ADOPTING A GRAPHICAL PERSPECTIVE IN INTERACTIVE INFORMATION RETRIEVAL RESEARCH

by

MATTHEW MITSUI

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Computer Science

Written under the direction of

Chirag Shah

and approved by

New Brunswick, New Jersey

OCTOBER, 2018

ABSTRACT OF THE DISSERTATION

Adopting a Graphical Perspective in Interactive Information Retrieval Research

by Matthew Mitsui

Dissertation Director: Chirag Shah

Previous work in interactive information retrieval (IIR) has explored the relationships between individuals' search behavior, the characteristics of their search tasks, and their perceptions of their tasks, such as perceived topic familiarity and task difficulty. This work ultimately serves goals like personalization and search satisfaction. It is believed that predictions of task characteristics or searcher characteristics from observed behavior can help tailor search experiences to support task completion and search satisfaction. Often, research examines changes in behaviors when one or two characteristics change at a time. It applies methods such as t-tests, ANOVAs, and multivariate regression. This dissertation shows the limitations of this empirical framework. The contribution of this dissertation is in demonstrating that task characteristics, user characteristics, and behaviors should be empirically studied as a network of dependencies. It expands empirical work using graphical modeling, which can uniquely capture phenomena such as mediation and conditional independence. Research questions regarding mediation and conditional independence can hence now be answered with this different framework. This dissertation empirically shows when knowledge about behavior and certain task characteristics can be used to learn about other aspects of the task. It shows how task and user characteristics simultaneously affect behavior while potentially affecting each other. Specifically applying path analysis and Bayesian structure learning, results are shown to agree well with past literature and to also extend our understanding of the information seeking process. This dissertation discusses and shows the benefits and challenges of this modeling approach.

Acknowledgements

I'd like to think that life – albeit brief for me thus far – has imparted some pieces of wisdom to me over the years. The first is that you sometimes meet people who eventually know you better than you know yourself. The second is that even with incredible perseverance and individual effort, luck and circumstance can be bigger influences in our lives than we realize. I'd like to take the time to thank the good people I've been lucky to meet over the years, some of whom know me better than I know myself and all of whom have pushed me to be a better person. And without them, this PhD certainly would not have been possible.

First and foremost, I would like to sincerely thank my dissertation advisor, Prof. Chirag Shah, for his guidance and rigorous support. "Like a hawk" would do well to describe his acuity, perception, and diligence, but like all fledglings, PhD students must eventually leave the nest. He is the first listed here who has perhaps come to know me better than I know myself, and he helped push me toward success. I am grateful to him for opening me to the world of information retrieval research that is such a large part of my life today. I am grateful for his encouragement throughout all stages of the PhD process, as this dissertation took many shapes and forms and as my path went through its ups and downs. I am grateful to him for encouraging me to ask the right questions, in order to enable me to succeed in the world of research.

I would like to thank my committee members, Prof. Amélie Marian, Prof. Yongfeng Zhang, and Prof. Emine Yilmaz for their invaluable guidance. All of them have provided useful and encouraging advice which helped to inform and solidify the dissertation presented here. Thank you.

I would like to thank the various professors with whom I've interacted extensively in the Department of Computer Science. This includes our past and present Graduate and Department chairs, including Prof. William Steiger, Prof. Matthew Stone, and Prof. Thu Nguyen. As 2 years of my life were also dedicated to teaching, I would like to thank Prof. Martin Farach-Colton, Prof. Kostas Bekris, and Prof. Brian Russell, under whom I was a Teaching Assistant. I can only hope to have fulfilled a fraction of my role here as an effective teaching assistant to the hundreds of students I've encountered. I would also like to thank the professors who encouraged me to even consider graduate school, which would not have even been an afterthought for a young undergraduate student from Southern New Jersey. Specifically, I would like to thank Prof. Alexander Schliep and the late Prof. Michael D. Grigoriadis in this regard.

I would likewise like to thank all sources of funding during my PhD. These include the Department of Computer Science, my advisor, the Rutgers Perceptual Science IGERT, the Institute of Museum and Library Services, the National Science Foundation, and various conference and summer school travel grants. Without these, the PhD would certainly have been impossible.

I would also like to thank various members of the InfoSeeking laboratory for their assistance throughout the years. Many people have come through this laboratory during my stay, and I apologize for missing anyone: Dr. Chathra Hendahewa, Dr. Roberto González-Ibañez, Dr. Dongho Choi, Dr. Erik Choi, Dr. Ziad Matni, Dr. Long Le, Ruoyuan Gao, Souvick Ghosh, Jiqun Liu, Soumik Mandal, Jonathan Pulliza, Manasa Rath, Shawon Sarkar, Shannon Tomlinson, Yiwei Wang, Abdurahman Sherif, SeoYoon Sung, Kevin Albertson, Anastasia Ryssiouk, SaraAnn Stanway, and Diana Soltani. I would like to thank people who have allowed me to use datasets to incorporate into my dissertation, including Souvick Ghosh, Jiqun Liu, and Yiwei Wang. I would like to give special thanks to Chathra Hendahewa for taking me under her wing during my first year in the InfoSeeking laboratory and assisting me with my transition process.

I would like to thank the administrative staff at both the School of Communication and Information and the Department of Computer Science for their assistance in various administrative matters, allowing me to basically have a home in two departments.

Friends have come and gone over the years. But I'd like to thank those whom I've met at the various summer schools and conferences for great conversations, tips, advice, commiserating, and generally making the whole PhD process more fun and more memorable. I'd also like to thank my friends outside of my field who have helped give me perspective. Without them, I could neither see the forest nor the trees. If I were to list everyone, that would take several pages, and going into adequate detail would merit another dissertation. But thank you for all the great conversations at Rutgers, Pino's, various conferences, summer schools, the ever omnipresent social media, in my travels domestic and abroad, and beyond. I look forward to seeing everyone at some point in life beyond the PhD.

I would like to thank my parents, who have also given me much unconditional support throughout the years, especially during the toughest times. I'd like to thank them for raising me and encouraging me to go beyond my limits.

And lastly, heartfelt gratitude goes toward my partner and fellow researcher, Dr. Stephanie Anglin, the last (but certainly not least) listed here who has come to know me better than I know myself. She has taught me about patience and love and has perhaps also taught me to how to be a better person. I enjoy and appreciate our candid but energized talks about research, and I cherish the love and support we share. Thank you for being my best friend.

Table of Contents

Chapter 1

Introduction

Google and Google-like search engines have helped define what it means to be a modern search engine. Google and its competitors have pushed paradigms like “top 10 blue links” and “knowledge graphs” into the common language and style of search engine result presentation. The top 10 results or “blue links” proceeding a user’s input query are a common staple of modern search engine result pages (SERPs), and “knowledge graphs” are summarized in cards that are interleaved into results [14]. Changes to search interfaces and algorithms are optimized to provide better results for specific queries as soon as possible. Decades of extensive research have further improved this style of search result presentation and optimization [41, 33, 145]. Google and its competitors have become the champions of answering individual queries, such as “atlanta flights”, “leonardo dicaprio’s age”, “usa current news”, and “wikipedia”.

Recent experience as of 2018 shows that engines offer varying levels of support for queries. A query like “What is the distance in kilometers from the Earth to the moon?” can yield an immediate answer at the top of the results page. Even broad but succinct information about some specific person or thing - such as “Leonardo DiCaprio” - is often provided in a brief summary displaying age, birthplace, family relations, related movies, and recent news.

As searches become more complex, the caliber of support begins to break down. As of the time this dissertation was written, a search for “trip to japan” may yield top cities to visit and initial flight offers, as well as some information on the country. It offers only an initial starting point for what will presumably be a final trip consisting of a flight plan, a list of hotels, and a list of tourist destinations in Japan. If a searcher must determine “whether it is worth getting a PhD in Business”, an engine eventually resort to only supplying the top 10 links. In addition, a person more experienced in planning trips to Japan may start with more specific queries, such as “shinkansen tickets japan”.

At a glance, perhaps engines are *good at answering, provided that we know what to ask*. Sometimes users issue poor queries, but thankfully, efforts in query recommendation assist users with such struggles [78, 129]. But there is an extent to which *the intrinsic search need may be difficult to operationalize.* Determining whether a PhD in Business is worthwhile cannot be satisfied in one query. *Concepts may need to be clarified*, for instance whether the PhD is a Doctorate in Business Administration or a Business PhD. *A searcher may be looking for very specific items*, such as the key events leading to America’s involvement in World War II, or *may have less well-defined goals*, as in planning a leisurely non-business trip. Moreover, a searcher *may not just want to locate facts.* Modern search engines can accommodate fact-finding. However, people may also look to produce new insights, as with a cost-benefit analysis. Such goals extend beyond what documents commonly provide explicitly but are related to the type of content that can be retrieved.

One way to view these cases is that users bring *tasks* to a search engine that vary in structure. Search engines optimize users’ queries, but users’ tasks - and not their queries - ultimately drive their behavior. Search engines do not directly assist in accomplishing the larger (possibly more complex) goal. Researchers of IR increasingly recognize that searching behavior is often influenced by task context. In addition, there is acknowledgment that other user-specific contexts also influence behavior. Some are related to the task, such as topic familiarity, and others only pertain to the user, such as general search expertise. The field of Interactive Information Retrieval (IIR) studies the relationship between search activity and context such as task.

From the above, we see that:

- Modern search engines optimize search results, given a query.
- Querying requires input and implicitly requires that a user knows what to search for. Current search engines require input (e.g., queries) to work best and to optimize results.

- Search engines rarely consider the type of task a person is working on when optimizing results or suggesting what actions a person should take [48].
- Tasks vary in their characterizations. One example trait is the degree to which information is known by the user. At one extreme are the most complex types of search tasks [21] or exploratory search tasks [132]
- Users also vary in their characterizations. They vary generally, such as whether they are more or less expert with using search engines generally. They also vary with respect to a task, for instance regarding their familiarity to a topic.
- These contexts are known by IIR researchers to influence behavior.

In addition, knowledge of context such as task can not only be leveraged to provide whole-task support but can also be used to improve traditional retrieval. [82] Hence, not only is research in search tasks and user characteristics worthwhile for academic purposes, but also for practical purposes in assisting searchers.

1.1 Task and Interactive Information Retrieval

It is hence at a premium to determine the structure of a user's task, and we know this can vary. How well-defined is the searcher's goal? Is the user simply looking to find a fact or set of facts, or is she synthesizing new insights? Perhaps users can be prompted about their task characteristics, but can the structure be inferred from their behaviors? Task research in IIR has largely examined relationships between behaviors and task characteristics. This has been done by determining whether single behaviors (or groups of behaviors) can distinguish different types of tasks. Similar methodology has been used to relate behaviors to other personal contextual characteristics.

- Knowledge of task complexity can help inform which information to provide to a user, or this knowledge can be used for ranking [82].
- Knowledge of general task type can also influence implicit feedback [135]. Even when controlling for topic, task type can influence relevance judgments, hence influence traditional ranking [61].

- Task behavior has been shown to be valuable in modeling search satisfaction [46].
- Task knowledge can lend itself to subtask recommendation interfaces [48].

Vakkari mentioned that among several limitations in task-related search, "the interaction of domain and systems knowledge as mediating factors between task and searching is neglected" [118]. Personal and contextual factors - while sometimes hard to control - should also be considered in how they affect behavior in addition to task's concurrent effect. For instance, both a task's complexity and interestingness could affect total time spent on a task. Interesting but simple tasks may consume as much time as uninteresting but complex tasks. Similarly, the number of pages viewed per search engine result page (SERP) is affected by whether the person is under time pressure. Similarly, topic familiarity - which is affected by topic - can in turn affect behaviors such as querying strategies.

1.2 Problem Definition and Contribution

Previous methodological approaches that compare task and behavior do not allow for this line of questioning. Studies have considered interaction effects between pairs of independent variables (e.g. task type and task difficulty) and their joint effect on behavior. Task research does not take a more holistic modeling approach, yet as this dissertation will show (and as Vakkari hinted [118]), such holistic modeling is implicit in information seeking theory. The main contribution of this dissertation is to bridge information seeking theory with empirical practice, showing how modeling complex relationships between variables is useful and essential in task-based research. This dissertation will bring graphical modeling (in particular, path analysis and Bayesian structure learning) to task research. It will also explain - both informally and mathematically - how adopting this modeling paradigm opens up new types of research questions in task research.

This dissertation demonstrates that task characteristics, user characteristics, and behaviors should be empirically studied as a network of dependencies, in particular with a method that conveys relationships such as mediation and conditional independence.

The dissertation empirically shows when knowledge about behavior and certain task characteristics can be used to learn about other aspects of the task. It shows how task and user characteristics simultaneously affect behavior while potentially affecting each other. The dissertation experimentally shows how this framework both confirms previous findings and gives insight to new ones.

The work presented here has several implications. The first is a framework for modeling the simultaneous interactions between task characteristics, user characteristics and behavior. The methodology and experiments presented here are first informed by past literature and theory, but they will also hopefully inform future studies, emphasizing the importance of considering several contextual features in IIR research. Hence, another implication is explicating the type of data required to model relationships between users and tasks, such as the importance of knowing about the time pressure felt by the user. The third implication lies in the results - both positive and negative. By comparing the results here to past literature, we can confirm or deny various claims about relationships between task and behavior. A fourth implication also lies in the results; the modeling used here suggests implications about the limitations of data collection. It is generally assumed that collecting more data is better for predicting contextual characteristics is better, but are there limits to what the data can tell? As a prediction algorithm learns more task or user characteristics, will the number of bookmarks or time spent on context pages still help predict unknown task characteristics, or do these things become independent?

Generally researching the relationship between task and behavior has several practical implications, particularly in a predictive context. Suppose a system can detect the type of task a person is working on. Knowledge of the task has been shown to improve performance of traditional information retrieval algorithms, such as implicit relevance feedback for boosting query performance [134, 82]. Dwell time, for instance, can be used to indicate the usefulness of pages, deciding usefulness based on dwell time should be tailored according to a searcher's task type [84]. When using decision rules to determine the usefulness of pages, the weight and importance of different behavioral features can vary depending on the task [82]. Additionally, a searcher's task can potentially influence the usefulness of various components of the traditional search interface, having real implications on the search options that should be offered to searchers. In government, library, and commercial domains, whether the user is engaging in a fact-finding/known-item task, information-seeking task to accomplish some other goal, or a learning task can affect the usefulness of information such as document date of creation and geographic location mentioned in the document [68]. Awadallah et al. previously suggested that in commercial search, there is potential for subtask recommendation, and they additionally acknowledged that current search systems do not provide adequate support for discovering aspects of a task that are worth exploring, and that further support for complex tasks is necessary [48].

1.3 Limitations

It is acknowledged that a particularly salient application is that of prediction. What aspects of the task can be predicted? Given data about some search sessions, can characteristics of future sessions be accurately inferred? Furthermore, can these characteristics be inferred mid-session, so as to assist the user? This is indeed an important and active line of research. Some work discussed in the Background of this dissertation has shown this to be a challenging and unsolved problem. The work in this dissertation will not directly address that problem and leaves it to future work. This dissertation focuses on the importance of contextual features in predicting task.

It should also be noted that this work largely relies on behavioral data collected in laboratory settings. Searchers were required to find information pertaining to carefully constructed tasks provided by researchers. Such studies have the benefit of being able to control task dimensions and collect rich data (e.g. surveys and eye tracking), but critics claim laboratory settings are not realistic. Log analysis, on the other side of the spectrum, involves realistic and large data, but several data cleaning steps and strong assumptions must be made to group users' behaviors into tasks [50]. Even then, rich information about task properties (such as task complexity whether the task is an exploratory task) is not possible to collect. There is no perfect study design, but this work relies on the laboratory study framework.

1.4 Methods

This dissertation will limit the scope of the task characteristics that are analyzed. Li and Belkin [79] have suggested that there are many facets along which to categorize a task, such as user knowledge, task complexity, the type of goal, and the number of sub-tasks. There are too many to discuss in a single dissertation and certainly too many to model in a single study or a small set of studies. We will focus on a subset of task characteristics (specifically task goal and task product). This dissertation will also only focus on a few of many types of user characteristics, such as task difficulty, time pressure, and topic familiarity. While this dissertation limits these variables, the graphical framework presented here broadens the types of questions that can be asked in IIR studies. It is also a framework that is generally applicable regardless of what aspects of task interest the reader. The reader may be interested in task complexity, task product, task goal, task open-endedness, and other variables not discussed here. Specific results may differ from the ones presented, but that in itself is a worthy subject of study.

1.5 Research Questions

The necessity and utility of graphical modeling in empirical IIR research will be justified by answering the following research questions:

1. **RQ1** - What is the nature of the influence between user background, user experience, task, and search behavior?

2. **RQ2** - To what extent do factors (including task) directly affect behavior versus indirectly through their effect on other session characteristics?

3. **RQ3** - To what extent does a more generalized modeling framework confirm or deny previous findings in task-based literature?

4. **RQ4** - What is the structure and size of data required for such an experimental framework?

1.6 Structure of the Dissertation

Chapter 2 of this dissertation contains the background. “Task” is an overloaded term, so it is necessary to discuss the many definitions of task in the literature. This section will also choose and justify the dissertation’s operating definition of task - in particular, the definition of Li and Belkin [79] that has been used in many IIR studies. The dissertation will discuss the aforementioned empirical work linking browsing behavior to task, either by statistical significance or prediction. This section will lastly cover empirical work elucidating the link between user characteristics and task/behavior, initiating the motivation for better models.

Chapter 3 will propose graphical modeling as a framework to bridge the shortcomings mentioned previously. It will open by discussing the common mathematical framework between the current empirical work and generalized linear models showing how the more complex models are a simple generalization of the current work. It will discuss the usefulness of graphical frameworks such as Bayesian modeling and structural equation modeling. It will specifically show how new questions about the mediation, independence, and separability of variables can be asked and answered with such a framework. The section will then discuss the methodology of designing such models, including data size requirements, how to determine relationships between variables, and data assumptions.

Chapter 4 will discuss the design of each experiment. There are 3 experiments in total. The first two complement each other and each answer RQ1, RQ2, and RQ3. The third experiment answers RQ4. 3 similar datasets collected in laboratory settings are used to facilitate replication, and each of these is discussed at the end of the section.

Chapter 5 discusses the results of each experiment, their implications, and how they relate to the research questions set forth in the introduction.

Chapter 6 concludes the dissertation by summarizing the main contributions and findings of this dissertation. Some cautionary notes are also provided regarding the interpretation of this work, as well as suggestions toward future work.

Chapter 2

Background

This chapter provides a review of related research in task classification and task prediction. First, we necessarily discuss the many definitions of task to acknowledge the breadth of literature, first with non-IR research and then with IR research. Then, one particular aspect of task will be selected for focus in this writing, namely a subset of Li and Belkin [79] that has been extensively studied in past IIR work. This follows with a review empirical work exploring the relationship between browsing behavior and task characteristics. This has been explored through statistical significance, classification, and prediction. This chapter will additionally describe other important user characteristics shown to be related to behaviors, such as topic knowledge, search experience, and time pressure. The chapter will describe the shortcomings of this literature. Namely, these works compare isolated user/task characteristics against changes in behavior. Such empirical work is theoretically incomplete in that it only captures a fraction of the theoretical conceptions of users and their tasks. This chapter hence shows that a more holistic modeling framework is not only necessary but desirable. The chapter serves as an overview of the literature, with an account of more specific details relegated to the Appendix.

2.1 Definitions of Task

2.1.1 Tasks in non-IR vs. IR settings

It is first worth acknowledging prior work on task outside of IR. “Task” is an overloaded term, but different disciplines have studied how characteristics of people’s tasks affect their behavior and task outcomes. Such research typically follows two guidelines. It typically defines different categories of tasks, either for all tasks or for tasks in a specific context. It also examines how variations in task type influence behavior. Interested readers may refer to the cited literature for a more thorough treatment of task in non-IR contexts.

Early abstractions of task began in the 1950s, starting with work such as that of Carter, Haythorn, and Howell. They constructed an activity-based classification of tasks, including tasks such as Intellectual construction, Mechanical assembly, and Discussion [25]. Hackman focused more specifically on reasoning tasks: Discussion tasks, Production tasks, and Problem-solving tasks. Hackman examined how these characteristics and the task difficulty affected performance of works in groups [44]. Hackman's typology of reasoning tasks was used in other work; Aronson, for instance, examined how these task types, environment, and group type affected people's reactions, social dynamics, and output quality [1]. Similarly, other researchers focused on very particular physical work environments, comparing behaviors across different task types. Whitley and Frost examined a research laboratory, in particular segmenting tasks into Responsibility, Extension, Development, and Research tasks. They examined "how task types influenced the selection of information sources and information dissemination channels in a research laboratory" [138]. Tushman [128] explored a R&D setting and in particular Basic research, Applied research, Development, and Technical service tasks. Tushman claimed these work characteristics affected technical communication, and communication and work characteristics in turn affect project effectiveness. Even outside of IR, task research centrally compares similarities and differences in behaviors across different types of tasks.

The above can be seen as *work tasks* that occur as part of employment or other real-life scenarios. Some involve no information seeking, such as the mechanical assembly of an item. When they involve some information seeking, they fall into the jurisdiction of IIR research. *Work tasks* that involve searching, according to Byström and Hansen, are comprised of *information seeking tasks*, for instance various information gathering tasks conducted for the sake of completing a work project at the office. Each IST, in turn, can be composed of several *information retrieval/search tasks*, with each *information* *retrieval task* comprising the satisfaction of some atomic information need (e.g., issuing a query and collecting relevant results) [20]. IIR research typically focuses on the *information seeking task*. Similarly to non-IR literature, tasks have undergone various categorizations to see how changes in task characteristics relate to changes in behavior. As in the previous task literature, *information seeking tasks* have been categorized along several dimensions by researchers. Task characteristics are implicitly understood as independent variables on which search outcomes depend [118].

2.1.2 A Holistic Taxonomy for Task

Several task characteristics have interested IIR researchers. Some examples include the following:

- **Fact-finding vs. Exploratory Tasks** - Bates acknowledged that not all searching consists of simple fact-finding [10]. Exploratory tasks are much more ill-defined and open-ended. They require multiple searches and even the clarification of goals and knowledge throughout the search process. Exploratory search has spawned several threads of research and workshops, advocating for new measures of evaluation and interfaces to support exploratory search [135, 137, 102, 26, 132, 115, 123].

- **Complexity** - The least complex search tasks can be highly automated, such as simple lookup searches or simple computations. The most complex tasks may require more deliberate unautomated decision making, and in these tasks, the required information or even expected result may not be known in advance [21], much like in exploratory search. Complexity may be categorized in terms of goal uncertainty. It may also be characterized by the necessary search paths to complete a task. Campbell characterized both the outcome and the multiplicity of paths to task completion, finding that complexity also increases information load, diversity, and information change [23]. The relationship between task complexity and information seeking behavior has largely studied in laboratory settings [65, 15].

- **User knowledge** - Searchers bear some personal characteristics relative to the task, which are also important. Several studies have examined how domain knowledge and general search expertise can affect a searcher's search strategies [100, 2, 122]. Other work has also examined how searchers' cognitive style, database search experience, and task type affect and Web search behavior and search outcomes [71, 72].
- Other discovered characteristics of tasks include the frequency of the task (one-shot or routine), the type of goal (explicit/concrete/specific or abstract/amorphous), the product (decision-making, producing new insights, or locating facts/data), and the source of motivation (internal/self-driven or external).

Li and Belkin observed that these various works on task - both IR-related and non-IR work - can be combined into a common comprehensive understanding of task. They performed a literature analysis and found that literature on task - both IR-related and non-IR related - only focused on a few task attributes at a time. Moreover, many works focused on common attributes, for instance, with several works focusing on open-ended tasks versus closed tasks [100, 70]. Roughly speaking, task characteristics can be organized into task characteristics such as this. Some are external to the user (e.g., source of the task, task doer, frequency of the task, duration, stage of the task, product/outcome of the task, the task process, and the type of goal), and some are user-dependent (e.g., task salience, urgency, difficulty, task complexity, knowledge of task topic, knowledge of task procedure). Each of these traits comprises of a set of possible values. For instance, the goal of the task may be very concrete or ill-defined, and a user can have varying degrees of topic knowledge. Li and Belkin's literature review provided a comprehensive classification of task encompassing all prior work on task classification [79], which has been cited in subsequent empirical work, described below. It is intended for use in classifying both *information seeking tasks* and *work tasks*. The full "faceted classification" is provided in Tables 2.1- 2.4.

2.2 Theory: Putting the User in Task

This characterization of task is more than a checklist. Several theoretical frameworks have been constructed showing how the task, work context, user, and even the search engine relate to each other. First, it is worth discussing "frameworks" generally in information retrieval. These often take the form of conceptual models. Conceptual models, according to Engelbart, require a specification of the following [37]:

- "Essential objects or components of the system to be studied.
- The relationships of the objects that are recognized.
- What kinds of changes in the objects or their relationships affect the functioning of the system - and in what ways.
- Promising or fruitful goals and methods of research." [55]

A plethora of information seeking models describe the information seeking process and its evolution from start to finish. To name a few conceptual models, Ellis enumerated different features of information seeking behavior, from the start of the process to the end. For instance, people may engage in browsing behavior, filtering through information sources, identifying relevant information, and even monitoring one's own performance. Other models have more broadly examined this as an iterative process [36]. Marcia Bates introduced the concept of berrypicking, where users' individual search queries are imagined as bushes and a searcher navigates between queries/bushes; this emphasized that future navigation behavior dependent on the user's state at any given moment [10]. Much more recent work graphically and probabilistically represented an iterative framework of the search process. For instance, Baskaya et al. modeled interactions of users as they formulate queries, scan result snippets, click links, read documents, judge relevance, and either stop the session or continue with another search [9]. An example of their framework can be found in Figure 2.1. Work by Maxwell further explicated this framework, adding the importance of "state" at each step of the process, including the importance of background knowledge, information need, and the user's subjective ideas of relevance at each step of issuing queries and evaluating results [103], though this was later used for the purposes of creating user simulations.

Several other frameworks instead focus on the relationship between the user, task, and other contexts. Ingwersen, for instance, developed a framework intending to emphasize the causal relationship between a user's information need, the problem state, the domain work task, and even task interest. While focusing largely on a cognitive perspective of information retrieval, this work emphasized mutual influence between the user's work environment, the structure of the problem, the user's information need, and ultimately behaviors such as querying [54]. Järvelin and Ingwersen later extended this model to discuss 9 different dimensions [58].

1. "The work task dimension covers the work task set by the organization, the social organization of work, collaboration between actors and the physical and system environment.
2. The search task dimension covers necessary seeking and retrieval practices, as understood collectively in organizational practice.
3. The actor dimension covers the actor's declarative knowledge and procedural skills, and other personal traits, such as motivation and emotions.
4. The perceived work task dimension covers the actor's perception of the work task: forming the task that is carried out.
5. The perceived search task dimension covers the actor's perception of the search task including information need types regarding the task and its performance process, and perceived information space.
6. The document dimension covers document contents and genres and collections in various languages and media, which may contain information relevant to the task as perceived by the actor.
7. The algorithmic search engine dimension covers the representation of documents or information and information needs. It also covers tools and support for query formulation and methods for matching document and query representations.

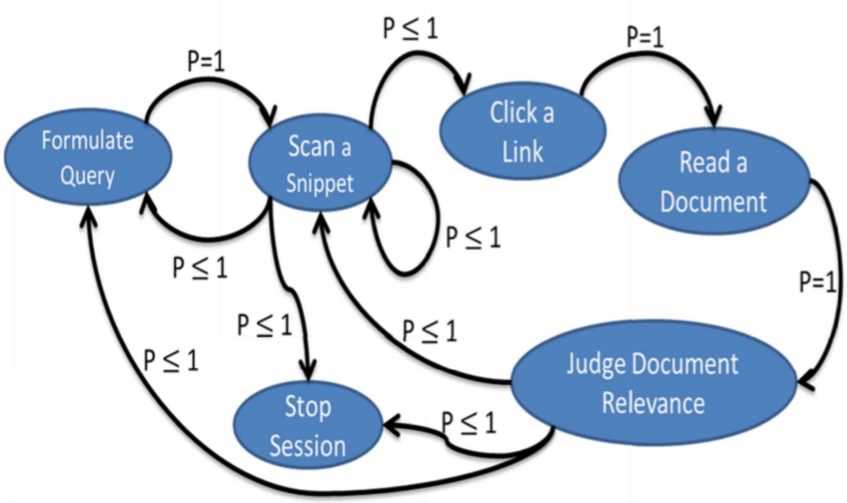

Figure 2.1: Process framework presented in [9].

8. The algorithmic interface dimension covers tools for visualization and presentation of information objects, collections and their organization.

9. The access and interaction dimension covers strategies of information access, interaction between the actor and the interface (both in social and in system contexts)."

They claimed that each dimension is composed of multiple variables, and that a subset should be explicated in any study depending on its aim. It moreover emphasized the importance of the complex interaction. Users exist in the context of the search systems they interact with and their socio-cultural context (e.g. their employment and the subsequent work tasks/information seeking tasks). But moreover, the socio-organizational context can influence the user, for instance through the user's interpretation of the task and understanding of the domain, which in turn influences behavior. This model is shown in Figure 2.2. In particular, this work stresses the importance of not only multiple variables to capture various aspects about the user, the search engine, the task context, and even the social/organizational context but also a complex relationship between them.

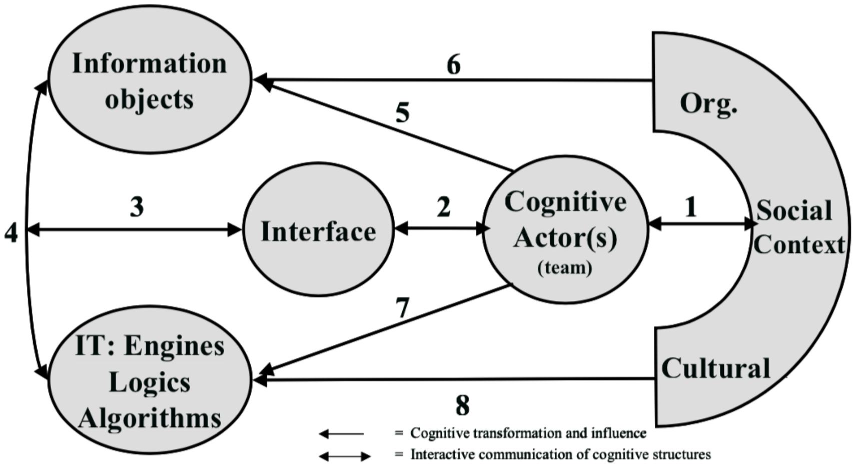
Information objects
Interface
IT: Engines Logics Algorithms
Cognitive Actor(s) (team)
Org.
Social Context
Cultural
1
2
3
4
5
6
7
8
= Cognitive transformation and influence
= Interactive communication of cognitive structures

Figure 2.2: Framework of actors in context, as in [56].

2.3 Characterizing Task from Behavior

Now that we have introduced frameworks in IIR, this chapter concludes with an overview of empirical work. It discusses empirical work relating task type to behaviors, followed by other work relating behavior to other characteristics like task complexity and user affect. It overviews the definitions of the constructs like complexity and task and notes the key accomplishments of this work. Details are left to the Appendix, but this section will conclude by briefly discussing the shortcomings of empirical work, with a more formal treatment in Chapter 3.

2.3.1 Task Type

Work examining the relationship between task type and behavior adopt different definitions of task for analysis. Much of the empirical work uses a subset of the taxonomy defined in Li and Belkin [79], which was shown in Tables 2.1- 2.4. Researchers manipulate a controlled set of tasks by manipulating their attributes. For instance, researchers have explicitly manipulated the goal - creating tasks with a well-defined goal, an abstract goal, or a mixture of the two [94, 91, 29, 30, 109, 110]. Researchers have likewise manipulated the task product - creating tasks about locating facts, about making a decision, or about producing new insights from facts [94, 91, 30, 109, 110]. Goal and product are common attributes, and some research has defined “task type” as the combination of goal and product, yielding 4-9 possible task types [94, 60, 30, 109, 110]. Other commonly manipulated attributes in the Li and Belkin taxonomy are the level of evaluation of a document (users must evaluate the whole document or only document segments) [94, 91, 29, 30] and the objective complexity of the task (many paths/high complexity or single path/low complexity) [94, 91, 30, 29]. Another body of work distinguishes task types by distinguishing between fact-finding and exploratory tasks or a similar distinction. Some work has presented users with tasks that are fact-finding/lookup or exploratory [4, 75]; other work examined fact-finding, information gathering, browsing, and transaction tasks [64, 119]. Another commonly manipulated task attribute is whether the task requires simple searches, hierarchical search patterns, or parallel search patterns on separate subtasks [83, 119]. Lastly, some work not only examined task characteristics in isolation but also their interactions with other important variables. One way this has been done is by modeling interaction effects [18]. Other work would determine whether a difference in behavior is significant under some condition but not another (e.g., whether “topic familiarity” is distinguished by behavior, but only within certain task types) [94].

Several behaviors were significantly different among manipulated task types. These include: the number of content pages and queries [119, 60, 91, 4, 109], the time spent on pages and queries [66, 60, 91, 4, 109], click and scroll depth [60, 4], query reformulation behaviors [83], several eye tracking features [91, 60, 30], and task completion time [119, 60]. Some of this work compared users’ behaviors over their entire sessions [60, 91], but other work compared their behaviors on the first query, to eventually aim for as-soon-as-possible task prediction [4]. A recent body of work has used more complex sequential features over the entire session. For instance, Cole et al. created a graphical representation of the user’s changes in cognitive activity through the duration of the search session; they distinguish tasks based on the properties of each task’s Markov graph of cognitive states [30]. Kotzyba et al. modeled sequential behavior in a session using Markov modeling and Hidden Markov modeling - where states were different pages types - to attempt task prediction [75].

One last recent development worth mentioning, then, is task prediction. Rather than checking for significance, can task characteristics such as goal and product be predicted from behaviors? Kotzyba et al.'s attempt demonstrated effective prediction distinguishing a session with a single exploratory task versus a session with multiple fact-finding tasks [75]. However, [109] used whole session browsing features similar to those in the previous work ([60, 91]), achieving statistically significant findings that entailed only marginally improved prediction performance. Such work combines several behaviors into a single predictive model, yet prediction results have been mixed.

The statistical tests used in the above works can be reduced to the following list: t-test, Mann-Whitney U test, one-way ANOVA, Kruskal Wallis, one-factor repeated measures ANOVA, and two-factor repeated measures ANOVA.

2.3.2 Complexity

Task complexity, while in part a subset of Li's task classification, has a separate thread of research that merits treatment. Complexity is an essential and important factor in IIR research [19, 130], though it has several definitions. All of them emphasize the objective aspect of a task's complexity; complexity is determined independently of the person doing the task. Campbell was one of the first to advocate for this model, specifically proposing that complexity is a function of: the number of possible paths to the outcome, the number of outcomes, interdependence among paths, and uncertainty between paths and outcomes [23]. Byström et al., in turn, was among the first to begin analyzing complexity at the level of individual tasks rather than larger projects that encompass tasks [21]. Works by Byström considered complexity as *a priori determinability*: how well information inputs, information search processes, and the task output can be known/determined in advance. Following this framework, her tasks in increasing order of complexity were: automatic information processing, normal information processing, normal decisions, known tasks, and genuine decision task [19]. Much more recent work by Capra et al. defined determinability in terms of whether certain aspects were specified. Capra et al. created tasks where users were to compare items along different dimensions; they manipulated whether a task description explicated the specific the items and/or dimensions [24]. Other work more specifically defines complexity by the multiplicity and structure of the paths to the outcome. The faceted classification above by Li lists highly complex tasks as involving many paths and low complexity ones as involving a single path [79]. Other complexity work is specifically outcome-focused, defining a simple search as a task satisfied by one piece of information, a hierarchical task as requiring multiple concepts structured in a nested hierarchy, and parallel task where the desired concepts are in the same level of the hierarchy rather than nested [83, 90, 126]. One last noteworthy definition of complexity is the amount of cognitive effort and learning required to complete a task; tasks in increasing order of complexity are organized into *remember, understand, analyze, evaluate*, and *create* tasks [15, 65].

Byström only performed qualitative analysis and found increases in determinability associated with: increase in complexity of information needed, increase in needs for domain information and problem solving information, increase in share of general-purpose sources, decrease in share of problem and fact-oriented sources, decrease in success of information seeking, decrease in internality of channels, and increase in number of sources [21]. Capra's determinability work found that tasks with a specified dimension had significantly greater queries, greater query length, fewer clicks per query, fewer bookmarks per query, smaller query log likelihood, greater number of unique queries. When both item and dimension are specified, there is greater time to first click and a greater number of unique URLs [24]. Research on high/low objective complexity found significant differences for time spent on a task, time spent per item selected, and total number of items selected (i.e., Word documents, PDFs, web pages, full-text papers, etc.) [80]. When comparing parallel vs. simple information needs, Liu et al. found that for query reformulations, specialization was most frequent in simple tasks and word substitution was more frequent in parallel tasks than simple ones [83]. Lastly, Brennan et al. showed that for Remember, Analyze, and Create tasks, there are significant differences in session length, total number of queries, query length, and total number of SERP clicks [15].

A couple works also examined interactions between complexity and other task characteristics. Toms et al. computed interaction effects between task complexity and task type on browsing behaviors, where task type was either decision making, fact finding, or information gathering and complexity was either parallel search or hierarchical [126]. Liu et al. conditioned on the task complexity (dependent vs. parallel task) and examined whether significant differences in dwell time for task stage and document usefulness varied according to the task complexity [90].

The statistical tests used in the above works can be reduced to the following list: Mann-Whitney U test, Kruskal-Wallus H test, one-factor repeated measures ANOVA, univariate ANOVA, multivariate ANOVA.

On a side note, complexity and difficulty may seem intertwined but can be distinguished. Complexity can be defined in terms of number of subtasks or steps, number of subtopics or facets, number of query terms and operators required, number of sources or items required, the indeterminate nature of the task, and the cognitive complexity of addressing the information need. Task difficulty involves attributes such as searcher performance, the match between terms in the task description and in the target page, the number of relevant documents in the collection, and the searchers' of experts' perceptions of difficulty. Complexity can be perceived in an objective sense independent of the task doer which has several primary dimensions. Multiplicity of steps or subtasks, multiplicity of facets where facets represent the concepts or types of concepts represented in the task (e.g. dates or genre), and some degree of indeterminability or uncertainty (i.e. for a complex task, the search process or outcomes cannot be determined in advance of completing the task) [139].

2.3.3 Task/Topic Familiarity and Expertise

There is less variability in the definitions of task familiarity, topic familiarity, and expertise than with task type of complexity. An early, often-cited work by Marchionini compared the search behaviors of 3rd/4th graders to 6th graders in a school setting, assuming the different grade levels had different levels of expertise using search engines (with 6th graders being more expert). The two groups used structurally different queries that suggested a different mental model for searching, and older searchers found required information more successfully, in less time [100]. This can be seen as a form of search expertise, and “expertise” often refers to expertise with search engines. Despite this common definition and despite long interest to the IR community, expertise has been difficult to measure. Bailey and Kelly, for instance, recently conducted an exploratory study to determine what constitutes search expertise [6]. One past effort by White and Morris defined expertise as a function of a searcher’s usage of advanced syntax operators [136], but this still an open area of research.

“Topic knowledge” or “domain knowledge” has been more deeply explored. This refers to someone’s general knowledge with respect to a broad topic domain or with respect to the topic of the particular search task at hand. A common method for measuring topic or domain knowledge is a self-assessment by the searcher before or after conducting a search task [67, 52, 143, 95, 89, 39]. A less common approach is to include a knowledge survey about the task topic before and/or after some training period, to measure an objective state of knowledge and changes in knowledge [140]. Another approach that does not require self-assessment is to monitor page behaviors of searchers over a period of time, assuming that domain or topic experts go to topically-specialized sites more often than novices [133]. Other researchers may recruit research participants according to field of general domain expertise, e.g., to compare psychologists and non-psychologists [112]. Some work also examined the possibility of interactions between familiarity and other task characteristics. Liu et al. compared whether the distinguishing difficult and easy tasks from behavior depended on whether the users were experts or novices [89]. Later work by Hienert et al. similarly found that whether there was a significant relationship between behavioral variables and perceived difficulty, perceived success, time pressure felt, and comprehension of the topic sometimes depended on the topic, and the two topics tested differed significantly in familiarity [51].

Topic knowledge and domain knowledge have been distinguished by several types of behaviors. The behaviors examined include content page and query browsing statistics [67, 95, 94], query reformulation patterns [52], and some eye tracking features [28]. They also include the amount of relevant information found [143] or search efficiency [67, 52], domain-specific word usage in queries [133], and the types of web domains that experts visited compared to non-experts [133]. These behaviors are often computed at the level of the whole session. One work compared the middle and end of a session: Liu et al. found that the mean first dwell time on a SERP for the middle and end of a session were significantly related to self-reported familiarity [95].

The statistical tests used in the above works can be reduced to the following list: t-test, Mann-Whitney U test, one-way ANOVA, Kruskal-Wallis H test, chi^2 test, Wilcoxon signed-rank test.

2.3.4 Task Difficulty

Task difficulty is to be distinguished from expertise and certainly to be differentiated from complexity. The previous section extensively treated *objective task complexity*. This section addresses the *subjective difficulty* experienced by searchers when searching [70]. As suggested, this is a subjective measure predominantly measured through a searcher's self-assessment. A common method of measurement is to ask a searcher about the difficulty experienced on a 5-point or 7-point likert scale. The scale is then divided into 3 or 2 points: "easy", "neutral", and "difficult" while possibly excluding the middle [43, 42, 87, 96, 85, 97]. Research in difficulty has also looked at "successful" and "unsuccessful" search sessions instead [5]. Other work determines the threshold between "easy" and "difficult" numerically, for instance by setting it as the mean in the data rather than the midpoint on a 1-5 scale [3, 89].

The behavioral features used in this research are largely similar to those in the previous sections. Some work examines behavioral features computed over the whole session, while other work - as previously - only uses features obtainable within the first query segment - including the pages and interactions before the second query. Some work also examines features per query segment; if a difficulty prediction cannot be made at the earliest step, perhaps it can be made in the middle. In whole session analysis, significant differences were found among various content page and query count statistics [43, 42, 97, 87, 3, 89, 5], query reformulation behaviors [83], mouse activity and viewing activity [3], time spent on content and queries [85, 43, 5, 87, 96], total task duration [42, 97, 3], bookmark statistics [42, 3, 96], and even eye tracking features on pages and queries [27]. Richer features worth mentioning include the linearity of a user's search path [43], the optimality of the search path [43], the richness of the searcher's query vocabulary [89], and the depth to which a user's queries explore an externally-controlled vocabulary [89]. First query segment features with significant differences include various content page and query count statistics [85, 96], time spent on content pages and query pages [85, 3, 96], mouse activities [3], viewing rank [3], and bookmark activities [3, 96]. Important per-query segment and mid-session features include: content page and query count statistics [97, 96], page and query dwell time statistics [97, 96], and query length statistics [5], and bookmark statistics [96].

A few researchers examined interaction. One work computed the interaction effect between difficulty and domain knowledge, finding significant differences for mean dwell time on content and the percentage of time on content pages [87]. Otherwise, this work conditions on the interacting variable and examines whether the significance of behavioral differences changes. One Liu et al. work found that the relationship between difficulty and behavior differed among different task types [92], and later work examined the effects of the difficulty on experts and novices separately, finding differences in query vocabulary richness, number of unique query terms, percent of terms from task description, and search session recall and F-score [89].

The statistical tests used in the above works can be reduced to the following list: Mann-Whitney U test, Kruskal-Wallis H test, logistic regression, 2nd order polynomial regression.

2.3.5 Time Pressure and Other Affective and Intentional Factors

There are several other search session characteristics worth mentioning. Namely, several affective components have been shown to affect user performance. Some of these are associated with changes in behavior, but others are associated with changes in other measures that in turn are known to affect behavior.

A recent one to consider is user engagement - the extent to which the user is engaged and invested in the search session. User engagement can be measured through a general User Engagement Scale questionnaire [114]. Much of this work is recent, but engagement has been associated with a few behavioral measures. Namely, it is associated with interest is associated with increase in SERP scrolling and time spent in query intervals [35] as well as increased task duration [34]. It is also associated with increases in prior knowledge [35], decreased perception of difficulty [35].

Recent work has also begun to explore user intentions, even though intentions in IIR are by no means a new concept. Early work by Marchionini categorized searchers' actions in a search session into types of tactics, moves, and strategies [101]. Xie later furthered this work by claiming that a search task leads to "interactive search intentions" and showed a relationship exists between strategies and high-level intentions such as locate a specific link and "learn domain knowledge" [142]. Descriptive non-inferential evidence from Mitsui et al. [111] suggested that these intentions are exhibited in different proportions among task types and perhaps can distinguish types. Rha et al. [120] showed that differences in reformulation strategies can be associated with differences the aforementioned intentions of [142]. Later work by Mitsui et al. showed that these intentions can be predicted at the query segment level, using machine learning methods with query browsing features as input. Mitsui et al. applied bookmark features, content page dwell time features, SERP dwell time features, query reformulation types, and query lengths, but the best approach was to generally use several browsing features all at once [107].

One last but important aspect is the amount of time pressure experienced by a user. This is often manipulated by the researcher in a laboratory setting. A researcher may impose 1 of 2 conditions on a searcher: a search task with no imposed time limit, or a search task with a very strict time limit. One such strict time limit in the literature is 5 minutes. Crescenzi et al. showed that imposing strict time constraints can increase feelings in time pressure as well as decrease task time, increase query rate, and facilitate shallower viewing in pages and documents per query, and less time is spent on each document and SERP [31]. Liu et al. found significant differences in whole session measures resulting in overall shallower behavior, such as a significant decrease in: number of content pages per query, number of unique content pages per query, number of SERPs per query, number of unique SERPs per query, total dwell time on content pages per query, total dwell time on SERPs per query, ratio of dwell time on content pages per query, average query interval time, average session time, and number of queries per session [88]. Time constraints can actually result in increased time spent on the first SERP [88].

The statistical tests used in the above works can be reduced to the following list: Mann-Whitney U test for engagement, machine learning techniques like regression and SVM for intentions, and mixed ANOVA and Mann-Whitney U test for time pressure.

2.4 Excursion: Log Task Extraction

There is another common body of IIR literature concerning task: namely, work on extracting coherent tasks from large-scale logs. It is only marginally related to the work here, but since it has been extensively studied, this work is worth mention and should also be acknowledged to make this review complete.

In some respects, the types of search queries and patterns found in the large scale logs and laboratory settings are similar. Just as various querying strategies have been found and studied in laboratory settings, searchers in large scale logs have been seen to conduct navigational, informational, and transactional searches [16]. Complex search tasks have also been identified in both laboratory and large-scale log settings, including exploratory search [102] and multi-step search [47].

Yet fundamental differences lie in the size and complexity of log data. Large-scale search logs provide important data that often dwarfs the size of laboratory study browsing behavior data. Yet log data is not just larger lab data. People may also be engaged in multiple tasks that are interleaved. Jones and Klinkner discovered that 17% of web search engine tasks were interleaved [62], and Lucchese et al. discovered that 75% of queries may involve multi-tasking activity [98]. Hagen et al. further distinguished three types of search episodes: *physical search sessions* (queries charcterized by time gaps), *logical search sessions* (consequtive queries for the same information need in a search sessions), and *search missions* (potentially disjoint logical sessions surrounding the same information need) [45]. A laboratory setting often provides a task and assumes that the searcher only works on that task, but research on log data explored methods of grouping queries into tasks in such noisy settings.

Jones and Klinkner, for instance, recognized that simple time-based cutoffs between queries to separate tasks were naïve and inaccurate; they combined temporal differences between queries, edit distances between queries, query co-occurrence features, and similarity between queries' search results to group queries into tasks and subtasks [62]. Other work created heterogeneous graph of topically related queries, topically related web pages, and connections between pages and queries indicating clicks [59]. Other work saw that users can perform similar tasks [99]. Future work took inspiration from Natural Language Processing. For instance, Li et al. combined topic modeling on queries using Latent Dirichlet Allocation) and temporal grouping from Hawkes processes to group temporally close queries on the same topic into tasks [77]. Verma and Yilmaz extracted named entities from queries using DBPedia and clustered queries hierarchically using this entity information [131]. Current state of the art uses more advanced NLP models, such as Mehrotra et al., which combined LDA, language models, and user-topic-task tensors to create models of task structure and users' task preferences [105]. Future work used word embeddings to determine tasks and sub-tasks [106, 104].

The goals and framework of this type of research tend to differ from those of a laboratory setting. One goal of such work is to properly extract tasks in a web-scale multi-tasking environment, with several of the previous works evaluating task and sub-task extraction on manually annotated data sets [62, 45, 106]. Another is to predict whether a task is likely to be continued in a future session [74]. Another is to predict and recommend future queries. Awadallah et al. used their task extraction model to predict future queries that would be useful to users engaged in a complex search task [48]. Mehrotra and Yilmaz used their framework to recommend queries from similar users, where similarity was based on users' task preferences [105]. Unlike the laboratory-based research described in previous sections, this research does not analyze or extract the search task characteristics. It is not known (or determined by an algorithm) whether these tasks have a simple or complex goal or whether the searcher is attempting to produce some new findings or retrieve simple facts. Both lines of research address complementary agendas. The web-scale work addresses the extraction of tasks and task-based recommendation. The other work mentioned extensively in this chapter creates and tests frameworks for understanding the nature of people's tasks, ultimately to understand the type of support that can be given in a search session, and it attempts to predict this type of information from behavioral data.

2.5 Summary

The takeaways from the above sections can be stated as follows:

1. There has been extensive interest in research relating task characteristics to behaviors and outcomes.
2. Tasks can be categorized along different dimensions. To name a few: goal, product, objective complexity, subjective difficulty, a user's topic familiarity, and time pressure
3. Each study focuses on a small set of these dimensions.
4. By varying these dimensions, the research often directly compares changes in behavior to determine if behaviors can distinguish different types of tasks or different user contexts.

The first thing that should be mentioned is the definition of task for this dissertation. In the empirical work conducted in this dissertation, the definitions of goal and product specified by Li and Belkin [79] will be used. It has already been shown above that these have been extensively studied and are clearly of interest to IIR researchers. But this work will also incorporate other task dimensions, such as subjective difficulty, topic familiarity, and time pressure. These will be clear in the proceeding sections.

As we will see in the next chapter, this outline presents something problematic. On one hand, the conceptual models researchers have developed for IIR are extremely complex, involving a network of relationships between a worker and their context or between different steps of the information seeking process. While the empirical work above has studied parts of the search context piece by piece, none has attempted to combine them into a broad model and test it. Not only should data validate the theory in conceptual models, but as data analysis techniques become more sophisticated, experimental frameworks should adjust in kind to ultimately catch up to complex theoretical work. The empirical literature has often ignored the potentially simultaneous effects on behavior, such as task expertise or time pressure can have on behavior. The next chapter will discuss this dissertation's contribution in this regard, proposing a modeling framework to fill this gap.

Table 2.1: Li's faceted classification of task [79]

	Facets	Sub-facets	Values	Operational definition
Generic facet of task	Source of task		Internal generated	A task motivated by a task doer. It is a self-motivated task.
			Collaboration	A task motivated through discussion of a group of people
			External assigned	A task assigned by task setters based on their individual purpose
	Task doer		Individual	A task conducted by one task doer
			Individual in a group	A task assigned and completed by different group members separately
	Time	Frequency	Unique	A task conducted at the first time
			Intermittent	A task conducted more than one time but assessed by task doer as not frequently conducted
			Routine	A task assessed by task doer as frequently conducted
		Length	Short-term	A task which could be finished within a short time period (e.g. less than one month)
			Long-term	A task which has to be finished within a long time period (e.g. more than one month)
		Stage	Beginning	A task which just launched
			Middle	A task which has been running for a while and in the middle way
			Final	A task which is almost done or has been completed
	Product		Physical (for WT)	A task which produces a physical product
			Intellectual (for WT and ST)	A task which produces new ideas or findings
			Decision/Solution (for WT)	A task which involves decision making or problem solving
			Factual information (for ST)	A task locating facts, data, or other similar information items in information systems
			Image (for ST)	A task locating images in information systems
			Mixed product (for ST)	A task locating different types of information items in information systems

Table 2.2: Li's faceted classification of task [79]

	Facets	Sub-facets	Values	Operational definition
	Process		One-time task	A task accomplished through one process without repeated procedures
			Multi-time task	A task accomplished through repeatedly engaging in the same or similar process
	Goal	Quality	Specific goal	A task with explicit or concrete goals
			Amorphous goal	A task with abstract goals
			Mixed goal	A task with both concrete and abstract goals
		Quantity	Multi-goal	A task with two or more goals
			Single-goal	A task with only one goal
Common attributes of task	Task characteristics	Objective task complexity	High complexity	A task which involves significantly more paths during engaging in the task
			Moderate	A task which may involve a few paths but not significantly more during engaging in the task
			Low complexity	A task which involves a single path during engaging in the task
		Interdependence	High interdependence	A task conducted through collaboration of a group of people (at least two people)
			Moderate	A task conducted by one task doer with suggestions or help from other people or group members
			Low	A task conducted by one task doer without any help from other people
	User's perception of the task	Salience of a task	High salience	A task assessed by the task doer as highly important
			Moderate	A task assessed by a task doer as moderate importance or the degree of salience of the task depends on specific situations
			Low salience	A task assessed by the task doer as unimportant

Table 2.3: Li's faceted classification of task [79]

	Facets	Sub-facets	Values	Operational definition
		Urgency	Immediate (urgent)	A task assessed by a task doer as highly urgent
			Moderate	A task assessed by the task doer as moderately urgent or the degree of urgency of the task depends on specific situations
			Delayed (not urgent)	A task assessed by the task doer as no urgency
		Difficulty	High difficulty	A task assessed by a task doer as high difficulty
			Moderate	A task assessed by a task doer as moderate difficulty or the degree of difficulty or the task depends on specific situations
			Low difficulty	A task assessed by a task doer as no difficulty or easy to complete
		Subjective task complexity	High complexity	A task assessed by a task doer as highly complex
			Moderate	A task assessed by a task doer as moderately complex or the degree of complexity of the task depends on specific situations
			Low complexity	A task assessed by a task doer as simple
		Knowledge of task topic	High knowledge	A task assessed by a task doer as highly knowledgeable on the task-related topic
			Moderate	A task assessed by a task doer as moderately knowledgeable on the task-related topic or the degree of knowledge on the task topic depends on specific situations
			Low knowledge	A task assessed by a task doer as unknowledgeable on the task-related topic

Table 2.4: Li's faceted classification of task [79]

	Facets	Sub-facets	Values	Operational definition
		Knowledge of task procedure	High knowledge	A task assessed by a task doer as highly knowledgeable on the method of procedures for completing the task
			Moderate	A task assessed by a task doer as moderately knowledgeable on the method of procedures to completing the task or the degree of knowledge on the method of procedures depends on specific situations
			Low knowledge	A task assessed by the task doer as not knowledgeable on the method or procedures for completing the task

Chapter 3

Framework

The previous chapter reviewed relationships shown between different types of tasks, behaviors, and user characteristics. It briefly overviewed the statistical techniques used in empirical work. This chapter will discuss these techniques in more detail. It will show how a graphical framework is a more general framework that not only encompasses all of the prior techniques but also can answer broader research questions. It will conclude with a discussion of Bayesian networks - which have been used in a variety of settings in information retrieval generally - and structural equation modeling - which has more specifically begun seeing use in IIR settings. We will conclude with some mathematical details about these frameworks, which will be used to answer the research questions posed in this dissertation.

3.1 Overview of Prior Statistical Techniques

This section summarizes the mathematical approaches taken by the literature reviewed in the last chapter. The analyses conducted can be summarized into the following types: independent sample tests (univariate and multivariate), machine learning, and dependent sample tests (univariate and multivariate). Each of these will be explained in turn. This section will explain how seemingly disparate techniques can be reduced to a few equations and graphical models. A more thorough argument for the graphical framework is then presented in the next section. The findings here are based on insights from Hutcherson, Graham, and Pearl [53, 40, 117].

3.1.1 Independent sample tests: Usages

Independent sample tests are the most common tests in this line of work and most straightforward. Changes in one independent variable (univariate) or multiple (multivariate) are compared against changes in a dependent variable. Independent univariate tests are by far the most common tests in the work reviewed so far. Independent sample tests comprise a family of tests; the specific one to use depends on: 1) The data type of the independent variable (real valued, categorical, or ordinal), 2) whether a categorical dependent variable takes more than 2 values, and 3) the distribution of the dependent variables (e.g., whether they are normal). The reviewed work used the t-test, the Mann-Whitney U test, one-way ANOVA, the Kruskal-Wallis H test, and the less common χ^2 test. These have been used specifically to detect significant behavioral changes among multiple task types [91, 83, 119, 29, 64, 80], among two task facets (e.g. factual/intellectual product) [91], among two levels of complexity [126] or multiple levels [126, 15], among two levels of topic familiarity [94] or multiple levels [80], among two levels of task difficulty or multiple [96, 85, 5, 43, 42, 89, 86, 96, 29, 83, 87], and among two levels of time pressure [31, 88]. Some works looked at interaction to see if the relationship between the variable and behavior changed under the conditions of another (e.g. behavior on high/low familiarity conditioned on task type); they used the Mann-Whitney test [94, 89].

Multivariate tests were less common but still practiced, with the used tests being two-factor/multi-factor ANOVA and linear modeling. These were used to compare behaviors when the independent variables are complexity and task type [126], task stage and document usefulness [90], task complexity and task product [80], time pressure and system delay on behavior [31].

3.1.2 Independent sample tests: Equation and Graphical Forms

In the univariate test, given the independent variable Y and dependent variable X, the claim that a statistical relationship holds between the two can be written as $Y \sim X$. There is a relationship - nonparametric or otherwise - between a task characteristic X and a behavior Y. As an example, the t-test can be written using:

$$\begin{aligned} y &= f(x) + \epsilon, \\ y &= \beta x + \epsilon \end{aligned}$$

ϵ is the error associated with y. The t-statistic is commonly written as $t = \frac{\mu - \mu_0}{SE_\mu}$, so it is instead written as $t = \frac{\beta - \beta_0}{SE_\beta}$. For statistics such as Mann-Whitney which make no assumption about the distribution or assume alternate distributions, this equation can be generalized with no assumptions about functional form (e.g., linearity between x,y,ϵ):

$$y = f_Y(x, \epsilon)$$

In the multivariate case with multiple dependent variables X_1,X_2,...X_n, the equivalent equations are:

$$\begin{aligned} y &= f(x_1, x_2, ...) + \epsilon, \\ y &= \sum \beta_i x_i + \epsilon, \\ y &= f_Y(x_1, x_2, ..., \epsilon) \end{aligned}$$

A graphical representation of this can be seen in Figure 3.1-3.2. In the t-test, each directed edge represents a β coefficient, whereas this denotes the functional relationship f_Y in the more general form.

3.1.3 Machine Learning: Usages and Forms

Machine learning techniques tend to flip the problem in reverse, predicting task characteristics as dependent variables from a set of behaviors as independent variables. The

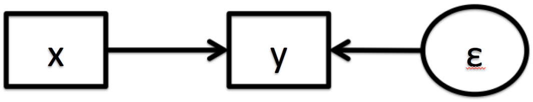

Figure 3.1: A graphical equivalent of independent univariate statistics. x is the independent variable, y is the dependent variable, and ϵ is the error term. β lies on the edge between x and y but is not shown.

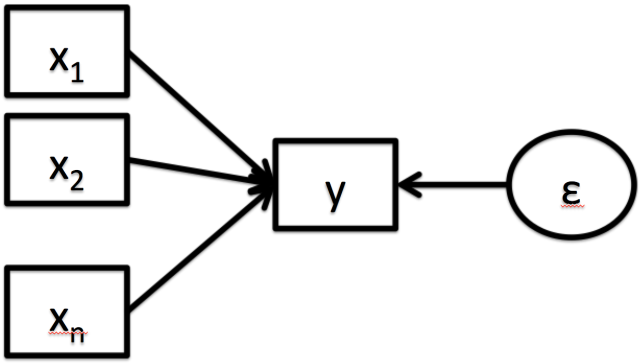

Figure 3.2: A graphical equivalent of independent multivariate statistics.

above notation would hence be converted to $x = f_X(y_1, y_2, ..., \epsilon)$. Prior work performed logistic regression to predict task difficulty [3], 2nd order polynomial regression to predict task success [5], and multiple machine learning models like SVM, multilayer perceptron, and to predict task type, goal, product, and query segment intentions [107, 109].

In these machine learning examples, the equations are identical, only flipping the direction of the relationship. The nonparametric form of the equation hence becomes $x = f_X(y_1, y_2, ..., \epsilon)$. This applies to nonlinear models like logistic regression and polynomial, which apply a logistic function after summing behavioral terms together. This even applies to more complex nonlinear models like multilayer perceptron and SVMs.

3.1.4 Dependent sample tests: Usages

In dependent sample tests, unlike independent sample tests, the key difference is that data points have a paired relationship. For instance, one might compare search behavior in one search session to behavior in a later session to detect a searcher's learning. Hence variants of the original tests like t-tests are used, where data points are paired. The previous work used one-factor and two-factor repeated measures ANOVA. These have been used to compare behaviors across multiple task types [81, 80], two task products [81], multiple levels of task complexity [81, 80, 15], and multiple levels of search task determinability [24].

3.1.5 Dependent sample tests: Equation and Graphical Forms

Univariate dependent data and multivariate dependent data are related as $Y \sim \Delta + X$ and $Y \sim \Delta + X_1 + X_2 + ...X_n$, respectively. Δ indicates the difference between paired points. The t-test in this instance is modified to $t = \frac{\mu_\delta - \mu_0}{SE_\delta}$. A similar functional form can be written as in the independent samples case, with data points being differences instead of raw values. The graphical form of univariate dependent samples t-test is somewhat modified and is shown in Figure 3.3. In this case, one can create independent variables for x_{i1} and x_{i2}, where 1 and 2 indicate the paired independent measurement and y_1 and y_2, respectively, and the double arrow indicates a correlation between x_{i1} and x_{i2}. Generalization is to be left to the reader.

3.2 A Case for Graphical Modeling

Anyone with passing knowledge in statistics can recognize that the generalizations listed above lead to Generalized Linear Modeling (GLM). The disparate statistical techniques have all been shown to be special cases of GLM, despite being developed independently [40]. Anyone with passing knowledge in machine learning can also see these models are not nearly as complex as the plethora of modern ones in machine learning. But despite this simplification, placing disparate statistical techniques in a larger framework is something that is often overlooked. More specifically, while the

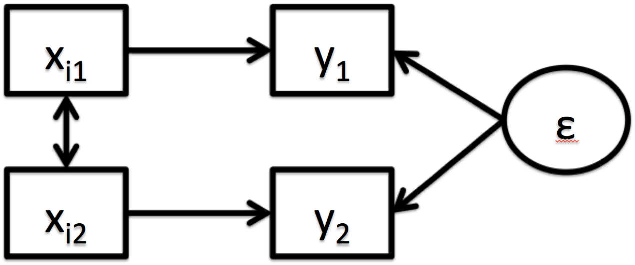

Figure 3.3: A graphical equivalent of dependent univariate statistics. The specific example here is shown for the t-test, where paired data points are represented by x_1,x_2.

cited literature has done significant work to advance IIR and our understanding of the user, little has been done to put the used techniques into a broader framework. This framework is broader than linear modeling and is graphical, and the argument for this shall now follow. Consider the following list of the types of questions that presently used techniques can answer:

- **Independent statistics** - Can variations in one or more task characteristics be distinguished by changes in behavior?
- **Machine learning (in IIR research)** - Can behaviors be used to accurately classify the task characteristics of unseen sessions?
- **Machine learning (in IIR research)** - Can behaviors be used to accurately classify the task characteristics of a known session as early as possible?
- **Dependent statistics** - Given a set of paired sessions, can variations in one or more task characteristics be distinguished from differences in behavior?
- **Interaction effects** - Does whether task complexity affects task completion time depend on the user's domain expertise?

The following cannot be answered by these techniques:

- **Conditional independence** - If a user's task familiarity is given, does learning about the task type still help predict task completion time?

- **Mediation (1)** - Does task in itself directly affect behavior? Or is this affect only through task familiarity?
- **Mediation (2)** - Even if task has no effect on behavior, is this because no effect exists? Or because this effect is being suppressed by some other factor?
- How does task affect behavior in the context of several other varying factors, such as a user's time pressure and general search expertise?

Not only can these not be answered with the currently used techniques, but a framework that can answer these questions is necessary to bridge the gap between conceptualization and practice.

3.2.1 Argument 1: Conditional Independence and Mediation

In statistics, there is a distinction between *moderation* and *mediation*, as pointed out by Baron and Kenny [8]. *Moderation* is identical with *interaction effect.* In contrast, Figure 3.4 shows an example of *mediation.* x_1 has two effects on y: one that is directly through x_1, as indicated by the arrows (call this arrow β_1) and one that is indirectly through x_2 (call these arrows β_2 and β_3, respectively). Baron and Kenny stated that "a moderator is a qualitative...or quantitative...variable that affects the direction and/or strength of the relation between an independent or predictor variable and a dependent or criterion variable." In contrast, "a given variable may be said to function as a mediator to the extent that it accounts for the relation between the predictor and the criterion." If that does not clarify the distinction, a simple linear example shows they are not identical. In an independent sample test with 2 independent variables, an equation including an interaction term $x1 \times x2$ for moderation is represented as:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \epsilon,$$

In contrast, the above mediation is represented by 2 equations:

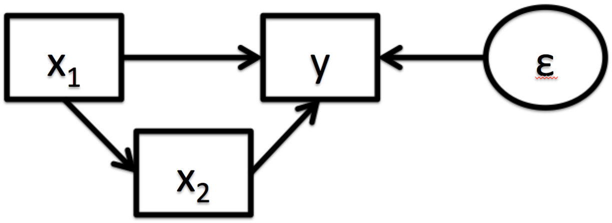

Figure 3.4: An example of mediation. x_1's effect on y is partially mediated through

$$\begin{aligned} x_2 &= \beta_2 x_1 + \epsilon_2, \\ y &= \beta_1 x_1 + \beta_3 x_2 + \epsilon_1, \\ &= \beta_1 x_1 + \beta_3 \beta_2 x_1 + \epsilon_1 \end{aligned}$$

A mediating relationship suggests that any effect of x_1 on y, if it exists, is entirely through x_1. Terminology for mediation includes: *partial mediation*, *total mediation*, *direct effect*, *indirect effect*, and *total effect*. *Partial mediation* occurs when $\beta_1 x_1$'s effect is significant but the overall *indirect effect* of the path $\beta_3\beta_2 x_1$ is also significant. In this case, the effect $\beta_1 x_1$ is the direct effect, $\beta_1 x_1 + \beta_3 x_2$ represents the total effect. If the effect $\beta_1 x_1$ on y is insignificant while there is an indirect effect, the relationship between x_1 and y is totally mediated through x_2. None of these relationships can be represented by the aforementioned tests, but a graph with mediation can be extended to include interaction effects by adding the variable $x_1 \times x_2$.

More generally, the question of mediation highlights the importance of conditional independence, thoroughly studied in graphical modeling frameworks such as Bayesian modeling. Consider again Figure 3.4. In a case of total mediation, x_1 is conditionally independent of y, if x_2 is given/known. In a case of only partial mediation, no claims about conditional independence can be made. It might be of interest to IIR researchers, for instance, to determine if task (x_1) genuinely has an effect on behavior (y) (as in

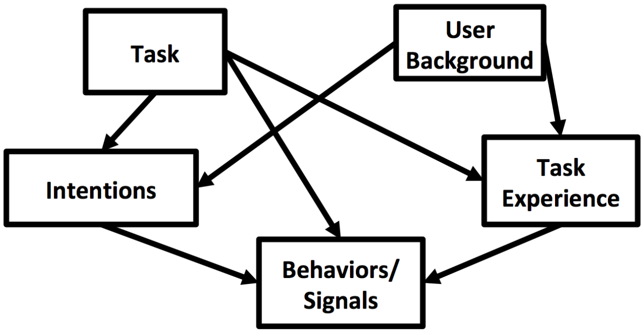

Figure 3.5: One possible graphical model depicting the relationship between objective task characteristics, user characteristics, and behaviors.

work like [81]) or if it is only through topic familiarity (x_2).

A formal definition of conditional independence is in order. We present the definition from Pearl [116] but leave burden of proof on the reader:

- *Definition (d-separation)* - A set S of nodes is said to block a path p if either (1) p contains at least one arrow-emitting node that is in S, or (2) p contains at least one collision node that is outside S and has no descendant in S.
- *Conditional independence* - If S blocks all paths from set X to set Y , it is said to “d-separate X and Y,” and then, it can be shown that variables X and Y are independent given S, written $X \perp\!\!\!\perp Y|S$.

Consider Figures 3.1-3.3. In these, the only noteworthy conditional independence relationship is in Figure 3.2: if y is known, each x_i,x_j are conditionally independent of each other. This graphical framework is commonly used in Naive Bayes modeling. In contrast, consider the tentative model in Figure 3.5. Suppose “Experience” includes situational variables such as topic familiarity, task expertise, and time pressure. It would be interesting to know whether there is a significant direct effect of task type on behaviors. If not, then the arrow Task→Behaviors is omitted, and we can make interesting claims such as, “If we know a user’s search intentions, task expertise, topic familiarity, and feelings of time pressure, then we can perhaps predict a user’s behavior. But task and other user background variables provide no additional information”.

3.2.2 Argument 2: Bridging the Theoretical and Practical

By now is should be clear that mathematical shortcomings in previous empirical work exist. Are these shortcomings of practical concern, and is expanding upon them worthwhile in IIR research? According to our prior literature review, this would seem to be the case. First, researchers try to control confounding factors that could affect the relationship between independent and dependent variables of interest, but some factors are hard to control. Researchers may attempt to control user factors by limiting variation in the population, and they control task factors by constructing search tasks to assign to research study participants. Yet some factors are still hard to control, for instance time pressure or general search expertise. If they are important, any work not taking them into account could lose valuable information. Second, recent work suggests that researchers are actively interested in a lack of control. For instance, He and Yilmaz gathered information on real-life Web search tasks in naturalistic settings [50]. Lastly, consider a model of information seeking behavior such as that by Ingwersen and Järvelin in Figure 3.6 or the previously shown Figure 2.2. Not only do these suggest highly nested relationships between user and task characteristics (and search engine traits), but Ingwersen and Järvelin suggest conditional independence between characteristics. For instance, the model suggests that interactions with the search engine are purely a function of the user. Knowing about the user's cognitive state is entirely sufficient to predict behaviors, with external task characteristics such as goal and product adding no value afterwards. It should hence be clear that expanding on the framework of modeling user interactions is not only of mathematical interest but of theoretical and practical interest to the IR community.

3.3 Graphical Modeling (and Structural Equation Models) in IR

What type of graphical modeling approach should be adopted, since "graphical modeling" refers to a large family of functions and relationships. Much recent work in IR has adopted various graphical models. One example includes applying topic modeling to a corpus of documents. Variants of Latent Dirichlet Allocation [12] - a Bayesian model

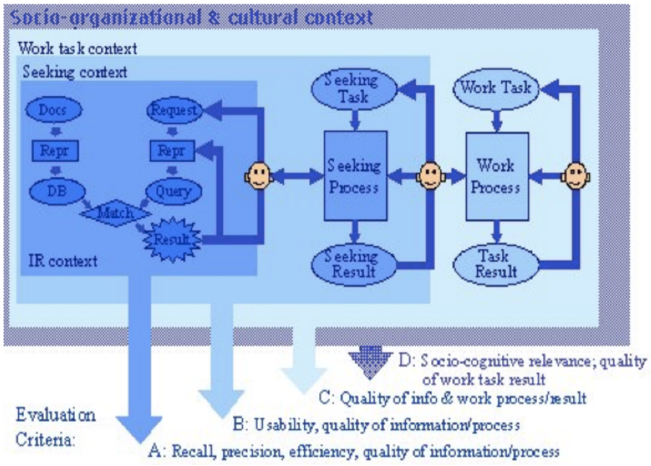
Socio-organizational & cultural context
Work task context
Seeking context
IR context
Docs
Repr
DB
Request
Repr
Query
Match
Result
Seeking Task
Seeking Process
Seeking Result
Work Task
Work Process
Task Result
Evaluation Criteria:
A: Recall, precision, efficiency, quality of information/process
B: Usability, quality of information/process
C: Quality of info & work process/result
D: Socio-cognitive relevance; quality of work task result

Figure 3.6: Nested contexts of information seeking and use, from [56].

with Dirichlet priors - are still used, for instance to model a corpus of tweets [38]. Deep learning research is extensive in IR, such as for classic retrieval problems [7]. Deep learning work is also graphical and shares properties such as d-separation. However, it is worth noting that such models are typically applied to large scale settings with mostly unstructured data. In several interactive IR settings - including the experiments in this dissertation - the provided data is completely opposite: structured, small, and only consisting of a few dozen sessions and several hundred queries.

Such a data setting lends itself to an older technique: namely structural equation modeling. Although an old technique, structural equation models (SEMs) are making a recent splash in the IR community. For instance, Ishita et al. [57] used a SEM to demonstrate the relationship between information retrieval skills and efficacies such as critical thinking, logical thinking, and formal internet training. Zhang et al. [144] used a SEM to understand the relationship between document relevance, document reliability, understandability, topicality, novelty, and scope. SEMs - like many other models - contain explicit and latent variables. SEMs without latent variables are called path models. Both path analysis and SEM were presented recently in a IR conference tutorial [63] so have garnered interest in the IR community.

In its commonly understood, unmodified form, SEM makes strong assumptions about the relationships between variables - such as linear relationships between variables as in the above equations. This runs counter to other modern techniques like LDA which use alternative distributions [12]. But Pearl indicated that this is an unnecessary restriction and that SEMs can have a nonparametric representation that does not commit to linear functions. In addition, he showed that SEMs share graphical properties such as d-separation [117]. The work here will largely use the frameworks of SEMs, and specifically path analysis.

3.3.1 Building Graphical Models - Path Models and Bayesian Networks

There are two commonly accepted methods of building graphical models. One approach for SEMs builds the model using empirical or theoretical findings from literature review. Other approaches in SEM and Bayesian literature are more data driven. Both types of approaches will be discussed, as both are used in this dissertation's experiments. Issues of required sample size for such models are also discussed.

Model Specification through Literature Review

One approach to model specification starts with literature. Namely, given a data set and research questions, the researcher must manually specify relationships between variables. These are specified as one-directional influences (equations in the model) or as bidirectional relationships (covariance between two variables). The researcher makes model-building judgments through literature review, for instance literature suggesting that smoking causes cancer or that home environment affects standardized score success because it affects the amount of time a student spends studying. The researcher can also fix if the size of the effect between some variables is known. [125] If a relationship between a pair of variables has been found to be significant, this is one that should be estimated. This model specification technique is confirmatory, since the researcher is confirming previous findings. The researcher then assesses the model for how well it fits the data. This approach to model building has been taken in human-computer interaction work such as Khakurel et al [69].

Data-Driven Model Specification

Two alternative techniques do not rely on literature but learn structure from a given data set. First, for structural equation modeling, suppose that for a given data set: 1) one has multiple indicator variables (e.g. survey questions), 2) one wants to combine subsets of the variables into latent constructs, and 3) one would like to specify relationships between those latent constructs. This can be explored through an exploratory factor analysis (EFA) technique [49] and was done in Zhang et al. Specifically, they applied techniques called standard maximum likelihood and principle axis factoring to reduce a set of 15 questions into 5 latent variables. They verified whether they could use this EFA technique with Kaiser-Mayer-Olkin Measure of Sampling Adequacy and Bartlett's test of Sphericity [144]. This approach will not be used here, as we do not have the proper setting for constructing these latent constructs, but it should be mentioned.

The second approach lies in structure learning, from the Bayesian literature. Two types of algorithms are commonly applied in structure learning. Structure Learning is a NP hard problem, so the algorithms are approximation algorithms. The types of algorithms are known as constraint-based algorithms and score-based algorithms (and some hybrid ones). One of each type was used in the experiments, and each will be discussed in the next chapter. Yet generally speaking, one starts with a set of variables V and a dataset D. The goal is to construct a directed acyclic graph G that meets the requirements of the algorithm. In the score-based case, e.g., hill-climbing, the goal is to maximize the following function:

$$Score(G : D) \quad = \quad LL(G : D) - \phi(|D|)||G||$$

Here, $LL(G : D)$ is the log-likelihood of D under G. The subtracting term penalizes for model complexity. $|D|$ is the size of the data $||G||$ is the number of parameters in the graph, and ϕ refers to a complexity measure function, such as the Akaike information criterion ($\phi(|D|)$=1) or the Bayesian Information Criterion ($\phi(|D|) = \frac{log(|D|)}{2}$).

A popular family of scoring functions is the Bayesian Dirichlet score. If $P(\Theta_G)$ is the prior probability of the parameters, the scoring function is provided as:

$$P(D|G) \quad = \quad \int P(D|G,\Theta_G)P(\Theta_G|G)d\Theta_G$$

Lastly, it is worth noting that in graphical modeling there exists the notion of "equivalent models"; multiple graphical models that are equivalent. Pearl and Lee and Hershberger defined equivalence in terms of d-separation: two graphs that are equivalent must be identical in terms of conditional independence, even if the specification of equations is different [117, 76]. Sprites et al. further discussed the concept of covariance equivalence classes - a group of possible graphs over the same variables that recreate the same covariance matrix over over their data. They moreover proved several theorems, including: 1) If two graphs are directed acyclic graphs, they are covariance equivalent if and only if they are d-separation equivalent 2) Two directed acyclic graphs are d-separation equivalent if and only if they contain the same vertices, the same adjacencies, and the same unshielded colliders. D-separation equivalaence and covariance equivalence can hold for acyclic as well as cyclic graphs [124].

3.4 Determining Sample Size

For computing necessary sample size in structural equation models, some claim there is no strict rule to determining the required size. Several informal rules of thumb have been published, including: a minimum sample size of 100 or 200 [13, 73], 5 or 10 observations per estimated parameter/degree of freedom [11], and 10 cases per variable [113]. Wolf et al. indicated that these rules can lead to either overestimated or underestimated sample size requirements, increasing the changes of Type I and Type II errors, respectively [141]. Wolf et al. explained 3 proposed solutions for this problem: estimating power based on the amount of model misspecification [121], obtaining a confidence interval for goodness of fit metrics [22], and Monte Carlo simulations (chosen by Wolf) to show the power and bias for individual effects (i.e., single directed edges) in the model [141]. Our experiments will adopt the approach of MacCallum [22], who proposed sampling to product a confidence interval for measures of fit. MacCallum specifically used the confidence interval of RMSEA, which is an accuracy measure for path analysis and structural equation modeling. Sampling the confidence interval of RMSEA over different sample sizes can tell us how accuracy of a path model estimate improves, worsens, and becomes more or less confident over time. Another evaluation metric called the expected cross-validation index (ECVI) [17], when plotted over sample sizes, can tell us how well our estimate of model parameters would generalize to some external data set. Both will be used here to address RQ4 (regarding sample size), and mathematical details of both will be discussed in the next chapter.

Chapter 4

Experimental Design

Experiment 1 will use confirmatory analysis to explore how well relationships found in the literature are validated using path analysis. The literature-based model building approach mentioned previously will be used to first construct the path model. Namely, if a significant relationship is found between two variables, this is included as an edge in the graphical model, and its weight is to be estimated. Standard goodness-of-fit metrics in path analysis will be used to determine how well this theoretical model fits the data provided from one laboratory study. The model largely reflects the conceptual one shown in Figure 3.5. Afterwards, the model will be iteratively tweaked to determine the usefulness of certain sets of features. Specifically, it will systematically incorporate and remove variables and relationships for users' general search experience, background knowledge, affectional variables, and search intentions. Not only will the confirmatory fit be assessed, but it will be analyzed to determine whether new insights regarding mediation and indirect effect can be discovered with this modeling technique; these cannot be found using previous techniques. This experiment can be found in Mitsui and Shah [110].

Experiment 2 will apply Bayesian structural learning to learn the optimal structure of a graphical model from data, using both a score-based and constraint-based approach to structure learning. In contrast with the previous experiment where the relationships between variables (edges in a graph) were hand-constructed from literature review, several models will be built in this experiment using data-driven approaches. Models will first first be built on a combined dataset, merging the datasets from 3 separate user studies; behavioral data and survey data common to all 3 studies will be used. Then models will then be built on 3 datasets separately; in this case, models will be built from each data set using all the data available respective to that data set. A bootstrapping and sampling framework will be used to construct more valid results in both cases. The results will be combined to provide implications about replications and genuinely significant effects.

Experiment 3 will explore the relationship between data size and accuracy. The results from Experiments 1 and 2 will be used to construct graphs to analyze. This experiment will explore how data size affects the goodness of fit, both with respect to the accuracy of the fit and the confidence in the estimated model. As data collection is at a premium in laboratory studies, it is useful to know how much data is required to create a good model relating user characteristics, task characteristics, and behaviors - or at least how accuracy may change over varying data set sizes. This analysis will be performed once again on the 3 datasets on various sample sizes, randomly sampling to produce means and confidence intervals of evaluation metrics.

4.1 Experiment 1: Confirmatory Analysis

4.1.1 Experimental Design

The purpose of this experiment was to confirm how well the relationships found in literature were confirmed by analysis of path models. It was also meant to extract relationships of mediation, which could not be extracted using previous analysis frameworks. A confirmatory experiment using path analysis must start with a hand-built path model. This was constructed through literature review. Significant relationships between variables from literature indicate arrows/dependencies in the model. This set of relationships is constructed into a set of equivalent equations (discussed in the previous chapter) in which the weights β are estimated. Listed below are all relationships included in our model, along with the relevant literature. For a more thorough exploration of the literature, refer to the Background chapter or the Appendix. Unless otherwise specified, relationships listed below are one-way directed arrows, specifying equations rather than correlations.

Task → Behaviors - Task goal, product → *all behaviors* [91, 60, 4]. Every task and topic variable in this experiment has an arrow to every behavioral variable.

Task → Intentions → Behaviors - Task goal, product → intention groups [111]; intention groups→*all behaviors* [107]. Every task variable has an arrow directly to every intention variable.

Task/Topic → Search Experience - Task product, goal→search difficulty [93]; topic → topic familiarity.

Background → Search Experience - Search years → search difficulty; search frequency → search difficulty.

Background → Intentions - Search expertise → intentions [100, 120].

Experience → Behaviors - Topic familiarity → *all behaviors* [97, 51]; search difficulty → *Behaviors* [97, 3, 5]. These also have arrows going to every single behavior.

Within-category Correlations - Adequate time⟷task difficulty [32]; assignment experience ⟷ search difficulty [93]; task goal ⟷ task product (our data is not perfectly balanced); topic familiarity → search difficulty.

Figure 3.5 once again can be referenced for a summary. Each node in the figure indicates several variables. For instance, in this experiment the “Task” node indicates 3 binary variables: the task goal, the task product, and the task category. A path indicates that there is some one-way arrow between nodes in one set and nodes in another. Also note that henceforth we use “behaviors” and “signals” interchangeably.

To confirm whether claims in the literature lined up under this constructed model, the following had to be asked: 1) How well does the data fit the model overall? 2) Are certain features more useful than others in showing the relationship between task type, behavior, and other user characteristics? Both of these questions can be answered in the evaluation framework of path analysis, explained in the next subsection. The first question could be answered by comparing the model constructed above to the “saturated model”, where it is assumed that relationships between every pair of variables should be estimated (as nonzero correlations). The saturated model has a perfect fit according to evaluation metrics but is the most complex.

The second question requires a comparison of variations of the model, which need to

Model Name	$\beta_{T,E}$	$\beta_{T,I}$	$\beta_{B,E}$	$\beta_{B,I}$	$\beta_{E,S}$	$\beta_{I,S}$	$\beta_{T,S}$
Full Model	Y	Y	Y	Y	Y	Y	Y
IB	N	Y	N	Y	N	Y	Y
IE	Y	Y	N	N	Y	Y	Y
I	N	Y	Y	N	N	Y	Y
BE	Y	N	Y	N	Y	N	Y
E	Y	N	N	Y	Y	N	Y
Task Only	N	N	Y	Y	N	N	Y

Table 4.1: The different models tested, as well as whether the edges in each group are unconstrained (Y=Yes,N=No). Variables are Task (T), Experience (E), Intentions (I), Background (B), and Behavioral Signals (S).

be specified carefully. Suppose we group the features broadly: task characteristics, user background characteristics, user experience characteristics, intentions, and behaviors. Further suppose we want to know how useful each group of features is. We can construct the following algorithm.

- Select groups of nodes $G_1, ..., G_n$ (e.g., search expertise variables, search difficulty variables).
- Split the graph into two groups of variables: $V' \in G_1, ..., G_n$ and $V = G - V'$
- Constrain edges between V and V' to 0.

Rather than removing variables, we constrain paths to be 0 so we can compare evaluation metrics across variations of the graph. Path analysis evaluation metrics are relative to the number of degrees of freedom and number of variables, as shown in the next section. A table of the compared variations is shown in Table 4.1. Task and behaviors were kept in every variation; in addition to exploring the usefulness of variables, a subgoal was to explore the usefulness of variables in mapping the relationship between variables and task type.

In this experiment, we used dataset 1, explained in detail in the last sections of this chapter. The is because dataset 1 (as explained later in this chapter) contains almost all of the features each data set has, and additionally includes searchers' intentions. Searchers reported what their intentions were for each query segment. A side effect of this experiment was to see if intentions were useful and could be justifiably discarded in future experiments when mapping the relationship between task type and behavior. The data points in this experiment are the query segments of users. Each query segment is a vector of behavioral features, demographic characteristics (e.g., general search expertise), session characteristics (e.g., task difficulty), and task type characteristics. The specific features and their distributions are provided in Tables 4.3, 4.4, and 4.5. The number of data points at least meets the requirements of informal rules for SEM building (e.g., at least 200 points - there are over 600).

The SEM was constructed and evaluated using the SPSS AMOS Software, Version 25[1]. A linear path model was used, with estimation of the β parameters done using maximum likelihood estimation on 200 bootstrapped samples of the data.

4.1.2 Evaluation Methodology

Recall that a path model is a SEM without latent variables. A given path model combined with the data can be used to produce a covariance matrix S between variables X_i and X_j:

$$Sij = E[(X_i - \mu_i)(X_j - \mu_j)] \tag{4.1}$$

This is an approximation of the true covariance matrix of the data, Σ. Evaluation metrics for path models are largely based on goodness of fit of S to Σ. The saturated model recreates Σ perfectly. A fundamental evaluation metric is χ^2:

$$\chi^2 = \sum_{ij} \frac{(S_{ij} - \Sigma_{ij})^2}{\Sigma_{ij}} \tag{4.2}$$

A similar metric is the goodness of fit index (GFI).

$$GFI = 1 - \frac{Cov_{residual}}{Cov_{total}} \tag{4.3}$$

Where Cov_{total} is the total covariance of Σ, and $Cov_{residual}$ is leftover covariance from the error terms; higher scores are better.

---

[1] https://developer.ibm.com/predictiveanalytics/2017/09/07/whats-new-spss-amos-25/

Feature	Intentions	Learning	Expert
General Search Expertise	X		X
Years Spent Searching	X	X	
Search Frequency	X	X	
Professional Domain Expertise	X		
Topic Familiarity	X	X	X
Assignment Experience	X		
(Post) Search Difficulty	X	X	X
Adequate Time	X		X
Behaviors	X	X	X

Table 4.2: Features present (X) and absent in the intentions dataset (Intentions), searching as learning dataset (Learning), and expert opinion dataset (Expert).

Other scores adjust in favor of model simplicity. These penalize based on degrees of freedom, number of parameters, or the number of data points. Two such are the adjusted GFI (AGFI) and parsimonious GFI (PGFI). Another popular one, the root mean squared error (RMSEA), is provided by:

$$RMSEA = \sqrt{\frac{\chi^2 - df}{df(N-1)}} \tag{4.4}$$

Lastly, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) are shown below, similarly weighting with respect to the number of parameters and degrees of freedom (a higher score is worse). Given N data points, k parameters, and df degrees of freedom:

$$AIC = \chi^2 + k(k+1) + 2df \tag{4.5}$$

$$BIC = \chi^2 + ln(N)\left(\frac{k(k+1)}{2} - df\right) \tag{4.6}$$

4.2 Experiment 2: Structure Learning

4.2.1 Experimental Design

The results of Experiment 1 will show that there is room for improvement in the model constructed from literature review. The purpose of this experiment is to try to derive a model from data. Rather than just learning the parameters of a pre-specified graph, the structure of the graph will also be learned. To do this, structural learning algorithms will be applied to data to learn a graphical model from 3 comparable research study data sets.

Each of the 3 comparable data sets is explained in detail in the last parts of this chapter. The types of data collected are not completely identical but overlap. In addition to having comparable browsing data, each of the respective studies asked similar questions about users' background and about their experiences with the search tasks. Table 4.2 gives an account of the features available in each data set. There are some questions that all datasets have in common, and each data set has at least some one question belonging to every node in the conceptual Figure 3.5: behaviors, task, background, and experience, but *intentions* are only provided in the first data set. In Experiment 1, though, it was found that dropping intentions does not yield much loss in performance in terms of path model metrics. Intentions can hence be dropped in other analyses, specifically Experiment 2.

In this experiment, in an analysis combining datasets, only the common subset of questions will be used. However, in an analysis of each data set separately, all of the data sets' respective questions will be used, to extract any possible relationships that should be explored. The complete set of attributes collected for each data set are shown in the dataset-specific subsections in this chapter. There are some caveats.

- In analysis of the Expert Opinion dataset exlusively, search expertise was calculated as an average over 4 expertise-related questions.
- With the Expert Opinion dataset, task product will not be incorporated into building the model, as it is constant (Factual) for all 6 tasks.
- In analysis of the Searching as Learning dataset exclusively, task product is either Intellectual or Mixed (Factual+Intellectual). It will still be used here, treating Mixed as the lower of 2 levels, if analysis requires this.
- With the Searching as Learning dataset, task goal will not be incorporated, as it is constant (Specific) for all 4 tasks.

The structure learning framework in this experiment proceeded in two phases. In the first phase, the 3 datasets were combined into one. The Intentions, Searching as Learning, and Expert Opinion datasets contained 1274, 693, and 594 query segments, respectively, yielding a combined total of 2561 data points. The purpose of this phase was to determine possibly genuine relationships that are generalizable across data sets; pooling data can increase data set size and also provides justification for replicability, if strong relationships are found. However, the data sizes are skewed, so this phase of the experiment involved downsampling the first 2 datasets to 594 points each. Several model building trials were run with randomly sampled datasets of $594 = 1782$ points. The second phase performed model building on each of the datasets separately, this time using all of the data collected in each study rather than only the overlapping features. The models of the first and second phase could then be compared to each other. Even if there are genuinely important relationships extracted from the first phase, can these still be discovered in the second phase using more features? If not, what did the first phase miss?

Bayesian structure learning is a NP-hard problem, so two approximation algorithms were used to learn directed acyclic graph (DAG) structures. The first is a constraint-based algorithm called the Incremental Association Markov blanket (IAMB) [127]. A Markov blanket of a node is the set of nodes that keep it conditionally independent from the other nodes not in the blanket. The algorithm begins by initializing a graph with no edges. Then, an estimate of the Markov blanket B of a node is made. In a forward phase, variables belonging to the blanket (and possibly false positives) enter the blanket. Suppose that the mutual information of two nodes X and Y is given by $M(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log(\frac{p(x,y)}{p(x)p(y)})$. The forward step greedily adds edges to the node Y, specifically adding edges that maximize the score $M(X;Y|Blanket(Y))$. In the backward phase that follows, false positives are removed, pruning edges that violate the constraint $M(X;Y|Blanket(Y)) < \tau$ for some τ.

The second is a score-based approach known as hill-climbing (HC), which starts with an empty graph and incrementally adds, removes, and reverses edges until a (local) maximum score is reached. This greedy algorithm proceeds as follows[2]:

1. Start with an empty graph G, data D
2. While $Score(G)$ increases:
   (a) G' = add, delete, or reverse an edge of G, provided that the operation gives an acyclic graph G'
   (b) Compute $Score(G') = BIC(G'; D) = LL(G'; D) - \frac{log(m)}{2} dim(G')$
   (c) G = G'

The scoring function used in these experiments is the Bayesian Information Criterion score, where LL is the log likelihood of the graph given the data, m denotes the number of samples and G denotes the number of parameters in G. The Hill-Climbing algorithm constrains the final graph to be a directed acyclic graph (DAG), but the former does not. The former creates undirected edges in the case that an edge orientation cannot be decided. This not necessary for the current analysis under consideration but worth noting.

There is potentially a random component to the algorithm outcomes, and in the case of hill-climbing there is a likely chance that given a dataset, it will greedily optimize to a local optimum. Hence, several graphs were constructed several times using bootstrapped samples of the data, rather than the original data. In the first phase of this experiment where datasets were combined, a total of 2000 graphs built with 2000 bootstrapped samples – 1000 for IAMB and 1000 for HC. In the second phase where data sets were separated, 5000 bootstrapped graphs were constructed for each data set and for each algorithm, yielding 30000 total bootstrapped iterations. Experiments were run using the bnlearn library[3] of R. To meet with the requirements of the implementation, all variables were treated as real-numbered values (which can be justified even for a binary variable such as specific/amorphous goal, which is arguably a scale).

---

[2] https://www.stats.ox.ac.uk/lienart/gml15_bayesianet.html

[3] http://www.bnlearn.com/

4.2.2 Evaluation Methodology

As the purpose of this experiment is largely exploratory, the evaluation methodology is exploratory as well. A standard metric of the quality of a graph is not used, partly due to the random sampling nature of the algorithms but also due to a desire to also qualitatively evaluate the results of the graphs. Since several sample graphs are created, we can compute the probability that an edge will be included in a graph. High strength in the probability of inclusion of an edge gives us confidence that any graph will likely contain the edge. This analysis will be done on both the hill climbing results and IAMB results, on all data sets. From this, a single graph can be constructed - one which was actual output from the algorithm and contains a very high score on most of its edges, according to both algorithms. In particular, in analysis, we will examine one of the most commonly observed graphs constructed from phase one and examine it qualitatively - how well does it compare with intuition? It can also be examined quantitatively. How common are the proposed edges? Which edges from this proposed graph are also commonly present in the second phase, where more features are used for graph construction?

As a final step of evaluation, the anatomy of the proposed graph will be explored, in particular conditional independence relationships. What conditional independence relationships arise from the graph, and are these noteworthy? This entails the limitations of what data can tell us. Although it is generally agreed upon that more context is better for determining characteristics about a search session such as difficulty and task type, little work has explicitly explored the limitations of the data. The exploratory findings will be compared against literature but also post future directions for analysis. Are there any other relationships that are noteworthy? This will justify the use of the proposed graph in the final experiment.

4.3 Experiment 3: Sample Size

4.3.1 Experimental Design

Working graphical models are found in Experiment 1 and Experiment 2. But as these experiments deal are on laboratory study-sized data, a potential criticism is that the lack of data makes it difficult to produce any insights. It is hence good to determine how sample size can affect results. This experiment explores this question by exploring changes in goodness-of-fit, as a function of sample size.

This experiment can also be split into two parts. The first explored the data from Experiment 1. Taking the intentions data, it applies the data to the full model and background and experience measures (BE) model put forth in Experiment 1, examining changes in goodness-of-fit metrics across random samples of the dataset. The second phase similarly applies the model in Experiment 2 to the 3 data sets separately on their most common features. Occasionally, 1 variable needed to be omitted from analysis. For the searching as learning data, the goal had to be omitted, since it was always specific. For the expert opinion dataset, the product was omitted, as it was always factual.

Several random samples of the data were constructed for sample sizes of 30%, 40%, 50%, in increments of 10 up to 100%. As in Experiment 1, SPSS AMOS is used to approximate the parameters of a graphical model (in this case, a SEM) and determine goodness of fit. Maximum likelihood estimation was used once again, with 200 bootstrapped samples within each random sample.

4.3.2 Evaluation Methodology

As in Experiment 1, RMSEA is used be used to show how goodness of fit changes with sample size. An additional measure called the Expected Cross-Validation Index will also be used. ECVI assesses the discrepancy between the covariance matrix fitted with the sample and the expected covariance matrix from a sample of equivalent size. It performs cross-validation to examine the difference between the covariance matrix produced by the data versus the covariance produced. Smaller is better in both cases.

We are interested in how much the metrics improve over time as well as the confidence of those metrics. In addition to reporting raw metrics, we also report the 95% confidence interval for these metrics at each sample size. The equations for the metrics are given below:

$$
\begin{aligned}
RMSEA &= \sqrt{\frac{\chi^2 - df}{df(N-1)}} \\
ECVI &= E\{F(S_v, S_t|S_t)\}
\end{aligned}
$$

$F(S_v, S_c|S_c)$ is a discrepancy function (e.g. chi-square) between the correlation matrix produced by a model on a training set S_t and the model produced on an external validation set S_v. $E\{\}$ is the expected value of the discrepancy function.

4.4 Dataset 1: Information Seeking Intentions

4.4.1 Data Collection Design

This dataset was collected in a laboratory setting. Undergraduate journalism students were recruited from Rutgers University during school semester periods. Participants were required to have completed at least one course in news writing. Each participant's study session consisted of 2 search tasks. Each task was preceded by a pre-task interview, proceeded by a post-task interview, and subsequently followed with an "intention annotation task". Participation in the study began with a demographic questionnaire and ended with a verbal exit interview. All activity except for the exit interview was conducted at a desktop computer in a laboratory. Search activity was recorded in Firefox by the Coagmento browser plugin[4] [108], eye-fixation behavior by GazePoint[5], and annotatable video of the search by Morae[6]. In addition, Coagmento provided a bookmark annotation tool that participants used while conducting provided search tasks. Coagmento also recorded browsing activity and sent it to a remote server.

---

[4] http://coagmento.org

[5] http://www.gazept.com/

[6] https://www.techsmith.com/morae.html

Participants began by answering the demographic questionnaire and watching a tutorial video on using Coagmento, before beginning the search task. Then, participants read the task description and answered a short questionnaire on their familiarity with the topic and task as well as the anticipated difficulty. Participants then had 20 minutes to complete the search task; this was shown to be a sufficient amount of time in pilot tests, and the researchers needed to control for a constant time limit. They could finish before 20 minutes if they felt they completed their task early. Afterwards, participants answered a post questionnaire on the actual difficulty of the task. They were then asked to annotate their intentions for each query segment, first watching a video demonstrating how to conduct this "intention annotation task". Users were also given a handout of a short description of each intention. This was for further clarification and to also reduce variability in the data from differing interpretations of the intentions. They then completed the intention annotation task with no time limit. Participants repeated the process with more questionnaires, another search task, and another intention annotation task before the exit interview. The experimental session lasted about two hours.

For the intention annotation task, participants were asked to select which intentions applied to each *query segment* (all activity that occurred from one query to the next) in the search session. This was accomplished by playing the video of the search, segment by segment. They could select, from a displayed list, any number of intentions for a segment. For instance, if a participant knew nothing about coelacanths and issued the query "coelacanths" as the first query in a session, that person might mark "identify something to get started" and "learn domain knowledge". The participant was then asked to mark whether each of these checked intentions was satisfied. The participant may mark "yes" for "identify something to get started" but "no" for "learn domain knowledge". If a participant marks "no", she must then state why that intention was not satisfied. For example, while she found some new keywords to search, she may not have learned any knowledge that was required by the task description. If the participant had some other intention in addition to the 20 listed, the participant may also check "other", give a short description of that additional intention, and also mark whether it

Task	Product	Level	Goal	Named	$\|T\|$	$\|CO\|$	$\|CL\|$	$\|Q\|$
CPE	Factual	Segment	Specific	True	22	11	11	206
STP	Factual	Segment	Amorphous	False	18	9	9	108
REL	Intellectual	Document	Amorphous	True	18	9	9	155
INT	Intellectual	Document	Amorphous	False	22	11	11	224

Table 4.3: Task type and session characteristics for the Intentions data set. Task product, task goal, and other controlled task characteristics are specified, as well as the number of sessions in each task type ($|T|$), the number of coelacanth sessions ($|CO|$), the number of methane clathrates sessions ($|CL|$), and the number of query segments for each task type (Q).

was satisfied. They repeated this annotation process for each query segment separately. For the entire process, participants were incentivized with additional reward for being among the best performers. This provided incentive to issue good searches, instead of meeting the minimum requirements. Participants were told that "good performance" also included marking intentions well - i.e. marking all and only those that applied. In total, the data collected was for 48 participants.

4.4.2 Assigned Tasks

There were 4 possible tasks and 2 topics per task. Two task types are a copy editing task (CPE) and interview preparation task (INT), as specified in Cole et al [29]. The other tasks - Relationships (REL) and Story Pitch (STP) task - were novel to this study. One topic was "coelacanths", and another was "methane clathrates and global warming". The chosen topics were familiar enough to generate participant interest yet unfamiliar enough so participants would likely not know the requested information before arrival. We further gave a faceted classification of each task in Table 4.3. Task facets were assigned according to task goal, task product, level of document evaluation, and whether the task was named [29, 79]. Each user completed 2 search tasks and hence 2 annotation tasks. Task types were paired into 4 groups, based on differences in facet values. Each participant searched for 2 tasks in one of these 4 groups, each task on a different topic. Order of the 2 tasks and 2 topics in each group was flipped, yielding a total possible $4 \times 2 \times 2 = 16$ configurations. Descriptions of each task are as follows:

- **Task = Copy Editing (CPE). Topic = Coelacanths. Product = Factual. Goal = Specific.**

  Your Assignment: You are a copy editor at a newspaper and you have only 20 minutes to check the accuracy of the six italicized statements in the excerpt of a piece of news story below.

  Your Task: Please find and save an authoritative page that either confirms or disconfirms each statement.

  The coelacanth ("see-la-kanth") is a 'living fossil' fish, thought to be long extinct before a specimen was discovered in 1938 at East London, South Africa. Fourteen years later, the "discovery" of coelacanths was confirmed in the Comoro Islands, off the coast of Madagascar. Forty six years after that, a new population was identified from at least two specimens caught off of North Sulawesi, Indonesia. In 1987, a submersible crew finally managed to obtain film footage of live coelacanths.

  Coelacanths are the size of humans. They are slate-blue when alive, with white flecks on the thick scales that cover their bodies. They live in the gloaming, around 200-400 meters below the surface, where light barely penetrates and few creatures venture. They spend their days sheltering in rocky caves in small groups, coming up to feed at night as the water above them cools. Unlike most fish, they give birth to live young - small, perfectly formed baby coelacanths - and when disturbed they lift themselves into headstands, apparently using an electro-sensory organ in their snout to detect the presence of predators or prey.

  The handful of people who have seen them in their natural habitat talk of their glowing eyes and their gentle demeanor. They describe coelacanths moving with surprising grace, deploying their fanned fins in a diagonal formation - right fin in front, left trailing behind - that is similar to a lizard walking. Their blue and silver color provides excellent camouflage in the underwater cave homes covered in sponges and oysters that they prefer to eat.

- **Task = Story Pitch (STP). Topic = Coelacanths. Product = Factual.**

Goal = Amorphous.

Your Assignment: You are planning to pitch a science story to your editor and need to identify interesting facts about the coelacanth (“see-la-kanth”), a fish that dates from the time of dinosaurs and was thought to be extinct.

Your Task: Find and save web pages that contain the six most interesting facts about coelacanths and/or research about coelacanths and their preservation.

- **Task = Relationships (REL). Topic = Coelacanths. Product = Intellectual. Goal = Amorphous.**

  Your assignment: You are writing an article about coelacanths and conservation efforts. You have found an interesting article about coelacanths but in order to develop your article you need to be able to explain the relationship between key facts you have learned.

  Your Task: In the following there are five italicized passages, find an authoritative web page that explains the relationship between two of the italicized facts.

  “The coelacanth (“see-la-kanth”) is a lobe-finned fish related to lungfish and to primitive tetrapods (animals with four limbs, like us). They have a long fossil record from the Devonian to the Cretaceous - about 300 million years. But no coelacanth fossils have been discovered in sediments younger than the Cretaceous, about 70 million years ago. Yet a specimen was discovered in 1938 at East London, South Africa and a number have been caught in shallow open waters of less than 100 meters since then. The ’home’ of the coelacanths is the Comoros Islands, a young West Africa volcanic island chain only 6.5 million years old, but a few have been caught elsewhere in West Africa and at least two in Indonesia. It is estimated there are only a thousand or so coelacanths in the world.

  Coelacanths are the size of humans, are believed to live for 100 years or more, and are poor swimmers. They are slate-blue when alive, with white flecks on the thick scales that cover their bodies. They live in the gloaming, around 200-400 meters below the surface, where light barely penetrates and few creatures venture. They spend their days sheltering in rocky caves in small groups. Unlike most fish, they give birth to live young - small, perfectly formed baby coelacanths - and when disturbed they lift themselves into headstands, apparently using an electro-sensory organ in their snout to detect the presence of predators or prey.

The handful of people who have seen them in their natural habitat talk of their glowing eyes and their gentle demeanor. They describe coelacanths moving with surprising grace, deploying their fanned fins in a diagonal formation - right fin in front, left trailing behind - that is similar to a lizard walking. Their blue and silver color provides excellent camouflage in the underwater cave homes."

- **Task = Interview Preparation (INT). Topic = Coelacanths. Product = Intellectual. Goal = Amorphous.**

  Your Assignment: You are writing an article that profiles a scientist and their research work. You are preparing to interview Mark Erdmann, a marine biologist, about coelacanths and conservation programs.

  Your Task: Identify and save authoritative web pages for the following:

  Identify two (living) people who likely can provide some personal stories about Dr. Erdmann and his work.

  Find the three most interesting facts about Dr. Erdmann's research.

  Find an interesting potential impact of Dr. Erdmann's work.

- **Task = Copy Editing. Topic = Methane clathrates and global warming. Product = Factual. Goal = Specific.**

  Your Assignment: You are a copy editor at a newspaper and you have only 20 minutes to check the accuracy of the six italicized statements in the excerpt of a piece of news story below.

  Your Task: Please find and save an authoritative page that either confirms or disconfirms each statement.

  Arctic methane 'time bomb' could have huge economic costs

  "Scientists say that the release of large amounts of methane from thawing permafrost in the Arctic could have huge economic impacts for the world. The researchers estimate that the climate effects of the release of this gas could cost \$60 trillion, roughly the size of the global economy in 2012. The impacts are most likely to be felt in developing countries they say.

Scientists have had concerns about the impact of rising temperatures on permafrost for many years. Large amounts of methane are concentrated in the frozen Arctic tundra but are also found as semi-solid gas hydrates under the sea. Previous work has shown that the diminishing ice cover in the East Siberian Sea is allowing the waters to warm and the methane to leach out. Scientists have found plumes of the gas up to a kilometer in diameter rising from these waters.

In this study, the researchers have attempted to put an economic price on the climate damage that these emissions of methane could cause. Methane is a powerful greenhouse gas, even though it lasts less than a decade in the atmosphere. Using an economic model very similar to the one used by Lord Stern in his 2006 review of the economics of climate change, the researchers examined the impact of the release of 50-gigatonnes of methane over a decade. They worked out that this would increase climate impacts such as flooding, sea level rise, damage to agriculture and human health to the tune of \$60 trillion.

The researchers say their study is in marked contrast to other, more upbeat assessments of the economic benefits of warming in the Arctic region. It is thought that up to 30% of the world's undiscovered gas and 13% of undiscovered oil lie in the waters. Transport companies are looking to send increasing numbers of ships through these fast melting seas. According to Lloyds of London, investment in the Arctic could reach \$100 billion within ten years.

But according to the new work, these benefits would be a fraction of the likely costs of a large scale methane emission. The authors say a release of methane on this scale could bring forward the date when global temperatures increase by 2C by between 15 and 35 years. Some scientists have cautioned that not enough is known about the likelihood of such a rapid release of methane. Even though it has been detected for a number of years, it has as yet not been found in the atmosphere in large amounts. “We are seeing increasing methane in the atmosphere. When you look at satellite imagery, for instance the MeToP satellite, that’s gone up significantly in the last three years and the place where the increase is happening most is over the Arctic,” Prof Wadhams said.”

- **Task = Story Pitch. Topic = Methane clathrates and global warming. Product = Factual. Goal = Amorphous.**

  Your Assignment: You are planning to pitch a science and economics story to your editor and need to identify interesting facts about the economic impact of global warming on the Arctic region.

  Your Task: Find and save web pages that contain the six most interesting facts about the world economic impact of global warming on the Arctic region.

- **Task = Relationships. Topic = Methane clathrates and global warming. Product = Intellectual. Goal = Amorphous.** Your Assignment: You are writing an article about the Arctic and global warming. You have found an interesting article about the potential for global warming to both increase economic development of the Arctic but also to accelerate and magnify the world economic impact of global warming. To develop your article you need to be able to explain the relationship between key facts you have learned.

  Your Task: In the following, there are five italicized passages. Please identify a relationship between any two of the facts and find an authoritative page that explains the relationship between the two italicized facts.

  ”There are economic benefits of warming in the Arctic region. It is thought that up to 30% of the world’s undiscovered gas and 13% of undiscovered oil lie in the waters. Transport companies are looking to send increasing numbers of ships through these fast melting seas. According to Lloyds of London, investment in the Arctic could reach $100 billion within ten years.

  There are also economic costs. Scientists have had concerns about the impact of rising temperatures on permafrost for many years. Large amounts of methane are concentrated in the frozen Arctic tundra. Previous work has shown that the diminishing ice cover in the East Siberian Sea is allowing the waters to warm and the methane to leach out. Scientists have found plumes of the gas up to a kilometer in diameter rising from these waters.

Methane lasts less than a decade in the atmosphere. Some scientists have cautioned that not enough is known about the likelihood of such a rapid release of methane. Even though it has been detected for a number of years, it has as yet not been found in the atmosphere in large amounts. The release of 50-gigatonnes of methane over a decade would increase climate impacts such as flooding, sea level rise, damage to agriculture and human health to the tune of $60 trillion.

"We are seeing increasing methane in the atmosphere. When you look at satellite imagery, for instance the MeToP satellite, that's gone up significantly in the last three years and the place where the increase is happening most is over the Arctic," Prof Wadhams said."

- **Task = Interview Preparation. Topic = Methane clathrates and global warming. Product = Intellectual. Goal = Amorphous.**

  Your Assignment: You are writing an article that profiles a scientist and their research work. You are preparing to interview Igor Semiletov, an expert in chemical oceanography, about Arctic greenhouse gases and global warming.

  Your Task: Identify and save authoritative web pages for the following: Identify two (living) people who likely can provide some personal stories about Dr. Semiletov and his work.

  Find the three most interesting facts about Dr. Semiletov's research.

  Find an interesting potential impact of Dr. Semiletov's work.

4.4.3 Survey Data

Variables regarding general search experience were captured in the demographic questionnaire at the start of the study. Participants were asked about their frequency of searching, the number of years they have been doing online searching, and their expertise regarding searching and journalism. Task-related questionnaires were also provided before and after participants worked on each task. These included a pre-task question about topic familiarity and experience with the type of assignment, questions about anticipated (pre-) difficulty and actual (post-) difficulty, and whether the participant felt they had enough time. Unlike other datasets, this one uniquely collected survey data about participants' intentions for each query segments. For analyses, we treated each intention as a binary variable (1 if present, 0 if absent) and summed categories together (identify, learn, keep, evaluate), with 3 categories grouped together because of their similarity (find/access/obtain). Several questions were asked in each questionnaire, but the subset of questions chosen for analyses - as well as summary statistics - are provided in Table 4.4.

In addition to the demographic, pre-task, and post-task data, participants were asked about their intentions for each query segment. After completing each task, participants were provided a video of their prior search activity. For each query segment, participants were asked to report what intentions they had when issuing the query and examining subsequent information, including all activity up to and until the next query. They could choose any number of the following options per query segment, which were grouped into categories:

Identify: Identify something to get started - For instance, find good query terms. **Identify: Identify something more to search** - Explore a topic more broadly. **Learn: Learn domain knowledge** - Learn about the topic of a search. **Learn: Learn database content** - Learn the type of information/resources available at a particular website - e.g., a government database. **Find: Find a known item** - Searching for an item that you were familiar with in advance. **Find: Find specific information** - Finding a predetermined piece of information. **Find: Find items sharing a named characteristic** - Finding items with something in common. **Find: Find items without predefined criteria** - Finding items that will be useful for a task, but which haven't been specified in advance. **Keep: Keep record of a link** - Saving a good item or an item to look at later **Access: Access a specific**

Category	Variable	Description	Values	Summary
Background	Search Expertise	Please indicate your level of expertise with searching.	Likert: 1-Novice, 7-Expert	$\mu = 4.875, \sigma = 1.00$
Background	Search Years	How many years have you been doing online searching?	Numeric	$\mu = 10.65, \sigma = 3.01$
Background	Search Frequency	How often do you search using search engines or other online search tools?	Likert: Never, 5-11 times/year, 1-2 times/month, 1-2 days/week, 3-5 days/week, Once a day, several times a day	$\mu = 6.75, \sigma = 0.59$
Background	Journalism Searching	How often have you conducted online searching for journalism-related tasks?	Likert: Never, Once or twice, 3-5 times, More often	$\mu = 3.35, \sigma = 0.92$
Experience	Topic Familiarity	How familiar are you with the topic of this assignment?	Likert: 1-Not at all, 4-Somewhat, 7-Extremely	$\mu = 1.725, \sigma = 1.30$
Experience	Assignment Experience	How much experience do you have with this kind of assignment?	Likert: 1-Not at all, 4-Somewhat, 7-Extremely	$\mu = 3.05, \sigma = 1.83$
Experience	Task Difficulty (Post)	How difficult was it to find the information you need for this assignment?	Likert: 1-Not at all, 4-Somewhat, 7-Extremely	$\mu = 2.8, \sigma = 1.65$
Experience	Adequate Time	Did you have enough time to complete the assignment successfully?	Likert: Far too little, Too little, Barely enough, Enough, More than enough	$\mu = 4.1, \sigma = 1.03$
Intentions	Query-level intentions	The searchers' intentions during a query segment [142]	20 indicators: present or absent, in 5 groups (numeric count)	$(\mu_{frequency}, \sigma_{frequency}) =$ (21.05%11.15%)

Table 4.4: Statistics on survey data in the Intentions data set. Questions are grouped into broad categories. Full question text is provided, as well as the range of possible values and summary statistics.

Category	Variable	Description	Values	Summary
Behavior	# Pages	# Pages	Count	$\mu = 5.75, \sigma = 2.96$
Behavior	Total content dwell time	Total time on pages	Seconds	$\mu = 76.01, \sigma = 95.47$
Behavior	Total SERP dwell time	Total time on SERPs	Seconds	$\mu = 8.79, \sigma = 14.61$
Behavior	Query length	Query length	# words	$\mu = 4.97, \sigma = 3.83$

Table 4.5: Statistics on behavioral data used from the Intentions data set.

item - Go to some item that you already know about. **Access: Access items with common characteristics** - Go to some set of items with common characteristics. **Access: Access a web site/home page or similar** - Relocating or going to a website. **Evaluate: Evaluate correctness of an item** - Determine whether an item is factually correct. **Evaluate: Evaluate usefulness of an item** - Determine whether an item is useful. **Evaluate: Pick best item(s) from all the useful ones (EB)** - Determine the best item among a set of items. **Evaluate: Evaluate specificity of an item** - Determine whether an item is specific or general enough. **Evaluate: Evaluate duplication of an item** - Determine whether the information in one item is the same as in another or others. **Obtain: Obtain specific information** - Finding specific information to bookmark, highlight, or copy. **Obtain: Obtain part of the item** - Finding part of an item to bookmark, highlight, or copy. **Obtain: Obtain a whole item(s)** - Finding a whole item to bookmark, highlight, or copy.

4.4.4 Behavioral Features

Behavioral variables were collected passively through Coagmento and the GazePoint eye tracker as participants conducted the study. Coagmento logged page browsing activities, such as page visits, dwell time on pages, and queries issued to Google, Yahoo, and Bing. Local timestamps were collected to calculate dwell times on pages and search engine result pages (SERPs). GazePoint data was not used in these experiments, with eye tracking left to future work. The subset of variables used in analyses is listed in Table 4.5.

4.5 Dataset 2: Searching as Learning

4.5.1 Data Collection Design

This dataset was also conducted as part of a study, but while the former dataset was collected in a laboratory setting, this data was collected in a naturalistic setting. Participants were allowed to conduct the study activities in any arbitrary setting, not under the supervision of a study coordinator. The students recruited were general undergraduate students from Rutgers University, who were required to be at least second year undergraduates. They were also required to use Google Chrome to complete the study. The study could be completed over the course of 3 consecutive days. Participants were asked to complete 4 tasks without time limit. Each task was preceded by a pre-task interview, proceeded by a post-task interview, and subsequently followed with an annotation task. Participation with the study began with a demographic questionnaire and ended with a verbal exit interview. Participants were additionally asked to collect relevant information they found for the task in an online text editor Etherpad[7]. In the post-task questionnaires, participants were asked to informally comment on their search activities. Coagmento[8] [108] was installed on participants' Chrome browsers, which was used to direct people to the study's website so they could conduct the study. Coagmento also recorded browsing information and sent it to a remote server. For the entire process, participants were incentivized with additional reward for being among the best performers. This provided incentive to issue good searches, instead of meeting the minimum requirements. In total, the data collected in this study was for 30 participants.

4.5.2 Assigned Tasks

There were 4 tasks, each under the topic "cyber bullying", a topic that would likely be interesting to the general undergrad population. The task descriptions were broad, and the tasks were mostly intellectual (with one mixed factual+intellectual product),

---

[7] http://etherpad.org/

[8] http://coagmento.org

Table 4.6: Task type and session characteristics for the Searching as Learning dataset.

Task	Product	Goal	$\|T\|$	$\|Q\|$
Similar (SIM)	Mixed	Specific	30	168
Clementi (TYC)	Intellectual	Specific	30	187
Causes (CAU)	Intellectual	Specific	30	110
Strategies (STR)	Intellectual	Specific	30	129

as the purpose of the study was to see if learning occurred through the duration of the study (a concept not explored in this dissertation). A faceted classification of each task is provided in Table 4.6. Each user completed all 4 tasks. No rotations in task order were given, and tasks were completed in the order presented below:

- **Product = Mixed (Factual+Intellectual). Goal = Specific.** - What is cyberbullying? How is it similar or different to other types of harassments (e.g. cyberbullying vs. traditional bullying)? What are some long-term/short-term risks faced with cyberbullying?

- **Product = Intellectual. Goal = Specific.** - In 2010, Rutgers University has witnessed the tragic incident of Tyler Clementi, whose case raised concerns about cyberbullying. Find out more about this case, and possibly some other cases. What does this case(s) show you about some common characteristics of cyberbullying?

- **Product = Intellectual. Goal = Specific.** - Having heard some of the recent reports on cyberbullying, what seems to be the main cause of the bullying behavior online? How much are technology and use of electronic communication associated with cyberbullying? Why?

- **Product = Intellectual. Goal = Specific.** - How effective are some of the currently available strategies to mitigate cyberbullying at schools and university campuses? Why? Which strategy/method do you think is best and why?

Category	Variable	Description	Values	Summary
Background	Search Years	How many years of online search experience do you have?	Numeric	$\mu = 9.43, \sigma = 4.37$
Background	Search Frequency	How often do you search the Web per day?	Ordinal: 1-3 searches per day, 4-6 searches per day, 7-10 searches per day, 10+ searches per day	$\mu = 3.13, \sigma = 0.92$
Background	Search Expertise	How would you rate your level of online searching skills?	Likert: 1-Novice, 5-Expert	$\mu = 4.16, \sigma = 0.68$
Experience	Topic Familiarity	How much do you think your knowledge on this topic will help you with the task?	Likert: Not at all helpful, A little helpful, Somewhat helpful, Sufficiently helpful, I know a lot helpful	$\mu = 2.78, \sigma = 1.09$
Experience	Task Difficulty (Post)	Overall, how difficult was this task?	Likert: 1-Not at all difficult, 2-A little difficult, 3-Somewhat difficult, 4-Much difficult, 5-Extremely difficult	$\mu = 1.86, \sigma = 0.91$

Table 4.7: Statistics on survey data in the Searching as Learning data set.

4.5.3 Survey Data Used

Variables regarding general search experience were captured in the demographic questionnaire at the start of the study. Participants were asked about their online searching skills and general search experience. Task-related questionnaires were also provided before and after participants worked on each task. These included a pre-task question about topic familiarity and questions about anticipated (pre-) difficulty and actual (post-) difficulty. The subset of questions chosen for analyses - as well as summary statistics - are provided in Table 4.7.

4.5.4 Behavioral Features Used

Behavioral variables were collected passively through Coagmento as participants conducted the study. Coagmento logged page browsing activities, such as page visits, dwell time on pages, and queries issued to Google, Yahoo, and Bing. Local timestamps were collected to calculate dwell times on pages and search engine result pages (SERPs).

Category	Variable	Description	Values	Summary
Behavior	# Pages	# Pages	Count	$\mu = 2.25, \sigma = 3.31$
Behavior	Total content dwell time	Total time on pages	Seconds	$\mu = 74.32, \sigma = 134.09$
Behavior	Total SERP dwell time	Total time on SERPs	Seconds	$\mu = 13.80, \sigma = 26.69$
Behavior	Query length	Query length	# words	$\mu = 4.41, \sigma = 4.79$

Table 4.8: Statistics on behavioral data used from the Searching as Learning data set.

The subset of variables used in analyses is listed in Table 4.8.

4.6 Dataset 3: Expert Opinions

4.6.1 Data Collection Design

This dataset was collected in a mixed environment - in both a naturalistic and a laboratory setting. Rutgers undergraduate students were recruited from the general university population rather than a specific demographic. Participants were only required to use Google Chrome as their default browser. The study was split into 3 parts, with participants required to conduct 2 tasks in each part. The first part was conducted in a naturalistic environment, the second in a university laboratory, and the third in a naturalistic environment. Each task was preceded by a pre-task interview and proceeded by a post-task interview. Participation with the study began with a demographic questionnaire and ended with a verbal exit interview. Participants were asked to use the Chrome browser at all stages of the study, and users installed a Coagmento Chrome extension for their home activities.

Participants began by answering the demographic questionnaire. Then, in the naturalistic portions of the study, participants were asked to conduct 2 search tasks that were created by the researchers. For each task, participants read the task description and answered a short questionnaire on their familiarity with the topic and task as well as the anticipated difficulty. They then had 20 minutes to complete the search task. Afterwards, participants answered a post questionnaire on the actual difficulty of the

Table 4.9: Task type and session characteristics for the Expert Opinion dataset.

Task	Product	Goal	$\|T\|$	$\|Q\|$
Movie (MOV)	Factual	Specific	30	161
Retirement (IRA)	Factual	Amorphous	30	75
Book (BOK)	Factual	Specific	30	294
Beijing Airport (PEK	Factual	Amorphous	30	256
TV shoes (PIN)	Factual	Specific	30	249
PhD programs (PHD)	Factual	Amorphous	30	239

task. In the laboratory setting, participants were asked to answer one new questionnaire. Users reported about their feelings about their own personal search efficacies, such as whether they have enough knowledge to help others search, whether they remain calm when facing difficulties searching, whether they are confident they can deal with complex search tasks, and whether they expect they can search like an expert in their field. A laboratory session lasted about 70 minutes. The study then concluded with 2 more search tasks in a naturalsitic setting. For the entire process, participants were incentivized with additional reward for being among the best performers. This provided incentive to issue good searches, instead of meeting the minimum requirements. In total, the data collected in this study was for 30 participants.

4.6.2 Assigned Tasks

This experiment contained a total of 6 possible tasks, each with a distinct topic. All contained a factual product, with the goal either being specific or amorphous. We give a faceted classification of each task in Table 4.9, according to a subset of facets in Li [79]. For definitions of each facet, see [29, 79]. Each user completed all 6 tasks in the order below: the first 2 in a naturalistic setting, the second 2 in a laboratory setting, then the last 2 in a naturalistic setting.

- **Product = Factual. Goal = Specific.** - You saw a clip of a TV movie on a bus and want to watch the full movie. You don't know the name of the movie. You only saw that the heroine is an amateur sleuth who owns a bakery. In the clip you were watching, she was worried about her competitor stealing her business. You really like the actor who plays the heroine's friend, a dentist. Find out the name of that actor.

- **Product = Factual. Goal = Amorphous.** - Suppose you are an employee at a US non-profit organization. You are trying to decide whether to contribute money to your employer-offered 403(b) retirement plan or to a personal IRA (Individual Retirement Account). Find resources that could help you decide which is best for your needs. Please provide your decision, a brief description of your rationales, and the links to the resources which you believe can support your decision making.
- **Product = Factual. Goal = Specific.** - You are looking for a book titled "Suede to Rest" by Diane Vallere that is not owned by Rutgers University Library. You want to find the library closest to Rutgers, New Brunswick that owns the book and if it is currently available and how you can check it out.
- **Product = Factual. Goal = Amorphous.** - You plan to visit Beijing and want to find information about Beijing Capital Airport (PEK); what shuttles and public transportations connect the airport to downtown; what hotels are close to the airport; what rental car services do they have; and where cheap parking around the airport is. (Please note that you can only search in English and view English websites.)
- **Product = Factual. Goal = Specific.** - Your friend really likes a pair of shoes that he saw in the South Korean TV series Pinocchio, which was worn by the hero. You want to buy these shoes as a birthday gift for him. But the TV series did not specify the brand of the shoes (the pair in the middle [picture interleaved with task description]). Find out the brand and style of the shoes. (Please note that you can only search in English and view English websites.)
- **Product = Factual. Goal = Amorphous.** - You are applying for Ph.D. programs in the area of Library and Information Science (LIS). You want to find out what the best LIS programs in the U.S. are and which of them guarantee at least four years of funding for Ph.D. students (tuition remission + monthly stipend). You also want to consider their job placement within the last three

years. Did their graduates find tenure-track positions in academia? Taking these factors into consideration, find five LIS programs that you want to apply, and supply your reasons. Please note that the names of the departments may vary (do not have to be LIS), but LIS has to be one of the research areas in your target department.

4.6.3 Survey Data Used

Variables regarding general search experience were captured in the demographic questionnaire at the start of the study, though this demographic questionnaire did not contain survey data about search experience (only about English language proficiency). Task-related questionnaires were also provided before and after participants worked on each task. These included a pre-task question about topic familiarity, questions about anticipated (pre-) difficulty and actual (post-) difficulty, and whether the participant felt they had enough time. In the lab study, participants were asked twice (once before each task) about their general search efficacy. Some expertise variables here were of interest, and for these we took the average of both responses, as they did not differ significantly. The subset of questions chosen for analyses - as well as summary statistics - are provided in Table 4.10.

4.6.4 Behavioral Features Used

Behavioral variables were likewise collected passively through Coagmento as participants conducted the study, in the same fashion. The subset of variables used in analyses is listed in Table 4.11.

Category	Variable	Description	Values	Summary
Background	Search Expertise 1	I can solve most problems involving searching for information if I invest the necessary effort.	Likert: 1-Not at all true,2-Hardly true,3-Moderately true,4-Very true,5-Exactly true	$\mu = 4.1, \sigma = 0.65$
Background	Search Expertise 2	I believe that I have good amount of knowledge on searching to support my daily searching at home for non-work events (e.g., travel plan, housing).	Likert: 1-Not at all true,2-Hardly true,3-Moderately true,4-Very true,5-Exactly true	$\mu = 4.23, \sigma = 0.84$
Background	Search Expertise 3	I believe that I have good amount of knowledge on searching to support my daily searching for completing work task at workplace.	Likert: 1-Not at all true,2-Hardly true,3-Moderately true,4-Very true,5-Exactly true	$\mu = 4.16, \sigma = 0.58$
Background	Search Expertise 4	I believe that I have good amount of knowledge on searching to help others solve regular search-related problems.	Likert: 1-Not at all true,2-Hardly true,3-Moderately true,4-Very true,5-Exactly true	$\mu = 3.8, \sigma = 0.79$
Background	Averaged Search Expertise	(Average of the above)	Likert: 1-Not at all true,2-Hardly true,3-Moderately true,4-Very true,5-Exactly true	$\mu = 4.075, \sigma = 0.46$
Experience	Topic Familiarity	How knowledgeable are you on this topic?	Likert: 1-Not at all, 4-Somewhat, 7-Extremely	$\mu = 3.11, \sigma = 1.70$
Experience	Task Difficulty (Post)	How difficult was it to find the information you need for this task?	Likert: 1-Not at all, 4-Somewhat, 7-Extremely	$\mu = 3.83, \sigma = 1.83$
Experience	Adequate Time	Did you have enough time to complete the task successfully?	Likert: Far too little, Too little, Barely enough, Enough, More than enough	$\mu = 4.07, \sigma = 0.94$

Table 4.10: Statistics on survey data in the Expert Opinion data set.

Category	Variable	Description	Values	Summary
Behavior	# Pages	# Pages	Count	$\mu = 2.00, \sigma = 3.45$
Behavior	Total content dwell time	Total time on pages	Seconds	$\mu = 53.73, \sigma = 90.83$
Behavior	Total SERP dwell time	Total time on SERPs	Seconds	$\mu = 9.72, \sigma = 22.45$
Behavior	Query length	Query length	# words	$\mu = 5.76, \sigma = 3.96$

Table 4.11: Statistics on behavioral data used from the Expert Opinion data set.

Chapter 5

Results and Discussion

5.1 Experiment 1: Confirmatory Analysis

Results for goodness-of-fit of the constructed SEM can be shown in Table 5.1. Tables 5.2-5.3 also explore the number of times pairs of variables contained significant relationships, showing the most frequent relationships. Findings are as follows:

The best model for most metrics uses only background and experience measures. - The BE model does not have the smallest χ^2, which is an unadjusted fit metric. But the model has the smallest χ^2/df. It also ranks the best among all models for AGFI, which adjusts GFI for parsimony, and obtains the lowest RMSEA score. It has a relatively low χ^2 and many degrees of freedom. This also helps to explain that while our full model has the best AIC score, the BE model has the lowest BIC score (150 degrees of freedom vs. 110).

The best-fitting model uses all features, but it is not the simplest - While the full model performs best in χ^2 and unadjusted measures, it is one of the poorest performers for adjusted measures. It has the worst AGFI and PGFI, and χ^2/df is on a

Model Name	# Params	df	χ^2	χ^2/df	RMR	GFI	AGFI	PGFI	RMSEA	AIC	BIC
Full	100	110	**2241.363**	20.376	16.714	**.725**	.475	.380	.167	**2441.363**	2895.466
IB	77	133	2719.991	20.451	20.710	.697	.522	.442	.168	2873.991	3223.650
IE	86	124	2412.235	19.454	**16.666**	.719	.524	.425	.163	2584.235	2974.763
I	71	139	2686.312	19.326	20.746	.696	.540	.460	.163	2828.312	3150.725
BE	60	150	2403.537	**16.024**	17.306	.711	**.596**	.508	**.147**	2523.537	**2795.999**
E	66	144	2458.304	17.072	17.312	.711	.579	.488	.152	2590.304	2890.012
TaskModel	51	159	2773.078	17.189	17.992	.688	.587	**.521**	.153	2835.078	3066.670
Saturated	210	0	0	NA	0	1	NA	NA	0	420	1373.616
Independent	20	190	3857.457	20.302	17.453	.626	.587	.567	.167	3897.457	3988.278

Table 5.1: Experiment 1 - Goodness of fit measures for the path models. Best performance (aside from Saturated and Independent models) are boldfaced.

Path From	Path To	# Direct Paths	# Indirect	# Significant Direct	# Sig. Indirect	# Sig. Total
Goal	Content dwell time	7	6	7	0	7
Goal	Query Length	7	6	7	1	6
Goal	SERP dwell time	7	6	1	0	5
Goal	# pages	7	6	0	0	1
Goal	Find / Access / Obtain	4	0	4	NA	4
Product	Content dwell time	7	6	4	0	4
Product	Query Length	7	6	4	0	3
Product	SERP dwell time	7	6	0	0	1
Product	# pages	7	6	7	4	7
Product	Difficulty	4	0	4	NA	4
Topic	Query Length	7	4	7	4	7
Topic	Content dwell time	7	4	0	0	2
Topic	SERP dwell time	7	4	2	3	1
Topic	# pages	7	4	0	3	1
Topic	Topic Familiarity	4	0	3	NA	3
Topic	Difficulty	0	4	NA	1	1

Table 5.2: Experiment 1 - Significant pathways from task goal, product, topic to endogenous variables. Includes total number of graphs with direct, indirect, and total effects, as well as the number of graphs with significant such effects.

Path From	Path To	# Direct Paths	# Indirect	# Significant Direct	# Sig. Indirect	# Sig. Total
Int - Evaluate	# pages	4	0	3	NA	4
Int - Evaluate	Content dwell time	4	0	3	NA	4
Int - Find / Access / Obtain	# pages	4	0	4	NA	4
Int - Identify	# pages	4	0	4	NA	4
Int - Identify	SERP dwell time	4	0	4	NA	4
Int - Identify	Query length	4	0	2	NA	2
Int - Keep	# pages	4	0	1	NA	2
Intent - Keep	Content dwell time	4	0	2	NA	2
Rushed	Content dwell time	4	4	4	0	4
Rushed	Query length	4	4	0	0	4
Search Expertise	SERP dwell time	0	3	NA	0	1
Search Expertise	Content dwell time	0	3	NA	2	1
Search Expertise	Query length	0	3	NA	1	1
Journalism Expertise	# pages	0	3	NA	0	1
Topic Familiarity	# pages	4	4	4	0	4
Topic Familiarity	SERP dwell time	4	4	4	0	4
Topic Familiarity	Query length	4	4	4	0	4

Table 5.3: Experiment 1 - Significant pathways from other survey variables to behaviors.

par with the independent model assuming no relationships.

In general, intentions reduce χ^2 at the cost of goodness of fit - Keeping the background and experience constant, toggling the intents toggles the degrees of freedom by 20 to 40, with a small improvement in χ^2 (2403.537-2241.363=162.174, 2458.304-2412.235=46.069, 2773.078-2686.312=86.766). Also, each model with intentions performs worse in several parsimony-based metrics with respect to its counterpart without intentions. This happens universally for χ^2/df, AGFI, PGFI, RMSEA, and BIC.

Experience variables account for much variance - All other things held constant, removing the links to and from experience variables adds substantial χ^2 (2719.991-2241.363=478.628, 2686.312-2412.235=274.077, 2773.078-2458.304=314.774). GFI, AGFI, and PGFI improve when removing experience, but most other metrics worsen.

None of the models is a particularly good fit - The saturated baseline can indeed be achieved by connecting all pairs of variables, and it perfectly fits the data. For good-fitting models, ideal fits for χ^2/df, GFI, AGFI, PGFI, and RMSEA are 2-5,0.9,0.9,0.9, and 0.08, respectively. That said, while our models do make improvements over the saturated or independent models in most metrics, our models are far from the ideal range, including the full model. This suggests that there are many connections not covered in this full model that should be included. This suggests that either literature was missed in the literature review (an oversight on a parameter to estimate) or perhaps gaps in the literature exist.

Inasmuch as covered by this model, there are still direct paths from task type (and other user characteristics) to browser signals - There are very frequently total and direct effects from task goal, product, and topic to our browser features. This may be a genuine direct effect or due to some unrecorded variable. Similarly, consistent direct effects were found from time pressure to content dwell time and query length and search expertise to content dwell time.

Topic familiarity also plays an important role - Each time topic familiarity was included in our model, it had a significant effect on the browsing features. Moreover, topic was only linked to topic familiarity and had significant indirect effects to certain browsing features 3-4 times, particularly query length, SERP dwell time, and number of pages. Therefore, topic influences these not only directly but indirectly through a user's topic familiarity.

Product (not goal) directly and indirectly influences behavior - Specifically, task product has an indirect effect on the # of pages viewed through the difficulty experienced by the searcher, but it also has a significant direct effect on this behavior. This contrasts with task goal, which does not have this indirect effect.

Intentions can influence searchers' behavior, but influence from task type to intention was not found - Several direct effects from intentions to behaviors can be found in Table 5. However, only task goal influences find/access/obtain intentions, even though it does so in every model. While intentions may influence their respective search session, perhaps intentions of a single query segment do not neatly map to task types. Perhaps intentions aggregated over an entire session map neatly to task type but not within a single query segment (counter to [4]). In our data, this would make a difference: even though there are 693 query segment and 693 corresponding intention vectors, there are only 80 sessions on 2 task products, 2 task goals, and 2 topics.

There is some influence from a user's background - occasionally, a user's search expertise and journalism expertise affects browsing behaviors, as in Table 5, but are not affected by task.

What, are the final takeaways from these findings? First, the findings here give credence to this type of modeling. We first found several instances of **mediation**, which is a phenomenon that could not be modeled with previous frameworks. The results show that topic affects browsing behaviors directly but also intermediately through a user's familiarity. Additionally, while product and goal directly influence behavior, only product indirectly influences the # of pages indirectly through the difficulty experienced by a searcher. Second, this new approach to modeling additionally confirms several previous findings. Task still seems to have important direct effects on browsing behavior. Factors aside from task still certainly influence behaviors. The simultaneous effects of all these types of variables have rarely, if ever, been measured. Additionally, this simultaneous is perhaps one of task prediction's obstacles, as these simultaneous

Data set	Algorithm	$\lvert V\rvert$	$\lvert E\rvert$	Total	$\geq$ 75%	$\geq$ 50%	$\geq$ 25%
Intentions	HC	15	210	209	17	31	55
Intentions	IAMB	15	210	199	10	31	55
Learning	HC	11	110	90	6	11	19
Learning	IAMB	11	110	90	4	19	25
Expert	HC	9	72	72	2	13	25
Expert	IAMB	9	72	72	5	18	26
Merged	HC	9	72	72	6	10	19
Merged	IAMB	9	72	0	0	0	16

Table 5.4: Experiment 2 - Overview of the number of chosen edges by each algorithm. Includes the number of variables ($\lvert V\rvert$), the total number of possible edges ($\lvert E\rvert$=(V choose 2) × 2 for edge orientation), the total number of edges chosen with nonzero probability, and the number over particular probability thresholds.

effects aren't taken into account. Third, none of the models presented here are an ideal fit with respect to the data, meaning that there is perhaps a gap in the literature. It means that according to this data, there are some important links that are not drawn, because they have not been covered by our literature review (and perhaps the literature generally). There is important unconsidered influence between some of these variables. Lastly, the task model using just background and experience information seems to provide the best fit overall, even though intentions data still has an affect on browsing. This at least justifies the removal of intentions in future experiments and allows analysis with the other data sets, allowing for the methodology of Experiment 2. Intentions could be removed from Experiment 2 analysis to allow multiple data sets to be integrated. This provided an opportunity to address some of the gaps from the first experiment. Namely, what model structure would be good if the structure here was not necessarily sufficient?

5.2 Experiment 2: Structure Learning

The results for Experiment 2 are summarized in Tables 5.4-5.7. Table 5.4 gives a rough overview of the experiments regarding the sizes of the graphs. It shows the total number of possible edges that could have been included in the graph and the total edges with a nonzero probability. It also shows how many edges were included with a high probability. The scoring functions for HC and IAMB discount for parsimony, so not many edges were included in the graphs, and this table briefly illustrates that.

For phase one of the study – in which all 3 datasets were combined in a bootstrapped process of graph construction – the common proposed graph can be found in Figures 5.1-5.2. This example was found using hill climbing. Superimposed on the edges of the figures are the probabilities that each edge was included by the respective algorithm. In addition, the probabilities for these edges and several more edges can be found in Table 5.5. The edges included in the proposed graph are boldfaced in this table. Most of the edges are among the highest probability edges according to both algorithms. All 13 edges are included in the top 20 for HC, with the top 10 being included in the graph. 12 out of 13 are in the top 20 of IAMB, and all 13 are in the top 30 (the last one with probability .117). Some of the top edges according to IAMB not included in the graph include *Search Expertise* → *Topic Familiarity*, *Content Time* → *Goal*, and *SERP Time* → *Content Time*, and *Topic Familiarity* → $Difficulty$. It should be noted, though, that in these cases the reverse edge is included instead, hence IAMB makes different decisions about orientation but not necessarily about inclusion. Considering in addition that 3 separately collected datasets were used to create this graph this graph, this will be taken as an acceptable graph explaining the relationship between the variables. An interpretation of this graph will be given in the last remarks of this experiment.

We next proceed to phase two. It is noted here that results differ. To compare phase one to phase two, 9 out of the 13 edges were found among the top 20 in at least one of the experiments in phase two. While not terribly strong evidence, this should not be considered invalidation for the proposed structure in phase one. Rather, it should be noted that the strongest relationships in phase two mostly include variables that did not overlap between all 3 data sets. Such variables include general search experience, experience in a field (e.g. journalism), frequency of searching, the amount of years spent searching, and topic as a controlled variable. If anything, the results tease out other possible relationships that should be considered in the future. One such noteworthy one is the relationship between demographic characteristics (field expertise, general search experience) and adequate time, as well as session characteristics (topic,

IAMB			Hill-Climbing		
Search Expertise	**# Pages**	**0.479**	**Goal**	**Product**	**0.992**
Content time	**# Pages**	**0.477**	**Topic Familiarity**	**Product**	**0.980**
Goal	**Product**	**0.471**	**Search Expertise**	**# Pages**	**0.954**
Topic Familiarity	**Product**	**0.462**	**Difficulty**	**Product**	**0.944**
Search Expertise	Topic Familiarity	0.453	**Content time**	**# Pages**	**0.920**
Topic Familiarity	**Search Expertise**	**0.452**	**Topic Familiarity**	**Search Expertise**	**0.875**
Content time	Goal	0.445	**Goal**	**Content time**	**0.746**
Goal	**Content time**	**0.433**	**Difficulty**	**Topic Familiarity**	**0.732**
SERP time	Content time	0.414	**Difficulty**	**# Pages**	**0.728**
Topic Familiarity	Difficulty	0.406	**Content time**	**SERP time**	**0.658**
Difficulty	**Topic Familiarity**	**0.398**	Goal	Query Length	0.501
Difficulty	**Query Length**	**0.367**	**Difficulty**	**Query Length**	**0.485**
Content time	**SERP time**	**0.353**	**Product**	**Query Length**	**0.445**
Query Length	Difficulty	0.324	Query Length	Difficulty	0.334
Product	**Query Length**	**0.296**	Search Expertise	Goal	0.304
Product	Difficulty	0.268	Product	# Pages	0.297
# Pages	Difficulty	0.243	Topic Familiarity	# Pages	0.292
Difficulty	**Product**	**0.226**	Topic Familiarity	Difficulty	0.267
Query Length	Goal	0.224	Topic Familiarity	SERP time	0.259
Goal	Query Length	0.214	**Search Expertise**	**SERP time**	**0.241**

Table 5.5: Overview of the top 20 edges and their respective probabilities in the merged data set, for the IAMB and Hill-climbing algorithm.

topic familiarity, difficulty) and adequate time. Multiple data sets show a relationship between the two in some orientation. Namely: *Difficulty → Adequate Time*, *Topic → Adequate Time*, *Field Expertise → AdequateTime*, *Difficulty → Adequate Time*, *Search Experience → Adequate Time*. Moreover, a more complex relationship was shown between topic, product, and topic familiarity. A proposed construct that could be drawn in addition could be *Product ← Topic → Topic Familiarity → Behaviors* (viz., # pages and time spent on content and SERPS). For this, see the IAMB results for the intentions study and both results for the expert opinion study.

Some cautionary notes should be mentioned. First, each of the datasets is one third of the size of the datasets measured. Hence, the above results are tentative and speculative, with perhaps more data being necessary to have more stable results. Secondly, the orientation of the edges should be interpreted correctly. Statements of causation are not being made, nor can they of course be made through statistics alone. The relationships are interpreted only as statements of conditional independence.

With that, let us further explore the proposed graph in Figures 5.1-5.2. First,

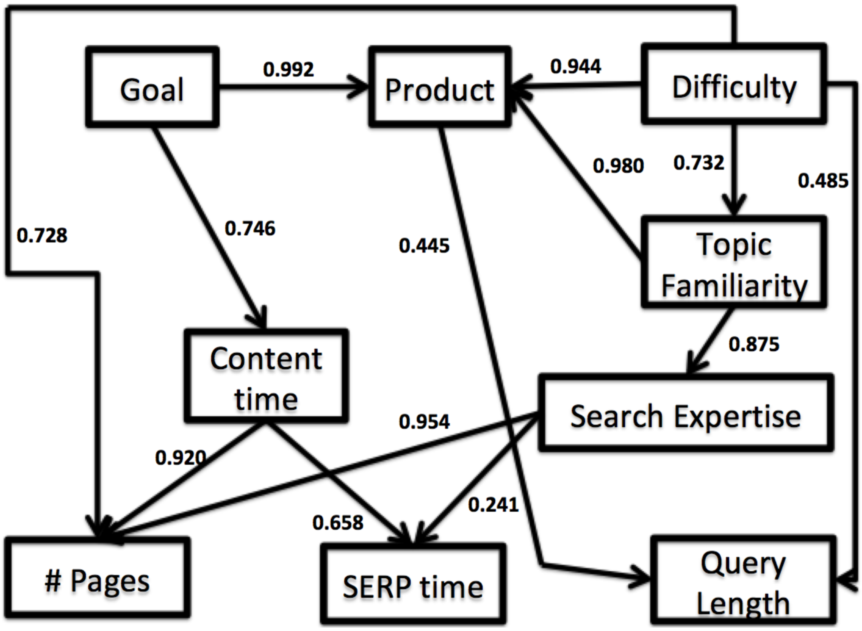

Figure 5.1: A common graphical model proposed by Hill-Climbing using the merged data set. Weights indicate the Hill-Climbing probabilities in Table 5.5

several results exist that agree with past literature. For instance, several session or task characteristics are directly related to behavior. For instance, product directly affects query length, goal affects content time, search expertise affects the number of pages and amount of time spent on SERPs, and difficulty affects query length. Content time, number of pages, and SERP time have some effects on each other, but this is understandably due to correlation (an issue which should be revisited in a future experiment). Further, these are affected by task and user characteristics, and not the other way around (as should be expected). Goal "affects" product, only due to correlation through experimental design. Goal is at the very top of the DAG along with difficulty, and behaviors are at the bottom, which is consistent with intuition. The graph structure is also very similar to the structure proposed in Figure 3.5, with perhaps some disagreement about the arrangement of difficulty, topic familiarity, and search expertise. Difficulty, in this model, affects topic familiarity and search expertise and product. While the former two can be seen as correlation between the variables, perhaps there is a genuine relationship between whether the task is fact-finding, whether the

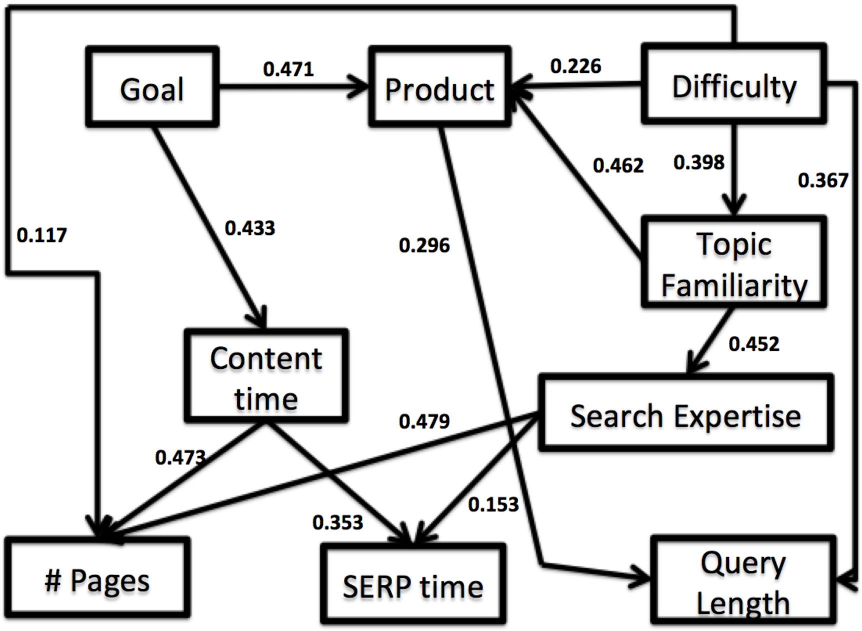

Figure 5.2: A common graphical model proposed by Hill-Climbing using the merged data set. Weights indicate the IAMB probabilities in Table 5.5

task is mixed, or whether it is intellectual, which would also agree with past literature.

Lastly, taking the graph at face value, some interesting relationships of independence and conditional independence can be drawn. These entail the limitations learning about task characteristics from data. $X \perp\!\!\!\perp Y$ – independence – states that information of X provides no information about Y (and vice versa). $X \perp\!\!\!\perp Y|S$ states that when S is known, information about X provides no information about Y (and vice versa). The most interesting are as follows:

- **Independence** - $Content\ Time \perp\!\!\!\perp Search\ Expertise$; $Goal \perp\!\!\!\perp Difficulty$;s $Goal \perp\!\!\!\perp Topic\ Familiarity$; Product not independent of Difficulty; SERP Time not independent of Search Expertise
- **Conditional Independence** - $Query\ Length \perp\!\!\!\perp Goal|Product$; $\#Pages \perp\!\!\!\perp Product|Goal$; $Content\ Time \perp\!\!\!\perp Product|Goal$; $SERP\ Time \perp\!\!\!\perp Product|Goal$, $Query\ Length \perp\!\!\!\perp Everything|Difficulty, Product$; $Content\ Time \perp\!\!\!\perp Topic\ Familiarity|Difficulty, Goal$; $Content\ Time \perp\!\!\!\perp Search\ Expertise|Difficulty, Goal$; Goal not independent of Difficulty given Product

Intentions			Learning			Expert		
From	To	P()	From	To	P()	From	To	P()
Dif	TimAd	0.981	Dif	SeExp	0.8776	QuLen	Dif	0.9026
FiExp	SeExp	0.9474	SFreq	SeExp	0.8736	Dif	TimAd	0.8964
Years	FiExp	0.8912	Prod	TopFam	0.8676	SeExp	SERP	0.8762
Dif	TasExp	0.8632	Cont	Page	0.8388	Cont	Page	0.829
Dif	SFreq	0.8522	Dif	TopFam	0.7514	TopFam	TimAd	0.7888
Topic	TimAd	0.8118	SFreq	SERP	0.6964	TopFam	Cont	0.7398
TopFam	Page	0.8022	Page	Cont	0.6938	Prod	SeExp	0.7384
QuLen	Topic	0.7928	SeExp	Dif	0.6926	SeExp	TopFam	0.7318
Topic	TopFam	0.7622	Years	SFreq	0.6904	TopFam	Page	0.7198
Cont	Page	0.7544	Dif	SFreq	0.6764	TopFam	SeExp	0.6834
Goal	Prod	0.7456	SFreq	Years	0.6708	SeExp	TimAd	0.6782
SFreq	SeExp	0.7144	SeExp	SFreq	0.663	SERP	SeExp	0.6746
Years	TopFam	0.7052	SFreq	Dif	0.6424	TimAd	Goal	0.6538
TopFam	TasExp	0.6868	Years	SERP	0.5822	TimAd	Dif	0.634
Dif	Prod	0.6704	Years	QuLen	0.5662	Dif	TopFam	0.6022
Topic	Prod	0.6362	QuLen	Years	0.5608	Goal	TopFam	0.5896
Prod	Dif	0.635	Years	Dif	0.554	Dif	QuLen	0.5574
Years	SeExp	0.608	Dif	Years	0.551	Cont	Goal	0.5292
SeExp	Years	0.5836	SERP	SFreq	0.5004	SERP	TopFam	0.4948
TasExp	SeExp	0.5762	SERP	Years	0.4862	TopFam	Goal	0.4868

Table 5.6: Overview of the top 20 edges and their respective probabilities in each dataset, for the IAMB algorithm.

For instance, if the task product and the difficulty the user is experiencing are given or discovered, these wholly determine the query length. No other information is needed nor useful, and conversely, query length can tell you nothing about additional task and user attributes. Similarly, given the product and query length, the query length can give no more additional information about the goal. The other behaviors can likewise tell nothing about the product once the goal is known. Goal is independent of difficulty, but interestingly product is not independent of difficulty. Time spent on content – as difficulty and task goal become known – can provide no information about the searcher's topic familiarity and search expertise. As mentioned, these entail the limitations of the data. As more information is discovered about the searcher's task, familiarity, and expertise, certain features become less and less useful. The nature of features' usefulness becomes explicit using this graphical modeling framework.

Intentions			Learning			Expert		
From	To	P()	From	To	P()	From	To	P()
Dif	TimAd	1.000	Prod	TopFam	0.9642	SeExp	TopFam	0.9466
Topic	TimAd	0.9996	SFreq	SeExp	0.9524	Page	Cont	0.762
FiExp	TimAd	0.999	Years	Dif	0.8876	SeExp	SERP	0.7486
SeExp	SFreq	0.9944	SFreq	Dif	0.8644	SeExp	TimAd	0.6902
Goal	Prod	0.9934	Years	SFreq	0.8004	QuLen	Dif	0.6658
Years	TimAd	0.9714	Page	Cont	0.7936	Cont	Goal	0.6386
Goal	TopFam	0.9694	SeExp	Dif	0.7384	Dif	TimAd	0.6098
SeExp	TasExp	0.9396	Dif	TopFam	0.6442	TimAd	Goal	0.5918
SeExp	FiExp	0.9142	Years	QuLen	0.626	TopFam	Dif	0.546
Dif	SFreq	0.8768	SFreq	SERP	0.5546	Page	TopFam	0.5398
SeExp	Years	0.8706	Years	SERP	0.4916	SeExp	Dif	0.5394
Prod	Dif	0.8546	Cont	SERP	0.4152	TopFam	SERP	0.5378
Topic	QuLen	0.8442	TopFam	QuLen	0.3958	Dif	TopFam	0.454
TopFam	TasExp	0.837	Prod	SFreq	0.3856	SERP	TopFam	0.4328
FiExp	SFreq	0.829	TopFam	Dif	0.3402	TopFam	Page	0.4238
Page	Cont	0.7954	SFreq	QuLen	0.3284	Goal	TopFam	0.412
Topic	SFreq	0.7934	SeExp	TopFam	0.2698	Goal	TimAd	0.4082
Years	TasExp	0.703	Dif	SeExp	0.2606	TopFam	Goal	0.4058
Years	TopFam	0.6856	SERP	SFreq	0.2568	TimAd	Dif	0.3902
TopFam	SFreq	0.6956	SeExp	QuLen	0.2572	TopFam	TimAd	0.3486

Table 5.7: Overview of the top 20 edges and their respective probabilities in each dataset, for the Hill-Climbing algorithm.

5.3 Experiment 3: Sample Size

The results for Experiment 3 can be summarized in Figures 5.3 - 5.4. Averages are shown as dots in the line plots, and bars on each dot are confidence intervals. Table 5.8 shows these results in tabular format. Comparing the 2 lines at the top (Experiment 1 data) to the 3 at the bottom (Experiment 2 data) should be done carefully. These different variable sets and different degrees of freedom. But we can still draw some general conclusions. First, regarding RMSEA, while there is some fluctuation in the score, it flatlines overall and neither significantly improves nor worsens over increase in data size. This suggests that even if we increase the amount of data, the extent to which the original covariance of the data can be captured does not change. Regarding ECVI, however, the story is different. Recall that ECVI shows how likely a model will recreate similar covariance matrices between the training data and some externally held data set. As data size increases, ECVI improves across the board. Smaller is better in this case, and the ECVI decreases significantly as the data size increases. This means that as sample size increases, the strength of relationships captured are truly genuine. The strength of the relationships between variables is externally validated on held-out data.

Why do we have this seemingly paradoxical result, and what do these mean, taken together? First, an increase in data suggests a more confident fit; more data can indeed be useful in confidently estimating the strength of relationships between variables. Yet this more confident fit does not always entail a better fit, as we see that the RMSEA score does not significantly change over increasing data size. How can we see improvements? While the answer does not seem to be in data size, perhaps it lies in the choice of variables or the relationships to estimate. Some insights can perhaps be drawn from the prior experiments. In Experiment 2, we saw that features excluded from the graph in Figures 5.1-5.2 contained some very strong relationships. Including these variables and their respective strong relationships can perhaps help improve goodness of fit. But as we saw with Experiment 1, haphazardly adding all features and additional relationships does not necessarily entail the best model. As seen in that case, a task-behavior model excluding intentions provided the best RMSEA. Yet the RMSEA score even using this graph (PA-be-int) does not improve RMSEA to the point where it matches the lower 3 lines. Adding the right variables would hopefully bring the RMSEA score of the top 2 lines closer to the bottom 3, yet this gap is due to adding either too many variables estimating extra relationships unnecessarily (hence skewing the number of degrees of freedom and number of parameters to estimate unnecessarily). In summary, variables must be chosen wisely in experimentation.

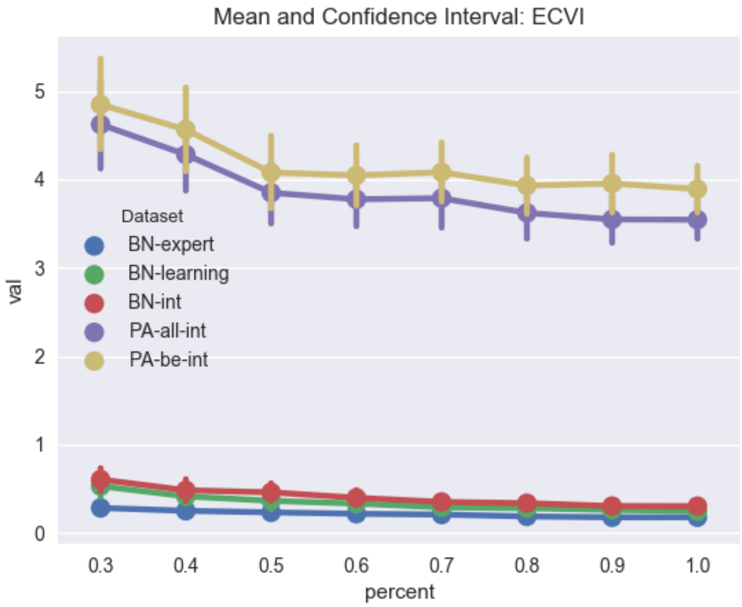

Figure 5.3: Experiment 3 - ECVI results. Means are shown in the line plot, and confidence intervals are bars. Results are shown for the full graph on Experiment 1 with intentions data (PA-all-int) and the graph omitting intentions (PA-be-int). Results are also shown for the Experiment 2 graph on the 3 datasets separately (BN).

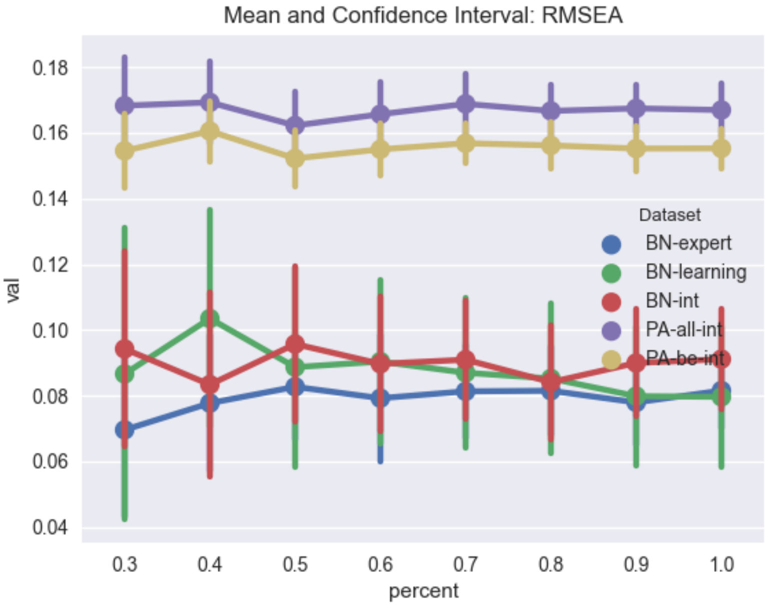

Figure 5.4: Experiment 3 - RMSEA results. Means are shown in the line plot, and confidence intervals are bars. See the previous figure for a key explanation.

	RMSEA				ECVI			
	PA - Intentions		BN - Learning		PA - Intentions		BN - Learning	
%	Mean	CI	Mean	CI	Mean	CI	Mean	CI
30	0.167	0.178	0.086	0.122	4.598	4.196	0.503	0.447
40	0.169	0.179	0.104	0.092	4.297	3.946	0.407	0.342
50	0.162	0.171	0.087	0.113	3.829	3.528	0.370	0.306
60	0.166	0.174	0.087	0.110	3.798	3.518	0.328	0.270
70	0.168	0.176	0.079	0.101	3.797	3.534	0.278	0.229
80	0.166	0.173	0.084	0.104	3.641	3.398	0.271	0.223
90	0.166	0.172	0.084	0.103	3.568	3.339	0.255	0.209
100	0.167	0.173	0.084	0.102	3.525	3.308	0.239	0.196

Table 5.8: Experiment 3 - Tabular form of some RMSEA and ECVI results, without loss of generality. Results are on the path analysis model constructed from Experiment 1 on the Intentions data, as well as on the Bayesian structure learned from Experiment 2 on the Searching as Learning data.

Chapter 6

Conclusion

The contribution of this dissertation is ultimately a statement: a statement that future interactive IR research should be conducted under a more general framework than presently done. This work makes a contribution towards bridging information seeking theory with interactive IR research practice. It explains how disparate statistical tests common to interactive IR can be combined into a single framework that more closely aligns with the vision of information seeking theory. It also introduces experimental methods for taking commonly collected laboratory study data and applying this new framework confirm old insights and extract new ones that could not previously be discovered. The final result of this work is not only a new conceptual framework but possible experimental methods to apply in the future.

6.1 Contributions

The contribution of this dissertation is in demonstrating that task characteristics, user characteristics, and behaviors should be empirically studied as a network of dependencies. It expands empirical work using graphical modeling, which can uniquely capture phenomena such as mediation and conditional independence. Research questions regarding mediation and conditional independence can hence now be answered with this different framework. This dissertation empirically shows when knowledge about behavior and certain task characteristics can be used to learn about other aspects of the task. It shows how task and user characteristics simultaneously affect behavior while potentially affecting each other. Specifically applying path analysis and Bayesian structure learning, results are shown to agree well with past literature and to also extend our understanding of the information seeking process.

The contribution of this dissertation is spread throughout the chapters. Even beginning with the literature review in Chapter 2, this dissertation helped set up the need for a common framework. Various literature has empirically shown that changes in tasks can affect user behavior. Other work has shown that changes in user characteristics such as time pressure or task difficulty can also be detected through changes in behavior. These insights are all pieces to the broader puzzle of how task characteristics affect users and how all other parts of the information seeking process affect each other while searching, and how these can be manifested through behavior. Some work attempted to address the limitations of only examining pairwise relationships in experiments, by examining interactions between behavior and two types of task traits and/or user characteristics. This was a necessary but incremental improvement.

Chapter 3 followed this by further explaining why we are not only in need of incremental improvements but in need of a paradigm shift. The chapter presented some theoretical considerations; information seeking researchers have always been conceptually interested in the whole picture of the information seeking process. Claims can be found in the literature regarding the directionality and type of influence each component has on the other in one grand picture. Some works even explicate which parts affect each other and which do not. But concepts like independence cannot even be tested in the current framework. Let alone testing the validity of the whole picture has only been pursued incrementally at best. Chapter 3 complemented this discussion by showing the practical and mathematical necessity of a shift. The current framework impedes itself in the types of questions it can ask. For instance, questions of conditional independence can be very interesting to researchers - such as whether query length can provide additional information once certain task or user characteristics become known. Yet it is shown that independence cannot even be addressed with the current framework, something that graphical models support. The chapter slowly builds to graphical modeling. It explains how this shift is not necessarily a clash of ideas against a former way of doing things. If anything, it is a generalization and of ideas that have been actively pursued (albeit in fractions) by the previous work.

Yet there does not exist an experimental framework to construct and test graphical models in interactive IR. Chapter 4 and 5 sought to explicate such a framework. It sought to address whether IIR experimental data could be used to make a graphical model mapping the relationships between variables. This included a combination of confirmatory analysis, goodness of fit tests, and matching findings with theory. These chapters applied several graphical techniques - including path analysis and Bayesian structural learning - to tease out insights that address the research questions originally posed. We may now come full circle and answer the research questions:

RQ1 - What is the nature of the influence between user background, user experience, task and search behavior? - Experiments 1 and 2 showed us that direct effects between task/user characteristics and behavior are not the only important relationships in an information seeking episode. They showed that things such as indirect effects, mediation, and conditional independence indeed occur.

RQ2 - To what extent do factors (including task) directly affect behavior versus indirectly through their effect on other session characteristics? - All have some degree of importance, inasmuch as they each affect behavior, as shown in Experiment 1. In cases where data gathering is a luxury, perhaps all of these should be collected together. But in cases where data gathering is at a premium, plenty of information can still be afforded by simple background questionnaires and pre- and post-task questionnaires. This result was validated by both Experiments 1 and 2. As shown in Experiment 1, there are some interesting indirect effects, such as topic solely affecting behaviors through topic familiarity in Experiment 1. But there are some direct effects as well, such as goal and product directly affecting behaviors. Experiment 2 shows several independence and conditional independence relationships. Many of these agree with past literature but also extend it.

RQ3 - To what extent does a more generalized modeling framework confirm or deny previous findings in task-based literature? We saw in Experiment 2 that a generalized model can confirm findings very well. An automatically learned model not only produced some consistent findings in itself but consistent findings that fit the purely proof-of-concept built from literature in Figure 3.5. In addition, the model could be used to answer some interesting questions about conditional independence that could not be explored previously, such as the possibility that only difficulty and product alone are sufficient to predict some behaviors such as query length.

RQ4 - What is the structure and size of data required for such an experimental framework? An increase in data size can significantly improve how confident we are in our estimation of parameters in a model. However, it cannot improve how well the model can recreate/fit the data. This can only be improved by tweaks to the structure of the model.

6.2 Important Open Questions

At least four open questions remain to be left to future work. While we have shown the usefulness and necessity of complex graphical modeling with respect to relating task type, user characteristics, and behaviors, the modeling of the information seeking process was not very complex, in a sequential or multidimensional sense. Recent work such as that by Maxwell and colleagues [103] created a conceptual and mathematical model of searchers issuing queries, scanning through pages, and saving results. This sequential, interactive, iterative process is absent from the work here - as each individual query segment is treated as an independent data point - due to practical constraints. Lab studies interested in designing and controlling replicable tasks tend to have small datasets. Incorporating interactive, sequential behavior would require too much data. Yet the work discussed [103] creates simulations rather than using large scale data on controlled task types - a current issue in IIR research today. Nevertheless, modeling of the sequential activities of users should should be pursued in the future. Perhaps this can lead directly to a model like Figure 3.6 that completely separates from the task and behavior and moves entirely into a cognitive domain.

It is also worth noting the obvious omission of latent variable modeling in this work. Latent variables were used by work such as Zhang et al [144]. In practice, latent variables are used to combine the signals from multiple measures into a common construct. For instance, multiple survey questions about education can be combined into a latent score regarding one's education level, or multiple survey questions about document usefulness can be combined into a latent usefulness score. As our study designs do not provide the data for latent constructs, they are not used here. Latent variable modeling is perhaps of interest here - as it is implicit in modern techniques like deep learning - but theory needs to be developed to justify the inclusion of latent variables in modeling the information seeking process.

While this dissertation has shown that graphical modeling is a worthy pursuit and that a graphical model can be estimated on laboratory study data using query segments as data points, there may still be value in using different features that require larger amounts of data. For instance, there may be value in using whole-session features instead of query segment features. Yet not enough data was used for this, and such a scale of data is difficult (and extremely rare) in laboratory settings where rich data is often collected. Future work would need to balance large scale data collection with the rich survey data collection of laboratory studies. Inspiration can be found in work such as He and Yilmaz [50], which additionally collects such data about naturalistic search tasks rather than imposed search tasks.

Lastly, on a related note, the claims here are only as good as the data collection, the experiments, and the theory. In particular, a direct relationship may have been observed between task and behaviors, which is consistent with the literature. But consider for a moment an experiment that gathers quality metrics on a child's home environment, the amount of time a child spends studying, and the child's scores on a standardized test. With a path model, the study finds significant direct effects between the quality of life metrics and the score, but it moreover finds a significant indirect effect through the child's study duration. If information about the child's study duration were never measured, a significant effect between quality of life metrics and standardized scores could still be measured. That said, direct effects can only be taken on assumption that the model is correct. It is our burden to not only derive accurate models but ensure that we've considered all possible variables before drawing conclusions.

6.3 Concluding Remarks

As a closing to this work, it is the hope of the author that this line of work is more actively pursued, both theoretically and in practice. It is hoped either that this framework will be used to frame standard experiments or that it will lead into novel data collection methods to get us closer to the vision of prior conceptual work in interactive IR. This work has aimed to provide the justification and building blocks of a graphical framework in interactive IR research.

Chapter 7

Appendix

7.1 Overview of Empirical Analyses

The following tables provide a more organized format of literature that has drawn significant relationships between behaviors and task or user characteristics. The tables contain the independent variables, the changes in behavior, and the relevant literature. These also show the statistical techniques applied in the cited work.

IV	Method	Results	Citation
Task Type (Product + Level + Complexity + Goal)	One-way ANOVA, Kruskal-Wallis	Task completion time, # pages, # sources, # queries, # search sources, decision time, read/scan ratio, saccade distance	[91]
Task Type (Product + Complexity)	One-factor repeated measures ANOVA	# IR systems, # result pages # items viewed, # items selected, # search engines consulted, # web results, # search portals visited, # web items viewed, # web items selected, # library resources consulted, # library result pages viewed, # library items viewed, # queries, # unique queries, mean query length, unique non-stop query terms.	[81]
Task Type (Complexity + Goal + Level)	Kruskal-Wallis	total SERP text acquired (pixels) over session, text acquired (pixels) on content pages per query, read-scan transitions	[29]
Task Type (Goal + Product)	(Not specified.)	total task time, # queries # unique clicks, % time viewing SERP, # SERP views per query, # unique clicks per query. # unique fixations, probability of moving up for SERP scanning, probability of sequential scanning, # unique fixations on a SERP, SERP view interval, average examined rank, max examined rank, # unique clicks per SERP view, % SERP views without clicks, probability of a click given that the result was examined, % clicks relevant, % clicks visited, average click rank, deepest click rank, first query vs. whole session: # SERP views for query, # unique fixations per query, # unique clicks per query, probability of click given examination, average examined rank, deepest examined rank, time of SERP viewing	[60]
Task Type (Fact-finding, Information Gathering, Browsing, Transactions	one-way ANOVA, Kruskal-Wallis	# windows opened during task, # pages loaded, proportions in navigation tools used (auto-complete, back button, bookmarks, google toolbar, hyperlinks, select URL, typed-in URL), time of day of the task, use of browser functions	[64]
Task Type (Conditioned on User)	one-way ANOVA	mean display time	[66]
Task Type (Fact-Finding/Information Gathering/Decision Making)	uni/multi-variate ANOVA	query length, amount overlap between query terms and terms in assigned task statements, # pages, # unique pages, # bookbag items, # tools used per query	[126]

Table 7.1: Review of empirical analyses. Provided are independent variables (IV), method used for analysis, significant results on dependent variables (-/+ for increase/decrease where applicable), and relevant literature citations.

IV	Method	Results	Citation
Task Type (Lo-Mid-Hi Complexity + Intellectual/Decision)	one-way ANOVA, one-factor repeated measures ANOVA, two-factor repeated measures ANOVA	time spent on task, time spent per item selected, number of items selected	[80]
Task Goal (Specific / Amorphous / Mixed)	Kruskal-Wallis	# sources, decision time, read/scan ratio, saccade distance	[91]
Task Level (Document / Segment), Objective Complexity (Low/High)	Mann-Whitney	Task completion time, # pages, # sources, # queries, # search sources, decision time (Level), read/scan ratio (Level), saccade distance (Complexity)	[91]
Task Product (Factual / Mixed)	Mann-Whitney	Task completion time, # pages, # sources	[91]
Task Product (Intellectual / Decision)	One-factor repeated measures ANOVA	# IR systems consulted, # result pages viewed, # search engines consulted, # web result pages viewed, # library resources consutled, # library result pages viewed, mean query length	[81]
Task Complexity (Low / Med / Hi)	one-way ANOVA, one-factor repeated measures ANOVA, two-factor repeated measures ANOVA	#IR systems consulted, # result pages viewed, # items viewed, # search engines consulted, # portals visited, # web result pages viewed, # web items viewed, # library resources consulted, # library result pages viewed, # library items viewed, # library items selected, # unique queries issued, # unique query terms issued, # unique non-stop query terms issued, time spent, # items selected (docx, pdf, full-text papers, etc.)	[81, 80]
Task Complexity (Simple / Hierarchical / Parallel)	Mann-Whitney, Kruskal-Wallis	Specialization query reformulation, word substitution, mean completion time, and mean number of queries	[83, 119]
Task Complexity (Parallel / Hierarchical)	uni/multi-variate ANOVA	time spent on queries, query length, # bookbag items, # tools used per query	[126]

Table 7.2: Continuation of the overview of empirical analyses.

IV	Method	Results	Citation
Task Complexity (Remember / Analyze / Create)	mixed factor ANOVA	session length, # queries, query length, # SERP clicks	[15]
Task Complexity (P / H) x Task Type (FF / IG / DM) Interaction	uni/multi-variate ANOVA	time on queries	[126]
Task Determinability (unspecified / specified items, unspecified / specified dimensions)	one-way repeated measures ANOVA	# queries, query length, clicks per query, bookmarks per query, query log likelihood, # unique queries, # unique URLs, time to first click	[15]
Stage, Usefulness, Stage x Usefulness (Conditioned on Task Complexity)	General Linear Model	decision time (stage x usefulness, combined tasks and parallel)	[90]
Search Engine Expertise	Cohen's d	# queries (-), # queries per day (+), query length (+), click depth (+), click probability (-), # repeated queries (+), # page revisits (-), time on search trails (-), time on documents (-)	[136]
Domain Expertise	Cohen's d, t-test	query length (+), # domain specific query terms (+), session length (+), # page views per session, # queries per session (+), probability of visiting non-commercial domains (e.g., gov and edu, +)	[133]

Table 7.3: Continuation of the overview of empirical analyses.

IV	Method	Results	Citation
Topic Familiarity	χ^2, Wilcoxon signed-rank	#bookmarks per page view (+), # query reformulation (-)	[67, 52]
Topic Familiarity	Mann-Whitney	ratio of content to SERP dwell time (+), average dwell time on unique SERPs (-), lower first dwell time on SERPs	[94]
Topic Familiarity (Conditioned on Amorphous Goal)	Mann-Whitney	average dwell time unique SERPs, mean first dwell time on SERPs	[94]
Topic Familiarity (Conditioned on Level)	Mann-Whitney	ratio of document time to all, ratio of SERP time to all	[94]
Topic Familiarity (Conditioned on Factual Product)	Mann-Whitney	task completion time, total time spent on SERPs, number of unique SERPs, number of viewed documents per query, number of saved documents per query, average dwell time of unique SERPs, mean first dwell time on all SERPs	[94]
Topic Familiarity (Conditioned on Low Complexity)	Mann-Whitney	average dwell time of unique SERPs, mean first dwell time on all SERPs	[94]

Table 7.4: Continuation of the overview of empirical analyses.

IV	Method	Results	Citation
Task Difficulty (Conditioned on Domain Knowledge)	Mann-Whitney	query vocabulary richness	[89]
Task Difficulty (Easy - Difficult scale) - whole session behavior	Mann Whitney, Kruskal Wallis	(+): # unique web pages, total # web pages, time per click, search linearity, total task duration, # SERP views on the first result page, # views on subsequent SERP result pages, # queries, # unique SERPs, # queries not leading to bookmarks, the ratio of queries not leading to bookmarks, lexical fixation duration excess (fixation time beyond minimum for lexical acquisition), mean LFDE, longest fixation duration LFDE, Generalization query reformulations (-): search optimality, # bookmarks, the ratio of queries leading to bookmarks, the ratio of document reading time to all, average query interval time, average time spent on documents, average dwell time on unique documents, # viewed documents per query, # unique viewed documents per query, the precision recall and F, reading speed (pixels/ms), inverse correlation with reading length of longest reading sequence (pixels) Repeat reformulations more frequently in medium difficulty tasks	[43, 42, 89, 86, 96, 29, 83]

Table 7.5: Continuation of the overview of empirical analyses.

IV	Method	Results	Citation
Task Difficulty (Easy / Difficult binary categories) - whole session behavior	logistic regression , Mann-Whitney	(+): total # web pages, total task duration, # query terms, # unique query terms, the ratio of token words to type words in queries, # clicks, the average click rank, the average viewrank, # page click, # and % of searches without clicks, the average bookmark rank, # queries with bookmarks, # of queries without bookmarks, % of queries without bookmarks, # clicks without bookmarks, % clicks without bookmark, # of views without bookmarks, % views without bookmark, total # mouseovers in a session, the average # mouseovers per query, the max mouseover rank per query, the average max mouseover rank per query, the total scroll distance in a session, the average number of scroll distance per query, the max scroll distance rank per query, the average max scroll distance rank per query, and the depth in the controlled vocabulary, % of queries with bookmarks, # queries (-): average dwell time on a landing page, query vocabulary richness, # query terms from the task description, # query terms from a controlled vocabulary	[3, 89]
Task Difficulty (Successful / Unsuccessful binary categories) - whole session behavior	t-test, 2nd order polynomial regression	(+): time spent on SERPs, the frequency of advanced query operators, proportion of time spent on SERPs, # question queries	[5]

Table 7.6: Continuation of the overview of empirical analyses.

IV	Method	Results	Citation
Task Difficulty - first query segment	Mann-Whitney, logistic regression	(+): the dwell time on the first SERP, the average click rank, the average page view rank, # page clicks, the average bookmark rank, # clicks without bookmarks, the percent views without bookmarks, the total mouseovers, the max mouseovers, and the total scroll distance, the max scroll position (-): # viewed pages, the total dwell time, the average dwell time, the duration, # bookmarks, the average query interval time, the mean dwell time of all documents, the average total dwell time of unique pages, # viewed documents, # unique viewed documents	[96, 3, 85]
Task Difficulty - per query segment	Mann-Whitney,t-test, 2nd order polynomial regression	(+): number of content pages per query, first dwell time, average dwell time. Regarding mid-session features, an increase difficulty has been associated with an increase in: mean dwell time of all SERPs, increase in average ist dwell time on SERPs, average query interval time, and with the longest query more likely to be earlier in the session (-): average rank of saved documents, # documents per query, # unique documents per query, # saved documents per query, # SERPs per query, # unique SERPs per query, average query interval	[96, 85, 5]
Task Difficulty (relative to task type)	Mann-Whitney	total dwell time on unique content pages, # content pages, and # unique content pages, # content pages per query, # unique content pages per query, first dwell time on unique content pages, first dwell time on unique SERPs, and mean dwell time on all SERPs	[92]
Difficulty x domain knowledge interaction effect	ANOVA , MANOVA	mean dwell time on content, % time on content pages	[87]

Table 7.7: Continuation of the overview of empirical analyses.

IV	Method	Results	Citation
Time pressure (Controlled strict time limit)	mixed ANOVA, Mann-Whitney	(+): time spent on the first SERP (-): time on content pages, time on SERPs, # content pages per query,# unique content pages per query, # SERPs per query, # unique SERPs per query, total dwell time on content pages per query, total dwell time on SERPs per query, ratio of dwell time on content pages per query, average query interval time, average session time, and # queries per session	[31, 88]
User Engagement (User Engagement Scale [114])	two-way repeated measures ANOVA	(+): SERP scrolling, time spent in query intervals, task duration, prior knowledge (-): perception of difficulty	[34, 35]
Intentions	logistic regression, multilayer perceptron, SVM	differences in reformulation strategies. prediction: bookmark features, content page dwell time features, SERP dwell time features, query reformulation types, and query lengths (best approach used all at once)	[107, 120]

Table 7.8: Continuation of the overview of empirical analyses.

References

[1] The relationships between task types and situational differences on the social interactions, affective reactions, and group products of family, ongoing, and ad hoc groups. *Unpublished dissertation*, 1982.

[2] Bryce Allen. Topic knowledge and online catalog search formulation. *The Library Quarterly*, 61(2):188–213, 1991.

[3] Jaime Arguello. Predicting search task difficulty. In *Proceedings of the 36th European Conference on IR Research on Advances in Information Retrieval - Volume 8416*, ECIR 2014, pages 88–99, New York, NY, USA, 2014. Springer-Verlag New York, Inc.

[4] Kumaripaba Athukorala, Dorota Glowacka, Giulio Jacucci, Antti Oulasvirta, and Jilles Vreeken. Is exploratory search different? a comparison of information search behavior for exploratory and lookup tasks. *J. Assoc. Inf. Sci. Technol.*, 67(11):2635–2651, November 2016.

[5] Anne Aula, Rehan M. Khan, and Zhiwei Guan. How does search behavior change as search becomes more difficult? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 35–44, New York, NY, USA, 2010. ACM.

[6] Earl Bailey and Diane Kelly. Developing a measure of search expertise. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16, pages 237–240, New York, NY, USA, 2016. ACM.

[7] Saeid Balaneshin-kordan and Alexander Kotov. Deep neural architecture for multi-modal retrieval based on joint embedding space for text and images. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 28–36, New York, NY, USA, 2018. ACM.

[8] R.M Baron and D.A. Kenny. The moderator-mediator variable distinction in social psychology research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51:1173–1182, 1986.

[9] Feza Baskaya, Heikki Keskustalo, and Kalervo Järvelin. Modeling behavioral factors ininteractive information retrieval. In *Proceedings of the 22nd ACM international conference on Conference on information &#38; knowledge management*, CIKM '13, pages 2297–2302, New York, NY, USA, 2013. ACM.

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

[11] PM Bentler and CH Chou. *Practical issues in structural modeling. Sociological Methods & Research.* 1987.

[12] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[13] A. Boomsma. Robustness of lisrel against small sample sizes in factor analysis models. pages 149–173, 1982.

[14] Horatiu Bota, Ke Zhou, and Joemon M. Jose. Playing your cards right: The effect of entity cards on search behaviour and workload. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16, pages 131–140, New York, NY, USA, 2016. ACM.

[15] Kathy Brennan, Diane Kelly, and Jaime Arguello. The effect of cognitive abilities on information search for tasks of varying levels of complexity. In *Proceedings of the 5th Information Interaction in Context Symposium*, IIiX '14, pages 165–174, New York, NY, USA, 2014. ACM.

[16] Andrei Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, September 2002.

[17] Michael W. Browne and Robert Cudeck. Alternative ways of assessing model fit. *Sociological Methods & Research*, 21(2):230–258, 1992.

[18] Georg Buscher, Ryen W. White, Susan Dumais, and Jeff Huang. Large-scale analysis of individual and task differences in search result page examination strategies. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 373–382, New York, NY, USA, 2012. ACM.

[19] Katriina Bystrom. Information and information sources in tasks of varying complexity. *Journal of the American Society for Information Science and Technology*, 53(7):581–591.

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

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

[22] Robert C. MacCallum, Michael W. Browne, and Hazuki M. Sugawara. Power analysis and determination of sample size for covariance structure modeling. 1:130–149, 06 1996.

[23] D. J. Campbell. Task complexity: A review and analysis. *Academy of Management Review*, 13(1), 1988.

[24] Rob Capra, Jaime Arguello, and Yinglong Zhang. The effects of search task determinability on search behavior. In *Advances in Information Retrieval - 39th European Conference on IR Research, ECIR 2017, Proceedings*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 108–121, Germany, 1 2017. Springer Verlag.

[25] L. Carter, W. Haythorn, and M. Howell. A further investigation of the criteria of leadership. *The Journal of Abnormal and Social Psychology*, 45(2):350 – 358, 1950.

[26] Marc-Allen Cartright, Ryen W. White, and Eric Horvitz. Intentions and attention in exploratory health search. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 65–74, New York, NY, USA, 2011. ACM.

[27] Michael J. Cole, Jacek Gwizdka, Chang Liu, and Nicholas J. Belkin. Dynamic assessment of information acquisition effort during interactive search. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–10, 2011.

[28] Michael J. Cole, Jacek Gwizdka, Chang Liu, Nicholas J. Belkin, and Xiangmin Zhang. Inferring user knowledge level from eye movement patterns. *Information Processing & Management*, 49(5):1075 – 1091, 2013.

[29] Michael J. Cole, Jacek Gwizdka, Chang Liu, Ralf Bierig, Nicholas J. Belkin, and Xiangmin Zhang. Task and user effects on reading patterns in information search. *Interacting with Computers*, 23(4):346 – 362, 2011. Cognitive Ergonomics for Situated Human-Automation Collaboration.

[30] Michael J. Cole, Chathra Hendahewa, Nicholas J. Belkin, and Chirag Shah. User activity patterns during information search. *ACM Trans. Inf. Syst.*, 33(1):1:1–1:39, March 2015.

[31] Anita Crescenzi, Diane Kelly, and Leif Azzopardi. Time pressure and system delays in information search. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 767–770, New York, NY, USA, 2015. ACM.

[32] Anita Crescenzi, Diane Kelly, and Leif Azzopardi. Impacts of time constraints and system delays on user experience. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16, pages 141–150, New York, NY, USA, 2016. ACM.

[33] W. Bruce Croft. Incorporating different search models into one document retrieval system. In *Proceedings of the 4th Annual International ACM SIGIR Conference on Information Storage and Retrieval: Theoretical Issues in Information Retrieval*, SIGIR '81, pages 40–45, New York, NY, USA, 1981. ACM.

[34] Ashlee Edwards and Diane Kelly. How does interest in a work task impact search behavior and engagement? In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16, pages 249–252, New York, NY, USA, 2016. ACM.

[35] Ashlee Edwards and Diane Kelly. Engaged or frustrated?: Disambiguating emotional state in search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 125–134, New York, NY, USA, 2017. ACM.

[36] DAVID ELLIS. A behavioural approach to information retrieval system design. *Journal of Documentation*, 45(3):171–212, 1989.

[37] Douglas C. Engelbart. Augmenting human intellect: A conceptual framework. Stanford Research Institute, 10 1962.

[38] Anjie Fang, Craig Macdonald, Iadh Ounis, and Philip Habel. Examining the coherence of the top ranked tweet topics. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 825–828, New York, NY, USA, 2016. ACM.

[39] Debanjan Ghosh. Effects of topic familiarity on query reformulation strategies. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16, pages 261–264, New York, NY, USA, 2016. ACM.

[40] James M. Graham. The general linear model as structural equation modeling. *Journal of Educational and Behavioral Statistics*, 33(4):485–506, 2008.

[41] Manish Gupta and Michael Bendersky. Information retrieval with verbose queries. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 1121–1124, New York, NY, USA, 2015. ACM.

[42] Jacek Gwizdka. Revisiting search task difficulty: Behavioral and individual difference measures. *Proceedings of the American Society for Information Science and Technology*, 45(1):1–12.

[43] Jacek Gwizdka and Ian Spence. What can searching behavior tell us about the difficulty of information tasks? a study of web navigation. *Proceedings of the American Society for Information Science and Technology*, 43(1):1–22.

[44] J.Richard Hackman. Effects of task characteristics on group products. *Journal of Experimental Social Psychology*, 4(2):162 – 187, 1968.

[45] Matthias Hagen, Jakob Gomoll, Anna Beyer, and Benno Stein. From search session detection to search mission detection. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 85–92, Paris, France, France, 2013. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE.

[46] Ahmed Hassan, Yang Song, and Li-wei He. A task level metric for measuring web search satisfaction and its application on improving relevance estimation. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 125–134, New York, NY, USA, 2011. ACM.

[47] Ahmed Hassan and Ryen W. White. Task tours: Helping users tackle complex search tasks. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1885–1889, New York, NY, USA, 2012. ACM.

[48] Ahmed Hassan Awadallah, Ryen W. White, Patrick Pantel, Susan T. Dumais, and Yi-Min Wang. Supporting complex search tasks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 829–838, New York, NY, USA, 2014. ACM.

[49] Larry Hatcher. *Advanced statistics in research: Reading, understanding, and writing up data analysis results*. ShadowFinch Media, LLC, 2013.

[50] Jiyin He and Emine Yilmaz. User behaviour and task characteristics: A field study of daily information behaviour. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pages 67–76, New York, NY, USA, 2017. ACM.

[51] Daniel Hienert, Matthew Mitsui, Philipp Mayr, Chirag Shah, and Nicholas J. Belkin. The role of the task topic in web search of different task types. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, CHIIR '18, pages 72–81, New York, NY, USA, 2018. ACM.

[52] Rong Hu, Kun Lu, and Soohyung Joo. Effects of topic familiarity and search skills on query reformulation behavior. In *Proceedings of the 76th ASIS&T Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries*, ASIST '13, pages 66:1–66:9, Silver Springs, MD, USA, 2013. American Society for Information Science.

[53] Graeme Hutcherson. Statistics and the generalized linear model.

[54] PETER INGWERSEN. Cognitive perspectives of information retrieval interaction: Elements of a cognitive ir theory. *Journal of Documentation*, 52(1):3–50, 1996.

[55] Peter Ingwersen and Kalervo Järvelin. On the holistic cognitive theory for information retrieval: Drifting outside the border of the laboratory framework. *Studies in the Theory of Information Retrieval.*

[56] Peter Ingwersen and Kalervo Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Springer-Verlag New York, Inc., 2005.

[57] Emi Ishita, Yosuke Miyata, Shuichi Ueda, and Keiko Kurata. A structural equation model of information retrieval skills. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pages 317–320, New York, NY, USA, 2017. ACM.

[58] Kalervo J:arvelin and Peter Ingwersen. Information seeking research needs extension towards tasks and technology. 10(1), 2004.

[59] Ming Ji, Jun Yan, Siyu Gu, Jiawei Han, Xiaofei He, Wei Vivian Zhang, and Zheng Chen. Learning search tasks in queries and web pages via graph regularization. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 55–64, New York, NY, USA, 2011. ACM.

[60] Jiepu Jiang, Daqing He, and James Allan. Searching, browsing, and clicking in a search session: Changes in user behavior by task and over time. In *Proceedings of the 37th International ACM SIGIR Conference on Research &#38; Development in Information Retrieval*, SIGIR '14. ACM, 2014.

[61] Jiepu Jiang, Daqing He, Diane Kelly, and James Allan. Understanding ephemeral state of relevance. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pages 137–146, New York, NY, USA, 2017. ACM.

[62] Rosie Jones and Kristina Lisa Klinkner. 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*, CIKM '08, pages 699–708, New York, NY, USA, 2008. ACM.

[63] Markus Kattenbeck and David Elsweiler. Estimating models combining latent and measured variables: A tutorial on basics, applications and current developments in structural equation models and their estimation using pls path modeling. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, CHIIR '18, pages 375–377, New York, NY, USA, 2018. ACM.

[64] Melanie Kellar, Carolyn Watters, and Michael Shepherd. A field study characterizing web-based information-seeking tasks. *J. Am. Soc. Inf. Sci. Technol.*, 58(7):999–1018, May 2007.

[65] Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wan-ching Wu. Development and evaluation of search tasks for iir experiments using a cognitive complexity framework. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, ICTIR '15. ACM, 2015.

[66] Diane Kelly and Nicholas J. Belkin. Display time as implicit feedback: Understanding task effects. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 377–384, New York, NY, USA, 2004. ACM.

[67] Diane Kelly and Colleen Cool. The effects of topic familiarity on information search behavior. In *Proceedings of the 2Nd ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '02, pages 74–75, New York, NY, USA, 2002. ACM.

[68] Kristof Kessler, Luanne Freund, and Richard Kopak. Does the perceived usefulness of search facets vary by task type? In *Proceedings of the 5th Information Interaction in Context Symposium*, IIiX '14, pages 267–270, New York, NY, USA, 2014. ACM.

[69] Jayden Khakurel, Antti Knutas, Mika Immonen, and Jari Porras. Intended use of smartwatches and pedometers in the university environment: An empirical analysis. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, UbiComp '17, pages 97–100, New York, NY, USA, 2017. ACM.

[70] Jeonghyun Kim. Task as a predictable indicator for information seeking behavior on the web. *unpublished dissertation*, 01 2006.

[71] Kyung-Sun Kim. Information-seeking on the web: Effects of user and task variables. *Library & Information Science Research*, 23(3):233 – 255, 2001.

[72] Kyung-Sun Kim and Bryce Allen. Cognitive and task influences on web searching behavior. *Journal of the American Society for Information Science and Technology*, 53(2):109–119, 2002.

[73] R.B. Kline. *Principles and Practice of Structural Equation Modeling*. Methodology in the social sciences. Guilford Publications, 2011.

[74] Alexander Kotov, Paul N. Bennett, Ryen W. White, Susan T. Dumais, and Jaime Teevan. Modeling and analysis of cross-session search tasks. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 5–14, New York, NY, USA, 2011. ACM.

[75] Michael Kotzyba, Tatiana Gossen, Johannes Schwerdt, and Andreas Nürnberger. Exploration or fact-finding: Inferring user's search activity just in time. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pages 87–96, New York, NY, USA, 2017. ACM.

[76] Sang-Zoo Lee and Scott Hershberger. A simple rule for generating equivalent models in covariance structure modeling. *Multivariate behavioral research*, 25 3:313–34, 1990.

[77] Liangda Li, Hongbo Deng, Anlei Dong, Yi Chang, and Hongyuan Zha. Identifying and labeling search tasks via query-based hawkes processes. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 731–740, New York, NY, USA, 2014. ACM.

[78] Ruirui Li, Ben Kao, Bin Bi, Reynold Cheng, and Eric Lo. Dqr: A probabilistic approach to diversified query recommendation. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 16–25, New York, NY, USA, 2012. ACM.

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

[80] Yuelin Li and Nicholas J. Belkin. An exploration of the relationships between work task and interactive information search behavior. *Journal of the American Society for Information Science and Technology*, 61(9):1771–1789, 2010.

[81] Yuelin Li and Nicholas J. Belkin. An exploration of the relationships between work task and interactive information search behavior. *Journal of the American Society for Information Science and Technology*, 61(9):1771–1789, 2010.

[82] Chang Liu, Nicholas J. Belkin, and Michael J. Cole. Personalization of search results using interaction behaviors in search sessions. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 205–214, New York, NY, USA, 2012. ACM.

[83] Chang Liu, Jacek Gwizdka, and Nicholas J. Belkin. Analysis of query reformulation types on different search tasks. iConference '10, 2010.

[84] Chang Liu, Jingjing Liu, Nicholas Belkin, Michael Cole, and Jacek Gwizdka. Using dwell time as an implicit measure of usefulness in different task types. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–4.

[85] Chang Liu, Jingjing Liu, and Nicholas J. Belkin. Predicting search task difficulty at different search stages. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 569–578, New York, NY, USA, 2014. ACM.

[86] Chang Liu, Jingjing Liu, and Nicholas J. Belkin. Predicting search task difficulty at different search stages. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 569–578, New York, NY, USA, 2014. ACM.

[87] Chang Liu, Jingjing Liu, Michael Cole, Nicholas J. Belkin, and Xiangmin Zhang. Task difficulty and domain knowledge effects on information search behaviors. *Proceedings of the American Society for Information Science and Technology*, 49(1):1–10.

[88] Chang Liu and Yiming Wei. The impacts of time constraint on users' search strategy during search process. In *Proceedings of the 79th ASIS&T Annual Meeting: Creating Knowledge, Enhancing Lives Through Information & Technology*, ASIST '16, pages 51:1–51:9, Silver Springs, MD, USA, 2016. American Society for Information Science.

[89] Chang Liu, Xiangmin Zhang, and Wei Huang. The exploration of objective task difficulty and domain knowledge effects on users' query formulation. *Proceedings of the Association for Information Science and Technology*, 53(1):1–9.

[90] Jingjing Liu and Nicholas J. Belkin. Personalizing information retrieval for multi-session tasks: The roles of task stage and task type. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 26–33, New York, NY, USA, 2010. ACM.

[91] Jingjing Liu, Michael J. Cole, Chang Liu, Ralf Bierig, Jacek Gwizdka, Nicholas J. Belkin, Jun Zhang, and Xiangmin Zhang. Search behaviors in different task types. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, pages 69–78, New York, NY, USA, 2010. ACM.

[92] Jingjing Liu, Jacek Gwizdka, Chang Liu, and Nicholas J. Belkin. Predicting task difficulty for different task types. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47*, ASIS&T '10, pages 16:1–16:10, Silver Springs, MD, USA, 2010. American Society for Information Science.

[93] Jingjing Liu and Chang Suk Kim. Why do users perceive search tasks as difficult? exploring difficulty in different task types. In *Proceedings of the Symposium on*

Human-Computer Interaction and Information Retrieval, HCIR '13, pages 5:1–5:10, New York, NY, USA, 2013. ACM.

[94] Jingjing Liu, Chang Liu, and Nicholas Belkin. Examining the effects of task topic familiarity on searchers' behaviors in different task types. *Proceedings of the American Society for Information Science and Technology*, 50(1):1–10.

[95] Jingjing Liu, Chang Liu, and Nicholas J. Belkin. Predicting information searchers' topic knowledge at different search stages. *Journal of the Association for Information Science and Technology*, 67(11):2652–2666, 2016.

[96] Jingjing Liu, Chang Liu, Michael Cole, Nicholas J. Belkin, and Xiangmin Zhang. Exploring and predicting search task difficulty. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1313–1322, New York, NY, USA, 2012. ACM.

[97] Jingjing Liu, Chang Liu, Jacek Gwizdka, and Nicholas J. Belkin. Can search systems detect users' task difficulty?: Some behavioral signals. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 845–846, New York, NY, USA, 2010. ACM.

[98] Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri, and Gabriele Tolomei. Identifying task-based sessions in search engine query logs. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 277–286, New York, NY, USA, 2011. ACM.

[99] Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri, and Gabriele Tolomei. Discovering tasks from search engine query logs. *ACM Trans. Inf. Syst.*, 31(3):14:1–14:43, August 2013.

[100] Gary Marchionini. Information-seeking strategies of novices using a fulltext electronic encyclopedia. *Journal of the American Society for Information Science*, 40(1):54–66, 1989.

[101] Gary Marchionini. *Information seeking in electronic environments*. Number 9. Cambridge university press, 1997.

[102] Gary Marchionini. Exploratory search: From finding to understanding. *Commun. ACM*, 49(4):41–46, April 2006.

[103] David Maxwell and Leif Azzopardi. Agents, simulated users and humans: An analysis of performance and behaviour. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 731–740, New York, NY, USA, 2016. ACM.

[104] Rishabh Mehrotra, Prasanta Bhattacharya, and Emine Yilmaz. Deconstructing complex search tasks: a bayesian nonparametric approach for extracting subtasks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 599–605. Association for Computational Linguistics, 2016.

[105] Rishabh Mehrotra and Emine Yilmaz. Terms, topics &#38; tasks: Enhanced user modelling for better personalization. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, ICTIR '15, pages 131–140, New York, NY, USA, 2015. ACM.

[106] Rishabh Mehrotra and Emine Yilmaz. Extracting hierarchies of search tasks & subtasks via a bayesian nonparametric approach. *CoRR*, abs/1706.01574, 2017.

[107] Matthew Mitsui, Jiqun Liu, Nicholas J. Belkin, and Chirag Shah. Predicting information seeking intentions from search behaviors. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 1121–1124, New York, NY, USA, 2017. ACM.

[108] Matthew Mitsui, Jiqun Liu, and Chirag Shah. Coagmento: Past, present, and future of an individual and collaborative information seeking platform. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, CHIIR '18, pages 325–328, New York, NY, USA, 2018. ACM.

[109] Matthew Mitsui, Jiqun Liu, and Chirag Shah. The paradox of personalization: Does task prediction require individualized models? In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, CHIIR '18, pages 277–280, New York, NY, USA, 2018. ACM.

[110] Matthew Mitsui and Chirag Shah. The broad view of task type using path analysis. In *Proceedings of The 8th International Conference on the Theory of Information Retrieval*, ICTIR '18, New York, NY, USA, 2018. ACM.

[111] Matthew Mitsui, Chirag Shah, and Nicholas J. Belkin. Extracting information seeking intentions for web search sessions. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 841–844, New York, NY, USA, 2016. ACM.

[112] Sophie Monchaux, Franck Amadieu, Aline Chevalier, and Claudette Marin. Query strategies during information searching: Effects of prior domain knowledge and complexity of the information problems to be solved. *Information Processing & Management*, 51(5):557 – 569, 2015.

[113] JC Nunnally. *Psychometric theory.* McGraw-Hill, New York, NY, 1967.

[114] Heather L. O'brien and Elaine G. Toms. Examining the generalizability of the user engagement scale (ues) in exploratory search. *Inf. Process. Manage.*, 49(5):1092–1107, September 2013.

[115] Emilie Palagi, Fabien Gandon, Alain Giboin, and Raphaël Troncy. A survey of definitions and models of exploratory search. In *Proceedings of the 2017 ACM Workshop on Exploratory Search and Interactive Data Analytics*, ESIDA '17, pages 3–8, New York, NY, USA, 2017. ACM.

[116] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[117] Judea Pearl. The causal foundations of structural equation modeling. pages 68–92, 2012.

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

[119] Peng Qu, Chang Liu, and Maosheng Lai. The effect of task type and topic familiarity on information search behaviors. In *Proceedings of the Third Symposium on Information Interaction in Context*, IIiX '10, pages 371–376, New York, NY, USA, 2010. ACM.

[120] Eun Youp Rha, Matthew Mitsui, Nicholas J. Belkin, and Chirag Shah. Exploring the relationships between search intentions and query reformulations. *Proceedings of the Association for Information Science and Technology*, 53(1):1–9, 2016.

[121] Albert Satorra and Willem E. Saris. Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, 50(1):83–90, Mar 1985.

[122] S. Shute and P. Smith. Knowledge-based search italics. *Information Processing & Management*, 29:29 – 45, 1993.

[123] Georg Singer, Ulrich Norbisrath, and Dirk Lewandowski. Ordinary search engine users carrying out complex search tasks. *Journal of Information Science*, 39(3):346–358, 2013.

[124] Peter Sprites, Thomas Richardson, Chris Meek, Richard Scheines, and Clark Glymour. Using path diagrams as a structural equation modelling tool. *Technical Report*, 1997.

[125] Joseph Stevens. Structural equation modeling (sem) - workshop presentation.

[126] Elaine G. Toms, Heather O'Brien, Tayze Mackenzie, Chris Jordan, Luanne Freund, Sandra Toze, Emilie Dawe, and Alexandra MacNutt. Task effects on interactive search: The query factor. In Norbert Fuhr, Jaap Kamps, Mounia Lalmas, and Andrew Trotman, editors, *Focused Access to XML Documents*, pages 359–372, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[127] Ioannis Tsamardinos, Constantin F. Aliferis, and Alexander R. Statnikov. Algorithms for large scale markov blanket discovery. In *FLAIRS Conference*, 2003.

[128] Michael Tushman. Technical communication in r&d laboratories: Impacts of project work characteristics. *Academy of Management Journal*, 21(4):624 – 645, 1978.

[129] Hossein Vahabi, Margareta Ackerman, David Loker, Ricardo Baeza-Yates, and Alejandro Lopez-Ortiz. Orthogonal query recommendation. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 33–40, New York, NY, USA, 2013. ACM.

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

[131] Manisha Verma and Emine Yilmaz. Category oriented task extraction. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16, pages 333–336, New York, NY, USA, 2016. ACM.

[132] Ryen White and Resa Roth. *Exploratory Search.* Morgan & Claypool Publishers, 2008.

[133] Ryen W. White, Susan T. Dumais, and Jaime Teevan. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 132–141, New York, NY, USA, 2009. ACM.

[134] Ryen W. White and Diane Kelly. 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*, CIKM '06, pages 297–306, New York, NY, USA, 2006. ACM.

[135] Ryen W. White, Bill Kules, and Ben Bederson. Exploratory search interfaces: Categorization, clustering and beyond: Report on the xsi 2005 workshop at the human-computer interaction laboratory, university of maryland. *SIGIR Forum*, 39(2):52–56, December 2005.

[136] Ryen W. White and Dan Morris. Investigating the querying and browsing behavior of advanced search engine users. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 255–262, New York, NY, USA, 2007. ACM.

[137] Ryen W. White, Gheorghe Muresan, and Gary Marchionini. Report on acm sigir 2006 workshop on evaluating exploratory search systems. *SIGIR Forum*, 40(2):52–60, December 2006.

[138] Richard Whitley and Penelope Frost. Task type and information transfer in a government research laboratory. *Human Relations*, 26(4):537–550, 1973.

[139] Barbara Wildemuth, Luanne Freund, and Elaine G. Toms. Untangling search task complexity and difficulty in the context of interactive information retrieval studies. *Journal of Documentation*, 70(6):1118–1140, 2014.

[140] Barbara M. Wildemuth. The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology*, 55(3):246–258, 2003.

[141] Erika J Wolf, Shaunna L Clark, and Mark W Miller. Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. pages 913—934, 1982.

[142] Hong (Iris) Xie. Patterns between interactive intentions and information-seeking strategies. *Information Processing & Management*, 38(1):55 – 77, 2002.

[143] Xiangmin Zhang, Jingjing Liu, and Michael Cole. Task topic knowledge vs. background domain knowledge: Impact of two types of knowledge on user search performance. In Álvaro Rocha, Ana Maria Correia, Tom Wilson, and Karl A. Stroetmann, editors, *Advances in Information Systems and Technologies*, pages 179–191, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[144] Yinglong Zhang, Jin Zhang, Matthew Lease, and Jacek Gwizdka. Multidimensional relevance modeling via psychometrics and crowdsourcing. In *Proceedings of the 37th International ACM SIGIR Conference on Research &#38; Development in Information Retrieval*, SIGIR '14, New York, NY, USA, 2014. ACM.

[145] Wei Zheng, Xuanhui Wang, Hui Fang, and Hong Cheng. An exploration of pattern-based subtopic modeling for search result diversification. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, JCDL '11, pages 387–388, New York, NY, USA, 2011. ACM.