Taking Search to Task

Chirag Shah
University of Washington
Seattle, USA
chirags@uw.edu

Ryen W. White
Microsoft Research
Redmond, USA
ryenw@microsoft.com

Paul Thomas
Microsoft
Canberra, Australia
pathom@microsoft.com

Bhaskar Mitra
Microsoft Research
Montréal, Canada
bmitra@microsoft.com

Shawon Sarkar
University of Washington
Seattle, USA
ss288@uw.edu

Nicholas Belkin
Rutgers University
New Brunswick, USA
belkin@comminfo.rutgers.edu

ABSTRACT

The importance of tasks in information retrieval (IR) has been long argued for, addressed in different ways, often ignored, and frequently revisited. For decades, scholars made a case for the role that a user's task plays in how and why that user engages in search and what a search system should do to assist. But for the most part, the IR community has been too focused on query processing and assuming a search task to be a collection of user queries, often ignoring if or how such an assumption addresses the users accomplishing their tasks. With emerging areas of conversational agents and proactive IR, understanding and addressing users' tasks has become more important than ever before. In this paper, we provide various perspectives on where the state-of-the-art is with regard to tasks in IR, what are some of the bottlenecks in deriving and using task information, and how do we go forward from here. In addition to covering relevant literature, the paper provides a synthesis of historical and current perspectives on understanding, extracting, and addressing task-focused search. To ground ongoing and future research in this area, we present a new framing device for tasks using a tree-like structure and various moves on that structure that allow different interpretations and applications. Presented as a combination of synthesis of ideas and past works, proposals for future research, and our perspectives on technical, social, and ethical considerations, this paper is meant to help revitalize the interest and future work in task-based IR.

CCS CONCEPTS

• **Information systems → Information retrieval**.

KEYWORDS

Tasks; Contextual search; Proactive search

1 INTRODUCTION

Scholars have long argued for the importance of considering task information in information retrieval (IR) for truly helping people with complex, unexpressed, or unclear needs [17, 45]. Over the decades, the concept of task has been studied by many researchers who have produced notable theoretical and practical outcomes. Several attempts have been made to understand search tasks, characterize and extract them, and use task knowledge to better provide support in search and recommendation applications. There are several small and practical successes along the way, including search services incorporating spatial and temporal information in understanding or expanding a query, as well as using the current context and history activity to provide contextual recommendations. However, these efforts can be limiting at best and harmful at worst as they fail to regard user intents or goals as a way to model the ongoing task.

Can we meaningfully connect operationalization of search to conceptualization of task (take search to task)? How do we create a framing device with tasks with the explicit purpose of applying it to various IR applications? What do we gain (and lose) if we are successful with this? These are some of the core questions that triggered our investigations – some theoretical, some empirical, and others simply thought experiments – resulting in this perspective paper. Thus, the purpose of this perspective paper is to shine the light, once again, on this very important area of IR and provide a new foundation built with current understanding and future possibilities that include emerging domains of conversational agents, multi-device search, and proactive recommenders to guide users to complete their tasks step by step.

The remainder of the paper is organized as follows. The next section reviews some of the most important and transformative research on tasks in IR over the last few decades. We also list several recent events and activities to demonstrate the importance of this area and emphasize the scholarly interest. In Section 3, we present a framing device to think through possibilities and challenges for capturing task-related information in IR. Section 4 extends this by providing paths and perspectives as we move forward, specifically focusing on task representation and using such representations in IR applications. Some of such applications that are taking shape now and are important in the future of IR are outlined in Section 5. In Section 6, we briefly discuss methods and metrics for evaluating task-based applications. Finally, we conclude in Section 7 with our

thoughts on task futures, along with a discussion of ethical considerations regarding using tasks in search and other applications.

2 A BRIEF HISTORY OF TASKS IN IR

A task is generally considered as a set of connected physical, cognitive, and affective actions through which individuals try to accomplish some goals in their work or everyday lives [26, 168]. In the context of IR, the concept of task has taken on explicit meanings related to understanding and supporting information seeking and searching. In this section, we give an overview of the ways in which task has been understood in previous IR-related research, beginning with a general survey of different approaches, then considering some specific aspects of task that have been investigated, followed by discussion of some significant attempts to apply knowledge of task in IR, and concluding with a discussion of recent workshops concerning task in IR.

2.1 Overview of Approaches to Task in IR

Some of the earliest prior research in IR related to task can be traced back to the cognitive perspective in IR [18], which was centrally concerned with understanding what motivated a person to engage in information seeking and searching. This perspective influenced works by Vakkari [167] and Ingwersen and Järvelin [77], which consider tasks in the design of IR systems to find out for what purposes the system is used [139] and thus provides implications for IR system design to personalize information search according to the task at hand. Based on a series of empirical works, Vakkari [167] developed a framework of task-based information searching comprising three stages: *pre-focus*, *focus formulation*, and *post-focus*.

Tasks are often considered multi-level information seeking processes in which people need information to achieve a goal [e.g., 29, 139, 143, 166]. Many existing task models [e.g., 34, 88, 99] have investigated and identified searchers' tasks as static and overarching goals that motivate search actions, but this is not always desired as the task evolves with time and changing cognitive states. Conversely, different characteristics or facets of tasks [99] influence people's interaction with intelligent systems, such as search engines [107]. Search tasks are influenced by the work task or everyday life task that drives them to seek information or are associated with a problematic situation [28].

Identification of task, at various levels, has been an area of focus. Broder [23] proposed that a person's intent or goal in engaging with a search engine could be one of three types: *informational*, *transactional*, and *navigational*. This scheme has been successfully used in a great deal of research, to classify tasks according to search behaviors and for study and support of search according to type. Rose and Levinson [137] extended Broder's scheme by specifying types of information goals, and adding a new goal type (*resource*). They tested their scheme of motivating goals (or tasks), by classifying search engine queries.

Many early works investigated and identified various aspects of task which could influence a variety of search behaviors, including task complexity [e.g., 29], task difficulty [e.g., 92] and work context [e.g., 58]. Others considered the interactive and dynamic nature of search tasks themselves [e.g., 13].

Apart from task, existing studies in IR segment information seeking behaviors into various levels of explicit and implicit signals.

While performing tasks, searchers' actions are also driven by intentions and can be well-defined or ill-defined [77]. These studies have indicated that there is a close association between searchers' performance of a task and the information need, the search strategies employed, and the assessment of document relevance and utility.

Beyond search, tasks permeate almost every aspect of our daily work and personal lives [5]. They involve different activities, have different constraints, and take different amounts of time to complete. Users of task management applications would benefit from assistance with many aspects of task management, especially task planning [20] and prioritization [128]. There has been recent progress in task intelligence, in areas such as discovering digital assistant capabilities [176], estimating task durations [181], and automatically tracking task status [182].

It should be noted that not all research concerned with tasks in IR has been explicitly about modeling or using task. Some such examples include research on task trails [101], personalized search [178], trail recommendation [150], cross-session tasks [174], task continuation [1], and cross-device tasks [175].

2.2 Task Levels

According to Byström and Hansen [27, 28], task contexts in information practices can be represented by a nested model consisting of three levels (from outer level to inner level): work task, information seeking task, and search task. Specifically, work tasks are separable parts of a person's duties in his or her workplace [28]. Everyday life tasks that emerge from non-work scenarios can also lead to active information seeking and searching practices (e.g., search for and book a hotel for travel) [2].

In addition to Byström and Hansen's nested model of task, Xie [190] also explored the multilevel nature of user goals and tasks and developed a four-level hierarchical framework of goals. This four-level typology covers a wide range of user goals and tasks (from long-term task-independent goals to local goals behind specific search tactics) and was verified via user studies [103, 190].

2.3 Task Facets

Focusing on different dimensions or task taxonomies, previous research has examined the impacts of task *types* and *facets* on search interactions from different perspectives. Liu et al. [105] and Jiang et al. [82] examined the associations between user behaviors and objective task features (i.e., task product, task goal, task type) and discussed to what extent these behavioral features can help disambiguate search tasks of different types. Capra et al. [34] found that manipulating task *a priori* determinability via modifying task items and dimensions can significantly affect users' perceived task difficulty and choices of search strategies.

With respect to task-user combined features, Wildemuth [186] argued that in task-based information search, search tactics are influenced by users' topical knowledge. Liu et al. [106] demonstrated that both whole-session level and within-session search behaviors are affected by task difficulty, and that the dynamic relationships between search behavior and task perception are influenced by task type. Similarly, Aula et al. [7] investigated search behavioral variations under tasks of different levels of difficulty, and found more query variance, more usage of advanced syntax, and longer time on search engine result pages (SERPs) with more difficult tasks.

Li and Belkin [99] developed a faceted approach to conceptualizing tasks in IR based on related literature on task classification as well as their empirical studies on task-based information searching [97, 98]. The faceted framework provides a holistic approach to exploring the nature of tasks and conceptually supported a series of empirical studies on task-based search interactions.

2.4 Task Stages

Task process is an aspect of task which differs from static task properties or facets [99] (e.g., predefined task goal, task product). When conceptualizing tasks from the process-oriented perspective, we are essentially looking at the process of *doing* or *performing* tasks. The core argument here is that in the context of information seeking, we cannot define or study a task without examining *how* the task was actually completed (or failed). Therefore, to fully understand a task, we need to explore both the objective task features and users' responses to the evolving task environments at multiple levels (e.g., behavioral, cognitive, emotional).

Many search process models focus on behavioral aspects and examine the transitions of information seeking and search actions. For instance, to describe the general process of information seeking, Ellis [53] studied the information seeking patterns of academic social scientists and broke it down into six characteristics: starting, chaining, browsing, differentiating, monitoring, and extracting. Wilson [188] suggests that in some circumstances, Ellis' "characteristics" can be organized as a sequence of information seeking stages. Ellis' model clearly identifies the features of information seeking patterns and has been modified and tested empirically [e.g., 54, 55]. However, this model only describes the behavioral level of task-based information seeking. It does not consider the interaction between the information seeker and the multi-dimensional context in which task states and information seeking activities evolve.

2.5 Applying Task Knowledge in IR

Applications of task knowledge to IR have demonstrated that task representations can be used to provide users with better query suggestions [8], build user models for improved personalized search [115, 178] and recommendation [199], and help in satisfaction prediction [68, 173]. Mehrotra et al. [114] used a tensor-based approach, representing each user as a combination of their topical interests and their search task behaviors for personalization. Other works have developed various novel task context embeddings to represent queries via search logs to provide task-based personalization, query suggestion, and re-ranking [115, 119]. Tolomei et al. [160] investigated the concept of task flows and analyzed query logs to generate task-based query suggestions. Baraglia et al. [12] introduced the notion of search shortcuts and offered query suggestions to drive goal attainment.

Vu et al. [171] has also used tasks to model user interests in search. In a similar vein but in other contexts, several scholars have leveraged task information to provide long-term support for task completion [e.g., 1, 84, 181]. Cai et al. [31] used task models to improve the ranking of retrieved search results to provide task-based support to users. Tasks help users achieve their search goals and understand and evaluate a system's competency in helping users do so. Hassan et al. [68] used search task constructs to predict satisfaction. White and Kelly [185] used them to improve relevance

feedback. Song and Guo [152] demonstrated that task information could help to automate tasks to reduce user burden.

Other researchers have focused on assistive systems in terms of tours or trails to lead users through their search process [70, 120, 132], predicting users' next search action based on the current actions, either by predicting the next result click [32] or by predicting short-term interests based on task topic information [177].

## 2.6 Recent Research, Development, Activities

There continues to be significant interest and activity surrounding tasks from the research community. Several workshops have been held on task-based IR, focusing on search interactions, searcher intents, and tasks in information search. This includes the SIGCHI 2012 workshop on *End-user Interactions with Intelligent Systems* [156], and the *Second Strategic Workshop on Information Retrieval in Lorne (SWIRL)* [4]. The *Task-based and Aggregated Search* workshop held in 2012 [96] focused on the challenges of task-based and aggregated search, such as the mismatch between search interface and specialized task-based functionalities, the lack of homogeneous systems to support different tasks, and so on. In the same year, the SIGIR 2012 workshop entitled *"Entertain Me" Supporting Complex Search Tasks* [19] focused on fostering potential solutions to problems faced by searchers with complex information needs.

An NSF-sponsored workshop on *Task-Based Information Search Systems*, held in 2013, discussed challenges in developing systems and tools to support tasks and user needs Kelly et al. [87]. The SIGIR 2013 workshop on *Modeling User Behavior for Information Retrieval Evaluation* [39], examined ways to model search intent based on queries. Workshops on *Supporting Complex Search Tasks* held in 2015 [61] and 2017 [16] initiated interdisciplinary dialog on many task-related open research questions, including evaluation and the role of context. The WSDM 2018 workshop on *Learning from User Interactions* [113], focused on task-based intelligent systems, more specifically on six related topics – user needs and task understanding, user modeling and personalization, metrics and evaluation, user interaction processes and context, intelligent interface design and applications. The WSDM 2019 workshop on *Task Intelligence* [71], focused on tasks in the context of system development, including areas such search assistance, personalization, and recommendation. Shah and White [146] also delivered a well-attended tutorial on this topic at SIGIR 2020.

3 TASK COMPOSITION AND SUPPORT

These decades of work have led to many different mechanisms for representing tasks, which we can divide into two sets: explicit and implicit representations. Explicitly represented tasks are often presented as hierarchies, trees, or lists of aspects. These are explainable and readily interpretable. Implicit representations often use a probability distribution (over latent aspects of tasks) or encoded vectors. Such representations are usually not meant to be interpreted by humans, but they can offer more flexibility.

We have experimented with both of these representations over the years [e.g., 42, 107, 121], recognizing their advantages and disadvantages. However, we have started to converge on ideas that offer the best of both worlds—providing the interpretability of an explicit representation, with the scalability of an implicit representation. For example, we focused on task completion, defining three stages
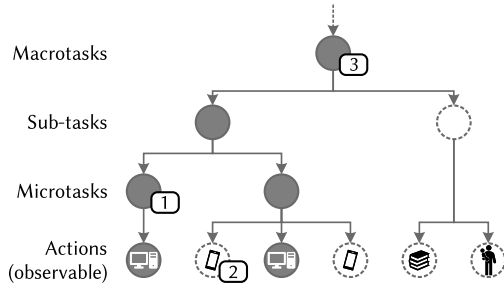
Figure 1: An abstract task "tree". Larger tasks may be decomposed into smaller tasks, and ultimately to actions. Some of these may be unobservable (dotted lines). Task support needs us to move "up", "down", and "across" the tree. See text for notes ①, ②, and ③.

of a task: *beginning*, *continuing/exploration*, and *ending/terminal*. For a given task and its stage, we also attempted to identify the kinds of support the user could use. Such support may include query or document suggestions, snippets or answers, as well as external tools. The goal here was to do manual (explicit) annotations for many search sessions with known tasks to then learn a model that could create an implicit representation (e.g., vector embeddings) of a task with respect to some application, such as next query prediction.

However, as we worked with several real-world datasets of search sessions, we realized that our coding scheme for task stage and support annotations was not as comprehensive or robust as we had hoped. We need a better framework that offers both comprehensive representation of a task as well as enough flexibility to be able to accommodate various applications and datasets. We discuss a possible approach next.

3.1 Tasks as Trees

We now consider some mechanisms and times for a search system to support a searcher's tasks. Simplifying the nested model of Byström and Järvelin [29] and the hierarchies of Xie [190], we can say that a task (also called a "macrotask" [37]) is composed of sub-tasks, sub-sub-tasks, and so on. For example, "arrange a vacation in Austin" may consist of "find the best dates" and "make bookings"; "make bookings" might be composed of "book flights" and "book a hotel"; etc. Each of these sub-tasks could be at any of Byström and Järvelin's levels (Figure 1). At the lowest level, a simple task is instead composed by "actions": the observable things people (or as we will discuss later, systems) might do. These could be instances of queries, or clicks; but could also be reading books, conversing with friends, or other moves (bottom level of Figure 1). In some cases this structure will be explicit, as in a project plan or a hierarchical to-do list, but more often it will not be. The structure might not be mapped out at the start, will certainly be dynamic in all but the simplest cases, and different strategies will be useful at different points. As searchers may be simultaneously engaged in multiple tasks, the corresponding hierarchies of sub-tasks and actions may also interleave in interesting and dynamic ways. It is our perspective that hierarchical representations are key to task modeling that is supported by a body of existing literature [28, 29, 37, 80, 154, 163, 190, 191].

In principle, a search system can offer support at each level of this hierarchy, although in practice search support tends to be small-scale. For example, actions are supported by techniques such as query auto-completion (supporting the current action) or query suggestion (supporting the next action), and these supports are relatively well-studied [30]. Some low-level tasks are also supported in search systems: for example, major web search engines offer booking widgets for flights and hotels, directly supporting these small transactional tasks. Mid-level tasks can be supported by, for example, recognizing a flight booking and offering to book a hotel and transport. Although only partially search applications, airline websites routinely offer this. High-level tasks, such as planning an entire vacation, are not at all well supported in software but are routinely supported by (human) agents and delegates.

3.2 Moves

The tree of (sub)tasks and actions also suggests certain moves that competent software should make. To move *left to right* in the tree is to predict or suggest the next thing in a sequence. To move *up the tree*, action to task or sub-task to super-task, is to recognize a more complex task, having recognized its constituents [e.g., 82, 108]. Finally, to move *down the tree* is to decompose a task [e.g., 73, 198].

For example, by re-ranking search results, Bennett et al. [21] consider short-term and long-term context information for personalization which in our framework corresponds to moving *left to right* for short and longer distances but without explicitly modeling the hierarchy. Similarly, Mitra [120] considers sessions as paths in query embedding spaces, again moving *left to right* without specifically modeling the hierarchical relationships. Finally, Sordoni et al. [153] use a hierarchical recurrent encoder-decoder architecture to simultaneously model the sequential relationship between terms in a query and between queries in a session. While they do not consider higher level relationships between search sessions, sub-tasks, and tasks, it may be natural to employ such methods to model task hierarchies.

3.3 Challenges

This model illustrates some challenges we face, if we are to build task-aware search complement in long run. First, some moves around the tree are easier than others. For example, at the time of writing, popular web search engines support small, transactional tasks—such as booking a flight—only when the most-recent query looks promising. Research on building longer-term task models is still limited [95, 174], even at the scale of consecutive searches [100, 178], meaning this move is currently only possible when there is a 1:1 correspondence between task and action (point ① in Figure 1).

Some actions are also unobserved, or unobservable, from software, even in practice (②). For example, a web search engine will most likely be unaware of a searcher's other activity online; all online services will be blind to a face-to-face conversation.

Finally, observed actions are sparse signals and more than one task will have similar steps, so moving up the tree is more difficult than moving down or sideways. We can easily imagine support for decomposing tasks, and can also imagine going across the tree at any level: for example, we could predict the next action given a sequence of actions, or we could predict the next microtask given

a sequence of microtasks. It is harder to imagine getting from observed actions to the uppermost (macro-) task or goal ③, especially when observations are incomplete ②. We must also note that the data searchers give us will be bound by the affordances we give them; in practice, that means that searchers will express themselves in short keyword phrases ("lhr lax flights") rather than explain a task ("I need to get to the LA office for Wednesday's big meeting").

Challenges for supporting tasks in search therefore include:

(1) *Representing tasks* in ways that allows the system to take actions. This representation needs to handle tasks at different granularity, with different topics and strategies, and tasks which persist over time.
(2) *Observing more task-relevant context*, to better identify and track tasks as they happen. This needs to include tracking across different devices and different timescales, so we can better identify tasks from actions and "move up" the tree.
(3) *Developing task-oriented interfaces* that encourage descriptions of task, not need and not short queries; and which support tasks as they happen, either in the search interface or elsewhere.

4 TASK MODELING

There are different ways we can extract, represent, and apply task information to address the challenges discussed in Section 3.3. In this section, we review some possible approaches we could take in modeling and extracting complex task structures composed of any number of tasks or sub-tasks.

4.1 Task Representation

In the model shown in Figure 1, tasks can be defined at different granularity levels. This flexibility provides ways to represent tasks from different theoretical and methodological perspectives. At the same time, it asks for a far-reaching representation capable of modeling work at multiple levels of abstraction [135]. Task descriptions can range from a high level of abstraction to a concrete, granular action-oriented level with precise information need strongly associated to the task. As mentioned by Paterno [134], to build an intelligent task-aware search system, it is necessary to support tasks at each level of the task hierarchy not only from top to down but also from left to right. There are many possibilities to instantiate our task framework by applying diverse supervised and unsupervised techniques depending on the availability of search interaction signals. Assuming that there may be multiple sub-tasks associated with a user's information need and that these sub-tasks could be interleaved across different sessions, a bare tree extraction algorithm has the potential to extract a hierarchical representation of tasks/sub-tasks embedded in search processes as considered by Mehrotra and Yilmaz [117] (e.g., decomposing a macro task into microtasks as moving down the tree in Figure 1). The approach allows us to go across the tree at any level.

Another possible approach could be a vector representation of tasks implicit in search behaviors (i.e., points ① and ③ in Figure 1) by triangulating observable search events with other situational and contextual information related to the search process. This abstract representation of tasks can especially be helpful in search scenarios where searchers' tasks are not clearly expressed or manifested. For example, existing research has shown how such signals indicate the nature of the task being done [e.g., 38, 107, 108, 122].

To move up and down the task hierarchy, action to task or sub-task to macro-task, it is crucial to know the connections among the contextual components of the search session. Based on the idea that in a real-world information network, proximal nodes in the network structure tend to be similar or related to one another, it is intuitive to visualize user-system interactions initiated by a specific task as a complex graph network structure of users' actions (i.e., query submission, clicks on a document) and systems' reactions (i.e., analyze, retrieve, and display relevant related items). Similarly, queries issued and actions performed by a user and documents viewed within a short time period are more likely to be different stages of the same task, sub-tasks, or sub-sub-tasks; therefore, the search state can be extracted based on similar node representation patterns. Therefore, a sequential heterogeneous graph embedding-based task model [e.g., 60] could potentially capture the structural features of interactive search sessions and represent tasks from observable behavioral signals. This way, the model can represent the macro-task (moving up in the tree) or the next microtask given a sequence of microtasks (moving down or right in the tree).

We have seen several attempts to model search sessions as Markov Decision Proceses [e.g., 36, 194], Hidden Markov Models [e.g., 33, 50] or Partially Observed Markov Models [e.g., 193]. Taking the idea further, we could apply reinforcement learning approaches to learn to predict or suggest the next action/task given a sequence of actions or tasks. This is similar to search intent prediction by Yao et al. [195].

4.2 Inferring Tasks from Observable Events

Many studies used lexical and content-based features, such as the lexical content of queries, for determining topical and task change in the sequence of query formulations. For example, Verma and Yilmaz [169] tried to identify entities and clusters of terms related to entities in queries (e.g., using tagging, TF-IDF scoring, term filtering, category terms) to represent a task as a set of terms related to an entity. Other studies have used latent search interaction events to infer tasks (query-based features: query term cosine similarity; URL-based features: URL domain clicked, Jaccard coefficient between clicked URL sets; session-based features: same session and the number of sessions in between, query reformulations, click entropy, query length, post-click actions, and session lengths; temporal features: dwell time for action events). Studies have shown how such signals indicate the nature of the task being performed, even when there is no explicit statement [107–109, 122, 175]. Depending on the availability of search interaction features at a given time, we could exploit several clustering algorithms to extract tasks.

5 APPLICATIONS OF TASK IN SEARCH

Task information applications can pave the way for simulating, developing, and evaluating task-aware support. Although existing search systems have improved incredibly and support users with specific factual information tasks, their support is still lacking for complex and exploratory search tasks. Given the nature of these tasks, they need to be decomposed into multiple actionable sub-tasks (i.e., move down the task tree shown in Figure 1). They may require numerous rounds of interaction (queries/clicks, from

a search engine perspective) to complete those tasks [7]. Tracking and completing those sub-tasks increases cognitive demands, regardless of user experience level. The task tree can be applied to decompose exploratory and complex tasks into smaller goals, hence reducing cognitive load. This can also help narrow the focus of the assistance offered to the specific task at hand, which could be represented in a semantic space (the so-called "implicit representations" referenced earlier) to better identify the task and more fully capture the user's underlying goals and intentions.

In this section, we examine four applications where such considerations of task-based knowledge are valuable.

5.1 Contextual Search

Searches are performed within a situational context. Understanding and modeling this context, especially the current task, is vital for search systems in finding the most relevant information. Task models derived from recent queries and clicks (i.e., the observable actions in the leaf nodes of Figure 1) within the current session can be applied to improve search engine performance [148, 189]. These task representations can assume many forms, including distributions over topical categories [21] or semantic vectors [118].

As we try to model tasks in a short-term search context, we often find ourselves discussing *sessions* (sequences of interactions demarcated by topic or time [84]), which are not exactly the same as tasks (especially given multi-tasking [155]) but are a reasonable proxy for task in a search setting and are a valuable source of tasks data [100, 101]. Task models must evolve over time as more evidence is collected about user interests and intentions (implicitly, explicitly, or both) and ideally be transferable across sessions as tasks are suspended and resume over time [1]. Other search-related applications of task models that span the leaves of our task tree include personalizing search results [116] and generating query suggestions [62].

5.2 Multi-device Search

Complex tasks can span both time and space. Another way that the leaves on the task tree can be related is in terms of the devices used. As mentioned in the previous section, there has been some focus in IR on supporting cross-session tasks [1]. Cross-device searching [126, 175], where people initiate a task at one time and/or on one device and resume it later, perhaps on a different device, is related to cross-session and may be simply because of necessity, but also the device capabilities (e.g., larger display, availability during commute). Supporting both types of searching requires a task representation that is transferable between devices (something more abstract and consistent than a sequence of observable actions). This involves moving up in our task tree, from actions to micro-tasks, sub-tasks, and so on, stopping at the point where the device space can be most fully represented without being so broad that the task representation is meaningless. Multi-device experiences capitalize on the strengths of multiple devices simultaneously to support complex tasks (e.g., recipe preparation, home or auto repair) [180]. For example, we can combine a smart speaker such as an Amazon Echo with a tablet such as an Apple iPad capitalizes on the far-field speech recognition capabilities of the speaker and the high-resolution display of the tablet. In these experiences, the evolving

task representation (implicit, explicit, or both) plays a central role in connecting the devices and providing dynamic context.

In multi-device scenarios, as with many other task scenarios, task assistance can be offered to users at different stages of the task (e.g., proactively searching for resources related to the current action [130]) depending on an understanding of the task and the affordances available. This multi-device paradigm can also apply directly to a search context, where, for convenience, people can pose natural language questions to smart speakers via voice, obtain quick answers, and use their smartphones or tablet devices to review supporting information (videos, websites, documents, etc.). For example, a child getting quick responses from a digital assistant (e.g., an answer to a math question) on a smart speaker or smart watch can also be shown explanatory information on a larger display device. Supporting the use of combinations of devices in multi-device search can provide a way for people to maximize the quality and diversity of the information that they utilize. More fully representing tasks, and their dynamism and context sensitivity, is critical in supporting these multi-device behaviors.

5.3 Conversational Agents

One of the active areas of application for task-based IR is conversational agents. One can imagine the following conversation happening with an agent over voice using, for example, a smart speaker or a smartphone.

> **User**: I think I would like to go do some outside activity today. Do I need to wear a face mask if I go running?
> **Agent**: It depends where you are running, but if you are concerned about safety or compliance and still want an outdoor activity, may I suggest biking?
> **User**: Oh.. ya, sure, that could work. Do I need to know anything?
> **Agent**: While you don't need to wear a mask while biking, you should still bring one with you. There is also a chance of some rain showers, so plan for that. And yes, definitely carry some water.

Now let us examine what may be going on here. There are four distinct capabilities that we see the agent exhibiting.

- *Understanding the intention behind a user seeking information.* The agent understands that the user wants to do outdoor activity while being safe. This understanding enables the agent to make other recommendations beyond simply answering the question.
- *Addressing the effects of unknown unknowns (i.e., "people don't know what they don't know").* The user asked "what do I need to know if I go biking?", indicating their lack of knowledge about even what may be the right questions to ask. This often happens in human-human interactions. Here, the agent understands the situation (task), as well as the intention behind that question and responds with relevant suggestions.
- *Zero-query recommendations.* The user does not ask about weather, but the agent deems it important to convey that information as it may affect the outdoor activity. Also, given the nature of the activity (biking), the agent also recommends carrying water. These are examples of *zero-query recommendations*, in which an answer is provided without there being a clear question. Again, doing

something like this requires a deep understanding of the situation (task), the user, and their intentions.

- *Proactive recommendations.* The conversation starts by the user asking a question about running, but rather than completely answering that question, the agent makes a different suggestion (biking), which turns out to be a better one. This is a case of the agent being proactive. In order to go beyond the user's need (at least the expressed need) and provide a relevant and compelling answers or recommendations, an agent needs to be able to understand the purpose behind the potential task, the user's intention behind asking a question, and the world knowledge about how different tasks are executed.

In short, to create an intelligent agent like the one envisioned in the scenario above, we need to bring in the following capabilities:

- Abstracting out from a query or a question or even an observation to the task and/or context.
- Leveraging world knowledge (in this case, public health guidelines and mask mandates).
- Generating recommendations from that task/context and weighing whether that would outperform query/question-based recommendation.
- Learning how to perform a task.

As one can see, much of what we need revolves around tasks. This is just a simple example of a short conversation. Imagine having discussions (and even debates) about health, politics, and more. Imagine carrying out such conversations across multiple sessions, multiple devices, and multiple people. There are tremendous possibilities here for a giant leap for IR systems. We believe at its core is the notion of task and ways to capture, represent, and address it.

5.4 Proactive Search and Recommender Systems

The ability to identify and automatically extract and represent tasks accurately has implications for search or recommender systems in understanding users' information needs at different task levels as well as supporting people in task completion. Therefore, it is crucial to understand how to utilize this knowledge about tasks behind the request to improve a system's offerings to its users. Also, the ability to model users' tasks from their observable actions (at different levels per Figure 1) unlocks new directions for solving many problems and improving user engagement and satisfaction for building intelligent and proactive systems that can retrieve and recommend information implicitly without requiring explicit queries or other interactions [49]. This is important because research has shown that people often struggle to get their tasks done due to a lack of knowledge, motivation, or information literacy [142].

The observable actions covered earlier are primarily those taken by the user on their initiative, but this need not always be the case. In mixed-initiative systems, these actions can be prompted by the system or even taken by the system on the user's behalf [74], i.e., new leaf actions in the task tree can be proposed or created automatically. The notion of proactive search systems is not new. Letizia [102] was one of the earliest applications that provided proactive recommendations during web browsing. Commercially deployed proactive, intelligent systems such as Google Now and Microsoft Cortana can model short-term and long-term search intents and tasks based on search log history [64]. In recent times, Song and Guo [152]

proposed proactive recommendations to the user at specific times based on repeated pattern recognition over time. Incorporating task understanding into a proactive system could support users in each task stage and help enable task completion. A task-aware intelligent system could proactively identify potential problems in users' search paths and guide users at various task levels by providing help recommendations or what actions could be executed next to avoid future problems. The aforementioned task representation can be incorporated into various sequence-to-sequence models, probabilistic, or Markov decision-based reinforcement learning models to generate proactive recommendations.

6 EVALUATING TASK-BASED APPLICATIONS

Evaluation is central in IR [85] and this is no different in task-based search and recommendation systems. Many of the same *methodologies* (user studies, simulations, etc.) used in IR to evaluate system performance can be used to evaluate systems to support tasks in search and recommendation settings. Non-task-based IR systems tend to focus on ad hoc retrieval and consider each query independently. Task-based systems consider tasks holistically, spanning multiple queries and/or sessions, the associated context, and task outcomes. The *metrics* used to determine task-based system performance deserve special attention given the focus of these systems on supporting full task processes (not individual queries) and attaining task completion (not only result relevance). We now offer a perspective on methods and metrics for task-based evaluation.

6.1 Methodologies

Many standard evaluation methods (user study protocols, instruments, etc.) apply to the evaluation of task-based systems [85]. In IR, the Cranfield experiments [41] and TREC [170] have driven considerable progress, including in tasks research [197]. Beyond Cranfield and TREC, evaluation in IR must now take a broader view on tasks, users, and context [83], to improve experimental realism and the reliability of conclusions drawn. Methods such as living laboratories [91] bridge user- and system-centered research via resources, tools, and infrastructure for collaborative experimentation [11]. Mixed methods studies can provide a more complete picture of task performance, albeit with more complexity and greater cost than single-method studies. As mentioned earlier, tasks can extend over time and be part of larger macrotasks. This additional context should also factor into task-based evaluation [47].

6.2 Metrics

Evaluating systems on the basis of search task performance has been explored for decades [72]. All metrics make assumptions about task behavior, which must be validated [51]. Conceptualizing tasks and creating task models are important in determining appropriate task-based evaluation metrics. It is insufficient to solely target system functionality (or even more narrowly: specific components) when systems and users must collaborate to complete tasks successfully [14]. We should evaluate task-based systems *holistically* to reach actionable conclusions and understand system performance [10]. We discuss that now, targeting task processes and task outcomes.

6.2.1 Task Processes. Process metrics are focused on how people attempt to complete the task, regardless of the task outcome. They

include: (1) *Task completion time*, both actual time and perceived time. Time has been used in search evaluation [57, 192]. Task has been shown to affect document dwell times [89, 185]. Smucker and Clarke [151] studied time from the perspective of gain per unit time. Perceived time can differ from stopwatch time per factors such as attentional demand [44]; (2) *Effort expended* to complete the task (e.g., the number of actions taken, recommendations reviewed, dialog turns). In search, effort typically describes the number of searches or clicks [9, 43]. Kelly [86] discussed the relationship between expected and experienced effort (e.g., if experienced effort is less than expected, the task is considered easy). Effort underlies many user models in IR evaluation [e.g., 79, 125]. Kiseleva et al. [94] showed that user satisfaction is negatively correlated with the amount of effort to complete a task: more effort means less user satisfaction; (3) *Engagement* covers the connection between the user and the system, spanning emotional, cognitive, and behavioral aspects [78]. It is affected by many factors, including user and task characteristics, user experience, and biases [131]. It can be a goal in task-based systems (e.g., in open-domain dialog [76]) but also a side effect (e.g., in task-oriented dialog systems [35]), and; (4) *Progress* through the task. Detecting task completion can be straightforward for some tasks, e.g., transactional tasks, but complex for others, e.g., learning tasks [183]. Progress can be tracked using dedicated tools [20] or inferred [182]. Recent research has built benchmarks for measuring task progress in digital assistants [104]. Task-oriented dialog systems, focus on metrics such as number of slots filled (x of y) [25]. These four popular metrics are broadly applicable, are easy to define in task-based search and recommendation settings, and can be computed at low-cost at large scale. There are other metrics including cognitive load [15], learning [136], affect [56], and usability [3], which are more challenging to define and measure.

6.2.2 Task Outcomes. Outcome metrics focus on the product of tasks, either a real outcome (e.g., task completion) or a user-perceived outcome (e.g., satisfaction). Salient examples include: (1) *Task utility*, denoting the value of information obtained to complete the task, e.g., relevance [123]. Relevance is affected by task stage [158] and relevance metrics help estimate support for task completion [124]. Relevance metrics are usually computed per query but session-level metrics must also be considered in task scenarios [110], as must task support beyond result pages [46]. Relevance is personal and situational [141] and task-based evaluation must consider that, e.g., during contextual search [21]; (2) *Satisfaction* with the outcome of the task and the process, often modeled at the task/session level [69]. Satisfaction is non-binary and impacted by task and user effects [89, 93, 185] and even query position in the session [81]. More observations of on-task behavior enable more accurate models of satisfaction [75, 94], and; (3) *Task success*, covering whether task objectives were accomplished. This relates to satisfaction but not entirely and can be modeled based on behavioral signals [67]. Completion events such as in-world activities may be unobservable to online systems, making it difficult to measure task success, although proxies e.g., conversions [24] may offer insight. Other task outcome metrics, including novelty and diversity [40], creativity [149], and adoption and retention, e.g., search engine switching [184] and sustained use over time [48], are promising but are also less well defined and require data that can be difficult to obtain.

6.3 Additional Considerations

There are many other metrics that can apply to task-based systems including robustness, privacy, adaptivity, and scalability [147]. In developing task-based metrics, we also must consider user models (e.g., personas) and task models (e.g., search strategies and goals). Task performance is affected by many factors, including intrinsic properties of the task (e.g., nature of the task [121], topic [112], difficulty [187], complexity [29]) as well as extrinsic properties such as user attributes (e.g., expertise [179], familiarity [90]), the situation [77, 80, 144], and other factors such as meta-cognitive skills in task planning and reflective assessment [22]. We must also understand the nature of the user experience, which impacts how metrics are defined and interpreted. Metrics also interact, e.g., effort affects satisfaction [196] and they trade off, e.g., time taken versus coverage [162]. Metrics must be contextualized, e.g., not all effort is detrimental and more effort could also mean more learning.

Task support systems also contain multiple connected components [128]. Evaluating per component performance has limited value in appraising what the user would experience [162]; hence our focus here on holistic metrics. However, the metrics may not be correlated [59]. Integrated metrics combine multiple variables [131, 157, 164], although these can be difficult to interpret. Sets of metrics are commonly employed in the evaluation of task-oriented dialog systems [172] and defining such a set of metrics that are agreed upon by the community could help evaluate task-based search and recommendation systems. Meta-analysis frameworks [6, 140] analyze the extent to which metrics capture key properties and align with user preferences; they may also be applicable here.

7 TASK FUTURES

Considering user tasks in IR is not a new idea, but every new generation of IR students and scholars seem to encounter it in a new light – sometimes leading to groundbreaking advancements, and other times redoing or incrementally adding to previous work. With the increasing attention to and importance of emerging IR applications, we believe the time is ripe for a new generation of scholars to not only rediscover task-based IR, but also take a conceptual and practical leap to finally realize the vision of supporting users in accomplishing their tasks, regardless of their information literacy or specificity in queries. We now consider some future directions and conclude by discussing key ethical considerations.

7.1 Research Threads and Directions

Here, we identify some big challenges, each suitable for one or more PhD dissertations or grant proposals:

- Task understanding
  - Formalize and validate various task representations (both implicit and explicit, as mentioned earlier), potentially tying them to different contexts or applications.
  - Investigate different ways to use contextual information (e.g., spatiotemporal signals, concurrent running applications) to better understand tasks.
  - Extend task understanding across multiple sessions and/or devices.
  - Attributing and aggregating observed actions into higher-level tasks (moving up the task tree).

- Task support
  - Make task a first class object in search support, e.g., surface guided tours in response to exploratory queries.
  - Provide support for task completion (not just providing search results), including recommending search as a means of task completion, where appropriate.
  - Integrate IR applications with existing task applications such as Microsoft To Do and Google Tasks, as well as email and calendar, to seamlessly surface task-related information and actions.
  - Better support complex tasks comprising multiple steps, including decomposing complex tasks into more manageable sub-tasks, and supporting search across multiple sessions and/or devices.
  - Support team tasks (direct collaboration, sub-task assignment, load balancing, etc.) in addition to individual tasks.
  - Cooperate with users directly, e.g., task-oriented dialog systems, to address tasks more explicitly and also to better educate users about the role of IR systems in solving tasks.
  - Explore task automation, starting with frequent or recurring tasks, e.g., travel planning, finding job opportunities, and researching a socio-political issue, including extending work on standing queries [127] and slow search [159].
- Task data and experimentation
  - Provide lightweight task capture mechanisms, as ground truth for machine learning models and to build trust in task assistance with users by giving them agency over what task-related information is shared with the system.
  - Find ways to uncover more unobservable events related to the task process (triangulate data sources, with user consent).
  - Create shared datasets and challenges, with user consent, to promote task-related research and mitigate risk of leaking sensitive data via methods such as differential privacy [52].

We believe the framing device presented in this paper (Figure 1) as well as our proposals for how such a device can be useful in modeling and using task in search applications (Sections 4 and 5) can help for at least some of these directions. For example, the task tree structure along with the formulations of various moves presented in Section 3 can be used to define a set of support actions (e.g., offer within-task query recommendations with traversal to a sibling node, suggest related tasks with a jump to a new parallel branch in the tree) in interactive search. This structure can be comprised of (1) identifying which part of this task tree the user is at a given moment; (2) deciding what could be the next set of sub/super/related tasks could be from this tree; and (3) making and revising recommendations based on user actions (moves).

7.2 Ethical Considerations

Capturing and representing tasks can have benefits, but at what cost? Many scholars have argued that low information literacy can lead to users not being able to fully utilize the available information or the tools to their most potential [138]. Even for users with reasonable or high information literacy, they often "don't know what they don't know" [17, 145]. In other words, if an IR system is relying on a user explicitly and at least partially expressing their information needs in order to provide them results or recommendations, it is likely to face challenges serving these populations of users. Extracting and using task information, and being proactive in search can help such users [185]. However, what is often ignored are the ethical considerations and responsibility of researchers and developers.

As we move toward systems that go beyond serving explicit requests from users, with task-based IR systems being one of the examples, there are dangers in how such systems could unduly influence user behaviors and nudge them in ways that perpetuate bias and a false sense of trust. With rapid development in artificial intelligence techniques that are being deployed in search systems, those systems become less and less *trustworthy*, even while usually remaining *trusted* [133]. The feedback loops created between systems recommending information and users selecting among recommendations make the selections less and less useful for training: we are no longer observing human behavior, but controlling it [111, 129]. This effect, along with other systemic effects, means that the datasets on which models are trained include significant biases [63, 161, 165].

This vicious cycle of a system getting ahead of user requests to recommend results and the users clicking on them as they either lack motivation or enough information literacy can be manifested in several ways. For instance, this proactive, task-based recommendation could lead to a search engine promoting its own services and tools simply because it has access to a lot more data and insights about those entities than those from their competitors.

Identifying and modeling tasks may call for more data collection from more people, even those who do not actively use the system. We need to balance the need for more data and the dangers of ubiquitous data collection such as surveillance capitalism and other forms of abuse [65, 66, 200].

As task modeling inherently necessitates predicting users' next actions/needs, we must consider the cost of false prediction (e.g., requiring user to perform even more actions to counter the system's false beliefs regarding user goals or intentions). A related question is how to recognize and respect user agency in their tasks and not overtly influence their course of action.

We should also not assume that a task modeling system can easily identify and address a singular objective or interest. When different stakeholder interests are involved, how do we balance across the different dimensions and control for unintended consequences? For example, a tool that makes it really easy to book a flight may unintentionally discourage users to do more research that may lead to cheaper tickets. Finally, if task modeling is inherently complex and resource intensive, it might mean that system designers need to prioritize which tasks they support, raising questions about fairness across different user populations. In short, explicating and using task information, while important and desired, must be done with ethical issues in mind. We should, in general, create a practice of integrating such considerations from the outset rather than trying to address them later or fix problems resulting from not considering them as a posthoc activity.

REFERENCES

[1] Eugene Agichtein, Ryen W. White, Susan T. Dumais, and Paul N. Bennett. 2012. Search, interrupted: understanding and predicting search task continuation. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 315–324.

[2] Denise E Agosto and Sandra Hughes-Hassell. 2006. Toward a model of the everyday life information needs of urban teenagers, part 2: Empirical model. *Journal of the American Society for Information Science and Technology* 57, 11 (2006), 1418–1426.

[3] William Albert and Thomas Tullis. 2013. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Morgan Kaufmann.

[4] James Allan, Bruce Croft, Alistair Moffat, and Mark Sanderson. 2012. Frontiers, challenges, and opportunities for information retrieval: Report from SWIRL 2012 the second strategic workshop on information retrieval in Lorne. *ACM SIGIR Forum* 46, 1 (2012), 2–32.

[5] David Allen. 2015. *Getting Things Done: The Art of Stress-free Productivity*. Penguin.

[6] Enrique Amigó, Damiano Spina, and Jorge Carrillo-de Albornoz. 2018. An axiomatic analysis of diversity evaluation metrics: Introducing the rank-biased utility metric. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 625–634.

[7] Anne Aula, Rehan M. Khan, and Zhiwei Guan. 2010. How Does Search Behavior Change As Search Becomes More Difficult?. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA). 35–44.

[8] Ahmed Hassan Awadallah, Ryen W. White, Patrick Pantel, Susan T. Dumais, and Yi-Min Wang. 2014. Supporting complex search tasks. In *Proceedings of the ACM CIKM Conference on Information and Knowledge Management*. 829–838.

[9] Leif Azzopardi, Diane Kelly, and Kathy Brennan. 2013. How query cost affects search behavior. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 23–32.

[10] Krisztian Balog. 2015. Task-completion Engines: A Vision with a Plan.. In *Proceedings of the Supporting Complex Search Task Workshop*.

[11] Krisztian Balog, Liadh Kelly, and Anne Schuth. 2014. Head first: Living labs for ad-hoc search evaluation. In *Proceedings of the ACM CIKM Conference on Information and Knowledge Management*. 1815–1818.

[12] Ranieri Baraglia, Fidel Cacheda, Victor Carneiro, Vreixo Formoso, Raffaele Perego, and Fabrizio Silvestri. 2009. Search shortcuts: Driving users towards their goals. In *Proceedings of the International Conference on the World Wide Web*. 1073–1074.

[13] Marcia J Bates. 1989. The design of browsing and berrypicking techniques for the online search interface. *Online Review* 13, 5 (1989), 407–424.

[14] Marcia J Bates. 1990. Where should the person stop and the information search interface start? *Information Processing and Management* 26, 5 (1990), 575–591.

[15] Micheline Beaulieu. 1997. Experiments on interfaces to support query expansion. *Journal of Documentation* 53, 1 (1997), 8–15.

[16] Nicholas Belkin, Toine Bogers, Jaap Kamps, Diane Kelly, Marijn Koolen, and Emine Yilmaz. 2017. Second workshop on supporting complex search tasks. In *Proceedings of the ACM CHIIR Conference on Human Information Interaction and Retrieval*. 433–435.

[17] Nicholas J Belkin. 1980. Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science* 5, 1 (1980), 133–143.

[18] Nicholas J Belkin. 1990. The cognitive viewpoint in information science. *Journal of information science* 16, 1 (1990), 11–15.

[19] Nicholas J Belkin, Charles LA Clarke, Ning Gao, Jaap Kamps, and Jussi Karlgren. 2012. Report on the SIGIR workshop on "entertain me" supporting complex search tasks. *ACM SIGIR Forum* 45, 2 (2012), 51–59.

[20] Victoria Bellotti, Brinda Dalal, Nathaniel Good, Peter Flynn, Daniel G Bobrow, and Nicolas Ducheneaut. 2004. What a to-do: Studies of task management towards the design of a personal task list manager. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 735–742.

[21] Paul N Bennett, Ryen W White, Wei Chu, Susan T Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. 2012. Modeling the impact of short-and long-term behavior on search personalization. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 185–194.

[22] John Biggs. 1988. The role of metacognition in enhancing learning. *Australian Journal of Education* 32, 2 (1988), 127–138.

[23] Andrei Broder. 2002. A taxonomy of web search. In *ACM Sigir forum*, Vol. 36. ACM New York, NY, USA, 3–10.

[24] Yuri M Brovman, Marie Jacob, Natraj Srinivasan, Stephen Neola, Daniel Galron, Ryan Snyder, and Paul Wang. 2016. Optimizing similar item recommendations in a semi-structured marketplace to maximize conversion. In *Proceedings of the ACM RecSys Conference on Recommender Systems*. 199–202.

[25] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. MultiWOZ-A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 5016–5026.

[26] Katriina Byström. 2007. Approaches to "task" in contemporary information studies. In *Proceedings of the International Conference on Conceptions of Library and Information Science*, Vol. 12. 1–10.

[27] Katriina Byström and Preben Hansen. 2002. Work tasks as units for analysis in information seeking and retrieval studies. In *Proceedings of the International Conference on Conceptions of Library and Information Science*. 239–251.

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

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

[30] Fei Cai and Maarten de Rijke. 2016. A Survey of Query Auto Completion in Information Retrieval. *Foundations and Trends in Information Retrieval* 10, 4 (2016), 273–363. https://doi.org/10.1561/1500000055

[31] Fei Cai, Shangsong Liang, and Maarten De Rijke. 2014. Personalized document re-ranking based on bayesian probabilistic matrix factorization. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 835–838.

[32] Huanhuan Cao, Daxin Jiang, Jian Pei, Enhong Chen, and Hang Li. 2009. Towards context-aware search by learning a very large variable length hidden markov model from search logs. In *Proceedings of the International Conference on the World Wide Web*. 191–200.

[33] Huanhuan Cao, Daxin Jiang, Jian Pei, Enhong Chen, and Hang Li. 2009. Towards context-aware search by learning a very large variable length hidden markov model from search logs. In *Proceedings of the 18th international conference on World wide web*. 191–200.

[34] Robert Capra, Jaime Arguello, Heather O'Brien, Yuan Li, and Bogeum Choi. 2018. The effects of manipulating task determinability on search behaviors and outcomes. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 445–454.

[35] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter* 19, 2 (2017), 25–35.

[36] Jia Chen, Yiqun Liu, Cheng Luo, Jiaxin Mao, Min Zhang, and Shaoping Ma. 2018. Improving session search performance with a multi-MDP model. In *Asia Information Retrieval Symposium*. Springer, 45–59.

[37] Justin Cheng, Jaime Teevan, Shamsi T Iqbal, and Michael S Bernstein. 2015. Break it down: A comparison of macro-and microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 4061–4064.

[38] Dongho Choi, Chirag Shah, and Vivek Singh. 2016. Probing the interconnections between geo-exploration and information exploration behavior. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 1170–1175.

[39] Charles LA Clarke, Luanne Freund, Mark D Smucker, and Emine Yilmaz. 2013. Report on the SIGIR 2013 workshop on modeling user behavior for information retrieval evaluation. *ACM SIGIR Forum* 47, 2 (2013), 84–95.

[40] Charles LA Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 659–666.

[41] Cyril Cleverdon. 1967. The Cranfield tests on index language devices. *Aslib Proceedings* 19, 6 (1967), 173–194.

[42] Michael J Cole, Chathra Hendahewa, Nicholas J Belkin, and Chirag Shah. 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*. 567–576.

[43] William S Cooper. 1968. Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation* 19, 1 (1968), 30–41.

[44] Mihaly Csikszentmihalyi. 1990. *Flow: The Psychology of Optimal Experience*. Harper and Row.

[45] Brenda Dervin. 1998. Sense-making theory and practice: an overview of user interests in knowledge seeking and use. *Journal of Knowledge Management* 2, 2 (1998), 36–46.

[46] Doug Downey, Susan Dumais, Dan Liebling, and Eric Horvitz. 2008. Understanding the relationship between searchers' queries and information goals. In *Proceedings of the ACM CIKM Conference on Information and Knowledge Management*. 449–458.

[47] Susan Dumais. 2009. Evaluating IR in situ. In *Proceedings of the SIGIR Workshop on the Future of IR Evaluation*. 2.

[48] Susan Dumais. 2013. Task-based search: a search engine perspective. In *Proceedings of the NSF Workshop on Task-Based Search*.

[49] Susan Dumais, Edward Cutrell, Raman Sarin, and Eric Horvitz. 2004. Implicit queries (IQ) for contextualized search. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. 594–594.

10

[50] Sebastian Dungs and Norbert Fuhr. 2017. Advanced hidden markov models for recognizing search phases. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. 257–260.

[51] Georges E Dupret and Benjamin Piwowarski. 2008. A user browsing model to predict search engine click data from past observations.. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 331–338.

[52] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*. Springer, 1–19.

[53] David Ellis. 1989. A behavioural approach to information retrieval system design. *Journal of Documentation* 45, 3 (1989), 171–212.

[54] David Ellis. 1993. Modeling the information-seeking patterns of academic researchers: A grounded theory approach. *The Library Quarterly* 63, 4 (1993), 469–486.

[55] David Ellis, Deborah Cox, and Katherine Hall. 1993. A comparison of the information seeking patterns of researchers in the physical and social sciences. *Journal of Documentation* 49, 4 (1993), 356–369.

[56] Henry A Feild, James Allan, and Rosie Jones. 2010. Predicting searcher frustration. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 34–41.

[57] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems* 23, 2 (2005), 147–168.

[58] Luanne Freund, Elaine G. Toms, and Charles L.A. Clarke. 2005. Modeling Task-Genre Relationships for IR in the Workplace. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 441–448.

[59] Erik Frøkjær, Morten Hertzum, and Kasper Hornbæk. 2000. Measuring usability: are effectiveness, efficiency, and satisfaction really correlated?. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 345–352.

[60] Xinyu Fu, Jiani Zhang, Ziqiao Meng, and Irwin King. 2020. Magnn: Metapath aggregated graph neural network for heterogeneous graph embedding. In *Proceedings of The Web Conference 2020*. 2331–2341.

[61] Maria Gäde, Mark M Hall, Hugo Huurdeman, Jaap Kamps, Marijn Koolen, Mette Skove, Elaine Toms, and David Walsh. 2015. Report on the first workshop on supporting complex search tasks. *ACM SIGIR Forum* 49, 1 (2015), 50–56.

[62] Darío Garigliotti and Krisztian Balog. 2017. Generating query suggestions to support task-based search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1153–1156.

[63] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. 2018. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231* (2018).

[64] Ramanathan Guha, Vineet Gupta, Vivek Raghunathan, and Ramakrishnan Srikant. 2015. User modeling for a personal assistant. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. 275–284.

[65] Drew Harwell. 2019. Doorbell-camera firm Ring has partnered with 400 police forces, extending surveillance concerns. https://www.washingtonpost.com/technology/2019/08/28/doorbell-camera-firm-ring-has-partnered-with-police-forces-extending-surveillance-reach/. Downloaded 2022-02-14..

[66] Paul Haskell-Dowland. 2021. How Apple's AirTag turns us into unwitting spies in a vast surveillance network. https://www.theguardian.com/technology/2021/may/17/how-apples-airtag-turns-us-into-unwitting-spies-in-a-vast-surveillance-network. Downloaded 2022-02-14..

[67] Ahmed Hassan, Rosie Jones, and Kristina Lisa Klinkner. 2010. Beyond DCG: user behavior as a predictor of a successful search. In *Proceedings of the ACM WSDM International Conference on Web Search and Data Mining*. 221–230.

[68] Ahmed Hassan, Xiaolin Shi, Nick Craswell, and Bill Ramsey. 2013. Beyond clicks: query reformulation as a predictor of search satisfaction. In *Proceedings of the ACM CIKM Conference on Information and Knowledge Management*. 2019–2028.

[69] Ahmed Hassan, Yang Song, and Li-wei He. 2011. A task level metric for measuring web search satisfaction and its application on improving relevance estimation. In *Proceedings of the ACM CIKM Conference on Information and Knowledge Management*. 125–134.

[70] Ahmed Hassan and Ryen W White. 2012. Task tours: helping users tackle complex search tasks. In *Proceedings of the ACM CIKM Conference on Information and Knowledge Management*. 1885–1889.

[71] Ahmed Hassan Awadallah, Cathal Gurrin, Mark Sanderson, and Ryen W White. 2019. Task Intelligence Workshop@ WSDM 2019. In *Proceedings of the ACM WSDM International Conference on Web Search and Data Mining*. 848–849.

[72] William Hersh, Jeffrey Pentecost, and David Hickam. 1996. A task-oriented approach to information retrieval evaluation. *Journal of the American Society for Information Science* 47, 1 (1996), 50–56.

[73] H Ulrich Hoppe and Franz Schiele. 1992. Towards task models for embedded information retrieval. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*.

[74] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159–166.

[75] Jeff Huang, Ryen W White, Georg Buscher, and Kuansan Wang. 2012. Improving searcher models using mouse cursor activity. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 195–204.

[76] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems* 38, 3 (2020), 1–32.

[77] Peter Ingwersen and Kalervo Järvelin. 2006. *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer.

[78] Richard David Jacques. 1996. *The nature of engagement and its role in hypermedia evaluation and design*. Ph. D. Dissertation. South Bank University.

[79] Kalervo Järvelin, Susan L Price, Lois ML Delcambre, and Marianne Lykke Nielsen. 2008. Discounted cumulated gain based evaluation of multiple-query IR sessions. In *Proceedings of the European Conference on Information Retrieval*. 4–15.

[80] Kalervo Järvelin, Pertti Vakkari, Paavo Arvola, Feza Baskaya, Anni Järvelin, Jaana Kekäläinen, Heikki Keskustalo, Sanna Kumpulainen, Miamaria Saastamoinen, Reijo Savolainen, et al. 2015. Task-based information interaction evaluation: The viewpoint of program theory. *ACM Transactions on Information Systems* 33, 1 (2015), 1–30.

[81] Jiepu Jiang and James Allan. 2016. Correlation between system and user metrics in a session. In *Proceedings of the ACM CHIIR on Conference on Human Information Interaction and Retrieval*. 285–288.

[82] Jiepu Jiang, Daqing He, and James Allan. 2014. Searching, browsing, and clicking in a search session: changes in user behavior by task and over time. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 607–616.

[83] Karen Sparck Jones. 2006. What's the value of TREC: is there a gap to jump or a chasm to bridge? *ACM SIGIR Forum* 40, 1 (2006), 10–20.

[84] Rosie Jones and Kristina Lisa Klinkner. 2008. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM conference on Information and knowledge management*. 699–708.

[85] Diane Kelly. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval* 3, 1-2 (2009), 1–224.

[86] Diane Kelly. 2015. When effort exceeds expectations: A theory of search task difficulty. In *Proceedings of the ECIR Workshop on Supporting Complex Search Tasks*.

[87] Diane Kelly, Jamie Arguello, and Robert Capra. 2013. NSF Workshop on Task-Based Search Systems. *SIGIR Forum* 47, 2 (2013), 116–127.

[88] Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wan-ching Wu. 2015. Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework. In *Proceedings of the ACM ICTIR International Conference on the Theory of Information Retrieval*. 101–110.

[89] Diane Kelly and Nicholas J Belkin. 2004. Display time as implicit feedback: understanding task effects. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 377–384.

[90] Diane Kelly and Colleen Cool. 2002. The effects of topic familiarity on information search behavior. In *Proceedings of the ACM/IEEE-CS JCDL Joint Conference on Digital Libraries*. 74–75.

[91] Diane Kelly, Susan Dumais, and Jan O Pedersen. 2009. Evaluation challenges and directions for information-seeking support systems. *Computer* 3 (2009), 60–66.

[92] Jeonghyun Kim. 2006. Task difficulty as a predictor and indicator of web searching interaction. In *CHI EA '06: CHI '06 Extended Abstracts on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 959–964.

[93] Youngho Kim, Ahmed Hassan, Ryen W White, and Imed Zitouni. 2014. Modeling dwell time to predict click-level satisfaction. In *Proceedings of the ACM WSDM International Conference on Web Search and Data Mining*. 193–202.

[94] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding user satisfaction with intelligent assistants. In *Proceedings of the ACM CHIIR Conference on Human Information Interaction and Retrieval*. 121–130.

[95] Alexander Kotov, Paul N Bennett, Ryen W White, Susan T Dumais, and Jaime Teevan. 2011. Modeling and analysis of cross-session search tasks. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 5–14.

[96] Birger Larsen, Christina Lioma, and Arjen de Vries. 2012. Report on TBAS 2012: workshop on task-based and aggregated search. *ACM SIGIR Forum* 46, 1 (2012), 71–77.

[97] Yuelin Li. 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 (2009), 275–291.

[98] Yuelin Li. 2010. An exploration of the relationships between work tasks and users' interaction performance. In *Proceedings of the ASIS&T Annual Meeting*. 1–9.

[99] Yuelin Li and Nicholas J Belkin. 2008. A faceted approach to conceptualizing tasks in information seeking. *Information Processing and Management* 44, 6

11

(2008), 1822–1837.

[100] Zhen Liao, Yang Song, Li-wei He, and Yalou Huang. 2012. Evaluating the effectiveness of search task trails. In *Proceedings of the 21st international conference on World Wide Web*. 489–498.

[101] Zhen Liao, Yang Song, Yalou Huang, Li-wei He, and Qi He. 2014. Task trail: An effective segmentation of user search behavior. *IEEE Transactions on Knowledge and Data Engineering* 26, 12 (2014), 3090–3102.

[102] Henry Lieberman et al. 1995. Letizia: An agent that assists web browsing. *IJCAI (1)* 1995 (1995), 924–929.

[103] Shin-Jeng Lin and Nicholas J Belkin. 2005. Validation of a model of information seeking over multiple search sessions. *Journal of the American Society for Information Science and Technology* 56, 4 (2005), 393–415.

[104] Jonathan Liono, Johanne R Trippas, Damiano Spina, Mohammad S Rahaman, Yongli Ren, Flora D Salim, Mark Sanderson, Falk Scholer, and Ryen W White. 2019. Building a benchmark for task progress in digital assistants. In *Proceedings of the WSDM Task Intelligence Workshop*.

[105] Jingjing Liu, Michael J Cole, Chang Liu, Ralf Bierig, Jacek Gwizdka, Nicholas J Belkin, Jun Zhang, and Xiangmin Zhang. 2010. Search behaviors in different task types. In *Proceedings of the Annual Joint Conference on Digital Libraries*. 69–78.

[106] Jingjing Liu, Jacek Gwizdka, Chang Liu, and Nicholas J. Belkin. 2010. Predicting Task Difficulty for Different Task Types. In *Proceedings of the ASIS&T Annual Meeting*. 1–16.

[107] Jiqun Liu, Matthew Mitsui, Nicholas J Belkin, and Chirag Shah. 2019. Task, information seeking Intentions, and user behavior: Toward a multi-level understanding of web search. In *Proceedings of the ACM CHIIR Conference on Human Information Interaction and Retrieval*. 123–132.

[108] Jiqun Liu, Shawon Sarkar, and Chirag Shah. 2020. Identifying and predicting the states of complex search tasks. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 193–202.

[109] Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri, and Gabriele Tolomei. 2011. Identifying task-based sessions in search engine query logs. In *Proceedings of the fourth ACM international conference on Web search and data mining*. 277–286.

[110] Jiyun Luo, Christopher Wing, Hui Yang, and Marti Hearst. 2013. The water filling model and the cube test: multi-dimensional evaluation for professional search. In *Proceedings of the ACM CIKM Conference on Information and Knowledge Management*. 709–714.

[111] Christoph Lutz. 2019. Digital inequalities in the age of artificial intelligence and big data. *Human Behavior and Emerging Technologies* 1, 2 (2019), 141–148.

[112] Rishabh Mehrotra, A Hassan Awadallah, Ahmed E Kholy, and Imed Zitouni. 2017. Hey Cortana! Exploring the use cases of a Desktop based Digital Assistant. In *Proceedings of the SIGIR Workshop on Conversational Approaches to Information Retrieval*.

[113] Rishabh Mehrotra, Ahmed Hassan Awadallah, and Emine Yilmaz. 2018. Learnir: Wsdm 2018 workshop on learning from user interactions. In *Proceedings of the ACM WSDM International Conference on Web Search and Data Mining*. 797–798.

[114] Rishabh Mehrotra, Prasanta Bhattacharya, and Emine Yilmaz. 2016. Characterizing users' multi-tasking behavior in web search. In *Proceedings of the ACM CHIIR Conference on Human Information Interaction and Retrieval*. 297–300.

[115] Rishabh Mehrotra and Emine Yilmaz. 2015. Terms, topics & tasks: Enhanced user modelling for better personalization. In *Proceedings of the ACM ICTIR International Conference on The Theory of Information Retrieval*. 131–140.

[116] Rishabh Mehrotra and Emine Yilmaz. 2015. Terms, topics & tasks: Enhanced user modelling for better personalization. In *Proceedings of the 2015 international conference on the theory of information retrieval*. 131–140.

[117] Rishabh Mehrotra and Emine Yilmaz. 2015. Towards hierarchies of search tasks and subtasks. In *Proceedings of the International Conference on the World Wide Web*. 73–74.

[118] Rishabh Mehrotra and Emine Yilmaz. 2017. Task embeddings: Learning query embeddings using task context. In *Proceedings of the 2017 ACM on conference on information and knowledge management*. 2199–2202.

[119] Rishabh Mehrotra, Emine Yilmaz, and Manisha Verma. 2014. Task-Based User Modelling for Personalization via Probabilistic Matrix Factorization. In *RecSys Posters*.

[120] Bhaskar Mitra. 2015. Exploring session context using distributed representations of queries and reformulations. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 3–12.

[121] Matthew Mitsui and Chirag Shah. 2018. The broad view of task type using path analysis. In *Proceedings of the ACM ICTIR International Conference on Theory of Information Retrieval*. 131–138.

[122] Matthew Mitsui and Chirag Shah. 2019. Bridging gaps: Predicting user and task characteristics from partial user information. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 415–424.

[123] Stefano Mizzaro. 1997. Relevance: The whole history. *Journal of the American Society for Information Science* 48, 9 (1997), 810–832.

[124] Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. 2017. Incorporating user expectations and behavior into the measurement of search effectiveness. *ACM Transactions on Information Systems* 35, 3 (2017), 1–38.

[125] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems* 27, 1 (2008), 1–27.

[126] George D Montanez, Ryen W White, and Xiao Huang. 2014. Cross-device search. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*. 1669–1678.

[127] Meredith Ringel Morris and Eric Horvitz. 2007. S 3: storable, shareable search. In *IFIP Conference on Human-Computer Interaction*. Springer, 120–123.

[128] Karen Myers, Pauline Berry, Jim Blythe, Ken Conley, Melinda Gervasio, Deborah L McGuinness, David Morley, Avi Pfeffer, Martha Pollack, and Milind Tambe. 2007. An intelligent personal assistant for task and time management. *AI Magazine* 28, 2 (2007), 47–47.

[129] Hugo Neri and Fabio Cozman. 2020. The role of experts in the public perception of risk of artificial intelligence. *AI & SOCIETY* 35, 3 (2020), 663–673.

[130] Elnaz Nouri, Robert Sim, Adam Fourney, and Ryen W White. 2020. Proactive suggestion generation: Data and methods for stepwise task assistance. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1585–1588.

[131] Heather L O'Brien and Elaine G Toms. 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 (2008), 938–955.

[132] Brendan O'Connor, Michel Krieger, and David Ahn. 2010. Tweetmotif: Exploratory search and topic summarization for twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 4.

[133] Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. 2007. In Google we trust: Users' decisions on rank, position, and relevance. *Journal of computer-mediated communication* 12, 3 (2007), 801–823.

[134] Fabio Paterno. 1999. *Model-based design and evaluation of interactive applications*. Springer Science & Business Media.

[135] Fabio Paternò. 2004. ConcurTaskTrees: an engineered notation for task models. *The handbook of task analysis for human-computer interaction* (2004), 483–503.

[136] Soo Young Rieh, Kevyn Collins-Thompson, Preben Hansen, and Hye-Jung Lee. 2016. Towards searching as a learning process: A review of current perspectives and future directions. *Journal of Information Science* 42, 1 (2016), 19–34.

[137] Daniel E Rose and Danny Levinson. 2004. Understanding user goals in web search. In *Proceedings of the International Conference on the World Wide Web*. ACM, 13–19.

[138] Daniel M Russell. 2019. *The Joy of Search: A Google Insider's Guide to Going Beyond the Basics*. MIT Press.

[139] Miamaria Saastamoinen and Kalervo Järvelin. 2017. Search task features in work tasks of varying types and complexity. *Journal of the Association for Information Science and Technology* 68, 5 (2017), 1111–1123.

[140] Tetsuya Sakai. 2012. Evaluation with informational and navigational intents. In *Proceedings of the International Conference on the World Wide Web*. 499–508.

[141] Tefko Saracevic. 2007. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology* 58, 13 (2007), 2126–2144.

[142] Shawon Sarkar, Matthew Mitsui, Jiqun Liu, and Chirag Shah. 2020. Implicit information need as explicit problems, help, and behavioral signals. *Information Processing & Management* 57, 2 (2020), 102069.

[143] Reijo Savolainen. 2012. Expectancy-value beliefs and information needs as motivators for task-based information seeking. *Journal of Documentation* 68, 4 (2012), 492–511.

[144] Chirag Shah. 2017. *Social Information Seeking*. Springer.

[145] Chirag Shah. 2018. Information Fostering-Being Proactive with Information Seeking and Retrieval: Perspective Paper. In *Proceedings of the ACM CHIIR Conference on Human Information Interaction and Retrieval*. 62–71.

[146] Chirag Shah and Ryen W White. 2020. Tutorial on Task-Based Search and Assistance. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 2436–2439.

[147] Guy Shani and Asela Gunawardana. 2011. Evaluating recommendation systems. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer, 257–297.

[148] Xuehua Shen, Bin Tan, and ChengXiang Zhai. 2005. Context-sensitive information retrieval using implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. 43–50.

[149] Ben Shneiderman. 2000. Creating creativity: user interfaces for supporting innovation. *ACM Transactions on Computer-Human Interaction* 7, 1 (2000), 114–138.

[150] Adish Singla, Ryen White, and Jeff Huang. 2010. Studying trailfinding algorithms for enhanced web search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 443–450.

[151] Mark D Smucker and Charles LA Clarke. 2012. Time-based calibration of effectiveness measures. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval.* 95–104.

[152] Yang Song and Qi Guo. 2016. Query-less: Predicting task repetition for nextgen proactive search and recommendation engines. In *Proceedings of the International Conference on the World Wide Web.* 543–553.

[153] Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *proceedings of the 24th ACM international on conference on information and knowledge management.* 553–562.

[154] Ayah Soufan, Ian Ruthven, and Leif Azzopardi. 2021. Untangling the concept of task in information seeking and retrieval. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval.* 73–81.

[155] Amanda Spink, Minsoo Park, Bernard J Jansen, and Jan Pedersen. 2006. Multi-tasking during web search sessions. *Information Processing & Management* 42, 1 (2006), 264–275.

[156] Simone Stumpf, Margaret Burnett, Volkmar Pipek, and Weng-Keen Wong. 2012. End-user interactions with intelligent and autonomous systems. In *Proceedings of the ACM SIGCHI Extended Abstracts on Human Factors in Computing Systems.* 2755–2758.

[157] Jean Tague-Sutcliffe. 1992. Measuring the informativeness of a retrieval process. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval.* 23–36.

[158] Arthur R Taylor, Colleen Cool, Nicholas J Belkin, and William J Amadio. 2007. Relationships between categories of relevance criteria and stage in task completion. *Information Processing and Management* 43, 4 (2007), 1071–1084.

[159] Jaime Teevan, Kevyn Collins-Thompson, Ryen W White, and Susan Dumais. 2014. Slow search. *Commun. ACM* 57, 8 (2014), 36–38.

[160] Gabriele Tolomei, Salvatore Orlando, and Fabrizio Silvestri. 2010. Towards task-based search and recommender systems. In *Proceedings of the IEEE International Conference on Data Engineering Workshops.* 333–336.

[161] Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. 2017. A deeper look at dataset bias. In *Domain adaptation in computer vision applications.* Springer, 37–55.

[162] E Toms and H O'Brien. 2009. The ISSS Measurement Dilemma. *IEEE Computer* 42, 3 (2009), 48.

[163] Elaine G Toms. 2019. Information activities and tasks. *Information at work: information management in the workplace. London: Facet Publishing* (2019), 33–62.

[164] Elaine G Toms, Heather L O'Brien, Rick Kopak, and Luanne Freund. 2005. Searching for relevance in the relevance of search. In *Proceedings of the International Conference on Conceptions of Library and Information Sciences.* 59–78.

[165] Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In *CVPR 2011.* IEEE, 1521–1528.

[166] Pertti Vakkari. 1999. Task complexity, problem structure and information actions: Integrating studies on information seeking and retrieval. *Information Processing and Management* 35, 6 (1999), 819–837.

[167] Pertti Vakkari. 2001. A theory of the task-based information retrieval process: a summary and generalisation of a longitudinal study. *Journal of Documentation* 57, 1 (2001).

[168] Pertti Vakkari. 2003. Task-based information searching. *Annual Review of Information Science and Technology* 37, 1 (2003), 413–464.

[169] Manisha Verma and Emine Yilmaz. 2014. Entity oriented task extraction from query logs. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management.* 1975–1978.

[170] Ellen M Voorhees and Donna K Harman (Eds.). 2005. *TREC: Experiment and evaluation in information retrieval.* MIT Press, Cambridge.

[171] Thanh Vu, Alistair Willis, Son N Tran, and Dawei Song. 2015. Temporal latent topic user profiles for search personalisation. In *Proceedings of the European Conference on Information Retrieval.* 605–616.

[172] Marilyn Walker, Diane Litman, Candace A Kamm, and Alicia Abella. 1997. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and Conference of the European Chapter of the Association for Computational Linguistics.* 271–280.

[173] Hongning Wang, Yang Song, Ming-Wei Chang, Xiaodong He, Ahmed Hassan Awadallah, and Ryen W. White. 2014. Modeling action-level satisfaction for search task satisfaction prediction. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval.* 123–132.

[174] Hongning Wang, Yang Song, Ming-Wei Chang, Xiaodong He, Ryen W White, and Wei Chu. 2013. Learning to extract cross-session search tasks. In *Proceedings of the 22nd international conference on World Wide Web.* 1353–1364.

[175] Yu Wang, Xiao Huang, and Ryen W White. 2013. Characterizing and supporting cross-device search tasks. In *Proceedings of the sixth ACM international conference on Web search and data mining.* 707–716.

[176] Ryen W White. 2018. Skill discovery in virtual assistants. *Commun. ACM* 61, 11 (2018), 106–113.

[177] Ryen W White, Paul N Bennett, and Susan T Dumais. 2010. Predicting short-term interests using activity-based search context. In *Proceedings of the ACM CIKM Conference on Information and Knowledge Management.* 1009–1018.

[178] Ryen W White, Wei Chu, Ahmed Hassan, Xiaodong He, Yang Song, and Hongning Wang. 2013. Enhancing personalized search by mining and modeling task behavior. In *Proceedings of the 22nd international conference on World Wide Web.* 1411–1420.

[179] Ryen W White, Susan T Dumais, and Jaime Teevan. 2009. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the ACM WSDM International Conference on Web Search and Data Mining.* 132–141.

[180] Ryen W White, Adam Fourney, Allen Herring, Paul N Bennett, Nirupama Chandrasekaran, Robert Sim, Elnaz Nouri, and Mark J Encarnación. 2019. Multi-device digital assistance. *Commun. ACM* 62, 10 (2019), 28–31.

[181] Ryen W White and Ahmed Hassan Awadallah. 2019. Task duration estimation. In *Proceedings of the ACM WSDM International Conference on Web Search and Data Mining.* 636–644.

[182] Ryen W White, Ahmed Hassan Awadallah, and Robert Sim. 2019. Task Completion Detection: A Study in the Context of Intelligent Systems. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval.* 405–414.

[183] Ryen W White and Jeff Huang. 2010. Assessing the scenic route: measuring the value of search trails in web logs. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval.* 587–594.

[184] Ryen W White, Ashish Kapoor, and Susan T Dumais. 2010. Modeling long-term search engine usage. In *Proceedings of the International Conference on User Modeling, Adaptation, and Personalization.* 28–39.

[185] Ryen W White and Diane Kelly. 2006. A study on the effects of personalization and task information on implicit feedback performance. In *Proceedings of the 15th ACM international conference on Information and knowledge management.* 297–306.

[186] Barbara M Wildemuth. 2004. The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology* 55, 3 (2004), 246–258.

[187] Barbara M Wildemuth, Luanne Freund, and Elaine G Toms. 2014. Studies of search task complexity or difficulty. *Science and Technology* 62, 9 (2014), 1676–1695.

[188] Tom D Wilson. 1999. Models in information behaviour research. *Journal of Documentation* 55, 3 (1999), 249–270.

[189] Biao Xiang, Daxin Jiang, Jian Pei, Xiaohui Sun, Enhong Chen, and Hang Li. 2010. Context-aware ranking in web search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval.* 451–458.

[190] Iris Xie. 2008. *Interactive Information Retrieval in Digital Environments.* IGI Global.

[191] Iris Xie. 2009. Dimensions of tasks: influences on information-seeking and retrieving process. *Journal of Documentation* (2009).

[192] Ya Xu and David Mease. 2009. Evaluating web search using task completion time. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval.* 676–677.

[193] Grace Hui Yang, Xuchu Dong, Jiyun Luo, and Sicong Zhang. 2018. Session search modeling by partially observable Markov decision process. *Information Retrieval Journal* 21, 1 (2018), 56–80.

[194] Hui Yang, Dongyi Guan, and Sicong Zhang. 2015. The query change model: Modeling session search as a markov decision process. *ACM Transactions on Information Systems (TOIS)* 33, 4 (2015), 1–33.

[195] Jing Yao, Zhicheng Dou, Jun Xu, and Ji-Rong Wen. 2021. RLPS: A Reinforcement Learning–Based Framework for Personalized Search. *ACM Transactions on Information Systems (TOIS)* 39, 3 (2021), 1–29.

[196] Emine Yilmaz, Manisha Verma, Nick Craswell, Filip Radlinski, and Peter Bailey. 2014. Relevance and effort: An analysis of document utility. In *Proceedings of the ACM CIKM Conference on Information and Knowledge Management.* 91–100.

[197] Emine Yilmaz, Manisha Verma, Rishabh Mehrotra, Evangelos Kanoulas, Ben Carterette, and Nick Craswell. 2015. Overview of the TREC 2015 Tasks Track.. In *Proceedings of the Text Retrieval Conference.*

[198] Yi Zhang, Sujay Kumar Jauhar, Julia Kiseleva, Ryen White, and Dan Roth. 2021. Learning to decompose and organize complex tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 2726–2735.

[199] Yongfeng Zhang, Min Zhang, Yiqun Liu, Chua Tat-Seng, Yi Zhang, and Shaoping Ma. 2015. Task-based recommendation on a web-scale. In *Proceedings of the IEEE International Conference on Big Data.* 827–836.

[200] Shoshana Zuboff. 2019. Surveillance capitalism and the challenge of collective action. In *New labor forum*, Vol. 28. SAGE Publications Sage CA: Los Angeles, CA, 10–29.

13