-based IIR. Understanding the features and limitations of these studies can help researchers make more informed decisions in the future.

Another limitation of this study is the potential reliability of paper coding. While most features and factors such as paper publication year, paper venue, and authors were straightforward, some other features were not always obvious. For instance, in methodology sections, different researchers often use different labels and unique measures for similar concepts, which made it difficult to identify appropriate categories for grouping similar concepts and features together. Also, in the result section of some papers, it was difficult to determine what measures were significantly associated with the identified task dimensions and which features were actually included in prediction models. This suggests that our analysis of these measures might not be as replicable as the analysis of most quantitative measures. In addition, similar to Kelly and Sugimoto (2013), our analysis was also restricted by researchers' study designs and reporting practices. For example, although many IIR studies claimed that their findings can be of help for predicting task nature, most of them did not directly involve the process of feature engineering or ML prediction. Also, some papers did not fully report the tools and techniques employed in their studies, which made it harder to interpret and evaluate the reported findings.

6. Conclusion

This review synthesized the progresses we have made in understanding and predicting search tasks in twenty years of IIR studies and discussed the challenges we are facing at multiple levels. We explained how different task dimensions have been conceptualized, measured and predicted in empirical research, and highlighted the research gaps among three separate problems, namely characterizing search tasks, predicting task dimensions, and supporting task-based personalization. Addressing these gaps will result in great benefits for the research community.

In contrast to standard tests and Cranfield experiments, every IIR study often turns out to be unique in one or multiple aspects, such as conceptualization, study environment, tools and devices, as well as measurements and analysis. The variations and inconsistencies among specific empirical studies call for a systematic review that captures the common themes among them and summarizes the collective knowledge learned from seemingly disconnected individual user studies, experiments, interviews, and field observations. The accumulated knowledge and insights from these studies can help achieve the "maturation of the IIR specialty" (p.768, Kelly & Sugimoto, 2013) and push the IIR field to further specialize. Our study synthesizes the knowledge of search task conceptualization, characterization, and prediction and provides a guidance for further explorations and the accumulation of new evidences.

Now with the task dimensions, measures, and study limitations being documented and discussed, the research community is in a better position to determine how it wants to move forward in terms of task conceptualization (e.g., identifying new dimensions and factors), feature extraction, task prediction and task-aware interactive system design. Also, we will see more and more complex information acquiring and decision-making scenarios inviting new modalities of search interactions, such as conversational search and reality-based search. The knowledge about tasks and predictors synthesized in our review can be used as an intellectual benchmark for future studies on new forms of user-task interactions.

Author contribution

Jiqun Liu is responsible for all relevant roles in this article, including: conceptualization, methodology, data curation, investigation, writing, reviewing and editing.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ipm.2021.102522.

References

Agichtein, E., White, R. W., Dumais, S. T., & Bennet, P. N. (2012). Search, interrupted: Understanding and predicting search task continuation. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 315–324). https://doi.org/10.1145/2348283.2348328.

Ahn, J. W., Brusilovsky, P., He, D., Grady, J., & Li, Q. (2008). Personalized web exploration with task models. In *Proceedings of the 17th International Conference on World Wide Web* (pp. 1–10). https://doi.org/10.1145/1367497.1367499.

Aliannejadi, M., Harvey, M., Costa, L., Pointon, M., & Crestani, F. (2019). Understanding mobile search task relevance and user behaviour in context. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval* (pp. 143–151). https://doi.org/10.1145/3295750.3298923.

Arguello, J. (2014). Predicting search task difficulty. In *European Conference on Information Retrieval* (pp. 88–99). Cham: Springer. https://doi.org/10.1007/978-3-319-06028-6_8.

Arguello, J., Wu, W. C., Kelly, D., & Edwards, A. (2012). Task complexity, vertical display and user interaction in aggregated search. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 435–444). https://doi.org/10.1145/2348283.2348343.

Aula, A., Khan, R. M., & Guan, Z. (2010). How does search behavior change as search becomes more difficult?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 35–44). https://doi.org/10.1145/1753326.1753333.

Awadallah, A. H., Gurrin, C., Sanderson, M., & White, R. W. (2019). Task Intelligence Workshop @ WSDM 2019. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (pp. 848–849). https://doi.org/10.1145/3289600.3291374.

Belkin, N. J. (2015). Salton award lecture: People, interacting with information. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1–2). https://doi.org/10.1145/2766462.2767854.

Bin, G. (2009). Moderating effects of task characteristics on information source use: An individual-level analysis of R&D professionals in new product development. *Journal of Information Science, 35*(5), 527–547. https://doi.org/10.1177/0165551509105196.

Borlund, P., & Schneider, J. W. (2010). Reconsideration of the simulated work task situation: A context instrument for evaluation of information retrieval interaction. In *Proceedings of the Third Symposium on Information Interaction in Context* (pp. 155–164). https://doi.org/10.1145/1840784.1840808.

Brennan, K., Kelly, D., & Arguello, J. (2014). The effect of cognitive abilities on information search for tasks of varying levels of complexity. In *Proceedings of the 5th Information Interaction in Context Symposium* (pp. 165–174). https://doi.org/10.1145/2637002.2637022.

Broder, A. (2002). A taxonomy of web search. In *, 36. ACM SIGIR Forum* (pp. 3–10). ACM. https://doi.org/10.1145/792550.792552.

Büschel, W., Mitschick, A., & Dachselt, R. (2018). Here and now: Reality-based information retrieval: Perspective paper. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (pp. 171–180). https://doi.org/10.1145/3176349.3176384.

Buscher, G., White, R. W., Dumais, S., & Huang, J. (2012). Large-scale analysis of individual and task differences in search result page examination strategies. In *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 373–382). https://doi.org/10.1145/2124295.2124341.

Byström, K., & Hansen, P. (2005). Conceptual framework for tasks in information studies. *Journal of the American Society for Information science and Technology, 56*(10), 1050–1061. https://doi.org/10.1002/asi.20197.

Byström, K., & Kumpulainen, S. (2020). Vertical and horizontal relationships amongst task-based information needs. *Information Processing & Management, 57*(2), Article 102065. https://doi.org/10.1016/j.ipm.2019.102065.

Capra, R., Arguello, J., O'Brien, H., Li, Y., & Choi, B. (2018). The effects of manipulating task determinability on search behaviors and outcomes. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 445–454). https://doi.org/10.1145/3209978.3210047.

Capra, R., Arguello, J., & Zhang, Y. (2017). The effects of search task determinability on search behavior. In *European Conference on Information Retrieval* (pp. 108–121). Cham: Springer. https://doi.org/10.1007/978-3-319-56608-5_9.

Choi, B., Ward, A., Li, Y., Arguello, J., & Capra, R. (2019). The effects of task complexity on the use of different types of information in a search assistance tool. *ACM Transactions on Information Systems (TOIS), 38*(1), 1–28. https://doi.org/10.1145/3371707.

Cole, M. J., Gwizdka, J., Liu, C., Bierig, R., Belkin, N. J., & Zhang, X. (2011). Task and user effects on reading patterns in information search. *Interacting with Computers, 23*(4), 346–362. https://doi.org/10.1016/j.intcom.2011.04.007.

Cole, M. J., Hendahewa, C., Belkin, N. J., & Shah, C. (2014). Discrimination between tasks with user activity patterns during information search. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 567–576). https://doi.org/10.1145/2600428.2609591.

Cole, M. J., Hendahewa, C., Belkin, N. J., & Shah, C. (2015). User activity patterns during information search. *ACM Transactions on Information Systems (TOIS), 33*(1), 1–39. https://doi.org/10.1145/2699656.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2019). *The handbook of research synthesis and meta-analysis*. New York, NY: Russell Sage Foundation. https://doi.org/10.7758/9781610448864.
Dedema, M., & Liu, C. (2019). Examination of online information search stopping behaviors and stopping rules by task type. *Proceedings of the Association for Information Science and Technology, 56*(1), 631–633. https://doi.org/10.1002/pra2.114.
Diaz, F., Guo, Q., & White, R. W. (2016). Search result prefetching using cursor movement. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 609–618). https://doi.org/10.1145/2911451.2911516.
Edwards, A., & Kelly, D. (2016). How does interest in a work task impact search behavior and engagement?. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval* (pp. 249–252). https://doi.org/10.1145/2854946.2855000.
Edwards, A., & Kelly, D. (2017). Engaged or frustrated? Disambiguating emotional state in search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 125–134). https://doi.org/10.1145/3077136.3080818.
Feild, H. A., Allan, J., & Jones, R. (2010). Predicting searcher frustration. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 34–41). https://doi.org/10.1145/1835449.1835458.
Ferro, N., & Kelly, D. (2018). SIGIR initiative to implement ACM artifact review and badging. In *, 52. ACM SIGIR Forum* (pp. 4–10). ACM. https://doi.org/10.1145/3274784.3274786.
Fuhr, N. (2018). Some common mistakes in IR evaluation, and how they can be avoided. In *, 51. ACM SIGIR Forum* (pp. 32–41). ACM. https://doi.org/10.1145/3190580.3190586.
Ghosh, S., Rath, M., & Shah, C. (2018). Searching as learning: Exploring search behavior and learning outcomes in learning-related tasks. In *Proceedings of the 2018 ACM SIGIR Conference on Human Information Interaction and Retrieval* (pp. 22–31). https://doi.org/10.1145/3176349.3176386.
Gwizdka, J. (2010). Distribution of cognitive load in web search. *Journal of the American Society for Information Science and Technology, 61*(11), 2167–2187. https://doi.org/10.1002/asi.21385.
Gwizdka, J., & Spence, I. (2006). What can searching behavior tell us about the difficulty of information tasks? A study of Web navigation. *Proceedings of the American Society for Information Science and Technology, 43*(1), 1–22. https://doi.org/10.1002/meet.14504301167.
He, J., Bron, M., & de Vries, A. P. (2013). Characterizing stages of a multi-session complex search task through direct and indirect query modifications. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 897–900). https://doi.org/10.1145/2484028.2484178.
He, J., & Yilmaz, E. (2017). User behaviour and task characteristics: A field study of daily information behaviour. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (pp. 67–76). https://doi.org/10.1145/3020165.3020188.
Hienert, D., Mitsui, M., Mayr, P., Shah, C., & Belkin, N. J. (2018). The role of the task topic in web search of different task types. In *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval* (pp. 72–81). https://doi.org/10.1145/3176349.3176382.
Hong, T. (2006). The Internet and tobacco cessation: The roles of Internet self-efficacy and search task on the information-seeking process. *Journal of Computer-Mediated Communication, 11*(2), 536–556. https://doi.org/10.1111/j.1083-6101.2006.00026.x.
Hu, R., Lu, K., & Joo, S. (2013). Effects of topic familiarity and search skills on query reformulation behavior. *Proceedings of the American Society for Information Science and Technology, 50*(1), 1–9. https://doi.org/10.1002/meet.14505001062.
Huang, J., White, R. W., & Dumais, S. (2011). No clicks, no problem: Using cursor movements to understand and improve search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1225–1234). https://doi.org/10.1145/1978942.1979125.
Jiang, J., He, D., & Allan, J. (2014). Searching, browsing, and clicking in a search session: Changes in user behavior by task and over time. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 607–616). https://doi.org/10.1145/2600428.2609633.
Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval, 3*(1-2), 1–224. https://doi.org/10.1561/1500000012.
Kelly, D., Arguello, J., & Capra, R. (2013). NSF workshop on task-based information search systems. In *, 47. ACM SIGIR Forum* (pp. 116–127). ACM. https://dl.acm.org/doi/book/10.5555/3079761.
Kelly, D., Arguello, J., Edwards, A., & Wu, W. C. (2015). Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval* (pp. 101–110). https://doi.org/10.1145/2808194.2809465.
Kelly, D., & Cool, C. (2002). The effects of topic familiarity on information search behavior. In *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 74–75). https://doi.org/10.1145/544220.544232.
Kelly, D., & Sugimoto, C. R. (2013). A systematic review of interactive information retrieval evaluation studies, 1967–2006. *Journal of the American Society for Information Science and Technology, 64*(4), 745–770. https://doi.org/10.1002/asi.22799.
Kessler, K., Freund, L., & Kopak, R. (2014). Does the perceived usefulness of search facets vary by task type?. In *Proceedings of the 5th Information Interaction in Context Symposium* (pp. 267–270). https://doi.org/10.1145/2637002.2637039.
Kim, K. S. (2001). Information-seeking on the Web: Effects of user and task variables. *Library & Information Science Research, 23*(3), 233–255. https://doi.org/10.1016/S0740-8188(01)00081-0.
Koolen, M., Kamps, J., Bogers, T., Belkin, N., Kelly, D., & Yilmaz, E. (2017). Report on the second workshop on supporting complex search tasks. In *, 51. ACM SIGIR Forum* (pp. 58–66). ACM. https://doi.org/10.1145/3020165.3022163.
Kulahcioglu, T., Fradkin, D., & Palanivelu, S. (2017). Incorporating task analysis in the design of a tool for a complex and exploratory search task. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (pp. 373–376). https://doi.org/10.1145/3020165.3022156.
Li, Y. (2009). Exploring the relationships between work task and search task in information search. *Journal of the American Society for information Science and Technology, 60*(2), 275–291. https://doi.org/10.1002/asi.20977.
Li, Y., & Belkin, N. J. (2008). A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management, 44*(6), 1822–1837. https://doi.org/10.1016/j.ipm.2008.07.005.
Li, Y., Chen, Y., Liu, J., Cheng, Y., Wang, X., Chen, P., & Wang, Q. (2011). Measuring task complexity in information search from user's perspective. *Proceedings of the American Society for Information Science and Technology, 48*(1), 1–8. https://doi.org/10.1002/meet.2011.14504801092.
Liu, C., Liu, J., & Belkin, N. J. (2014). Predicting search task difficulty at different search stages. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (pp. 569–578). https://doi.org/10.1145/2661829.2661939.
Liu, C., Liu, J., Cole, M., Belkin, N. J., & Zhang, X. (2012). Task difficulty and domain knowledge effects on information search behaviors. *Proceedings of the American Society for Information Science and Technology, 49*(1), 1–10. https://doi.org/10.1002/meet.14504901142.
Liu, J., Liu, C., Cole, M., Belkin, N. J., & Zhang, X. (2012). Exploring and predicting search task difficulty. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (pp. 1313–1322).
Liu, C., Liu, Y. H., Gedeon, T., Zhao, Y., Wei, Y., & Yang, F. (2019). The effects of perceived chronic pressure and time constraint on information search behaviors and experience. *Information Processing & Management, 56*(5), 1667–1679. https://doi.org/10.1016/j.ipm.2019.04.004.
Liu, J., & Belkin, N. J. (2015). Personalizing information retrieval for multi-session tasks: Examining the roles of task stage, task type, and topic knowledge on the interpretation of dwell time as an indicator of document usefulness. *Journal of the Association for Information Science and Technology, 66*(1), 58–81. https://doi.org/10.1002/asi.23160.
Liu, J., Cole, M. J., Liu, C., Bierig, R., Gwizdka, J., Belkin, N. J., Zhang, J., & Zhang, X. (2010). Search behaviors in different task types. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries* (pp. 69–78). https://doi.org/10.1145/1816123.1816134.
Liu, J., Gwizdka, J., Liu, C., & Belkin, N. J. (2010). Predicting task difficulty for different task types. *Proceedings of the American Society for Information Science and Technology, 47*(1), 1–10. https://doi.org/10.1002/meet.14504701173.
Liu, J., Liu, C., & Belkin, N. J. (2016). Predicting information searchers' topic knowledge at different search stages. *Journal of the Association for Information Science and Technology, 67*(11), 2652–2666. https://doi.org/10.1002/asi.23606.
Liu, J., Liu, C., & Belkin, N. J. (2019). Personalization in text information retrieval: A survey. *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.24234.

Liu, J., Liu, C., & Belkin, N. J. (2020). Personalization in text information retrieval: A survey. *Journal of the Association for Information Science and Technology, 71*(3), 349–369. https://doi.org/10.1002/asi.24234.
Liu, J., Sarkar, S., & Shah, C. (2020). Identifying and predicting the states of complex search tasks. In *Proceedings of the 2020 International ACM SIGIR Conference on Human Information Interaction and Retrieval* (pp. 193–202). https://doi.org/10.1145/3343413.3377976.
Liu, J., & Shah, C. (2019a). Investigating the impacts of expectation disconfirmation on web search. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval* (pp. 319–323). https://doi.org/10.1145/3295750.3298959.
Liu, J., & Shah, C. (2019b). Interactive IR user study design, evaluation, and reporting. *Synthesis Lectures on Information Concepts, Retrieval, and Services, 11*(2). https://doi.org/10.2200/S00923ED1V01Y201905ICR067. i-93.
Liu, J., & Shah, C. (2019c). Proactive identification of query failure. *Proceedings of the Association for Information Science and Technology, 56*(1), 176–185. https://doi.org/10.1002/pra2.15.
Lopatovska, I. (2010). Searching for good mood: Examining relationships between search task and mood. *Proceedings of the American Society for Information Science and Technology, 46*(1), 1–13. https://doi.org/10.1002/meet.2009.1450460222.
Mehrotra, R., & Yilmaz, E. (2017). Extracting hierarchies of search tasks & subtasks via a Bayesian nonparametric approach. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 285–294). https://doi.org/10.1145/3077136.3080823.
Mitsui, M., Liu, J., Belkin, N. J., & Shah, C. (2017). Predicting information seeking intentions from search behaviors. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1121–1124). https://doi.org/10.1145/3077136.3080737.
Mitsui, M., Liu, J., & Shah, C. (2018). How much is too much? Whole session vs. first query behaviors in task type prediction. In *the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 1141–1144). https://doi.org/10.1145/3209978.3210105.
Moshfeghi, Y., Rothfeld, R., Azzopardi, L., & Triantafillou, P. (2017). A task completion engine to enhance search session support for air traffic work tasks. In *Proceedings of European Conference on Information Retrieval* (pp. 278–290). Cham: Springer. https://doi.org/10.1007/978-3-319-56608-5_22.
Moshfeghi, Y., Triantafillou, P., & Pollick, F. E. (2016). Understanding information need: An fMRI study. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 335–344). https://doi.org/10.1145/2911451.2911534.
Mostafa, J., & Gwizdka, J. (2016). Deepening the role of the user: Neuro-physiological evidence as a basis for studying and improving search. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval* (pp. 63–70). https://doi.org/10.1145/2854946.2854979.
O'Brien, H. L., Arguello, J., & Capra, R. (2020). An empirical study of interest, task complexity, and search behaviour on user engagement. *Information Processing & Management, 57*(3), Article 102226. https://doi.org/10.1016/j.ipm.2020.102226.
O'Brien, H. L., & Toms, E. G. (2008). What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American society for Information Science and Technology, 59*(6), 938–955. https://doi.org/10.1002/asi.20801.
Pharo, N., & Krahn, A. (2011). The effect of task type on preferred element types in an XML-based retrieval system. *Journal of the American Society for Information Science and Technology, 62*(9), 1717–1726. https://doi.org/10.1002/asi.21587.
Qu, C., Yang, L., Croft, W. B., Zhang, Y., Trippas, J. R., & Qiu, M. (2019). User intent prediction in information-seeking conversations. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval* (pp. 25–33). https://doi.org/10.1145/3295750.3298924.
Ruthven, I. (2008). Interactive information retrieval. *Annual Review of Information Science and Technology, 42*(1), 43–91. https://doi.org/10.1002/aris.2008.1440420109.
Sakai, T. (2016). Statistical significance, power, and sample sizes: A systematic review of SIGIR and TOIS, 2006-2015. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 5–14). https://doi.org/10.1145/2911451.2911492.
Sanchiz, M., Chin, J., Chevalier, A., Fu, W. T., Amadieu, F., & He, J. (2017). Searching for information on the web: Impact of cognitive aging, prior domain knowledge and complexity of the search problems. *Information Processing & Management, 53*(1), 281–294. https://doi.org/10.1016/j.ipm.2016.09.003.
Şendurur, E., & Yildirim, Z. (2015). Students' web search strategies with different task types: An eye-tracking study. *International Journal of Human-Computer Interaction, 31*(2), 101–111. https://doi.org/10.1080/10447318.2014.959105.
Song, Y., & Guo, Q. (2016). Query-less: Predicting task repetition for nextGen proactive search and recommendation engines. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 543–553). https://doi.org/10.1145/2872427.2883020.
Thatcher, A. (2006). Information-seeking behaviours and cognitive search strategies in different search tasks on the WWW. *International Journal of Industrial Ergonomics, 36*(12), 1055–1068. https://doi.org/10.1016/j.ergon.2006.09.012.
Toms, E. G., O'Brien, H., Mackenzie, T., Jordan, C., Freund, L., Toze, S., Dawe, E., & Macnutt, A. (2007). Task effects on interactive search: The query factor. In *International Workshop of the Initiative for the Evaluation of XML Retrieval* (pp. 359–372). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-85902-4_31.
Urgo, K., Arguello, J., & Capra, R. (2019). Anderson and Krathwohl's two-dimensional taxonomy applied to task creation and learning assessment. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval* (pp. 117–124). https://doi.org/10.1145/3341981.3344226.
Vakkari, P. (2020). The usefulness of search results: A systematization of types and predictors. In *Proceedings of the 2020 ACM SIGIR Conference on Human Information Interaction and Retrieval* (pp. 243–252). https://doi.org/10.1145/3343413.3377955.
Vtyurina, A., & Fourney, A. (2018). Exploring the role of conversational cues in guided task support with virtual assistants. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–7). https://doi.org/10.1145/3173574.3173782.
Walhout, J., Oomen, P., Jarodzka, H., & Brand-Gruwel, S. (2017). Effects of task complexity on online search behavior of adolescents. *Journal of the Association for Information Science and Technology, 68*(6), 1449–1461. https://doi.org/10.1002/asi.23782.
White, R. W. (2018). Skill discovery in virtual assistants. *Communications of the ACM, 61*(11), 106-113.10.1145/3185336.
White, R. W., & Awadallah, A. H. (2019). Task duration estimation. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (pp. 636–644). https://doi.org/10.1145/3289600.3290997.
Wildemuth, B. M. (2004). The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology, 55*(3), 246–258. https://doi.org/10.1002/asi.10367.
Wildemuth, B., Freund, L., & Toms, E. G. (2014). Untangling search task complexity and difficulty in the context of interactive information retrieval studies. *Journal of Documentation, 70*(6), 1118–1140. https://doi.org/10.1108/JD-03-2014-0056.
Wilson, M. L., Chi, E. H., Reeves, S., & Coyle, D. (2014). RepliCHI: The workshop II. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems* (pp. 33–36). https://doi.org/10.1145/2559206.2559233.
Xie, I., & Joo, S. (2012). Factors affecting the selection of search tactics: Tasks, knowledge, process, and systems. *Information Processing & Management, 48*(2), 254–270. https://doi.org/10.1016/j.ipm.2011.08.009.
Yoo, C. Y. (2009). The effects of persuasion knowledge on click-through of keyword search ads: Moderating role of search task and perceived fairness. *Journalism & Mass Communication Quarterly, 86*(2), 401–418. https://doi.org/10.1177/107769900908600209.
Yuan, X., & Belkin, N. J. (2010). Investigating information retrieval support techniques for different information-seeking strategies. *Journal of the American Society for Information Science and Technology, 61*(8), 1543–1563. https://doi.org/10.1002/asi.21314.
Zhang, Y., Chen, X., Ai, Q., Yang, L., & Croft, W. B. (2018). Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (pp. 177–186). https://doi.org/10.1145/3269206.3271776.
Zhang, X., Liu, J., Cole, M., & Belkin, N. (2015). Predicting users' domain knowledge in information retrieval using multiple regression analysis of search behaviors. *Journal of the Association for Information Science and Technology, 66*(5), 980–1000. https://doi.org/10.1002/asi.23218.
Zhang, Y., Sun, Y., & Kim, Y. (2017). The influence of individual differences on consumer's selection of online sources for health information. *Computers in Human Behavior, 67*, 303–312. https://doi.org/10.1016/j.chb.2016.11.008.