Bridging Gaps: Predicting User and Task Characteristics from Partial User Information

Matthew Mitsui, Chirag Shah
Rutgers University
New Brunswick, NJ 08901, USA
{mmitsui,chirags}@rutgers.edu

ABSTRACT

Interactive information retrieval (IIR) researchers often conduct laboratory studies to understand the relationship between people seeking information and information retrieval systems. They develop extensive data collection methods and tools create new understanding about the relationship between observable behaviors, searcher context, and underlying cognition, to better support people's information seeking. Yet aside from the problems of data size, realism, and demographics, laboratory studies are limited in the number and nature of phenomena they can study. Hence, data collected in laboratories contains different searcher populations and collects non-overlapping user and task characteristics. While research analyses and collection methods are isolated, how can we further IIR's mission of broad understanding? We approach this as a structure learning problem on incomplete data, determining the extent to which incomplete data can be used to predict user and task characteristics from interactions. In particular, we examine whether combining heterogeneous data sets is more effective than using a single data set alone in prediction. Our results indicate that adding external data significantly improves predictions of searcher characteristics, task characteristics, and behaviors, even when the data does not contain identical information about searchers.

CCS CONCEPTS

• **Information systems** → **Task models**; **Retrieval tasks and goals**; *Personalization*; • **Mathematics of computing** → *Multivariate statistics*; **Bayesian networks**; • **Theory of computation** → **Bayesian analysis**; *Structured prediction.*

KEYWORDS

Task prediction; Task classification; Interactive information retrieval; Task type; Searcher behavior; Structure learning; Bayesian networks

ACM Reference Format:
Matthew Mitsui, Chirag Shah. 2019. Bridging Gaps: Predicting User and Task Characteristics from Partial User Information. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3331184.3331221

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331221

1 INTRODUCTION

One of the goals of interactive information retrieval (IIR) research is to understand the relationship between searchers, their search goals, and their observable activities and to better support the accomplishment of the goals of their search. There has not only been a desire to determine which environmental or personal characteristics affect search behavior, but there has also been a desire to determine which characteristics are worth factoring into a personalized or contextualized search experience. Several such important characteristics include a searcher's task [20, 24], her general knowledge about the search topic [12], and time pressure experienced at a moment [6], among others. Determining a searcher's task has been shown to be useful in improving query recommendation [24] and ranking [16, 36]. Further, a suite of studies have shown how differing topic familiarity can affect querying strategies, which in turn suggests different conceptual models of searchers and perhaps in turn different methods for recommendation [12]. Moreover, with dynamic search tools on the rise, such as conversational search assistants [29], it will not only be important for an algorithm to infer such characteristics for a personal search experience but to do so quickly.

Researchers typically attempt to build understanding of searcher, task, and behaviors piecemeal, by collecting the search activity of a few dozen participants in a controlled environment like a laboratory. Researchers control and manipulate the tasks assigned to searchers or even the characteristics of the searchers (e.g., the time pressure experienced). In turn, researchers examine whether behavior is affected by these changes or inversely whether the characteristics can be predicted from behaviors. Since assigned tasks and a laboratory setting are occasionally criticized for a lack of realism, recent work expanded this paradigm to a more naturalistic setting, with a small set of users self-reporting the nature of their real tasks [10]. But both data collection paradigms still pervade current research practice. In either setup, the value in this data is in the ability to derive searcher and task characteristics from human annotations, which currently cannot be derived from large scale Web logs.

To create a completely personal picture of one particular searcher, several aspects about her may need to be extracted, including time pressure, the difficulty experienced with the task, the intentions for issuing particular queries, and so on. Additionally, she may need to be observed in the context of each possible type of task. First, it is unfeasible to ask for this quantity of annotation from a single searcher, let alone for even one of her search sessions. Secondly,

an exhaustion of laboratory resources is likely, as the number of searchers required for a study would increase linearly at best as more searcher and task variables are added as desired predictions. Lastly, more task characteristics are discovered and studied even today, as shown in recent work [4].

Hence, in practice, studies are limited to analyzing a specific isolated phenomenon, and it is not clear how to practically combine and validate these insights to further IIR's broader mission. How should the insights from any individual study be reconciled with the user and task factors they do not consider? While different laboratory studies may address different research questions, they often share several characteristics. Common sets of behavioral features may be used to make inferences about searcher and task. One study may find a strong relationship between task type, topic familiarity, and query strategies. It may not examine the relationship between the same query strategies and time pressure, while another may discover such an important relationship. If insights from one study can be transferred to the other, this could perhaps alter the nature of the findings of the study or even improve predictions of task and searcher characteristics from behavior. Perhaps the partial and sometimes heterogeneous information of multiple studies can be used in conjunction with each other to practically further the larger goals of IIR, all while maintaining today's data collection paradigms. We hence ask the following research question: Can predictions of the relationships between task characteristics, searcher characteristics, and behavior on a data set be improved when adding data from additional studies, rather than the single data set alone? In particular, we frame this as a structured prediction problem, predicting the user and task characteristics of query segment-based interactions and training on heterogeneous data (namely, data with non-overlapping missing values).

The rest of this paper is organized as follows. Section 2 discusses a more extensive treatment of IIR research practice and the aforementioned limitations. Section 3 discusses the structured prediction framework, as well as the novel application of a prior structured prediction framework in this IIR setting. Section 4 discusses 3 data sets from different laboratory studies combined in the experiments. Section 5 discusses the experimental framework combining multiple data sets and reports the results. Section 6 concludes by discussing implications for IIR as well as room for future work. In summary, we discovered that both on average and with respect to particular features, predictions can be improved by incorporating external data from other studies. Moreover, the studies do not necessarily need to include identical features or even the predictors to see such improvement.

2 BACKGROUND

Here, we discuss the necessary background: common and accepted data collection practices in IIR studies, characterizations of searcher and task, and the current practice in predicting relationships between searcher, task, and behavior.

2.1 Data collection in IIR

It should be acknowledged that some research pertaining to search tasks applies to large scale search logs. A large body of the work, for instance, has developed task extraction algorithms to transform otherwise unstructured logs into cohesive search tasks [14, 22, 34]. The goals for such large scale Web work typically involve improving upon standard recommendation problems, for instance using searchers' task preferences to create effective cohort-based query recommendation [24] or to recommend future useful queries specifically for complex search tasks [9]. In Web search logs, the research often creates latent structure from messy data to directly improve recommendations.

Laboratory studies address complementary research problems. Data often consists of a few dozen searchers with one or two search sessions instead of millions of users with several search sessions. In the Web, people may multitask, work on a task intermittently over several days, and otherwise have search performance influenced by other external factors. The laboratory reduces the effects of such confounding factors; searchers often focus on a single task curated by the researchers. The lab is additionally used as an opportunity to capture richer characteristics about a searcher, often through provided surveys. Deriving rich self-reports from a publicly available commercial search engine is unrealistic, impossible, or otherwise rarely done if ever.

A common criticism of labs is a potential lack of realism, but researchers have responded to this in several ways. Early work by Borlund and Ingwersen developed the notion of the *simulated work task*, in which a search task, while designed by a researcher, elicits a 'simulated information need' by describing to the participant "the source of the information need, the environment of the situation, the problem which has to be solved, and also serves to make the study participant understand the objective of the search" [1]. This type of work task is meant to instill a controlled information need in the searcher, balancing the desire for control yet also for realism, and this framework has been used extensively [2]. Furthermore, tasks from the general public and from work organizations have been studied in decades of research. A seemingly infinite number of search tasks have been condensed into simpler core attributes which all allegedly share. One such notable categorization includes that of Li and Belkin [15]; subsequent research has since designed search tasks according to this schema and studied their influence on searchers [11, 13, 20].

He and Yilmaz [10] addressed the lack of realism with a different approach. They had participants work on their own search tasks and hand-annotate them, although this nevertheless was a small research effort conducted by a small number of searchers. Such small settings – whether in a laboratory or naturalistic setting – have become a staple of IIR research, with such data collection spanning decades and even today [4]. A large body tasks assigned in laboratories – and their associated papers – can be found at the Repository of Assigned Search Tasks[1].

2.2 Task and Searcher Relationships

Classifications of search, including that of Li and Belkin, span dozens of user and task characteristics. Li and Belkin include the product of the task – whether the task is about finding facts, producing insights, making a decision, or creating a product – and the goal of the task – whether it is well-defined, ill-defined, or

[1] https://ils.unc.edu/searchtasks/search.html

somewhere in between [15]. Another heavily researched dimension is the task's objective complexity, proposed by Campbell and described as searcher-independent and comprised of the number of possible paths to the outcome, the number of outcomes, interdependence among paths, and uncertainty between paths and outcomes. A large number of these attributes are solely dependent on the task description and independent of the searcher [3]. Additionally, the number of task characteristics of interest continues to grow. Recent work by Capra et al. defined "determinability" as whether task aspects were explicitly defined, distinguishing differences in behavior among different levels of determinability [4].

Moreover, researchers are also interested in attributes of the searchers, some of which are dependent on their task and some of which are independent. One task-independent feature includes general search expertise; Marchionini shows that experts and non-experts have different mental models for searching, with experts being more efficient [23]. Another task-independent feature is time pressure, shown to similarly affect searching and browsing behaviors [6, 19]. In contrast, a popular task-dependent feature is topic familiarity, which has been associated with a suite of differences in behavior – not only search efficiency [12] but also eye movement patterns [5] and querying behavior [35]. These properties describe task or searcher properties of an entire search episode. Yet some properties of interest characterize search session dynamics and are only capable of being captured in the midst of a search session. One such example is "interactive search intentions", proposed by Xie and describing the activity on a single query and subsequent pages, including intentions to "locate a specific link" or to "learn domain knowledge" [37]. Hence, interests in characterizing the search session span multiple levels of granularity.

2.3 Characterization and Prediction in IIR

The purpose of such characterization is to ultimately provide tailored predictions of behavior and tailored recommendations. Traditionally, researchers examine how changes in searcher and task characteristics affect commonly observable behaviors – e.g., how querying patterns differ between fact-finding and exploratory tasks, between topic experts and non-experts, or each of these 4 combinations. Results can then be used in predictive models, using behaviors observed during a search session to predict the state of the searcher and the nature of the task. Prior work has examined such a predictive relationship directly between tasks and behaviors. For instance, Liu et al. and Jiang et al. showed differences in browsing behaviors and even eye tracking patterns between tasks with differing goals and products [13, 20]. Differing task complexity has similarly been shown to affect query reformulation patterns [17] and other observable behaviors. Prior work has also characterized direct relationships between searcher characteristics and behavior. Topic knowledge and domain knowledge affect query reformulation patterns, search efficiency [12], eye movement patterns [5], types of domains visited, and word usage in queries [35]. Increased time pressure specifically decreases task time and overall results in shallower page reading and shallower searching behavior [6, 19].

Intuitively, characteristics like familiarity with a task or topic should be a function of the task itself, and general search expertise may not affect behavior when a task is simple or fact-finding. Prior work has explored interactions between task and searcher characteristics on behaviors. Relationships have been discovered between task complexity, task type, and browsing behaviors [33]; task difficulty, domain knowledge, and time spent on content pages [18]; task type, task difficulty, and behaviors [21]; and task topic, behaviors, and perceived difficulty and success [11]. To another extreme, the same patterns used to identify tasks have been shown to more reliably predict who is searching rather than the task, suggesting a possible interaction between task and personal characteristics generally [25]. While each isolated finding regarding task, user, and behavior can be seen as a component to a larger network, few works have explored this possibility in practice. We previously applied structural equation modeling in such a fashion, using a combination of meta-analysis and parameter learning to estimate a network between task characteristics, searcher characteristics, and browsing behaviors. They showed that such network-based estimation is empirically necessary, as their findings complemented old results but discovered new patterns. Accounting for more task and user characteristics simultaneously could better help explain searcher behaviors [26].

Therefore, estimating the influence of several searcher and task contexts simultaneously on behavior is necessary. But as a practical limitation, no studies attempt to inventory a complete-as-possible list of searcher and task characteristics, except in meta-analyses. A laboratory study participant would need to complete extensive pre-task, post-task, and in-session surveys comprised of dozens of questions. Hence, several studies will contain non-overlapping information about participants. If one study inventories the effects of time pressure on behavior, how can this be reconciled with the findings of task versus behavior from another study that did not consider time pressure? Can the findings from one study be used to improve prediction of user and task characteristics from the findings in another? Therefore, we must learn the complex interaction between searcher, task, and behavior with incomplete information. This work therefore addresses a problem largely unexplored in IIR: 1) learning a structure on relationships between searcher, task, and behavior, and 2) learning such a structure from heterogeneous data – namely from multiple data sets recording non-overlapping searcher and task characteristics.

3 METHODOLOGY

Here, we discuss and motivate the necessary framework for understanding structure learning, the main approach used in prediction. We consider cases where data is both complete and incomplete in our experiments, with both cases discussed below.

3.1 Structure Learning

One formulation of a relationship between task, searcher, and behavior is of a graph $G = (V, E)$, where the characteristics are variables/vertices V and each edge represents the strength and direction of influence between variables, e.g., a unit increase in time pressure decreases result dwell time by 3 seconds. *Structure learning* is decomposed into the joint problems of learning 1) which edges to include and 2) the strength of these edges (*parameter learning*). Prior work investigated whether human cognitive modeling could

benefit from a structure learning approach [32]. Consider a Gaussian Bayesian Network as an example. Given n data points D each consisting of o observations (e.g., logged behaviors) and i inputs, our data is broken into $Y \in \mathbb{R}^{n\times o}$, $X \in \mathbb{R}^{n\times i}$, and triangular matrix $B \in \mathbb{R}^{i\times o}$, and error terms $\epsilon \in \mathbb{R}^{n\times o}$ that maximizes $P(B|G, D)$ such that:

$$y_o = \sum_i \beta_{i,o} x_i + \epsilon_o \tag{1}$$

$$P(Y_o|Parents(Y_o)) \sim N\left(\beta_{Y_o} + \sum_{X_i \in Parents(Y_o)} \beta_{i,o} X_i; \sigma_o^2\right) \tag{2}$$

One can engage in *parameter learning* without structure learning when one can justifiably skip the first step and manually specify a graph – e.g., when one has expert knowledge about which searcher traits and behaviors affect each other. Social applications managing the safety behavior of employees [27] and decision support for medical procedures [28] have benefited from parameter learning. The IIR work discussed in Section 2.3 can be seen as parameter learning, hand-selecting which interactions of variables to measure. We previously performed a meta-analysis to hand-construct structural equation models for analyses [26]. Yet for a more accurate determination of how variables affect each other, it can be beneficial to engage in both steps of *structure learning*. Structure learning has been shown to be more accurate than parameter learning (i.e., expert construction with parameter estimation) in applications such as lung cancer decision support [31].

Structure learning is NP hard so is generally approached with three types of approximation algorithms: constraint-based, score-based, and hybrid. Constraint-based algorithms determine the edges between nodes based on conditional independence relationships. Algorithms such as the PC algorithm first connect dependent nodes directly with an edge, then prune away edges and orient them. Score-based algorithms attempt to maximize a network score representing the goodness of fit of the graph to the data. For instance, hill climbing – a greedy approach – greedily adds edges that maximize the following score:

$$Score(G : D) = LL(G : D) - \phi(|D|)||G|| \tag{3}$$

$LL(G : D)$ is the log-likelihood of data D under G. $|D|$ is the size of the data $||G||$ is the number of parameters in the graph, and ϕ refers to a complexity measure function, essentially used for regularization. Hybrid algorithms borrow techniques from both constraint-based and score-based approaches, proceeding in two phases. The first phase applies a constraint-based technique to limit the number of graphs to explore. The second phase uses a score-based technique to find the optimal graph within this space [30].

If we consider the learned graph to be a directed acyclic graph (DAG), G is a graph consisting of the $i + o$ variables as vertices, and directed, weighted edges are represented by B. In our specific experiments, the i inputs are task and searcher characteristics, and the o outputs are searching behaviors. A DAG is suitable for our problem setting, which assumes the following hierarchical relationship:

- Tasks are pre-specified independently of the searcher
- Searchers bear some characteristics independently of the task (e.g., general search expertise)
- Searchers bear some characteristics relative to the task (e.g., topic familiarity)
- Searchers subsequently exhibit outward behaviors.

The relationship between searcher, task, and behavior is understandably more complex, which will be discussed in Section 6.

3.2 Structure Learning with Missing Data

Common structure learning methods work well for cases where data is complete. In our case, this means that every single behavior, user, and task characteristic is recorded. Suppose we are to examine data from multiple laboratory studies. If they each take place on a desktop computer while browsing the general Web (a likely scenario), identical browsing features can be derived from each. Yet as mentioned above, not every single aspect about a searcher's context will be reported in each study. One study may design tasks with differing goals and products and collect information about time pressure. Another may only vary task goals while asking about topic familiarity. That is to say that these combined data sets are heterogeneous, inasmuch as there are missing values.

A structure learning algorithm which considered missing values was derived by Friedman [7]. Specifically, if we consider D to be our data set, we can consider D^O to be the set of observed values and D^M to be the set of missing values. Similarly to the above approaches, this attempts to maximize the following score on graph (G, B), which includes terms for log-likelihood and regularization on the observations:

$$S_{D^O} = logP(D^O|G, B) - \phi(G, B, D^O) \tag{4}$$

When all data is provided, this score can be computed directly. However, in the case of partially missing data, we look at the expected values of the missing data D^M, maximizing an expectation score instead:

$$Q(G, B|G^*, B^*) = E[logP(D^O, D^M|G, B) - \phi(G, B, D^O)] \tag{5}$$

$$= \textstyle\sum_{d\in D^M} p_d [logP(D^O, d|G, B) - \phi(G, B, D^O)] \tag{6}$$

Where $p_d = P(d|D^O, G^*, B^*)$. Therefore, an expectation maximization algorithm can be derived as follows [7]:

(1) Choose G^0 and B^0 randomly
(2) Loop for $n = 0, 1, \ldots$ until convergence
  (a) Find a model G^{n+1} that maximizes $Q(., B^n|G^n, B^n)$
  (b) $B^{n+1} = argmax_\beta Q(G^{n+1}, \beta|G^n, B^n)$

3.3 Evaluation of Structure Learning

The primary method of evaluating a Bayesian network is with scores derived from the log-likelihood $LL(G : D)$. They include some penalty for model complexity, such as the Bayesian Information Criterion (BIC) defined as: $BIC(G : D) = LL(G : D) - \frac{log(|D|)}{2}||G||$. In all cases a higher score is preferred. These scores can be used to evaluate prediction performance on individual nodes or also on the entire graph (as the sum of the scores on the nodes). We adopt both approaches here, specifically applying the unregularized log-likelihood score, as the number of included and excluded variables varies greatly between experiments.

Table 1: Task type and session characteristics for intentions (INT), searching as learning (SAL), and expert opinion (EOP) data sets.

Data Set	Task #	Product	Goal	$\|T\|$	$\|Q\|$
INT	1	Factual	Specific	22	206
INT	2	Factual	Amorphous	18	108
INT	3	Intellectual	Amorphous	18	155
INT	4	Intellectual	Amorphous	22	224
SAL	5	Mixed	Specific	30	168
SAL	6	Intellectual	Specific	30	187
SAL	7	Intellectual	Specific	30	110
SAL	8	Intellectual	Specific	30	129
EOP	9	Factual	Specific	30	161
EOP	10	Factual	Amorphous	30	75
EOP	11	Factual	Specific	30	294
EOP	12	Factual	Amorphous	30	256
EOP	13	Factual	Specific	30	249
EOP	14	Factual	Amorphous	30	239

4 DATA SETS

This section details three independent laboratory studies conducted by the researchers, including the experimental setup and the user and task characteristics recorded in each. We show where the respective data sets overlap and where they differ.

4.1 Data: Laboratory and Naturalistic Logs

Our data consists of search logs collected from three independently conducted IIR studies, both in laboratory settings and naturalistic settings. Participants in the studies were undergraduates recruited from a university. In each study, participants conducted several search tasks designed by the researchers. While having different topics and task descriptions, the tasks were manipulated to have specific task goals and products according to the classification of Li and Belkin [15]. Each task required multiple queries to accomplish. A summary of each data set is provided in Table 1. In all studies, participants were incentivized to perform well with an additional bonus payment that would be given to "best participants".

The first data set (intentions – INT) centered around search behavior for journalism-type tasks. The purpose of this experiment was to measure the search intentions of participants. Participants (undergraduates) conducted 2 consecutive search sessions in a laboratory, and the tasks for each participant were rotated to reduce ordering effects. A participant's session began with a demographic questionnaire, and each search session was preceded by a pre-task interview and proceeded by a post-task interview. Participants had two 20 minute search tasks but could choose to finish each early. Search activity was recorded in a Firefox plugin. The search tasks were on the topics "coelacanths" (a type of fish) and "methane clathrates and global warming". The study consisted of 40 searchers who conducted 693 queries over 8 different tasks.

The second data set (searching as learning – SAL) centered around tasks that required learning. The purpose of this study was to observe learning effects of searchers conducting a task over the course of several days. This study was conducted in a naturalistic setting. Participants downloaded a Chrome extension that allowed them to interact with study materials (e.g., questionnaires), view the task descriptions, and record search activity. Participants were hence allowed to conduct the study activities in any arbitrary setting, not under the supervision of a research coordinator. Participants were general undergraduate students required to be at least in their second year of study. They were also required to use Google Chrome to complete the study. While participants were asked to complete 4 search tasks over three consecutive days of their choosing, participants otherwise had no imposed time limits. The topic of each task was cyber bullying. Participants were similarly asked demographic, pre-task, and post-task questionnaires. Multitasking for unrelated tasks was considered when cleaning data for analyses in this paper. In the INT data, this was of no concern, since all activities were conducted in a controlled laboratory setting. In total, the study consisted of 40 searchers who conducted 594 queries over 4 different tasks.

The final data set (expert opinion – EOP) was collected in a mixed environment – participants conducted search activity in both naturalistic and laboratory settings. The purpose of this study was to measure differences in behavior and search performance after participants were exposed to expert advice on search strategies, peer advice, or nothing at all. General undergraduates were also recruited for this population, with the only requirement being the use of Chrome for the study. For the naturalistic setting, participants similarly downloaded a Chrome extension, which was also used in the laboratory. The study was split into three parts, with participants required to conduct two tasks in each part. Participants in this study also conducted demographic, pre-task and post-task questionnaires. This study contained more heterogeneous topics, such as in travel, retirement, and entertainment. In total, the study consisted of 40 searchers who conducted 1274 queries over 6 different tasks.

As previously mentioned, participants answered several questionnaires, both general questionnaires and ones with respect to the task. Each study had some similar or identical questions that all studies asked. The first was general search expertise: INT and SAL asked for this in a demographic questionnaire on a 1 to 5 scale. EOP, due to its design around expert advice, asked 4 questions regarding general search expertise on a 1 to 5 scale; for analyses here, we combined these in an average score. In the pre-task, each study asked how familiar the participant was with the topic, on a 1 to 5 or 1 to 7 scale. The features, overviewed in Table 2, were transformed accordingly.

Each study also collected a common set of browsing behaviors, courtesy of the installed Chrome and Firefox extensions, also shown in Table 2, which also includes statistics on each of these variables. INT and SAL asked about the number of years participants spent conducting online searching and the frequency with which participants use search engines or other online search tools. INT exclusively asked how often participants conducted online searching with respect to their domain of expertise (journalism), on a 1 to 4 scale. It also exclusively asked pre-task how much experience the participant has with the type of assignment (e.g., copy editing), not the topic. INT and EOP asked post-task whether the participant had sufficient time to complete the task successfully, on a 1 to 5 scale.

Table 2: Features absent (–) present in the intentions (INT), searching as learning (SAL), and expert opinion (EOP) data sets. Statistics on each feature are also provided, as well as the number of query segments $|Q|$ per data set.

Feature	INT ($\|Q\|$=693)	SAL ($\|Q\|$=594)	EOP ($\|Q\|$=1,274)
General Search Expertise	$\mu = 4.875, \sigma = 1.00$	$\mu = 4.16, \sigma = 0.68$	$\mu = 4.075, \sigma = 0.46$
Years Spent Searching	$\mu = 10.65, \sigma = 3.01$	$\mu = 9.43, \sigma = 4.37$	–
Search Frequency	$\mu = 6.75, \sigma = 0.59$	$\mu = 3.13, \sigma = 0.92$	–
Professional Domain Expertise	$\mu = 3.35, \sigma = 0.92$	–	–
Topic Familiarity	$\mu = 1.725, \sigma = 1.30$	$\mu = 2.78, \sigma = 1.09$	$\mu = 3.11, \sigma = 1.70$
Assignment Experience	$\mu = 3.05, \sigma = 1.83$	–	–
(Post) Search Difficulty	$\mu = 2.8, \sigma = 1.65$	$\mu = 1.86, \sigma = 0.91$	$\mu = 3.83, \sigma = 1.83$
Adequate Time	$\mu = 4.1, \sigma = 1.03$	–	$\mu = 4.07, \sigma = 0.94$
# Pages	$\mu = 5.75, \sigma = 2.96$	$\mu = 2.25, \sigma = 3.31$	$\mu = 2.00, \sigma = 3.45$
Total content time	$\mu = 76.01, \sigma = 95.47$	$\mu = 74.32, \sigma = 134.09$	$\mu = 53.73, \sigma = 90.83$
Total SERP time	$\mu = 8.79, \sigma = 14.61$	$\mu = 13.80, \sigma = 26.69$	$\mu = 9.72, \sigma = 22.45$
Query length	$\mu = 4.97, \sigma = 3.83$	$\mu = 4.41, \sigma = 4.79$	$\mu = 5.76, \sigma = 3.96$

The unit of analysis in our experiments is the query segment – the behavior conducted starting from when a person issued a query up until the next query. While our data is comprised of only 100 search sessions, it is comprised of 2,561 query segments.

5 EXPERIMENTS AND RESULTS

Each experiment follows a similar template: a structure learning algorithm is applied on a training set and validated on a test set. Log-likelihood on the test set is used in all cases to indicate prediction performance. It is used to both determine fit on a specific variable and an entire graph. The specific choice of training data, test data, and learning algorithm depend on the goal of the experiment, as well as whether the training and test data are partial or complete. In experiments with complete data, we apply max-min hill climbing (MMHC) for structure learning, which is a popular, effective hybrid algorithm for structure learning [8], even for large data sizes. In experiments with incomplete data, we apply the previously mentioned EM-based method from Friedman [7].

5.1 Experiment 1: Control

Our first experiment is a benchmark to consider the performance of algorithms when a single data set from one laboratory study is used for both training and testing. First, how does performance change as the ratio of training data to test data changes on a single data set? Second, how does the amount of missing data in training affect test performance? This will allow us to determine the effects of the amount of missing data and amount of training data (absolute and relative) on future experiments. To this end, we first take each data set individually (i.e., INT, SAL, EOP) and apply MMHC on the set of features present in each. Training and test sets are split by $x\%$ and $100 - x\%$, respectively, where $x = 10, 20, \ldots, 80$. Next, we focus on a 80%/20% split of training and test, instead randomly deleting 10%, 20%, ...60% values before training.

Results are provided in Figures 1-2. Figure 1 shows the results of varying the ratio of training to test for each laboratory data set. The log-likelihood on the test set increases linearly with the number of points in the test set, with the smallest log-likelihood at 90% training. But the log-likelihood function is a sum of the log-likelihood of each

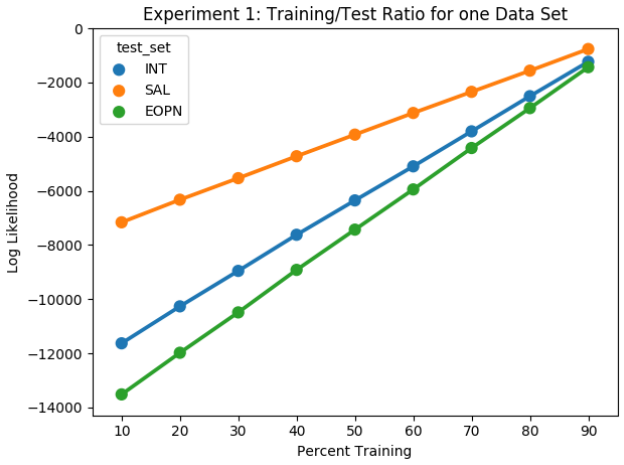

Figure 1: Experiment 1: Log-likelihood when varying ratio of training to test data.

data point in the test set. Therefore, the log-likelihood effectively stays constant, regardless of ratio of training to test. Figure 2 shows the effect of omitting random values from the training set. The results are once again effectively constant for different percentages, and therefore no predictive accuracy is lost with the missing data. In summary, for a single data set, the algorithms converge to a roughly constant log-likelihood per data point. Any differences in log-likelihood in future experiments, therefore, should be attributed to other characteristics of the training data used, not purely training data size or the percent of missing values.

5.2 Experiment 2: Combining Data Sets

Our second and main experiment determines whether predictions of searcher characteristics, task characteristics, and behavior on data from one study can be improved from data in other studies that collects different information about its users. Refer back to Table 2 for the list of user characteristics that are present and

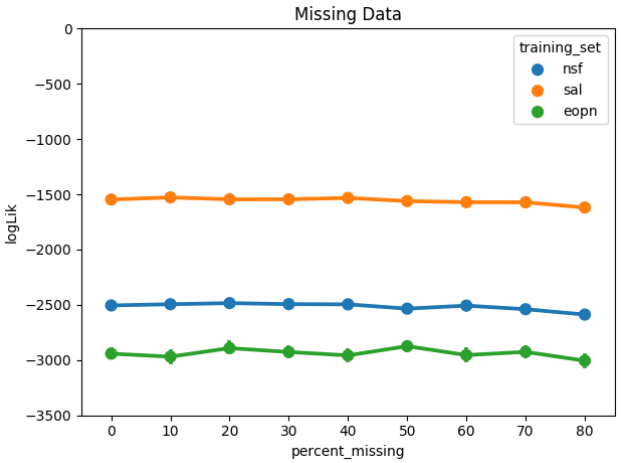

Figure 2: Experiment 1: Log-likelihood when varying the percent of missing data on a test set.

absent in each laboratory data set. This experiment's results can help determine the utility of combining heterogeneous data sets. To this end, suppose we have external data sets $D_{ext} = \{d_1, d_2, ...\}$ (e.g., SAL, INT) and our data set of interest $d_{int} \notin D_{ext}$ (e.g., EOP). We train using 100% of the data from D_{ext} and 80% or the data from d_{int}, with 20% held out for test. When training and testing with one data set alone ($D_{ext} = \emptyset$), MMHC learns a graph on the features present in d_{int}. When external data is added for training, the EM-based algorithm is used to account for missing values in training and testing. Afterwards, we conducted the same experiment while varying the amount of additional training data D_{ext}, ranging from 10% to 100%. All features were used and estimated, unless the feature is present in neither the training nor test data. Log-likelihood over the general graphs shows the general improvement of graphs in terms of their predictive power. Log-likelihood over specific nodes shows the level of improvement in predicting specific variables. We also examine structural hamming distance as a measure of graph similarity, which is the number of edge changes required (additions/removals/reversals) to convert one graph into another.

Our first results can be found in Table 3, which agree with the general premise guiding this study. For the log-likelihood score of an entire graph, it is generally better to rely on incorporating external data, rather than just using a single laboratory data set alone, with all improvements being significant. In particular, best performance was achieved across the board when using all 3 data sets rather than 1 or 2. Although the results for EOP seem to indicate otherwise (-1934.02 versus -1995.01), these two figures were not significantly different ($p = 0.39$). Figures 3- 5 suggest this result is invariant to the amount of training data used from the external training set. The combined data set often performs at least as well as some other data set and is among the best data sets. In the case of EOP, INT data is as good as SAL and INT combined, but both are better than using SAL. In the case of INT, all are fairly close. For SAL, INT and the combination of INT and EOP are better than EOP.

Table 3: Experiment 2: Log-likelihood scores on entire graph. Columns indicate the data sets used for test and training, as well as the log-likelihood scores. ** shows results significantly better than when using one data set alone ($p < .01$).

Test	Training	LL
EOP	EOP	-2944.34
	EOP,INT	-1934.02**
	EOP,INT,SAL	-1995.01**
	SAL	-2613.91**
INT	INT	-2430.50
	EOP,INT	-2504.07**
	INT,SAL	-2374.66**
	EOP,INT,SAL	-2381.30**
SAL	EOP,SAL	-1543.46
	INT,SAL	-1381.66**
	EOP,INT,SAL	-1168.96**
	SAL	-1561.56**

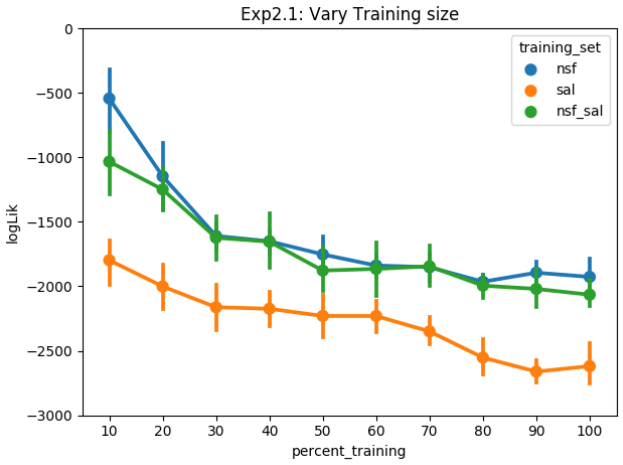

Figure 3: Experiment 2: Log-likelihood as a function of the percent of the data from the external data sets, for EOP.

A more striking result occurs when looking at the specific details, shown in Tables 4- 6. As generally expected and with a few exceptions, adding data improves predictions on specific features. Yet surprisingly and encouragingly, this improvement can occur even when some or all of the training data does not contain the feature that is improved. For the EOP data set (Table 4), for instance, the user's feeling of whether they had adequate time to complete the task was improved by using all data combined, even when SAL did not include such data. The INT data set (Table 5) contains all features. Yet improvements can be seen for professional domain expertise and assignment experience, when neither of the other data sets contained such information. For SAL (Table 6), the years spent searching could be better predicted using all data, even when EOP contained no such information, and search frequency could be better predicted with EOP.

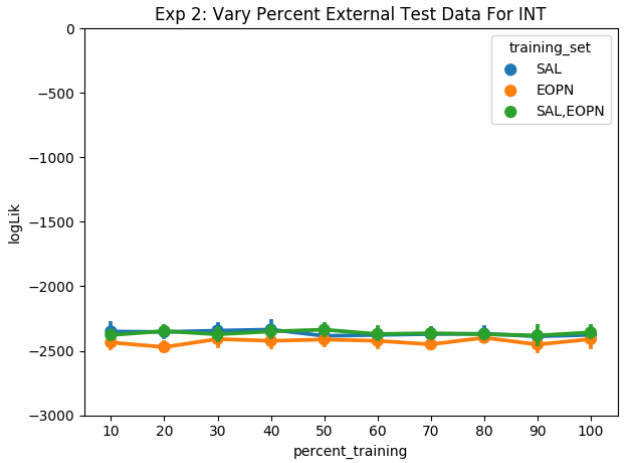

Figure 4: Experiment 2: Log-likelihood as a function of the percent of the data from the external data sets, for INT.

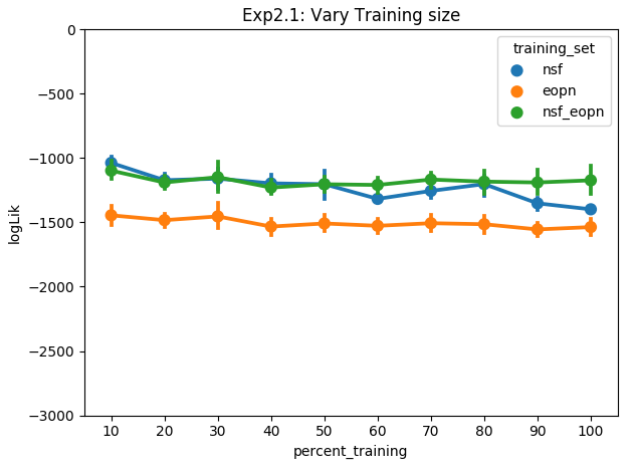

Figure 5: Experiment 2: Log-likelihood as a function of the percent of the data from the external data sets, for SAL.

Moreover, results are overall stable in ways that one would expect, further demonstrated in Table 7. Smaller distance entails that the graphs are similar. Graph structure tends to vary the least on graphs formed with under the same experimental setup. In addition, when using a set of data $d_1, ..., d_n$, the data set d_i used for testing does not affect the structural hamming distance much. Therefore, log-likelihood scores – such as those in Table 3 – can be attributed to the structures learned in those graphs.

Some curious results perhaps entail limitations in the data or in the analysis method used here. First were some results for the EOP data set. It is the only one whose likelihood changed as the external data used to create the graph increased, demonstrated in Figure 3. It moreover showed the markedly worst log-likelihood in Experiment 1 when used alone but showed the most improvement when external data was included. Yet EOP also had the least number

Table 4: Experiment 2: Log-likelihood of features present in EOP data. Columns indicate the feature, the log-likelihood when only using EOP data, the training data that yielded the best log-likelihood, and the respective score.

Feature	LL	Best Data	Best LL
Task Goal	-151.03	SAL,EOP	150.77
Topic Familiarity	-457.44	INT,EOP	25.05
General Search Expertise	-148.58	INT,SAL,EOP	311.62
(Post) Search Difficulty	-503.08	INT,SAL,EOP	13.02
Adequate Time	-334.04	INT,SAL,EOP	-176.81
# Pages	-340.14	SAL,EOP	-304.70
Total content time	-256.38	INT,SAL,EOP	-301.15
Total SERP time	-349.41	SAL,EOP	-338.39
Query length	-355.20	SAL,EOP	-192.30

of user characteristics. Performance increased as more such features were incorporated into the training data, with the largest increases when INT is included (which contained the complete features). Similarly, SAL showed the least number of user characteristics, with performance increasing marginally when EOP is included (no new features) and substantially when other features are included. Adding more features in training may hence considerably increase accuracy.

Moreover, our experimental approach hit some ceilings. For one, despite invariance in the log-likelihood, graphs still varied even when the combination of training and test set sizes are held constant, suggested by small but non-negligible structural hamming distance. Also, consider that the structural hamming distance is often small between configurations where all 3 data sets are used. Nevertheless, there is still a substantial significant difference between log-likelihoods shown in Table 3 on the combined data, when testing on EOP, INT, and SAL (-1995.01, -2381.30, and -1168.96, respectively). We take this to mean that each data set has its own characteristics. Why are the extracted graphs not completely stable with a hamming distance of 0? We interpret this as a limitation of the experimental setup, as query segments are treated as independent points. For a tighter fitting model less prone to variance, perhaps longer sequences of data are required, but this requires more data and is the realm of future work. Also, why does the log-likelihood vary so much when the training sets are nearly identical and the test set changes? This seems to suggest that there are intrinsic limitations, due to some characteristics of the data set. But from this experiment, an analogy can be drawn to crowdsourcing – even with differing and sometimes imperfect data sources, in some scenarios crowdsourcing can draw better results. Similarly in this case, combining multiple data sets of incomplete and otherwise different data can in the end improve prediction when combined.

6 CONCLUSION

In this paper, we advocated for a need for an experimental setup to address the larger problem of IIR. How in practice can one combine laboratory studies that address separate research questions to holistically characterize the relationship between searcher, task, and behavior? This paper shows that this can be done experimentally,

Table 5: Experiment 2: Log-likelihood of features in INT.

Feature	LL	Best Data	Best LL
Task Goal	-83.12	INT	-83.12
Task Product	-53.93	SAL,INT	-38.01
General Search Expertise	-163.00	INT,EOP	-159.41
Years Spent Searching	-326.47	INT,EOP	-318.22
Domain Expertise	-199.12	INT,SAL	-185.20
Topic Familiarity	-236.13	INT,SAL,EOP	-204.13
Assignment Experience	-241.75	INT,SAL,EOP	-219.94
(Post) Search Difficulty	-250.00	INT,EOP	-217.64
Adequate Time	-141.09	INT	-141.09
# Pages	-194.56	EOP	-141.21
Total content time	-140.91	INT,SAL	-99.01
Total SERP time	-199.01	INT,SAL,EOP	-151.62
Query length	-191.23	INT,SAL,EOP	-167.61

Table 6: Experiment 2: Log-likelihood of features in SAL.

Feature	LL	Best Data	Best LL
Task Goal	-72.28	INT,SAL	-63.35
Task Product	20.60	INT,SAL	30.65
General Search Expertise	-90.93	INT,SAL	-19.20
Years Spent Searching	-321.47	INT,SAL,EOP	-172.42
Search Frequency	-149.69	SAL,EOP	106.69
Topic Familiarity	-158.45	INT,SAL	-85.02
(Post) Search Difficulty	-159.41	INT,SAL	-14.96
# Pages	-150.74	SAL	-150.74
Total content time	-140.59	INT,SAL	-127.71
Total SERP time	-157.86	SAL	-157.86
Query length	-159.42	SAL	-159.42

even under the common practical constraints where various studies collect different types of data about their searchers.

Specifically, we discovered that combining data sets to predict user and task characteristics is most often better than using a data set alone. We discovered this both generally and with respect to specific features (Experiment 2). This is not simply in virtue of collecting more data for one particular laboratory study (Experiment 1). Rather, it is the data from other studies that improves prediction, more so than the addition of data itself (Experiment 2). Part of this is due to the features each study covers. As more features are covered by the training data, this can tend to boost prediction. But a study that covers all features is not necessary; even if a feature is largely missing in training data from other studies, this training data is still useful in even predicting that particular feature. Despite differences between data sets, the framework explored here behaves much like crowdsourcing. Combining laboratory studies to boost performance (in our case, on task/user/behavior prediction) is better than using any data set alone, even when the laboratory studies show various differences.

One implication of this work is a potential diagnostic model when conducting laboratory studies. Suppose one study collects several user and task characteristics but forgets time pressure or for pragmatic reasons omits query-specific intentions. They are not observing the results they would expect. How much of this can be attributed to time pressure? Did these searchers perhaps exhibit different intentions for their queries than the researchers would have assumed? Applying a learned graphical model to this can help diagnose possible causes of unexpected variations in data.

But with this work comes discussions predominantly revolving around privacy. This work entails that it is necessary to combine multiple data sets to achieve the goal of IIR. Not every researcher has the luxury of housing multiple data sets that they can combine to conduct the research performed here or to subsequently aim for the broad goal of IIR. For an approach such as this to be deployed, data would need to be shared, and hence the privacy implications of such sharing would need to be considered. An alternative – which still requires privacy research – is to develop a hybrid model that aggregates smaller models learned on individual data sets, as with a decision forest or some other aggregation function. But this is the realm of future work, and we nevertheless think this is a promising step towards holistic IIR.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation (NSF) grant IIS-1717488. It also uses data generated as a part of the NSF grant IIS-1423239.

REFERENCES

[1] Pia Borlund and Peter Ingwersen. 1997. The development of a method for the evaluation of interactive information retrieval systems. *Journal of documentation* 53, 3 (1997), 225–250.

[2] Pia Borlund and Jesper W. Schneider. 2010. Reconsideration of the Simulated Work Task Situation: A Context Instrument for Evaluation of Information Retrieval Interaction. In *Proceedings of the Third Symposium on Information Interaction in Context (IIiX '10)*. ACM, New York, NY, USA, 155–164. https://doi.org/10.1145/1840784.1840808

[3] D. J. Campbell. 1988. Task Complexity: A Review and Analysis. *Academy of Management Review* 13, 1 (1988).

[4] Rob Capra, Jaime Arguello, and Yinglong Zhang. 2017. 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))*. Springer Verlag, Germany, 108–121. https://doi.org/10.1007/978-3-319-56608-5_9

[5] Michael J. Cole, Jacek Gwizdka, Chang Liu, Nicholas J. Belkin, and Xiangmin Zhang. 2013. Inferring user knowledge level from eye movement patterns. *Information Processing & Management* 49, 5 (2013), 1075 – 1091. https://doi.org/10.1016/j.ipm.2012.08.004

[6] Anita Crescenzi, Diane Kelly, and Leif Azzopardi. 2015. 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)*. ACM, New York, NY, USA, 767–770. https://doi.org/10.1145/2766462.2767817

[7] Nir Friedman. 1997. Learning Belief Networks in the Presence of Missing Values and Hidden Variables. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML '97)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 125–133. http://dl.acm.org/citation.cfm?id=645526.657145

[8] Maxime Gasse, Alex Aussem, and Haytham Elghazel. 2012. An Experimental Comparison of Hybrid Algorithms for Bayesian Network Structure Learning. In *Proceedings of the 2012th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I (ECMLPKDD'12)*. Springer-Verlag, Berlin, Heidelberg, 58–73. https://doi.org/10.1007/978-3-642-33460-3_9

[9] Ahmed Hassan Awadallah, Ryen W. White, Patrick Pantel, Susan T. Dumais, and Yi-Min Wang. 2014. Supporting Complex Search Tasks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14)*. ACM, New York, NY, USA, 829–838. https://doi.org/10.1145/2661829.2661912

[10] Jiyin He and Emine Yilmaz. 2017. User Behaviour and Task Characteristics: A Field Study of Daily Information Behaviour. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17)*. ACM, New York, NY, USA, 67–76. https://doi.org/10.1145/3020165.3020188

Table 7: Structural hamming distance between graphs. The key in rows and columns is represented by training sets on the left and test on the right (E=EOP, I=INT, S=SAL).

Data Sets	EI/I	ES/S	EI/E	IS/S	EIS/S	EIS/E	ES/E	IS/I	EIS/I
EI/I	15.28	44.65	19.54	42.33	38.67	40.31	44.48	42.60	38.07
ES/S	44.65	14.92	46.86	46.82	45.82	43.65	14.71	45.95	41.81
EI/E	19.54	46.86	16.91	42.92	41.71	42.97	47.48	43.94	41.51
IS/S	42.33	46.82	42.92	8.39	45.56	43.33	46.80	17.91	44.80
EIS/S	38.67	45.82	41.71	45.56	13.71	23.77	47.15	44.48	21.57
EIS/E	40.31	43.65	42.97	43.33	23.77	20.04	43.59	40.68	24.64
ES/E	44.48	14.71	47.48	46.80	47.15	43.59	10.55	45.38	42.73
IS/I	42.60	45.95	43.94	17.91	44.48	40.68	45.38	17.45	42.26
EIS/I	38.07	41.81	41.51	44.80	21.57	24.64	42.73	42.26	18.57

[11] Daniel Hienert, Matthew Mitsui, Philipp Mayr, Chirag Shah, and Nicholas J. Belkin. 2018. 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)*. ACM, New York, NY, USA, 72–81.

[12] Rong Hu, Kun Lu, and Soohyung Joo. 2013. 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)*. American Society for Information Science, Silver Springs, MD, USA, Article 66, 9 pages. http://dl.acm.org/citation.cfm?id=2655780.2655846

[13] Jiepu Jiang, Daqing He, and James Allan. 2014. Searching, Browsing, and Clicking in a Search Session: Changes in User Behavior by Task and over Time. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)*. ACM, New York, NY, USA, 607–616. https://doi.org/10.1145/2600428.2609633

[14] Liangda Li, Hongbo Deng, Anlei Dong, Yi Chang, and Hongyuan Zha. 2014. 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)*. ACM, New York, NY, USA, 731–740. https://doi.org/10.1145/2623330.2623679

[15] Yuelin Li and Nicholas J. Belkin. 2008. A Faceted Approach to Conceptualizing Tasks in Information Seeking. *Inf. Process. Manage.* 44, 6 (Nov. 2008), 1822–1837.

[16] Chang Liu, Nicholas J. Belkin, and Michael J. Cole. 2012. 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)*. ACM, New York, NY, USA, 205–214.

[17] Chang Liu, Jacek Gwizdka, and Nicholas J. Belkin. 2010. Analysis of Query Reformulation Types on Different Search Tasks *(iConference '10)*. 4.

[18] Chang Liu, Jingjing Liu, Michael Cole, Nicholas J. Belkin, and Xiangmin Zhang. [n. d.]. Task difficulty and domain knowledge effects on information search behaviors. *Proceedings of the American Society for Information Science and Technology* 49, 1 ([n. d.]), 1–10. https://doi.org/10.1002/meet.14504901142 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/meet.14504901142

[19] Chang Liu and Yiming Wei. 2016. 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)*. American Society for Information Science, Silver Springs, MD, USA, Article 51, 9 pages. http://dl.acm.org/citation.cfm?id=3017447.3017498

[20] Jingjing Liu, Michael J. Cole, Chang Liu, Ralf Bierig, Jacek Gwizdka, Nicholas J. Belkin, Jun Zhang, and Xiangmin Zhang. 2010. Search Behaviors in Different Task Types. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries (JCDL '10)*. ACM, New York, NY, USA, 69–78. https://doi.org/10.1145/1816123.1816134

[21] Jingjing Liu, Jacek Gwizdka, Chang Liu, and Nicholas J. Belkin. 2010. 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)*. American Society for Information Science, Silver Springs, MD, USA, Article 16, 10 pages. http://dl.acm.org/citation.cfm?id=1920331.1920355

[22] Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri, and Gabriele Tolomei. 2011. Identifying Task-based Sessions in Search Engine Query Logs. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM '11)*. ACM, New York, NY, USA, 277–286. https://doi.org/10.1145/1935826.1935875

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

[24] Rishabh Mehrotra and Emine Yilmaz. 2015. Terms, Topics & Tasks: Enhanced User Modelling for Better Personalization. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval (ICTIR '15)*. ACM, New York, NY, USA, 131–140.

[25] Matthew Mitsui, Jiqun Liu, and Chirag Shah. 2018. The Paradox of Personalization: Does Task Prediction Require Individualized Models?. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18)*. ACM, New York, NY, USA, 277–280. https://doi.org/10.1145/3176349.3176887

[26] Matthew Mitsui and Chirag Shah. 2018. The Broad View of Task Type Using Path Analysis. In *Proceedings of The 8th International Conference on the Theory of Information Retrieval (ICTIR '18)*. ACM, New York, NY, USA, 8.

[27] Iraj Mohammadfam, Fakhradin Ghasemi, Omid Kalatpour, and Abbas Moghimbeigi. 2017. Constructing a Bayesian network model for improving safety behavior of employees at workplaces. *Applied Ergonomics* 58 (2017), 35 – 47. https://doi.org/10.1016/j.apergo.2016.05.006

[28] R Neapolitan, X Jiang, DP Ladner, and B Kaplan. 2016. A Primer on Bayesian Decision Analysis With an Application to a Kidney Transplant Decision. In *Transplantation*. 489–96.

[29] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17)*. ACM, New York, NY, USA, 117–126. https://doi.org/10.1145/3020165.3020183

[30] G. A. P. Saptawati and B. Sitohang. 2005. Hybrid algorithm for learning structure of Bayesian network from incomplete databases. In *IEEE International Symposium on Communications and Information Technology, 2005. ISCIT 2005.*, Vol. 1. 741–744. https://doi.org/10.1109/ISCIT.2005.1566960

[31] M. Berkan Sesen, Ann E. Nicholson, Rene Banares-Alcantara, Timor Kadir, and Michael Brady. 2013. Bayesian Networks for Clinical Decision Support in Lung Cancer Care. *PLOS ONE* 8, 12 (12 2013). https://doi.org/10.1371/journal.pone.0082349

[32] Joshua B. Tenenbaum and Thomas L. Griffiths. 2001. Structure Learning in Human Causal Induction. In *Advances in Neural Information Processing Systems 13*, T. K. Leen, T. G. Dietterich, and V. Tresp (Eds.). MIT Press, 59–65. http://papers.nips.cc/paper/1845-structure-learning-in-human-causal-induction.pdf

[33] Elaine G. Toms, Heather O'Brien, Tayze Mackenzie, Chris Jordan, Luanne Freund, Sandra Toze, Emilie Dawe, and Alexandra MacNutt. 2008. Task Effects on Interactive Search: The Query Factor. In *Focused Access to XML Documents*, Norbert Fuhr, Jaap Kamps, Mounia Lalmas, and Andrew Trotman (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 359–372.

[34] Manisha Verma and Emine Yilmaz. 2016. Category Oriented Task Extraction. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval (CHIIR '16)*. ACM, New York, NY, USA, 333–336. https://doi.org/10.1145/2854946.2854997

[35] Ryen W. White, Susan T. Dumais, and Jaime Teevan. 2009. 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)*. ACM, New York, NY, USA, 132–141. https://doi.org/10.1145/1498759.1498819

[36] Ryen W. White and Diane Kelly. 2006. A Study on the Effects of Personalization and Task Information on Implicit Feedback Performance. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM '06)*. ACM, New York, NY, USA, 297–306. https://doi.org/10.1145/1183614.1183659

[37] Hong (Iris) Xie. 2002. Patterns between interactive intentions and information-seeking strategies. *Information Processing & Management* 38, 1 (2002), 55 – 77. http://www.sciencedirect.com/science/article/pii/S0306457301000188