The Broad View of Task Type Using Path Analysis

Matthew Mitsui
Department of Computer Science
Rutgers University, New Brunswick, NJ
mmitsui@rutgers.edu

Chirag Shah
School of Communication & Information
Rutgers University, New Brunswick, NJ
chirags@rutgers.edu

ABSTRACT

Many past research efforts have examined the relationships between observable search behaviors, the task that drives searching, and user characteristics such as search expertise and topic familiarity. These studies often look at pairwise relationships between characteristics to show that task characteristics or user characteristics can be distinguished by differences in browsing behavior. Recent work has moved toward detecting and predicting task from browsing behavior but has not additionally considered user or session characteristics that clearly affect browsing. To what extent should user characteristics and unlogged traits be considered when inferring task type from browsing behavior? This paper examines this complex relationship through the lens of path analysis, showing that such a holistic view should be considered in future task prediction research. Future task prediction work should consider direct links from task to behavior, indirect links through other factors, and other characteristics that affect browsing.

CCS CONCEPTS

• **Information systems → Task models**; **Retrieval tasks and goals**; *Personalization*; • **Mathematics of computing →** *Multivariate statistics*;

KEYWORDS

structural equation modeling; path analysis; task prediction; task classification; interactive information retrieval; task type; user behavior

ACM Reference Format:
Matthew Mitsui and Chirag Shah. 2018. The Broad View of Task Type Using Path Analysis. In *2018 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '18), September 14–17, 2018, Tianjin, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3234944.3234951

1 INTRODUCTION

It has been long known that searchers look for information to accomplish a task. Interactive Information Retrieval (IIR) research has long been interested in drawing relationships between searchers' behaviors and characteristics of the task they are trying to accomplish. Various work has shown that a system's knowledge about searchers' tasks can assist them in task accomplishment. Knowledge about task and topic can improve query prediction and query recommendation [22], and knowledge about task type can improve retrieval performance in session-based tasks [16]. Different types of tasks require different types of information, such as exploratory and fact-finding tasks [27]. Hence, predicting the task characteristics of a search session is useful for making such improvements real.

Some work has cast task prediction as query prediction [22], and other recent work has explored predicting general *task types*. The latter uses frequentist statistics to show that tasks with different goals, products, levels, and objective complexities (according to the schema of Li & Belkin [15]) can be characterized by significant differences in whole session behaviors [17]. Other research efforts have distinguished tasks using whole session eye tracking patterns and dynamic changes between pages [6]. Other work has explored whether first query behaviors can be used [2] and even compared how first measures differ from their whole session counterparts among various task types [10].

When generalizing to task prediction, results are more mixed. Task prediction involves predicting the task type of an unobserved session. Previous work used Markov Modeling and Hidden Markov Modeling to build models of users' transitions between pages, demonstrating effective prediction distinguishing a session with a single exploratory task versus a session with multiple fact-finding tasks [14]. However, other work has used whole session browsing features similar to those in the previous work ([10, 17]), achieving statistically significant findings that entailed only marginally improved prediction performance [24].

What bottlenecks could exist? The previous statistical work distinguished task types using individual behaviors, for instance with t-tests or chi-square statistics. Additionally, task prediction work typically combines behavioral characteristics into a single model. Yet other work emphasizes the importance of other user or task characteristics. Query prediction research such as Mehrotra et al. 2015 [22] models a task as a combination of topics. Hienert et al. 2018 [8] also suggested that topic influences search behavior. Also, some user characteristics influence behavior but are difficult to control in experimental settings. Age can affect the types of search strategies used [20], as can critical thinking skills [9]. Task difficulty [1] and time pressure [7] also significantly impact behaviors. Therefore, while task type in itself may significantly alter behavior, additional factors should be considered.

In short, current understanding of modeling the relationship between task and behavior is insufficient. Factors like age and time pressure clearly affect behavior. Even if these can be modeled or controlled, it may be simplistic to assume these are all independently related to behavior. This is an assumption of typical linear models with independent variables like linear regression, which do

ICTIR '18, September 14–17, 2018, Tianjin, China

ACM ISBN 978-1-4503-5656-5/18/09…$15.00
https://doi.org/10.1145/3234944.3234951

not capture layers of relationships. As an example of a nested relationship, variables like *task difficulty* and *topic familiarity*, which may affect behavior, are in turn affected by the task and topic.

Little work has examined these holistic - and possibly hierarchical - interactions of multiple variables. For instance, does the task in and of itself affect behaviors, or is this only due to the familiarity of the task? There are therefore two gaps in knowledge. First, little has been done to create such holistic models to study the relationship between task type, user, and behaviors. Second, little has hence been done to weigh the utility of different features – such as users' search expertise or topic familiarity – in such models.

We propose path analysis as a useful tool to bridge this gap, linking several variables holistically into one model. Using path analysis, we address the following research questions:

- To what extent are user background, subjective task properties, and search intentions useful in distinguishing task types when given search behaviors?
- To what extent do task characteristics directly or indirectly affect search behaviors?

We will first review discoveries made in IIR studies between pairs of variables. We will then describe path analysis as an improvement over these pairwise analyses. We will use our literature review to construct a path model linking tasks, user characteristics, and behavioral signals and will apply our model to a laboratory data set. We show that some indirect effects indeed are important, and while task type seems to directly affect browsing behaviors, other factors should additionally be considered in prediction research.

2 BACKGROUND

We begin this section with an overview of the relationships between task, other user characteristics, and behaviors observed in literature. We group these findings in Section 2.1 and then summarize the theoretical assumptions drawn from combining these findings. In Section 2.2, we will motivate the need for more complex modeling, describing how this complexity can be accomplished with graphical models like path analysis. Section 2.3 will conclude with some path analysis terminology

2.1 Task Type, Behaviors, and User Characteristics

Task → Behaviors - Several works have shown that some behaviors significantly differ between different task types. [10, 17] showed that task completion time and total number of pages and queries differ among types. Significantly different numbers of viewed pages can distinguish between different task goals. [2] showed significant differences between lookup and exploratory tasks for query length, query segment duration, query dwell time, and document dwell time for the first query. Lastly, [4] demonstrated significant differences in whole session behaviors among tasks of differing objective complexity (e.g., *analyze*, and *create* tasks).

Task → Intentions → Behaviors - [21] categorized searchers' actions in a search session into types of tactics, moves, and strategies. [28] 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". Descriptive non-inferential evidence from [25] suggested that these intentions are exhibited in different proportions among task types and perhaps can distinguish types. Later, [23] showed that these intentions can be predicted at the query segment level, using machine learning methods with query browsing features as input. [23] 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.

Task → Search Experience - [19] controlled tasks by product, goal, and complexity, demonstrating a significant difference in both pre-task and post-task difficulty among the task types.

Background → Intentions - [20] showed that searchers with varying expertise used differing strategies in an encyclopedia task, for instance differing query reformulation strategies. [26] subsequently showed that differences in reformulation strategies can be associated with differences the aforementioned intentions of [28].

Experience → Behaviors - Various work has shown that searching experience can affect behaviors. For instance, difficulty affects search behavior [1, 3], and the nature of this relationship can even differ among task types [18]. Topic and topic familiarity can also affect behaviors [8**?**].

Within-category Correlations - [7] found that time pressure correlates with task difficulty, and assignment experience has also been found to correlated with search difficulty [19].

Motivation - Already within this review, it can be seen that some variables affect each other in a complex manner. Even if relationships are mathematically linear, these relationships look more like a nested set of linear equations than one equation mapping behaviors and user characteristics to task type. For instance, topic familiarity can affect behaviors, but topic familiarity is a function of the user and the topic. Similarly, assignment experience correlates with difficulty - which in turn affects behavior - and is also a function of the task and the user - i.e., how familiar a user is with a particular type of task - suggesting three to four layers of effects. This suggests that modeling complex relationships between variables may be necessary in capturing the relationship between task and behavior. We express this more mathematically below.

2.2 Path Analysis - Motivation

For illustrative analogy, let us consider linear regression in its typical form:

$$\beta_1 x_1 + \beta_2 x_2 + \dots \beta_n x_n = y \tag{1}$$

y is a dependent variable - in our case the task type (e.g., binary "specific" or "amorphous" goal). Each x_i is an independent variable, for instance a behavioral feature like query length. Moreover, the variables are independent of each other, that is:

$$\forall_{i,j} x_i \perp x_j | \emptyset \tag{2}$$

This assumption is relaxed in path analysis:

$$\begin{gathered} \beta_1 y = x_1 \\ \beta_2 y = x_2 \\ \dots \\ \beta_k z + \beta_n y = x_n \end{gathered} \tag{3}$$

In one regard, path analysis is multiple regression that focuses on causality. Regression would combine multiple browser signals to infer task type (Equation 1), while a path model suggests that the task type gives rise to browser behaviors (Equation 3).

In another regard, path analysis allows for nesting - referred to as *recursion*. The output of one equation can be the input to another. This allows us to relax the assumption of Equation 2. Path analysis provides a tool for bridging the exact gap stated in the previous sections. Namely, we can simultaneously model relationships such as $topic\_familiarity \not\perp topic|\emptyset$, $topic\_familiarity \not\perp behaviors|\emptyset$, $topic \not\perp behaviors|\emptyset$.

2.3 Path Analysis - Terminology

Path models contain several important components. Variables are either *exogenous variables*, *endogenous variables*, or error terms, all of which are connected to each other. Exogenous and endogenous variables are defined in the data set, and error terms are residuals from estimating the path model (explained below). Variables are connected to each other directly by a single arrow (*direct effects*) or indirectly by unidirectional paths (*indirect effects*). The combination of all connections between two variables is their *total effect*.

Consider our example in Figure 1, borrowed from [5]. *Age* is the only exogenous variable. Age has direct effects $\beta_{age,aut}$ and $\beta_{age,inc}$, on *Autonomy* and *Income*, respectively. *Age* has an indirect effect on income through *Autonomy*, and its total effect is influenced by the weights $\beta_{age,inc}$,$\beta_{age,aut}$,and $\beta_{aut,inc}$.

Typically, such models are *overidentified*, namely there is not a unique solution to the set of equations provided. The coefficients β are estimated using a variety of techniques, such as ordinary least squares or maximum likelihood estimation. In linear regression, we try to maximize the following maximum likelihood function:

$$F = \sum_{i=1}^{n} \log p(y_i|x_i; \beta_1, ...\beta_j) \quad (4)$$

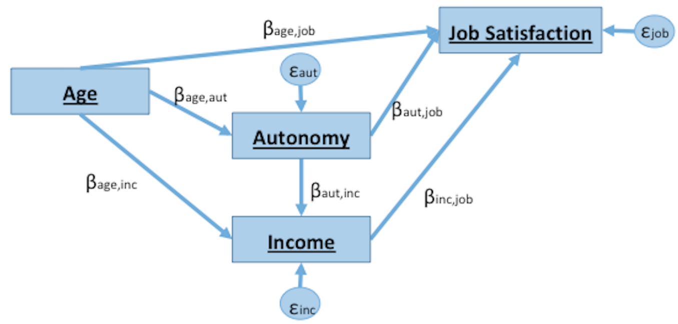
Where β_j are the parameters to estimate in the regression and (x_i, y_i) are datapoints. In path analysis, it is instead the following:

$$F = \ln|\Sigma| - \ln|S| + tr(S\Sigma^{-1}) + (z-\mu)^T\Sigma^{-1}(z-\mu) - (p+q) \quad (5)$$

Where Σ and S are respectively the covariance matrix of the input data and the covariance matrix of the model to estimate (i.e., the covariance matrix that results from estimating β's), tr($\cdot$) and $|\cdot|$ are the trace and determinant of matrices, μ and z are their respective mean vectors, and $p + q$ is the number of variables in the estimated model. A linear regression would estimate one row or column in Σ at best, where the row represents the target dependent variable y and each entry represents β. Regression does not account for dependencies among variables, even covariance between the independent variables themselves. As in Equation 4, path model estimation includes covariance between variables and adds levels of modeling that are perhaps necesssary.

A more general form of path analysis, structural equation modeling (SEM), has been used in recent information retrieval studies. For instance, it was used in [9] to demonstrate the relationship between information retrieval skills and efficacies such as critical thinking, logical thinking, and formal internet training. [29] used it

Figure 1: An example path model.



to understand the relationship between document relevance, document reliability, understandability, topicality, novelty, and scope. Both path analysis and SEM were presented recently in a tutorial in CHIIR 2018 [11]. SEM requires the creation of hidden latent variables that mediate those that exist in the data, and there must be justification for latent variables. We choose the simpler model that does not assume hidden variables.

Path models were constructed and run using SPSS AMOS.[1]

3 DATA SET

For applying path analysis, we use queries from an IIR laboratory study, as detailed in Table 2. 40 undergraduate journalism students each conducted 2 search tasks and were given 20 minutes to complete each task. After each task, they annotated the intentions of each query segment without time limit. Query segment intentions indicate the goals a searcher wants to accomplish when issuing a search query, reading SERPs, and clicking on the subsequent results. For instance, was the searcher attempting to keep record of a specific link or comparing multiple pieces of information? We used the 20 intentions from [25], a subset of those in Xie [28]. Participants also completed a general background questionnaire about their search expertise and pre- and post-task questionnaires regarding task knowledge and task difficulty. The questions we use in our analysis are listed under *Background* and *Experience* in Table 1.

IIR studies are challenged by data size, which is a serious consideration in path analysis. While there is no mathematical formula, typical data size suggestions for path models and SEM include a minimum 200 data points in a data set [13]. Our data set provides 80 sessions, 80 task questionnaires, and 40 user background questionnaires. Nevertheless, we use 693 data points, where 693 is the number of query segments in our data. A query segment begins with a query – e.g., to Google – and ends with the next query. Our path model therefore contains variables at different levels (e.g. task variables with 80 instances and query variables with 693 instances), but we believe this to be a reasonable approach to path analysis, as we demonstrate results that agree with past literature.

[1] https://www.ibm.com/us-en/marketplace/structural-equation-modeling-sem

Table 1: Variables in the path model

Category	Variable	Description	Values	Summary
Task	Product	Task product [15]	{Specific, Amorphous}	% Specific=27.5%
Task	Goal	Task goal [15]	{Factual, Intellectual}	% Factual=50%
Topic	Topic	Task topic	{Coelacanths, Climate Change}	% Climate Change=50%
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	Search Difficulty	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 [28]	20 indicators: present or absent, in 5 groups (numeric count)	$(\mu_{frequency}, \sigma_{frequency}) =$ (21.05%11.15%)
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 2: Task characteristics.

Task	Product	Goal	$\lvert T \rvert$	$\lvert Q \rvert$
CPE	Factual	Specific	22	206
STP	Factual	Amorphous	18	108
REL	Intellectual	Amorphous	18	155
INT	Intellectual	Amorphous	22	224

4 EXPERIMENTS AND METRICS

4.1 Model Building

There are two methods to developing path models and SEMs. The first begins with exploratory factor analysis to discover the optimal number of latent variables in a SEM and the strengths of relationships between variables. This is followed by confirming the model's goodness of fit on external or held-out data. Such an approach was taken to model relationships between document reliability, understandability, topicality, novelty, and scope [29].

The second approach is to build a model from literature review. Significant relationships between variables from literature indicate dependencies/equations in the model (as in Equation 3). This approach has been taken in works like [12]. Since much literature has explored the relationship between task, topic, browser signals, and other user characteristics, we adopt the latter approach, later examining our findings for confirming evidence of our model choice.

Recall Section 2.1. Below we list all relationships included in our model (as directed and two-way arrows). Below we list citations for relationships where significant differences have been found, taking the opportunity to list additional relationships to test. All these features were included in our most complex path model, as discussed in the next section.

Exogenous variables - Our exogenous variables are task goal, task product, topic, and *Background* variables.

Behaviors/Signals - # pages viewed, total content page dwell time, total SERP dwell time, and query length for a query segment.

Task → Behaviors - Task goal, product → *Behaviors* [2, 10, 17].

Task → Intentions → Behaviors - Task goal, product → intention groups [25]; intention groups→*Behaviors* [23].

Task/Topic → Search Experience - Task product, goal→search difficulty [19]; topic → topic familiarity.

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

Table 3: The different models tested, as well as whether there are edges between each group (Y=Yes,N=No).

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

Background → Intentions - Search expertise → intentions [20, 26].

Experience → Behaviors - Topic familiarity → *Behaviors* [8, 18]; search difficulty → *Behaviors* [1, 3, 18].

Within-category Correlations - Adequate time⟷task difficulty [7]; assignment experience ⟷ search difficulty [19]; task goal ⟷ task product (our data is not perfectly balanced); topic familiarity → search difficulty.

See Figure 2 for a summary of the full model. Each node of the model indicates several variables. For instance, the "Task" node indicates 3 binary variables: the task goal, the task product, and the task category. And a path indicates that there is some dependency between them. Also note that henceforth we use "Behaviors" and "Signals" interchangeably.

4.2 Model Variations

A path analysis begins with two basic models: the *saturated* model and *independent* model. The saturated model assumes that all variables are correlated with each other. That is, when given n variables there are $\frac{n(n+1)}{2}$ paths. The independent model, in contrast, assumes no variables are connected to each other and that variables' values are only manifest through their error variance.

These two models are compared to the models the researcher creates. We will henceforth delineate the model constructed in the previous section as our *full model*. We derive several models from the full model as follows: 1) Select categories of variables $C_{excl} = \{C_1, ..., C_n\}$ (e.g. {*Background*, *Experience*}). 2) Select edges from C_{excl} that would directly or indirectly connect it to task properties or browser signals, and constrain the edges to 0. Figure 2, in addition to showing the full figure, shows an example which disconnects the *Intentions* from task and signals. In addition to its direct connections, its connection to background is also severed so that it does not influence the background variables.

One could remove variables from the path model entirely, but the evaluation metrics for path analysis are relative to the saturated model, dependent on the covariance matrix, and therefore dependent on the number of variables. We therefore constrained path values as above. See Table 3 for a summary of these variations.

4.3 Evaluation Metrics

For evaluation, we ask: How important is each variable category in affecting task and/or behaviors? In path analysis, this is equivalent to: How well do different path constraints explain covariance in our data? The data supplied to path analysis is a covariance matrix. That is, a square matrix Σ where each index Σ_{ij} is:

$$\Sigma_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)] \tag{6}$$

Evaluation metrics for path models are largely based on goodness of fit, with respect to recapturing Σ. The saturated model recreates this covariance matrix perfectly, while other models create an imperfect covariance matrix S. A fundamental evaluation metrics χ^2, which compares S to Σ:

$$\chi^2 = \sum_{ij} \frac{(S_{ij} - \Sigma_{ij})^2}{\Sigma_{ij}} \tag{7}$$

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

$$GFI = 1 - \frac{Cov_{residual}}{Cov_{total}} \tag{8}$$

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

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:

$$RMSEA = \sqrt{\frac{\chi^2 - df}{df(N-1)}} \tag{9}$$

Lastly, the Aikake information criterion (AIC) and Bayesian information criterion (BIC) are provided as follows:

$$AIC = \chi^2 + k(k+1) + 2df \tag{10}$$

$$BIC = \chi^2 + ln(N)\left(\frac{k(k+1)}{2} - df\right) \tag{11}$$

Where N is the number of data points, k is the number of parameters and df is the number of degrees of freedom.

5 RESULTS

We compared the saturated model, independent model, and those listed in Table 3. Recall that in entries listed as N (No), factor loadings β were constrained to 0, assuming these variables were unimportant in the model. We evaluate models on two levels. First, we check their goodness of fit and examine possible reasons metrics could fluctuate. Second, we look at significant factor loadings, namely significant *direct effects*, *indirect effects* and *total effects*. We arrive at the following conclusions:

The best model for most metrics uses only background and experience measures. - While not having the smallest χ^2 among the tested models, the BE model has the smallest χ^2/df. It also ranks the highest for adjusted 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 in terms of adjusted

Figure 2: The full path model used in our analyses. Blue paths indicate all connections used in the full path model. The red doted lines indicate paths that are omitted when intentions are omitted from analysis.

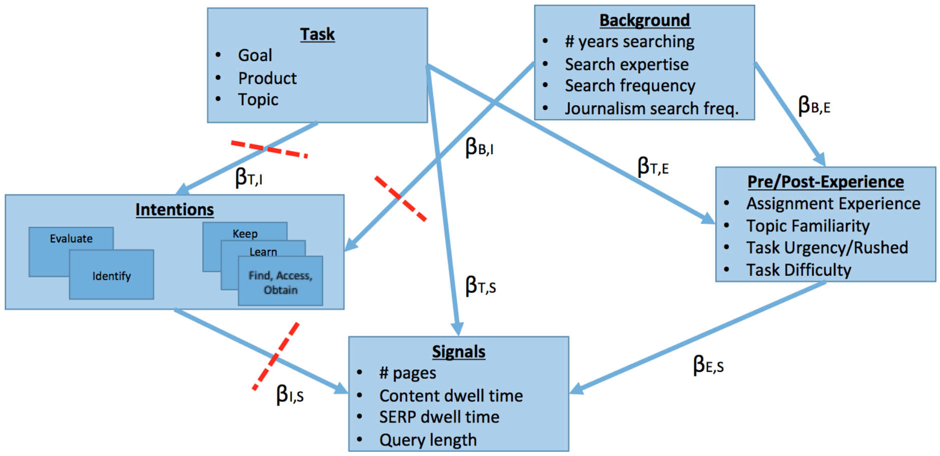


Table 4: Goodness of fit measures for the path models. Best performers (aside from Saturated and Independent) are boldfaced.

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

measures. It has the worst AGFI and PGFI, and χ^2/df is on a 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, 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 potential gaps in the literature.

Table 5 shows all of the important effects influencing or influenced by tasks or browser signals. We arrive at the following conclusions.

Inasmuch as covered by this model, there are still direct paths from task type to browser signals - There are very frequently total and direct effects from task goal, product, and topic to our browser features, as shown in Table 5. This may be a genuine direct effect or due to some unrecorded variable.

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.

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

Table 5: Significant pathways: 1) From task to endogenous variables. 2) From endogenous variables to browser signals

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
Intent - Evaluate	# pages	4	0	3	NA	4
Intent - Evaluate	Content dwell time	4	0	3	NA	4
Intent - Find/Access/Obtain	# pages	4	0	4	NA	4
Intent - Identify	# pages	4	0	4	NA	4
Intent - Identify	SERP dwell time	4	0	4	NA	4
Intent - Identify	Query length	4	0	2	NA	2
Intent - 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

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 [2]). 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, intermediate effects cannot be ignored. For instance, topic affects browsing behaviors. It affects them in and of itself but also through a user's familiarity. Models that account for intermediate effects are hence necessary. Second, task – as discussed here – still has important direct effects on browsing behavior. Yet factors aside from task certainly influence behaviors, and this is perhaps one of task prediction's obstacles. Third, none of the models presented here are a good 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, task model using just background and experience information seems to provide the best fit overall, but intentions data still has an affect on browsing.

6 CONCLUSION AND FUTURE WORK

In this study, we proposed path analysis as a comprehensive solution to examining the relationship between task type, user characteristics, and behavioral signals simultaneously in web search sessions. We learned from this study that not only is such a comprehensive model necessary, but our current understanding of how these variables relate to each other is perhaps incomplete. Specifically, we learned that task type seems to directly affect browsing, but this effect is somewhat mediated by user factors like topic familiarity. We further learned there are other variables that - agreeing with previous literature - clearly affect browsing. The effectiveness of browsing behavior to predict task type will ultimately be affected by variance in things like task difficulty, time pressure, and intentions, which are difficult to control but should be accounted for. We expanded on previous literature with our complex path analysis yet still found several findings that agree with past work.

Our first most obvious limitation lies in our choices in path model design. First, we chose paths from the literature rather than performing exploratory analysis. We similarly chose only 4 browser signal variables out of many possible and chose to combine 20 search intentions into 5 categories. This was done to limit the number of free variables, as 5-10 data points are recommended per degree of freedom [13]. Manually selecting paths also served to confirm (or reject) past findings. Yet we found that there is much variance unexplained in our current models, because we did not draw relationships we did not find in the literature. As such, perhaps more exploration - and even replications - are required to determine what other relationships exist and if they are real.

We similarly acknowledge that we use 693 query segments and query segment browsing behaviors, though they are grouped among 80 sessions and questionnaires for 40 users. To remove any imbalances that would come from this, it would be good to have hundreds of sessions, one user per session, and demographic surveys and task questionnaires from each. This would provide rich data for path analysis but would be difficult to obtain. Similarly, additional behaviors could be analyzed, such as eye tracking behaviors. Yet we believe this complex analysis is a solid step in the right direction for task modeling and prediction. We have demonstrated that analyzing complex relationships is both useful and necessary.

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] Jaime Arguello. 2014. Predicting Search Task Difficulty. In *Advances in Information Retrieval*, Maarten de Rijke, Tom Kenter, Arjen P. de Vries, ChengXiang Zhai, Franciska de Jong, Kira Radinsky, and Katja Hofmann (Eds.). Springer International Publishing, Cham, 88–99.

[2] Kumaripaba Athukorala, Dorota G, Giulio Jacucci, Antti Oulasvirta, and Jilles Vreeken. 2016. Is Exploratory Search Different? A Comparison of Information Search Behavior for Exploratory and Lookup Tasks. *J. Assoc. Inf. Sci. Technol.* 67, 11 (Nov. 2016), 2635–2651.

[3] Anne Aula, Rehan M. Khan, and Zhiwei Guan. 2010. How Does Search Behavior Change As Search Becomes More Difficult?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 35–44.

[4] Kathy Brennan, Diane Kelly, and Jaime Arguello. 2014. The Effect of Cognitive Abilities on Information Search for Tasks of Varying Levels of Complexity. In *Proceedings of the 5th Information Interaction in Context Symposium (IIiX '14)*. ACM, New York, NY, USA, 165–174.

[5] Alan Bryman and Duncan Cramer. 1994. *Quantitative Data Analysis for Social Scientists.* Routledge, New York, NY, 10001.

[6] Michael J. Cole, Chathra Hendahewa, Nicholas J. Belkin, and Chirag Shah. 2015. User Activity Patterns During Information Search. *ACM Trans. Inf. Syst.* 33, 1, Article 1 (March 2015), 39 pages.

[7] Anita Crescenzi, Diane Kelly, and Leif Azzopardi. 2016. 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)*. ACM, New York, NY, USA, 141–150.

[8] 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.

[9] Emi Ishita, Yosuke Miyata, Shuichi Ueda, and Keiko Kurata. 2017. A Structural Equation Model of Information Retrieval Skills. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17)*. ACM, New York, NY, USA, 317–320.

[10] 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.

[11] Markus Kattenbeck and David Elsweiler. 2018. 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)*. ACM, New York, NY, USA, 375–377.

[12] Jayden Khakurel, Antti Knutas, Mika Immonen, and Jari Porras. 2017. 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)*. ACM, New York, NY, USA, 97–100.

[13] R.B. Kline. 2011. *Principles and Practice of Structural Equation Modeling.* Guilford Publications.

[14] Michael Kotzyba, Tatiana Gossen, Johannes Schwerdt, and Andreas Nürnberger. 2017. 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)*. ACM, New York, NY, USA, 87–96.

[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] 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.

[18] 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

[19] Jingjing Liu and Chang Suk Kim. 2013. 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)*. ACM, New York, NY, USA, Article 5, 10 pages.

[20] Gary Marchionini. 1989. Information-seeking Strategies of Novices Using a Full-text Electronic Encyclopedia. *J. Am. Soc. Inf. Sci.* 40, 1 (Jan. 1989), 54–66.

[21] Gary Marchionini. 1995. *Information Seeking in Electronic Environments.* Cambridge University Press, New York, NY, USA.

[22] 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.

[23] Matthew Mitsui, Jiqun Liu, Nicholas J. Belkin, and Chirag Shah. 2017. 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)*. ACM, New York, NY, USA, 1121–1124.

[24] 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.

[25] Matthew Mitsui, Chirag Shah, and Nicholas J. Belkin. 2016. 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)*. ACM, New York, NY, USA, 841–844.

[26] Eun Youp Rha, Matthew Mitsui, Nicholas J. Belkin, and Chirag Shah. 2016. Exploring the Relationships Between Search Intentions and Query Reformulations. 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 48, 9 pages. http://dl.acm.org/citation.cfm?id=3017447.3017495

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

[28] Hong (Iris) Xie. 2002. Patterns Between Interactive Intentions and Information-seeking Strategies. *Inf. Process. Manage.* 38, 1 (Jan. 2002), 55–77.

[29] Yinglong Zhang, Jin Zhang, Matthew Lease, and Jacek Gwizdka. 2014. Multidimensional Relevance Modeling via Psychometrics and Crowdsourcing. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)*. ACM, New York, NY, USA, 435–444.