

## RESEARCH ARTICLE

# A Probabilistic Model Toward How People Search to Build Outcomes

YULI ZHAO<sup>1</sup>, YUYANG BAI<sup>1</sup>, YIN ZHANG<sup>1</sup>, BIN ZHANG<sup>1</sup>, AND PERTTI VAKKARI<sup>2</sup><sup>1</sup>Software College, Northeastern University, Shenyang, Liaoning 110000, China<sup>2</sup>Faculty of Information Technology and Communication Sciences, Tampere University, 33014 Tampere, Finland

Corresponding authors: Yuyang Bai (2210534@stu.neu.edu.cn) and Yin Zhang (zhangyin@mail.neu.edu.cn)

This work was supported in part by the Key Project of National Natural Science Foundation of China under Grant U1908212, in part by the Central Government-Guided Local Science and Technology Development Fund Project under Grant 1653137155953, in part by the Liaoning Province "Takes the Lead" Science and Technology Research Project under Grant 2021jh1/10400006, and in part by the Liaoning Provincial Natural Science Foundation of China under Grant 2022-MS-124.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Review Committee of the Intelligent Information Service Lab, Software College, Northeastern University, Shenyang, China.

**ABSTRACT** Although increased attention is being given to understanding how people search to build task outcomes, a formal model of the relation between how people search and how people build task outcomes is still lacking. This paper proposes a unified probabilistic model of how people search to build outcomes. The model involves 3 types of searcher behaviors (i.e., query submission, document selection, and information transformation) to model the effect of the information collected during search, and uses the item response theory to capture the ternary relations between the ability to transform information, the information collected, and the probability of successfully building task outcomes. We evaluate the proposed model in the task of identifying searchers' proficiencies under the assumption that high proficiency searchers would have high abilities to transform information. The results obtained high accuracies and F1 scores, which could reflect the effectiveness of the proposed model. The model contributes to the formal understanding of how people search to build task outcomes, and provides new possibilities for personalized and session-based information retrieval research.

**INDEX TERMS** Item response theory, searching as learning, outcomes.

## I. INTRODUCTION

People are confronted with a large number of tasks in their daily work. To complete these tasks, people need to satisfy the information requirements imposed by these tasks [1]. Since one's knowledge about a task may not be enough to meet the information requirements, retrieving additional information is a commonly used way to fill this knowledge gap. As search engines have become one of the most popular information retrieval systems, more and more people use search engines to collect information needed to complete tasks.

Today, search engines have been optimized to return information that meets certain information requirements.

The associate editor coordinating the review of this manuscript and approving it for publication was Kah Phooi Seng<sup>id</sup>.

Nevertheless, accessing the information required to complete tasks does not necessarily imply completing tasks. In most cases, people need to transform the collected information into task outcomes further to complete tasks. This fact has pushed search engine researchers to switch their focus from understanding how people search to understanding how they search to build task outcomes [2].

Recent years have seen increased attention being given to understanding how people search to build task outcomes. For example, in the domain of aerospace, how people search to build task outcomes has been used to guide the design of spacecraft systems [3]. In the domain of medical science, how people search to build task outcomes has been used to guide the fight against COVID-19 [4]. In the research domain of Searching as Learning [5], how people search to

build task outcomes has been deeply studied to understand how humans learn. As two examples, Kathryn et al. studied how engineering students accessed, used, and understood information when solving engineering problems [6], while Chi studied how well health consumers can search and learn in online health information seeking [7].

Nonetheless, the topic of how people search to build task outcomes is still under research and discussion. Especially, the relation between how people search and how people build task outcomes is still unclear [8]. Some studies try to understand the relation from the perspective of searching proficiency, but they failed to reach consistent conclusions. Hersh found that the proficiency has no effect on task outcome building [9]. Wildemuth found that the proficiency has negative effect [10] and Vakarri found that the proficiency has positive effect [2]. Some studies understand the relation from the perspective of efforts in search process, but their findings are also highly varied. Vakarri & Huuskonen found that efforts in search process improve task outcomes [2], while Bron found that efforts in search process have no effect on task outcomes [11]. Collins Thompson suggested that the main reason for the highly varied findings might be the lack of a more appropriate perspective for understanding the relation between how people search and how people build task outcomes [12]. The highly varied results may also imply that the relation between how people search and how people build task outcomes may be quite complex and the understanding of such complex relation is still lacking.

In [13], Hill suggested understanding the relation between how people search and how people build task outcomes from the perspective of the ability to transform information. Hill showed that searchers with a low ability to transform information might fail to complete tasks even when they have found related information. Similar results have also been observed in [14] and [15]. In [16], JR Hill and MJ.Hannafin further designed an experiment to explore the relation between searchers' ability to transform information and processes of searching to build task outcomes. In their experiment, they invited some students to build some task outcomes and provided them with personalized assistance based on their level of ability to transform information into outcomes. Experimental results showed that providing assistance that matches their level increases their probability of successfully building task outcomes. This finding reveals the possibility that if a searcher's ability to transform the information can be measured, personalized assistance in building task outcomes could be given. Thus, it is necessary to build a measurable model for helping building task outcomes. However, although the importance of the ability to transform information has been emphasized, a measurable model of how such ability affects the process of searching to build task outcomes is still lacking.

The purpose of this paper is to propose a probabilistic approach to formally model how the ability to transform information affects the process of searching to build task outcomes. Our key idea is to consider the ability to transform

information as a latent variable that parameterizes the probability of successfully building task outcomes. Meanwhile, since the transforming of information happens only based on collecting related information, the collected information also affects the probability of successfully building task outcomes. Such facts push us to design a unified probabilistic model to reflect the ternary relation between the ability to transform information, the information collected, and the probability of successfully building task outcomes.

Our main contributions are stated as follows:

- A probabilistic model of how people search to build task outcomes is proposed from the perspective of the ability to transform information. The model parameterized the probability of successfully building task outcomes according to the ability to transform information and the information collected during search.
- The item response theory (IRT) is adopted as the theoretical framework to capture the ternary relation between the aforementioned three aspects.
- Three experiments were designed and conducted to evaluate the proposed model. In the first experiment, our proposed model achieved an accuracy of 95.48% and a F1 score of 0.9554 in the task of identifying searchers' proficiencies. In the second experiment, the result of our proposed model shows an accuracy of 95.52% and a F1 score of 0.9600. In the third experiment, our proposed model is compared to state-of-the-art methods for identifying searchers' proficiencies and our proposed model outperformed than other methods. The performance reflected the effectiveness of the proposed model in measuring searchers' ability to transform information and identifying searchers' proficiencies.

## II. RELATED WORKS

### A. SEARCHING TO BUILD TASK OUTCOMES

When carrying out complex tasks through an information retrieval (IR) system, searchers would hope the IR system can provide information that helps build task outcomes [2]. Conforming to such needs, Yen et al. developed an intelligent state machine to support searchers build task outcomes [17]. Garigliotti and Balog et al. also developed a query suggestion system to support searchers in building task outcomes [18]. These tools affect how searchers build task outcomes by influencing how they search. However, the relation between search processes and task outcomes remains unclear [8].

One of the reasons for the unclearness may be the lack of an appropriate model of the use of information. The use of information, which is considered the link between search processes and task outcomes [8], is usually modeled as the "text reuse" between documents that emerged in the search process and task outcomes. However, the use of information is not just text reuse (e.g., text duplication, reformulation, or partially rewrite) but refers to a more complex process that requires searchers to transform information into task outcomes through behaviors such as understanding and learning.

Some studies proposed to understand this complex process from the perspective of searchers' ability to transform information. For example, Hill found that naïve searchers with low ability to transform information fail to solve problems even when they found helpful information [13]. However, although the importance of searchers' ability to transform information has been emphasized, few studies tried to propose qualitative models to link the ability to transform information and task outcomes. In this paper, we proposed a quantitative model based on item response theory to fill this gap. Our quantitative model reveals the possibility to understand how people search to build task outcomes from the perspective of searchers' ability to transform the information into outcomes. Furthermore, through our quantitative model, searchers' ability to transform the information can be obtained quantitatively. Based on the quantitative ability, more personalized and more effective assistance can be provided to help searchers build task outcomes.

### B. SEARCHING AS LEARNING

Searching is widely agreed to be a cognition-related activity, and there have been lots of studies trying to understand searching from the perspective of cognition. Vakkari emphasized the concept of Searching as Learning in [5]. He found that searchers' cognitive structure will change in their search process, and this change allows searchers to solve problems that were difficult for searchers before the search. Since the change in cognitive structure is difficult to be observed directly, task outcomes that can reflect the change in cognitive structure have been concerned in research on Searching as Learning. For example, Ghosh et al. evaluate how different search behaviors affect learning [19].

Searching as Learning has attracted more attention in recent years as the focus of searching is shifting to task outcome. Many studies apply the perspective of Searching as Learning to solve problems in fields that are closely related to outcomes, such as engineering and medicine. The concept of Searching as Learning has been used to understand how engineering students search to solve engineering problems [6] and how patients search for disease information or medical advice [7].

Searching as Learning is also combined with artificial intelligence (AI) methods in recent years. Change in cognitive structure provides AI with a perspective for understanding searching activities. AI methods also allow Searching as Learning to accurately capture changes in cognitive structure. For example, Tibau et al. studied the application of knowledge graphs in Searching as Learning situations [20]. Tang et al. proved the possibility of applying reinforcement learning to Searching as Learning [21].

In this paper, we proposed a measurable model based on the theory of Searching as Learning. The model proposed by us describes the relation between searchers' ability to transform information into outcomes and their probability of successfully build task outcomes. Furthermore, we designed a classification experiment to verify the performance of

our model in estimating searcher's ability to transform the information into outcomes. The high accuracy achieved by our model in classification experiment proves that Searching as Learning is a suitable perspective of understanding how people search to build task outcomes.

### C. SEARCHING AND SEARCHERS' ABILITIES

As searching for information on the Internet has gradually become one of the main methods to solve problems, the ability to search for information attracts more and more attention. It is widely agreed that 1) searchers with high abilities are easier to solve problems than those with low abilities, and 2) when faced with the same problem, searchers with different abilities may form different information needs and search strategies.

Many studies have noticed that the differences in abilities would be reflected in search behaviors. Search behavior patterns of novice and experts were extracted in [22] and [23]. Both studies pointed out the differences between the behavior patterns of searchers with high and low abilities. Such differences could be a possible basis for identifying searchers' abilities.

Domain expertise is perhaps the most studied ability in the context of searching. White et al. found that domain expertise contributed a lot to searching information [14]. Frerejean et al. also emphasized the effect of domain knowledge on searchers' problem solving [15]. They suggested that searchers with more domain knowledge have a higher ability to solve problems through searching.

More and more studies realized that searching ability is a cognitive skill in recent years. They thus tried to understand the ability to search for information from a cognitive perspective. Brand-Gruwel et al. suggested that searching for information is a complex cognitive skill [24]. Kalyani et al. observed searchers of different cognitive levels. They found that cognitive levels are positively correlated with the ability to search for information [25]. These findings support the use of cognitive tools to understand searching.

Although transforming information into task outcomes is an essential part of searching to build task outcomes [2], little attention has been paid on searchers' ability to transform the information. We fill this gap in this paper by proposing a model to describe the relation between searcher's ability to transform information into and their probability of successfully building task outcomes.

### D. IRT

IRT is widely used in test construction and test evaluation [26], [27]. IRT refers to a set of mathematic models. These mathematic models describe the relation between a person's response to a test item and their latent abilities. Compared with the other widely used test theories such as the classical test theory (CTT), IRT has the following advantages:

- 1) IRT provides the possibility of comparing the latent abilities of individuals of different statistical populations when they are submitted to the same test items.

This means that test takers' statistical population can be ignored when test items are determined. Thus, IRT is suitable for situations where test takers' statistical population is unknown.

- 2) In IRT, the standard error of ability measurement is regarded as a function of examinees' ability [28]. This means the standard error of IRT models can be assessed easily, and the accuracy of the results of IRT models can be evaluated with less effort.
- 3) IRT has a loose constraint on the completeness of the data used. Suppose some responses to test items given by a test taker are missing. In that case, IRT provides a likelihood-based method to calibrate test takers' latent abilities based on available data. This advantage makes IRT suitable for situations where data is incomplete.

These advantages extend IRT to many domains, such as education, medicine [29], [30], and psychology [31]. The possibility of applying IRT in the domain of IR has also been confirmed in recent years. Leng et al. applied IRT models to estimate the changes in the information collected in search processes [32]. Yarandi et al. applied IRT models to estimate changes in knowledge structure in processes of learning from searching [33]. These studies support the use of IRT in the domain of IR.

In this paper, we applied IRT models to describe the process of a searcher searching to build task outcomes, which extends the applicability of IRT models. Furthermore, we improved the IRT model to make it applicable for situations where the difficulty of a test item changes dynamically. This improvement makes it possible to model the influence of information collected on searchers' probability of successfully building task outcomes.

### III. MODEL

Based on [13] and [34], when a searcher searches to build a task outcome to complete a task, the probability of successfully building the outcome is related to the searcher's ability to transform information and the information collected by the searcher. Moreover, as stated in [5], the ternary relations between the ability to transform information, the information collected, and the probability of successfully building task outcomes can be generalized to all search tasks requiring learning as the relations are derived from Kuhlthau's task-independent information search process model [35]. To propose a model that captures the ternary relations, we need to answer the following two research questions:

Research Question 1 (RQ1): How to relate searcher's ability to transform the information collected and the probability of successfully building the outcome?

The relation between the two is rarely studied in the research domain of information retrieval. However, how a certain ability influences test takers' performance in an examination has been extensively studied. Some models have been proposed to describe this influence in the domain of test

theory, such as item response theory models [29] and classical test theory models [26]. These models reveal the possibility and provide theoretical bases to answer RQ1.

Research Question 2 (RQ2): How to relate the information collected in a search process and the probability of successfully building an outcome?

Exploratory search theories proposed by Marchionini have already provided a perspective to answer RQ2. In exploratory search theories, different information is thought to contribute differently to outcomes building [34]. By considering how certain information contributes to the successful building of a task outcome [34], we could classify the information collected in a search process into two categories: 1) information that can be transformed into the outcome or part of the outcome, and 2) information that cannot be transformed into (part of) the outcome but can lead to a correct search direction or can rule out wrong search directions. The contribution of the information collected in a search process could thus be used to capture the relation between information collected and task outcome.

We seek to answer the two research questions in the following subsections. The following sections provide a mathematical model to answer the research questions followed by the architecture and an intuitive justification of the model.

#### A. RELATING THE ABILITY TO TRANSFORM INFORMATION AND THE PROBABILITY OF SUCCESSFULLY BUILDING AN OUTCOME

In response to RQ1, we propose an IRT-based approach to express the relation between a searcher's ability to transform the information collected and the probability of successfully building an outcome. As one of the most popular modern test theories, IRT reveals the relation between a certain ability and the probability of solving a test item designed to measure the ability [26]. An IRT model is given in Equation (1).

$$P(Y_{ik} | \theta_i, b_k) = \frac{\exp(\theta_i - b_k)^{1-Y_{ik}}}{1 + \exp(\theta_i - b_k)} \quad (1)$$

where  $Y_{ik}$  is a Boolean variable and the value of  $Y_{ik}$  indicates the correctness of the answer to test item  $k$  given by test taker  $i$ .  $\theta_i$  indicates the ability measured by the test of searcher  $i$ .  $b_k$  represents the difficulty of test item  $k$ . Difficulty is a concept used by IRT. For a test item, the higher the difficulty means the lower the probability of being answered correctly. In particular, when a test taker's ability is equal to the difficulty of the test item, the probability of the test taker correctly answering the test item is 50%.

Equation (1) provides theoretical possibilities to relate the ability to transform information and the probability of successfully building an outcome. To apply IRT models to the process of searching to build task outcomes, we propose a possible solution. The solution is to consider the task of building an outcome based on collected information as a test item. We then consider whether the outcome is successfully built as the correctness of the test answer. In detail, we could

consider  $\theta_i$  as searcher  $i$ 's ability to transform information,  $b_k$  as the difficulty of building outcome  $k$  based on the information collected, and  $Y_{ik}$  as whether outcome  $k$  has been successfully build by searcher  $i$ . Then we could estimate  $\theta_i$  as:

$$\begin{aligned} P(\theta_i | Y_{ik}) &= \int \frac{\exp(\theta_i - b_k)^{1-Y_{ik}}}{1 + \exp(\theta_i - b_k)} * P(b_k) * \frac{P(\theta_i)}{P(Y_{ik})} db_k \\ &\propto \int \frac{\exp(\theta_i - b_k)^{1-Y_{ik}}}{1 + \exp(\theta_i - b_k)} * P(b_k) * P(\theta_i) db_k \end{aligned} \quad (2)$$

where the term  $P(\theta_i)$  represents the prior information about the ability of the searcher  $i$  to transform the information collected. The term  $P(b_k)$  represents the distribution of the difficulty of building outcome  $k$  based on information collected. The term  $P(Y_{ik})$  represents the probability of searcher  $i$  successfully building outcomes. In actual use,  $P(Y_{ik})$  term can be ignored since its value is fixed.

Estimating the ability to transform information using Equation (2) may seem feasible. Still, a critical difference between how a test taker answers a test item and how a searcher builds an outcome is hard to determine  $b_k$  in Equation (2). In the original IRT,  $b_k$ , although to be estimated, is a constant as the difficulty of test item  $k$  is considered as unchangeable [26]. Such an assumption holds only when test takers cannot freely collect information. Many studies on open-book tests have pointed out a fact that if test takers could collect information as they need, the difficulty of test item  $k$  will reduce [36]. In such a situation,  $b_k$  will become a variable and the original IRT (i.e., Equation (1)) will not hold.

Being different from answering test items, collecting information is essential for searching to build task outcomes. When searchers notice knowledge gaps in building task outcomes, they will search for additional information. Based on exploratory search theories [34], the information collected will reduce the difficulty of building task outcomes and thus increase the probability of successfully building task outcomes. Such an observation pushes us to improve the original IRT to consider the variable difficulty of building a task outcome as searchers collect new information. Meanwhile, such an observation also provides a possible way to resolve RQ2.

Our improvements to IRT consist of two parts: 1) modeling the effect of the information collected in a search process, and 2) modeling the reduction of difficulty caused by the information collected. We present the details in the following subsections.

## B. MODELING THE INFORMATION COLLECTED IN A SEARCH PROCESS

Since the information collected by a searcher cannot be observed directly, a commonly used way is to model the effect of the information collected as search behaviors. The search behaviors we consider in this paper are originated

from [37]. The first type of search behavior we consider is query submission:

**Query Submission (QS):** A searcher submits a query formulated based on their information needs to a search engine, and the search engine returns a “search engine result page” (SERP).

According to exploratory search theories [34], different information collected contributes to outcome in different ways. Based on [34], we could classify the information collected into two categories: 1) information that can be transformed into outcomes, and 2) information that cannot be transformed but can lead to a correct search direction or rule out wrong search directions. In order to model the different contributions, we further extended QS into two search behaviors:

**QS-Positive:** The SERP of the QS contains a link to information that can be transformed into outcomes.

**QS-Negative:** The SERP of the QS contains no link to information that can be transformed into outcomes.

We then consider how QS-Positive and QS-Negative affect the difficulty of building a task outcome. Since QS-Positive contains a link to information that can be transformed into outcomes, a searcher would have a chance to find the information. We thus consider QS-Positive would reduce the difficulty of building task outcomes.

While for QS-Negative, as it contains no link to information that can be transformed into outcomes, a searcher may still find information that can lead to a correct search direction or at least rules out the current search direction. We thus consider QS-Negative could slightly reduce the difficulty of building task outcomes. However, the scale of the difficulty reduced by QS-Negative would be significantly lower than that of QS-Positive.

The second type of search behavior we consider is document selection:

**Document Selection (DS):** A searcher selects (clicks) a document from a result page.

According to whether the document of DS contains information that can be transformed, we classify DS into DS-Positive and DS-Negative. DS-Positive/Negative also contributes to reducing difficulty as QS-Positive/Negative does.

The last behavior we consider is information transformation:

**Information transformation (IT):** A searcher transforms the information collected into task outcomes.

In our model, IT is the indicator of whether a task is completed. According to the completion status of a task, IT will be classified into IT-Positive/Negative. IT-Positive/Negative contributes to determine the value of  $Y_{ik}$  in Equation (2). When a searcher  $i$  transforms the information collected into a task outcome  $k$ , this transformation will be modeled as an IT. If this IT is classified into IT-Positive, the value of  $Y_{ik}$  in Equation (2) will be determined to be 1. Otherwise, if this IT is classified as IT-Negative, the value of  $Y_{ik}$  will be determined to be 0.

**C. MODELING THE REDUCTION OF DIFFICULTY CAUSED BY THE INFORMATION COLLECTED**

We introduced four variables  $\Delta b_{QS-Pos}$ ,  $\Delta b_{QS-Neg}$ ,  $\Delta b_{DS-Pos}$ , and  $\Delta b_{DS-Neg}$  to represent the reduction of difficulty caused by the information collected. All these variables in our model are estimated to follow truncated normal distributions and the range of the truncation is from 0 to  $+\infty$ . This estimation is based on the fact that whether the information collected can be transformed into outcomes is the main factor affecting the scale of difficulty reduced but not the only factor. Factors such as the form of the information collected [38], the amount of the information collected [2], and the obscurity of the information collected [39] will also slightly affect the scale of difficulty reduced. Based on [27], it is a feasible approach to model the reduction on difficulty caused by a main factor combined with many other factors as a variable that follows normal distribution. Take  $\Delta b_{QS-Pos}$  as an example:

$$\Delta b_{QS-Pos} \sim N(u_{QS-Pos}, \sigma_{QS-Pos}^2) I(0, \infty),$$

where  $u_{QS-Pos}$  and  $\sigma_{QS-Pos}^2$  are hyperparameters. Based on the hierarchical IRT [27], the hyperparameters in IRT models are considered to follow normal distributions:

$$u_{QS-Pos} \sim U(a_{QS-Pos}, c_{QS-Pos}),$$

$$\sigma_{QS-Pos}^2 \sim U(d_{QS-Pos}, f_{QS-Pos}).$$

We thus propose our improved IRT model as follows:

$$P(\theta_i | y_k) = \frac{\int P(y_k | \theta_i, b_{f(k)}, g(i)) * P(\Delta b_i) * P(\theta) * P(b_{f(k)}) db_{f(k)}}{P(y_k)},$$

$$J(x) = \begin{cases} QS - Pos & x = 1 \\ QS - Neg & x = 2 \\ DS - Pos & x = 3 \\ DS - Neg & x = 4 \end{cases},$$

$$g(i) = \sum_{p=1}^4 \sum_{n=1}^{n_{J(p)_i}} \frac{1}{n} * \Delta b_{J(p)_i n}$$

$$P(y_k | \theta_i, b_{f(k)}, g(i)) = \frac{\exp(b_{f(k)} + g(i) - \theta_i)^{1-y_k}}{1 + \exp(b_{f(k)} + g(i) - \theta_i)},$$

$$P(\Delta b_i) = \prod_{p=1}^4 \iint P(u_{J(p)}) * P(\sigma_{J(p)}^2) * \Theta,$$

$$\Theta = \prod_{n=1}^{n_{J(p)_i}} P(\Delta b_{J(p)_i n} | u_{J(p)}, \sigma_{J(p)}^2) du_{J(p)} d\sigma_{J(p)}^2,$$

where  $f(k)$  represents the outcome built through attempt  $k$ .  $y_k$  represents the result of the  $k$ th attempt.  $\theta_i$  represents the ability to transform the information of searcher  $i$ .  $b_i$  represents the difficulty of building the outcome of task  $i$ .  $n_{k_i}$  means the total amount of behavior  $k$  between attempt  $i$  and attempt  $i - 1$ .  $\Delta b_{k_i n}$  represents the reduction on  $b_i$  caused by the  $n$ th behavior  $k$  between attempt  $i$  and attempt  $i - 1$ .

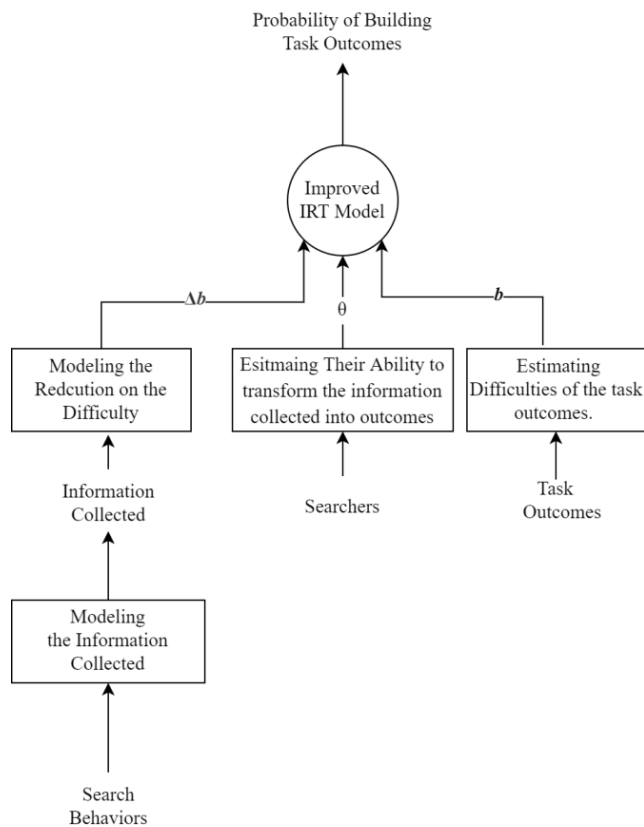


FIGURE 1. The overall structure of our proposed model.

A searcher usually makes multiple attempts in the process of building an outcome. The result of  $n$  attempts by a searcher is recorded as  $\vec{y}$ .  $\theta_p$  represents the estimation of the searcher's ability after the  $p$ th ( $p < n$ ) attempt. The expression of the searcher's ability after  $n$  attempts is given as follows:

$$P(\theta_n | \vec{y}) = P(\theta) * \prod_{i=1}^n \frac{p(y_i | \theta_i)}{P(y_i)}$$

**D. ARCHITECTURE**

After introducing our proposed model from a mathematical perspective, we summarize the architecture of our proposed model in this section. Figure 1 shows the overall structure of our proposed model.

Our proposed model is divided into four parts.

In the first part, as one of the inputs to our model, searchers' search behaviors will be used to model the information collected. Since the information collected will help searchers build task outcomes, the information collected will be further used to model the reduction on the difficulty of task outcomes. The reduction on the difficulty will be represented by the variable  $\Delta b$  in our proposed model.

In the second part, searchers' ability to transform the information  $\theta$  will be estimated. Based on the observation by Hill [13], searchers with high ability to transform the information collected into outcomes will have a higher probability of building task outcomes. This observation

implies the necessity of searchers' ability  $\theta$  in estimating the probability of building task outcomes.

In the third part, difficulties of the task outcomes  $b$  will be estimated by our proposed method. Traditional IRT models believe that outcomes with higher difficulty have a lower probability of being built. Thus, for estimating the probability of building task outcomes, it is necessary to estimate the difficulty  $b$  as one of the inputs to our model.

In the last part, the outputs of the above three parts:  $\Delta_b$ ,  $\theta$  and  $b$  will be used as inputs to our proposed model in Section III-C. Through this model, the probability of a searcher building task outcomes will be estimated as the output of our model.

### E. INTUITIVE JUSTIFICATIONS

The idea behind IRT is that the ability of a testee could be determined according to the answers to the test items made by the testee. The essential parts of IRT are as follows:

- Ep.t1) A testee who has a certain ability.
- Ep.t2) A test that is designed to evaluate the ability of a testee.
- Ep.t3) Test items that constitute the test. Each test item has a difficulty.
- Ep.t4) Answers made by the testee to the test items.

The relations between the ability and the difficulty of a test item are that:

- Ti.1) If the difficulty is high, the testee has a low probability of solving the item.
  - a) If the testee solves a high difficulty item, the ability of the testee may be high.
  - b) If the testee fails to solve a high difficulty item, the ability of the testee may not be high. However, we cannot say that the ability of the testee is low.
- Ti.2) If the difficulty is low, the testee has a high probability of solving the item.
  - a) If the testee solves a low difficulty item, the ability of the testee may be at least low. However, we cannot say that the ability of the testee is high.
  - b) If the testee fails to solve a low difficulty item, the ability of the testee may be low.

Our proposed model uses IRT to assess the ability of a searcher to transform information by observing how the searcher search to build outcomes. The essential parts of our proposed model are as follows:

- Ep.o1) A searcher who has the ability to transform information.
- Ep.o2) A task that is designed to evaluate the ability of the searcher. The task requires building an outcome.
- Ep.o3) Subtasks that constitute the task. Each subtask has a difficulty and requires building a suboutcome.
- Ep.o4) Suboutcomes made by the testee to the subtasks.

The relations between the ability to transform information and the difficulty of a subtask are that:

- St.1) If the difficulty is high, the searcher has a low probability of building the suboutcome:

- a) If the searcher builds a suboutcome with high difficulty, the ability of the searcher may be high.
  - b) If the searcher fails to build a suboutcome with high difficulty, the ability of the searcher may not be high. However, we cannot say that the ability of the searcher is low.
- St.2) If the difficulty is low, the searcher has a high probability of building the suboutcome:
    - a) If the searcher builds a low difficulty suboutcome, the ability of the searcher may be at least low. However, we cannot say that the ability of the searcher is high.
    - b) If the searcher fails to build a low difficulty suboutcome, the ability of the searcher may be low.

We use  $R_{\theta \& b}$  to refer to the relations between the ability to transform information and the difficulty of a subtask.

According to [30], we argue that the information collected during the search process of a subtask would affect the difficulty of the subtask:

If the searcher access information that can be transformed into the suboutcome, the difficulty is reduced.

If the searcher access information that cannot be transformed into the suboutcome, the information may still lead to a correct search direction or rule out wrong search directions. Then overall speaking, the information still slightly reduce the difficulty of the subtask.

We use  $\Delta_b$  to refer to how the information collected affects the difficulty. We use some examples to explain the joint effects of  $R_{\theta \& b}$  and  $\Delta_b$ . Suppose a searcher is working on a moderate difficulty subtask:

- Je.1) Suppose the searcher accesses related information (i.e., information that can be transformed into the suboutcome) and succeeds in building the suboutcome. We then estimate the ability of the searcher according to the reduced difficulty.
- Je.2) Suppose the searcher accesses related information and fails to build the suboutcome. Since the difficulty is already reduced and yet the searcher still fails, the ability of the searcher may be low.
- Je.3) Suppose the searcher access unrelated information (i.e., information that cannot be transformed into the suboutcome) and fails to build the suboutcome. Since the difficulty is only slightly reduced, we cannot say that the ability of the searcher is low.

As discussed in Section III-B, since the information collected during search cannot be observed, we use search behaviors to model the effect of the information collected. For query submission behavior, since a query corresponds to a SERP, and it is easy to tell whether a SERP contains related information, we use SERP to model the effect of the information collected during query submission. We also use the document selected to model the effect of the information collected during document selection. Combining the search behaviors,  $R_{\theta \& b}$ , and  $\Delta_b$ , we could come to more complex examples. Suppose a searcher is working on a high difficulty subtask:

If the searcher submitted an unrelated keyword or selected an unrelated document, the searcher may fail to build the suboutcome at this moment. As a result, the searcher may thus rule out an unrelated search direction. The difficulty is thus slightly reduced. However, if the searcher continues to make lots of unrelated tries, the difficulty will be reduced significantly. Then if the searcher fails to build the suboutcome at the end, as the difficulty has been largely reduced, the ability of the searcher may be low.

Suppose the searcher makes many unrelated tries and finally manages to build the suboutcome. In that case, since the difficulty has been largely reduced, we cannot say that the ability of the searcher is high.

#### IV. ESTIMATING PARAMETERS

In this section, we introduce methods to estimate the parameters and hyperparameters of our proposed model.

##### A. ESTIMATING THE DIFFICULTY OF BUILDING AN OUTCOME

In traditional IRT models, the commonly used method of estimating the difficulty of a test item consists of two steps: 1) determining a prior distribution of the difficulty based on previous experience, and 2) obtaining the posterior estimation of the difficulty. However, such a commonly used method is not applicable for our improved model. The reason is bifold:

- 1) Prior distributions significantly impact the effectiveness of posterior distributions. For traditional IRT models, strong experience on test items is required to ensure the correctness of prior distributions of difficulty. While for our proposed model, it would be hard to determine proper prior distributions for the difficulty of building task outcomes as the studies on related topics are quite insufficient.
- 2) The posterior estimation requires massive data. In traditional IRT models, answers to test items can be collected at a low cost. The low cost ensures the large amount of data required for posterior estimation. Meanwhile, the cost of collecting data of people searching to build outcomes is much higher. The high cost would limit the amount of data available for parameter estimation.

In response to the problems above, we propose a heuristic method to estimate the parameters of our proposed model. The heuristic method generates massive candidate data for practical parameter estimation. Compared with the commonly used method, our method is more suitable for situations where background experience is lacking, and the amount of data is limited. Moreover, the method is suitable for estimating both difficulties and hyperparameters. We present the details in the following subsections.

##### B. HEURISTIC ESTIMATION OF PARAMETERS

Our proposed estimation method starts from some heuristic rules on the value of parameters. For difficulties of building task outcomes, we apply the following rules:

- Dr1) The difficulties of building all the outcomes fall within an interval  $[b_{min}, b_{max}]$ , where  $b_{min}$  is the 0.25 quantile of the prior distribution of searchers' ability to transform information divided by 2 and  $b_{max}$  is the 0.75 quantile divided by 2. The estimation of  $b_{min}$  and  $b_{max}$  is based on the method of estimating the difficulty in traditional IRT models. To ensure that testees with different abilities can be distinguished through the test, difficulties of test items in the test usually fall within a range [27]. The lower bound of this range is the 0.25 quantile of the prior distribution of testees' ability divided by 2 and the upper bound is the 0.75 quantile of the prior distribution of testees' ability divided by 2.
- Dr2) The difficulties of all the task outcomes are sorted in descending order of their completion rates. The completion rate of task outcome  $i$  is defined as  $N_{c_i}/N_{a_i}$ , where  $N_{c_i}$  represents the number of searchers who successfully built the outcome and  $N_{a_i}$  represents the total number of searchers who tried to build the outcome.
- Dr3) Insert the median of the interval  $(b_{min} + b_{max})/2$  in the sorted difficulties. If the completion rate of an outcome is higher than 50%, the difficulty of building it is lower than  $(b_{min} + b_{max})/2$ . Conversely, if the completion rate of an outcome is lower than 50%, the difficulty of building it is higher than  $(b_{min} + b_{max})/2$ .

For hyperparameters, we apply the following rules:

- Hr1) Since all the hyperparameters in our model follow uniform distributions, it is necessary to estimate each parameter in the uniform distribution that each hyperparameter follows. We divide the hyperparameters of our model into two groups: 1) Means, including  $u_{QS-Pos}$ ,  $u_{QS-Neg}$ ,  $u_{DS-Pos}$ , and  $u_{DS-Neg}$ ; 2) Variances, including  $\sigma_{QS-Pos}^2$ ,  $\sigma_{QS-Neg}^2$ ,  $\sigma_{DS-Pos}^2$ , and  $\sigma_{DS-Neg}^2$ . We use PIDH to refer to the parameters in the uniform distribution that the hyperparameter follows.
- Hr2) The relation among PIDH within group Means is given by the following equations and inequalities:

$$\begin{aligned}
 &0 < \text{lowerbound of } u_{QS-Neg} \\
 &\text{lowerbound of } u_{QS-Neg} < \text{upperbound of } u_{QS-Neg} \\
 &\text{upperbound of } u_{QS-Neg} = \text{lowerbound of } u_{DS-Neg} \\
 &\text{lowerbound of } u_{DS-Neg} < \text{upperbound of } u_{DS-Neg} \\
 &\text{upperbound of } u_{DS-Neg} = \text{lowerbound of } u_{QS-Pos} \\
 &\text{lowerbound of } u_{QS-Pos} < \text{upperbound of } u_{QS-Pos} \\
 &\text{upperbound of } u_{QS-Pos} = \text{lowerbound of } u_{DS-Pos} \\
 &\text{lowerbound of } u_{DS-Pos} < \text{upperbound of } u_{DS-Pos} \\
 &\text{upperbound of } u_{DS-Pos} < b_{max} - b_{min}
 \end{aligned}$$

where  $b_{min}$  is the 0.25 quantile of the prior distribution of searchers' ability to transform information divided by 2 and  $b_{max}$  is the 0.75 quantile divided by 2. The estimation of the upper bound  $b_{max} - b_{min}$  is based on the observation that searchers who successfully

complete low difficulty outcomes fail to complete high difficulty outcomes even if they have collected a lot of information that can be transformed into outcomes [13]. This observation implies that the reduction on difficulty of building outcomes caused by information collected is limited, and it is almost impossible to reduce the highest difficulty to the lowest difficulty by collecting information. Since the highest difficulty of building an outcome is estimated as  $b_{max}$  and the lowest difficulty is estimated as  $b_{min}$ , the upper bound is estimated as  $b_{max} - b_{min}$ .

Hr3) The relation among PIDH within group Variance is given by the following equations and inequalities:

$$\begin{aligned}
 &0 < \text{lowerbound of } \sigma_{DS-Pos} \\
 &\text{lowerbound of } \sigma_{DS-Pos} < \text{upperbound of } \sigma_{DS-Pos} \\
 &\text{upperbound of } \sigma_{DS-Pos} = \text{lowerbound of } \sigma_{QS-Pos} \\
 &\text{lowerbound of } \sigma_{QS-Pos} < \text{upperbound of } \sigma_{QS-Pos} \\
 &\text{upperbound of } \sigma_{QS-Pos} = \text{lowerbound of } \sigma_{DS-Neg} \\
 &\text{lowerbound of } \sigma_{DS-Neg} < \text{upperbound of } \sigma_{DS-Neg} \\
 &\text{upperbound of } \sigma_{DS-Neg} = \text{lowerbound of } \sigma_{QS-Neg} \\
 &\text{lowerbound of } \sigma_{QS-Neg} < \text{upperbound of } \sigma_{QS-Neg} \\
 &\text{upperbound of } \sigma_{QS-Neg} < c
 \end{aligned}$$

where  $c$  is the variance of the prior distribution of the ability to transform information. The estimation of  $c$  is based on the observation by Hill [13] that the ability to transform information into outcomes is the main factor in building task outcomes. As the main factor, searchers' ability to transform the information into outcomes has a greater influence than other factors.

Following the above heuristic rules, the difficulty of building an outcome falls within a known interval. The value of each PIDH also falls within a known interval. We could then generate candidate data and estimate difficulties of outcomes or suboutcomes and hyperparameters. The steps of estimating these factors are as follows:

- S1) Generate  $n_b$  sets of vectors. The length of each vector is equal to the number of outcomes in the dataset. The value of the  $n$ th member in the vector is randomly selected in the possible interval of the difficulty of the  $n$ th outcome. The  $i$ th set is denoted as  $\vec{b}_i$ .
- S2) Generate  $n_{PIDH}$  sets of vectors. The length of each vector is equal to the number of PIDH in our model. The value of the  $n$ th member in the vector is randomly selected in the possible interval of the estimated value of the  $n$ th parameter. The  $i$ th set is denoted as  $\vec{PIDH}_i$ .
- S3) Denote  $n_s$  as the number of searchers on the dataset.
- S4) For  $i = 1$  to  $n_b$ :
  - a) For  $j = 1$  to  $n_s$ :
    - i) Create a group  $G_j$ . This group contains all the data in the dataset except the  $j$ th searcher's data.

ii) Define model M as:

$$\begin{aligned}
 &P(\theta_n | \vec{y}) \\
 &= P(\theta) * \prod_{i=1}^n \frac{P(y_i | \theta_i)}{P(y_i)} \\
 &P(\theta_i | y_k) \\
 &= \frac{\int P(y_k | \theta_i, b_{f(k)}) * P(\theta) * P(b_{f(k)}) db_{f(k)}}{P(y_k)}
 \end{aligned}$$

- iii) Take  $\vec{b}_i$  as  $\vec{b}$  in model M, then estimate the ability of  $n_s - 1$  searchers in  $G_j$  by using model M.
  - iv) Take the classification result given by domain experts as the standard. Classify the estimation results in the last step by using a Linear SVC (Support Vector Classifier) and the leave-one-out method. The classification accuracy of this classifier is recorded as  $ACC_j$ .
- b) Calculate the average classification accuracy of  $n_s$  groups:

$$\overline{ACC}_i = \frac{\sum_{i=1}^{n_s} ACC_i}{n_s}$$

- S5) For  $i = 1$  to  $n_b$ : Find the biggest ones of  $\overline{ACC}_1 \dots \overline{ACC}_{n_b}$ . If one of the biggest ones is  $\overline{ACC}_k$ , choose  $\vec{b}_k$  as one of the final estimates. Denote  $\vec{b}_k$  as  $\vec{b}_i^{chosen}$ , where  $i$  implies the number of  $\vec{b}$  that have been selected as the final estimate. Record the total number of  $\vec{b}$  chosen as final estimate as  $n_{chosen}$ .
- S6) For  $i = 1$  to  $n_{PIDH}$ :
  - a) Create a group  $G_j$ . This group contains all the data in the dataset except the  $j$ th searcher's data.
  - b) For  $k = 1$  to  $n_{chosen}$ :
    - i) Define model N as the equation can be derived, as shown at the bottom of next page;
    - ii) Take  $\vec{b}_i^{chosen}$  as  $\vec{b}$  in model N. Take  $\vec{PIDH}_i$  as the values of  $\vec{PIDH}$  in model N. Estimate the ability of  $n_s - 1$  searchers in  $G_j$  by using model N.
    - iii) Take the classification result given by domain experts as the standard. Classify the estimation results in the last step by using a Linear SVC and leave-one-out method. The classification accuracy of this classifier is recorded as  $ACC_{jk}$ .

### C. DETERMINING TERM $P(\Delta_{b_i})$

As a multidimensional integral, term  $P(\Delta_{b_i})$  is hard to be calculated directly. However, the known integrand of  $P(\Delta_{b_i})$  makes the use of Markov chain Monte Carlo sampling on the estimation of  $P(\Delta_{b_i})$  possible. The steps for estimating  $P(\Delta_{b_i})$  by MCMC sampling are as follows:

For  $i = 1$  to  $n$ :

- 1) Choose a Markov matrix  $Q(i, j)$ . The value of the element in the matrix at position  $(i, j)$  is the probability

of a normal distribution with a mean of  $i$  and a variance of 1 at position  $j$ :

$$Q(i, j) = \frac{1}{\sqrt{2\pi}} * \exp\left(-\frac{(j-i)^2}{2}\right).$$

2) Determine the stable distribution  $\pi(x)$ :

$$\pi(x) = p(u_{J(p)}) * p(\sigma_{J(p)}^2) * \prod_{n=1}^{n_{J(p)}} p(\Delta b_{J(p)_n} | u_{J(p)}, \sigma_{J(p)}^2) * P(x).$$

3) Determine the number of state transitions  $n_1$ . Determine the number of the samples required  $n_2$ .

4) Draw a sample  $x_0$  from any simple probability distribution.

5) For  $t = 0$  to  $n_1 + n_2 - 1$ :

a) Sample from the conditional probability distribution  $Q(x | x_t)$  to obtain a variable  $x_*$ :

$$Q(x | x_t) = \frac{1}{\sqrt{2\pi}} * \exp\left(-\frac{(x-x_t)^2}{2}\right)$$

b) Sample from a uniform distribution to obtain a variable  $u$ :

$$u \sim \text{uniform}[0, 1].$$

c) Determine the value of  $x_{t+1}$  according to the following expression:

$$\begin{cases} x_{t+1} = x_*, u < \min\left\{\frac{\pi(x_*) * Q(x_*, x_t)}{\pi(x_t) * Q(x_t, x_*)}, 1\right\} \\ x_{t+1} = x_t, u \geq \min\left\{\frac{\pi(x_*) * Q(x_*, x_t)}{\pi(x_t) * Q(x_t, x_*)}, 1\right\} \end{cases}$$

6) Take  $(x_{n_1}, x_{n_1+1}, \dots, x_{n_1+n_2-1})$  as a sample set from distribution  $\pi(x_t)$ . The members of this sample set can be considered to follow the distribution  $p(u_{J(p)}) * p(\sigma_{J(p)}^2) * \prod_{n=1}^{n_{J(p)}} p(\Delta b_{J(p)_n} | u_{J(p)}, \sigma_{J(p)}^2)$  when  $n_1$  is big enough. The estimation of  $p(\Delta b_i)$  is the average of the product of each member in the sample set and the integration area.

## V. EXPERIMENTAL EVALUATIONS

We evaluate the proposed model in the task of identifying searchers' proficiencies. The idea behind such an evaluation is that searchers with high proficiencies should have high abilities to transform information. Based on such an assumption, we estimate searchers' abilities to transform information and use the ability measures to identify searchers with high and low proficiencies.

The performance of our model was demonstrated on two datasets. The first dataset was built from scratch. We designed

a search task requiring learning and collected various search behaviors conducted by volunteers. The search task, the setups of our proposed model, and the experimental results are introduced in detail from Section V-A to Section V-F. The second dataset was from [40], and the results are presented in section V-G. In Section V-H, we present a comparative experiment. In this comparative experiment, our proposed model was compared with state-of-the-art methods on the task of identifying searchers' proficiencies.

### A. EXPERIMENTAL SETUPS

We recruited 55 undergraduate students to participate in the experiment. We have obtained informed consent from all the volunteers. Based on a reference check, 26 (47.27%) volunteers were considered to have high programming proficiencies, while 29 (52.72%) volunteers had low programming proficiencies.

We designed a programming task for our evaluation. The task consisted of 5 subtasks: 1) Click on a button and show a video capture interface for the user. The user will record a video. 2) Pop up a dialog for the user to choose a folder. 3) Save the recorded video to the chosen folder. 4) Click on another button and pop up a dialog for the user to choose the saved video file. 5) Play the video. Volunteers were asked to program the previous subtasks on the Universal Windows Platform (UWP). During programming, volunteers could search for any information they needed, but collaboration was prohibited. We built a controlled environment to log the querying, browsing, and programming behaviors (<https://github.com/zhangyinhub/thesallab.alllogger>).

After collecting these search behaviors, we tagged these search behaviors mentioned in Section III-B according to a set of key APIs for each of the 5 subtasks, as shown in Table 1.

After search behaviors were tagged, these tagged behaviors were used in three sets of experiments to evaluate our model and our methods. Section V-B presents the setups to evaluate the proposed difficulty estimation method. Section V-C presents the setups to evaluate the proposed PIDHs estimation method. Section V-D presents the setups to evaluate the performance of considering different combinations of search behaviors (i.e. QS, DS, and IT) in our proposed model. Section V-E presents the evaluations measures and Section V-F presents the results.

### B. SETUPS OF ESTIMATING DIFFICULTIES

Since the proposed method for estimating difficulty is based on heuristic rules, some parameters in these heuristic rules

$$P(\theta_n | \vec{y}) = P(\theta) * \prod_{i=1}^n \frac{p(y_i | \theta_i)}{P(y_i)}$$

$$P(\theta_i | y_k) = \frac{\int P(y_k | \theta_i, b_{f(k)}, g(i)) * P(\Delta b_i) * P(\theta) * P(b_{f(k)}) db_{f(k)}}{P(y_k)}$$

TABLE 1. Key apis of the subtasks.

Subtask	KeyAPIs
1	CameraCaptureUI, VideoSettings.Format, CameraCaptureUIVideoFormat.Mp4, CaptureFileAsync
2	FolderPicker, FileTypeFilter.Add, PickSingleFolderAsync
3	MoveAsync, DeleteAsync
4	FileOpenPicker, FileTypeFilter.Add, PickSingleFileAsync
5	MediaSource.CreateFromStorageFile, .Play

need to be determined first. We set the prior distribution of the ability to transform information as a normal distribution  $N(0, 2)$ . According to Section IV-B,  $b_{min} = -0.4769$ ,  $b_{max} = 0.4769$ . We set  $b_{min} = -0.5$ ,  $b_{max} = 0.5$  for convenient. By calculating the 55 volunteers' completion rates, we determined the order of the difficulties of the 5 subtasks. The order of difficulties is  $b_1 = b_4 < b_2 < (b_{min} + b_{max})/2 < b_5 < b_3$ . According to this order, we generated 30 possible vectors of  $\vec{b} = \langle b_1, b_2, b_3, b_4, b_5 \rangle$  according to the heuristic rules. These vectors of  $\vec{b} = \langle b_1, b_2, b_3, b_4, b_5 \rangle$  would be trained by the training method proposed in Section IV-B. The vector performed best among 30 vectors, the vector performed worst among 30 vectors, and all these 30 vectors would be considered as setups in the experiment.

For  $\vec{b}$ , the 5 setups in the experiments are as follows:

- 1)  $\vec{b} = \langle 0, 0, 0, 0, 0 \rangle$ . This was equal to estimating  $\theta$  according to IT-Positive. This setup was used as a baseline.
- 2)  $\vec{b} = \langle -0.2, 0.2, 0.3, 0, -0.2 \rangle$ , which was given by domain experts.
- 3)  $\vec{b} = \langle -0.1, 0, 0.2, -0.1, 0 \rangle$ , which was the one that performed the worst of the 30 generated vectors of  $\vec{b}$  in the training method proposed in Section IV-B.
- 4) The average results of the 30 generated vectors of  $\vec{b}$ .
- 5)  $\vec{b} = \langle -0.5, -0.2, 0.3, -0.5, 0 \rangle$ , which was the one that performed the best of the 30 generated vectors of  $\vec{b}$  in the training method proposed in Section IV-B.

These different setups are referred as  $\vec{b}_1$  to  $\vec{b}_5$ , and the results of these setups will be introduced in the following section.

### C. SETUPS OF ESTIMATING PIDHS

Since the proposed method for estimating PIDHs is based on the difficulty performed best in the training and the heuristic rules, the difficulty and some parameters in these heuristic rules needs to be determined first. The difficulty determined is  $\vec{b} = \langle -0.5, -0.2, 0.3, -0.5, 0 \rangle$  which is referred as  $\vec{b}_5$  in the previous experiments. Since the prior distribution of the ability to transform the information is set to  $N(0, 2)$ , the upper bounds of PIDHs were 0.9538 and were set to 1 for convenience. We then generated 30 possible vectors of PIDHs. These vectors would be trained through the method proposed in Section IV-B. The vector performed best among the 30 vectors, the vector performed worst among the

30 vectors and all these 30 vectors would be considered as setups in the experiment.

For PIDHs, the 4 setups considered in the experiment are as follows:

- 1) Ignoring PIDHs in our proposed model. This setup was used as a baseline.
- 2)  $\overrightarrow{PIDH} = \langle (u_{QS-Pos} = 0.22, \sigma_{QS-Pos}^2 = 1.73), (u_{QS-Neg} = 0.86, \sigma_{QS-Neg}^2 = 0.6), (u_{DS-Pos} = 0.39, \sigma_{DS-Pos}^2 = 1.59), (u_{DS-Neg} = 0.94, \sigma_{DS-Neg}^2 = 0.14) \rangle$ , which was the one that performed the worst of the 30 generated vectors of PIDHs in the training method proposed in Section IV-B.
- 3) The average results of the 30 generated vectors of PIDHs.
- 4) The average results of the three vectors that performed the best of the 30 generated vectors of PIDHs in the training method proposed in Section IV-B:
  - a)  $\overrightarrow{PIDH} = \langle (u_{QS-Pos} = 0.05, \sigma_{QS-Pos}^2 = 1.91), (u_{QS-Neg} = 0.88, \sigma_{QS-Neg}^2 = 0.82), (u_{DS-Pos} = 0.17, \sigma_{DS-Pos}^2 = 1.28), (u_{DS-Neg} = 0.9, \sigma_{DS-Neg}^2 = 0.79) \rangle$ ,
  - b)  $\overrightarrow{PIDH} = \langle (u_{QS-Pos} = 0.07, \sigma_{QS-Pos}^2 = 1.94), (u_{QS-Neg} = 0.52, \sigma_{QS-Neg}^2 = 0.64), (u_{DS-Pos} = 0.3, \sigma_{DS-Pos}^2 = 1.52), (u_{DS-Neg} = 0.56, \sigma_{DS-Neg}^2 = 0.61) \rangle$ ,
  - c)  $\overrightarrow{PIDH} = \langle (u_{QS-Pos} = 0.15, \sigma_{QS-Pos}^2 = 1.66), (u_{QS-Neg} = 0.91, \sigma_{QS-Neg}^2 = 0.5), (u_{DS-Pos} = 0.38, \sigma_{DS-Pos}^2 = 1.35), (u_{DS-Neg} = 0.99, \sigma_{DS-Neg}^2 = 0.07) \rangle$ .

These different setups are referred as  $\overrightarrow{PIDH}_1$  to  $\overrightarrow{PIDH}_4$ , and the results of these setups will be introduced in the following sections.

### D. SETUPS OF COMBINATIONS OF SEARCH BEHAVIORS

Our proposed model is a unified model of how searchers search and build outcomes. The model considers 3 types of search behaviors (i.e., QS, DS, and IT) to estimate the ability to transform information. To demonstrate the necessity of involving all the 3 types of search behaviors and the ternary relation in the proposed model, we considered the following

alternated approaches to estimate the ability to transform information:

- 1)  $P(\theta) = P(\theta_{IT} | Y_{ik}) = \frac{\exp(\theta_{IT})^{1-Y_{ik}}}{1+\exp(\theta_{IT})} * P(\theta_{IT})$ , namely the ability to transform information equals to the ability to generate IT-Positive.  $Y_{ik} = 1$  if the  $k$ th IT of searcher  $i$  is IT-Positive and  $Y_{ik} = 0$  if it is IT-Negative.  $P(\theta_{IT})$  is the prior distribution of  $\theta_{IT}$ . We assume  $\theta_{IT} \sim N(0, 2)$  in our experiment.
- 2)  $P(\theta) = P(\theta_{QS} | Y_{ik}) = \frac{\exp(\theta_{QS})^{1-Y_{ik}}}{1+\exp(\theta_{QS})} * P(\theta_{QS})$ , namely the ability to transform information equals to the ability to generate QS-Positive.  $Y_{ik} = 1$  if the  $k$ th QS of searcher  $i$  is QS-Positive and  $Y_{ik} = 0$  if it is QS-Negative.  $P(\theta_{QS})$  is the prior distribution of  $\theta_{QS}$ . We assume  $\theta_{QS} \sim N(0, 2)$  in our experiment.
- 3)  $P(\theta) = P(\theta_{DS} | Y_{ik}) = \frac{\exp(\theta_{DS})^{1-Y_{ik}}}{1+\exp(\theta_{DS})} * P(\theta_{DS})$ , namely the ability to transform information equals to the ability to generate DS-Positive.  $Y_{ik} = 1$  if the  $k$ th DS of searcher  $i$  is DS-Positive and  $Y_{ik} = 0$  if it is DS-Negative.  $P(\theta_{DS})$  is the prior distribution of  $\theta_{DS}$ . We assume  $\theta_{DS} \sim N(0, 2)$  in our experiment.
- 4)  $\theta = \langle \theta_{QS}, \theta_{IT} \rangle$ , namely the ability to transform information equals to the vector  $\langle \theta_{QS}, \theta_{IT} \rangle$ .
- 5)  $\theta = \langle \theta_{QS}, \theta_{DS} \rangle$ , namely the ability to transform information equals to the vector  $\langle \theta_{QS}, \theta_{DS} \rangle$ .
- 6)  $\theta = \langle \theta_{DS}, \theta_{IT} \rangle$ , namely the ability to transform information equals to the vector  $\langle \theta_{DS}, \theta_{IT} \rangle$ .
- 7)  $\theta = \langle \theta_{QS}, \theta_{DS}, \theta_{IT} \rangle$ , namely the ability to transform information equals to the vector  $\langle \theta_{QS}, \theta_{DS}, \theta_{IT} \rangle$ .

These alternated approaches are referred as  $\langle \theta_{IT} \rangle$ ,  $\langle \theta_{QS} \rangle$ ,  $\langle \theta_{DS} \rangle$ ,  $\langle \theta_{QS}, \theta_{IT} \rangle$ ,  $\langle \theta_{QS}, \theta_{DS} \rangle$ ,  $\langle \theta_{DS}, \theta_{IT} \rangle$ , and  $\langle \theta_{QS}, \theta_{DS}, \theta_{IT} \rangle$  accordingly.

## E. EVALUATION MEASURES

As the amount of data was limited in our experiment, we used leave-one-out cross-validation to evaluate the compared methods. For each fold of validation, the data of 54 volunteers were used to train the classifier, and the data of the last volunteer were used to validate the classification result. The evaluation was conducted 55 times. We refer to the high proficiency group as the positive class and the low proficiency group as the negative class. We then logged the following measures for all the 55 evaluations:

- 1) True Positive (TP): The number of predictions where the classifier correctly predicts the positive class as positive.
- 2) True Negative (TN): The number of predictions where the classifier incorrectly predicts the negative class as negative.
- 3) False Positive (FP): The number of predictions where the classifier incorrectly predicts the negative class as positive.
- 4) False Negative (FN): The number of predictions where the classifier incorrectly predicts the positive class as negative.

We then calculated the accuracy, precision, recall, and F1 measure as:

$$\begin{aligned} accuracy &= \frac{TP + TN}{TP + FN + FP + TN}, \\ precision &= \frac{TP}{TP + FP}, \\ recall &= \frac{TP}{TP + FN}, \\ F1 &= \frac{2 \times precision \times recall}{precision + recall}. \end{aligned}$$

It should be noticed that instead of 55 runs of cross-validation,  $\vec{b}_4$  and  $PIDH_3$  consisted of 1650 runs, while  $PIDH_4$  consisted of 165 runs. This was due to the results of  $\vec{b}_4$  and  $PIDH_3$  were the average of 55 generated candidates, and the results of  $PIDH_4$  were the average of the 3 best candidates.

## F. EXPERIMENTAL RESULTS ON THE FIRST DATASET

The results of the compared methods for estimating the difficulties on the first dataset are shown in Table 2, the results of the compared methods for estimating the difficulties on the first dataset are shown in Table 3 and the results of the compared methods for comparing the different combinations of search behaviors on the first dataset are shown in Table 4.

From the results of the different setups of  $\vec{b}$  in Table 2, we could observe that the accuracy of the main setup  $\vec{b}_5$  had reached 90.91%, which had a significant advantage than the accuracy that random classifiers could achieve ( $Z = -4.692, p < .001$ ).

Furthermore, the result in Table 2 showed that domain experts provided  $\vec{b}$  (i.e.,  $\vec{b}_2$ ) and the best generated  $\vec{b}$  (i.e.,  $\vec{b}_5$ ) outperformed all the other setups in accuracy. The same accuracy showed that the performance differences between  $\vec{b}_2$  and  $\vec{b}_5$  were not substantial. Considering the cost of obtaining the difficulties provided by domain experts, these results showed that the performance differences between the heuristic rules described in Section IV-B could be a practical approach to estimate the difficulties of tasks.

In addition, we could notice that the recall of the main setup  $\vec{b}_5$  was higher than the precision of  $\vec{b}_5$ . This result implied that the probability of estimating low-ability searchers as high-ability searchers is higher than the probability of estimating high-ability searchers as low-ability searchers under the main setup  $\vec{b}_5$ . The main reason for this result was that a small number of low-ability searchers have prior knowledge about task outcomes in the experiment. Although all the volunteers were restricted from obtaining information related to the task outcomes before the experiment, it was almost impossible to restrict all the information obtained by volunteers. According to Section III-B, all the information collected by searchers would reduce the difficulty of building task outcomes. Thus, some volunteers have a higher probability of building task outcomes and their estimated abilities are higher than their actual abilities. This overestimation leads to a higher recall of  $\vec{b}_5$ .

**TABLE 2.** Experimental results of the compared methods for estimating difficulties on the first dataset.

Setup / Method	Accuracy	Precision	Recall	F1
$\vec{b}_1$	85.45%	90.91%	76.92%	0.8330
$\vec{b}_2$	90.91%	92.00%	88.46%	0.9020
$\vec{b}_3$	87.27%	88.00%	84.62%	0.8651
$\vec{b}_4$	89.45%	92.51%	85.75%	0.8920
$\vec{b}_5$	<b>90.91%</b>	<b>88.00%</b>	<b>91.67%</b>	<b>0.8980</b>

**TABLE 3.** Experimental results of the compared methods for estimating pidhs on the first dataset.

Setup / Method	Accuracy	Precision	Recall	F1
$\overrightarrow{PIDH}_1$	90.91%	88.00%	91.67%	0.8980
$\overrightarrow{PIDH}_2$	90.91%	88.46%	92.00%	0.9020
$\overrightarrow{PIDH}_3$	91.37%	91.33%	90.43%	0.9088
$\overrightarrow{PIDH}_4$	<b>95.48%</b>	<b>93.75%</b>	<b>97.40%</b>	<b>0.9554</b>

While for the worst generated  $\vec{b}$  (i.e.,  $\vec{b}_3$ ), the performance under this setup was almost the same as the baseline (i.e.,  $\vec{b}_1$ ) ( $Z = -0.278, p < .1$ ). Such a result implied the necessity of choosing the best generated one from all the generated setups of  $\vec{b}$  in our proposed model.

Although the necessity of finding appropriate setups of  $\vec{b}$  was implied, high time complexity of the method proposed in Section IV-B makes choosing  $\vec{b}$  a tradeoff. The time complexity of the method for choosing the best  $\vec{b}$  proposed in Section IV-B is  $O(n^2)$ , where  $n$  refers to the number of generated groups. As  $n$  goes up, the cost of the proposed choosing method could become substantial. It is thus necessary to develop more efficient methods to reduce the time cost for applications on a larger scale.

Moreover, all the generated setups of  $\vec{b}$  performed better than the baseline  $\vec{b}_1$ . This result suggested that the difficulties of transforming information into different outcomes varied, and estimating difficulty was indispensable for modeling the relations between the ability to transform information, the information collected, and the probability of successfully building task outcomes.

From the results of different setups of  $\overrightarrow{PIDH}$  in Table 3, we could observe that the accuracy of the main setup  $\overrightarrow{PIDH}_4$  had reached 95.48%, which had a significant advantage than the accuracy that random classifiers could achieve ( $Z = -5.344, p < .001$ ).

Moreover, we could notice that the recall of the main setup  $\overrightarrow{PIDH}_4$  was higher than the precision of  $\overrightarrow{PIDH}_4$ . This result showed that the probability of estimating low-ability searchers as high-ability searchers is higher than the probability of estimating high-ability searchers as low-ability searchers under the main setup  $\overrightarrow{PIDH}_4$ . The main reason for this result was that few search behaviors were conducted by low-ability searchers before their failing to build task outcomes. Although collecting information is one of the most effective and common ways to reduce the difficulty of

building task outcomes, some low-ability searchers was even not able to describe their information needs clearly while building task outcomes. This makes it difficult for low-ability searchers to collect information through search behaviors such as submitting queries and selecting documents. In our model, few search behaviors implies that the difficulty of building a task outcome is slightly reduced and the searcher fails to build a task outcome with high difficulty. According to Section III-E, if a searcher fails to build a task outcome with high difficulty, we can't say that the ability of this search is low. Thus, the estimated abilities of some low-ability volunteers were higher than those given by domain experts, resulting in a higher recall.

Another observation from the results in Table 3 was that the average performance off all generated  $\overrightarrow{PIDH}$  (i.e.,  $\overrightarrow{PIDH}_3$ ) outperformed than the baseline  $\overrightarrow{PIDH}_1$ . This result implies that considering  $\overrightarrow{PIDH}$  did contribute to the improvements in performance.

Furthermore, compared with the accuracy reached by the baseline  $\overrightarrow{PIDH}_1$ , the accuracy reached by the main setup  $\overrightarrow{PIDH}_4$  had improved significantly ( $Z = -1.649, p < 0.05$ ). This result showed that choosing the best hyperparameters significantly contributed to the improvements in performance.

The time complexity of the heuristic method to choose a best  $\overrightarrow{PIDH}$  proposed in Section IV-B is also  $O(n^2)$ , where  $n$  refers to the number of generated groups. The time cost for large-scale applications could still be substantial. However, considering the significant improvements brought by generating and choosing  $\overrightarrow{PIDH}$ , the time consumption could be a valuable tradeoff even with the current estimation method.

From the results of different combinations of search behaviors in Table 4, we could notice that in terms of the abilities to generate positive QS, DS, and IT,  $\langle \theta_{QS} \rangle$  performed the worst,  $\langle \theta_{DS} \rangle$  performed moderate, while

**TABLE 4.** Experimental results of the compared methods for comparing different combinations of search behaviors on the first data set.

Setup / Method	Accuracy	Precision	Recall	F1
$\langle \theta_{IT} \rangle$	85.45%	90.91%	76.92%	0.8330
$\langle \theta_{QS} \rangle$	70.91%	76.19%	59.26%	0.6667
$\langle \theta_{DS} \rangle$	80.00%	83.33%	74.07%	0.7841
$\langle \theta_{QS}, \theta_{IT} \rangle$	85.45%	85.71%	85.71%	0.8571
$\langle \theta_{QS}, \theta_{DS} \rangle$	80.00%	83.33%	74.07%	0.7841
$\langle \theta_{DS}, \theta_{IT} \rangle$	85.45%	85.71%	85.71%	0.8571
$\langle \theta_{QS}, \theta_{DS}, \theta_{IT} \rangle$	85.45%	85.71%	85.71%	0.8571

**TABLE 5.** Experimental results of the compared methods for estimating difficulties on the second dataset.

Setup / Method	Accuracy	Precision	Recall	F1
$\vec{b}_1$	83.58%	90.63%	78.38%	0.8406
$\vec{b}_2$	89.55%	86.49%	94.12%	0.9014
$\vec{b}_3$	86.57%	88.89%	86.49%	0.8767
$\vec{b}_4$	88.06%	91.43%	86.49%	0.8889
$\vec{b}_5$	<b>91.04%</b>	<b>91.89%</b>	<b>91.89%</b>	<b>0.9189</b>

$\langle \theta_{IT} \rangle$  performed the best. The accuracy and F1 score of  $\langle \theta_{DS} \rangle$  was 80.00% and 0.7841, while for  $\langle \theta_{IT} \rangle$  they were 85.45% and 0.8330. These results coincided with the findings in related works saying that searchers with different domain expertise would have different search behaviors, and it was possible to identify searchers with high and low domain expertise by observing their search behaviors. Meanwhile, the cost of collecting QS and DS data is much lower than IT. Thus, a simple observation of DS may be a cost-effective estimation of proficiency.

Another observation from Table 4 was that precision of  $\langle \theta_{IT} \rangle$  was higher than the recall of  $\langle \theta_{IT} \rangle$ . Since searchers' ability to transform information into outcomes was considered to be equal to searchers' ability to generate IT-Pos, based on the definition of IT-Pos in Section III-B, this observation implied that it is inaccurate to estimate searchers' ability to transform information into outcomes without considering the information collected by searchers.

Furthermore, the precision of  $\langle \theta_{DS} \rangle$  is higher than the recall of  $\langle \theta_{DS} \rangle$ . Since searchers' ability to transform information into outcomes was considered to be equal to searchers' ability to generate DS-Pos, based on the definition of DS-Pos in Section III-B, this observation showed that it is inaccurate to estimate searchers' ability to transform information into outcomes as searchers' ability to collect useful information. Similarly, the precision of  $\langle \theta_{QS} \rangle$  was higher than the recall of  $\langle \theta_{QS} \rangle$ . Based on the definition of QS-Pos, this result also proved that estimating searchers' ability to transform information into outcomes as searchers' ability to collect useful information is inaccurate.

An interesting result that emerged in the results of the alternated approaches was that combining the abilities to generate positive QS, DS, and IT as vectors could not improve the performance. For example, the performance

of  $\langle \theta_{QS}, \theta_{DS} \rangle$  was the same as  $\langle \theta_{DS} \rangle$ . These results implied that the ternary relation between the ability to transform information, the information collected (with effect modeled as QS, DS, and IT), and the probability of successfully building task outcomes could not be modeled in a naïve approach.

### G. EXPERIMENTAL RESULTS ON THE SECOND DATASET

In [40], Li et al. built a search process dataset to verify the performance of their search task extraction method. The search task in this dataset consists of 5 subtasks. Searchers' behaviors and task outcomes were also collected. As a result, this dataset includes all the data needed for our model.

We involved two domain experts to evaluate the programming proficiencies of the 67 web learners involved in the dataset. 37 (55.22%) web learners were considered to have high programming proficiencies, while 30 (44.78%) web learners were considered to have low programming proficiencies. The performance of our proposed difficulty estimation method, the performance of our proposed PIDHs estimation method, and the performance of considering different combinations of search behaviors are evaluated in the same way as described in Section V-B, Section V-C and Section V-D. The results of the compared methods for estimating the difficulties on the second dataset are shown in Table 5. The results of the compared methods for estimating the difficulties on the second dataset are shown in Table 6. The results of the compared methods for comparing the different combinations of search behaviors on the second dataset are shown in Table 7.

From the results of the different setups of  $\vec{b}$  in Table 5, we could observe that the accuracy of the main setup  $\vec{b}_5$  had reached 91.04%, which had a significant advantage than

**TABLE 6.** Experimental results of the compared methods for estimating pidhs on the second dataset.

Setup / Method	Accuracy	Precision	Recall	F1
$\overrightarrow{PIDH}_1$	91.04%	91.89%	91.89%	0.9189
$\overrightarrow{PIDH}_2$	91.04%	89.74%	94.59%	0.9210
$\overrightarrow{PIDH}_3$	92.54%	92.11%	94.59%	0.9333
$\overrightarrow{PIDH}_4$	<b>95.52%</b>	<b>94.74%</b>	<b>97.30%</b>	<b>0.9600</b>

**TABLE 7.** Experimental results of the compared methods for comparing different combinations of search behaviors on the second data set.

Setup / Method	Accuracy	Precision	Recall	F1
$\langle \theta_{IT} \rangle$	83.58%	90.63%	78.38%	0.8406
$\langle \theta_{QS} \rangle$	65.67%	71.88%	62.16%	0.6667
$\langle \theta_{DS} \rangle$	77.61%	82.35%	75.68%	0.7887
$\langle \theta_{QS}, \theta_{IT} \rangle$	85.07%	84.62%	89.19%	0.8684
$\langle \theta_{QS}, \theta_{DS} \rangle$	79.10%	81.08%	81.08%	0.8108
$\langle \theta_{DS}, \theta_{IT} \rangle$	85.07%	88.57%	83.78%	0.8611
$\langle \theta_{QS}, \theta_{DS}, \theta_{IT} \rangle$	86.57%	88.89%	86.49%	0.8767

**TABLE 8.** Experimental results of the comparative experiment of identifying searchers' proficiencies.

Indicator	Classifier	Accuracy	Precision	Recall	F1
The quality of searchers' task outcomes	SVM	87.27%	86.21%	89.29%	0.8772
	Decision Tree	87.27%	86.21%	89.29%	0.8772
	Logistic Regression	87.27%	86.21%	89.29%	0.8772
The amount of searchers' query words	SVM	72.73%	71.43%	66.67%	0.6897
	Decision Tree	75.76%	76.92%	66.67%	0.7143
	Logistic Regression	69.70%	66.67%	66.67%	0.6667
The length of the sessions	SVM	78.79%	75.00%	80.00%	0.7742
	Decision Tree	72.73%	68.75%	73.33%	0.7087
	Logistic Regression	78.79%	75.00%	80.00%	0.7742
The quality of searchers' query words	SVM	48.48%	45.00%	45.00%	0.4500
	Decision Tree	60.61%	57.14%	53.33%	0.5517
	Logistic Regression	48.48%	45.00%	45.00%	0.4500
The relevance of searchers' search results	SVM	81.82%	76.47%	86.67%	0.8125
	Decision Tree	72.73%	71.43%	66.67%	0.6897
	Logistic Regression	78.79%	75.00%	80.00%	0.7742

the accuracy that random classifiers could achieve ( $Z = -5.192, p < .001$ ). Furthermore, the main setup  $\overrightarrow{b}_5$  also outperformed all the other setups in accuracy.

From the results of different setups of  $\overrightarrow{PIDH}$  in Table 6, we could observe that the accuracy of the main setup  $\overrightarrow{PIDH}_4$  had reached 95.52%, which had a significant advantage than the accuracy that random classifiers could achieve ( $Z = -5.867, p < .001$ ).

From the results of different combinations of search behaviors in Table 7, we could observe that the result is similar to the results shown in Table 4. The similarity validates our conclusions on the relation between the combination of search behaviors and the programming proficiency.

#### H. COMPARASIONS WITH OTHER APPROACHES

As described in the previous subsections, we consider searchers' ability measured by our proposed model as the indicator of their proficiencies. Experimental results showed that this approach achieved high classification accuracies and F1 scores in identifying searchers' proficiencies. The remarkable performance supported that our model can accurately measure the searchers' ability to transform information. Furthermore, since the ability to transform information is measured based on searchers' search behaviors, our proposed model provides a feasible perspective for understanding the relation between searchers' search behaviors and their proficiencies.

In recent years, many studies have attempted to identify searchers' proficiencies based on their search behaviors. Most of these studies focused on the quantity and quality of searchers' search behaviors. For example, in [41], Mao et al. explored the relation between searchers' proficiencies and the quality of their task outcomes. Their experimental results showed that proficient searchers built task outcomes with higher quality. Similar observations were also made in [42] by O'Brien et al. In addition to the quality of task outcome, the relation between searchers' proficiency and other search behaviors was also investigated in [43], [44], and [45]. The search behaviors investigated in these studies include the amount of query words, the length of sessions, the quality of query words, and searchers' judgement on the relevance of search results. However, the findings of these studies show that these search behaviors are weakly correlated or even uncorrelated with searchers' proficiencies. These weak correlations imply that searchers' proficiencies can hardly be identified accurately by the quantity and quality of most search behaviors.

Based on existing research, it shows that the quality of searchers' task outcomes is a state-of-the-art indicator of identifying searchers' proficiencies. To better understand the performance of identifying searchers' proficiencies with their ability to transform information, a comparative experiment on the first dataset was conducted. In the comparative experiment, searchers would be classified to have high proficiencies and low proficiencies under 15 different setups. Each setup consists of an indicator for identifying searchers' proficiencies and a classifier. All the five indicators used are listed as follows:

- 1) The quality of searchers' task outcomes [41] [42].
- 2) The amount of searchers' query words [43].
- 3) The length of search sessions [44].
- 4) The quality of searchers' query words [44].
- 5) The relevance of searchers' search results [45].

And the three classifiers used are listed below:

- 1) Linear support vector machines (SVM)
- 2) Decision trees
- 3) Logistic regression classifiers.

The accuracy, the precision, the recall, and the f1 score under each setting are shown in Table 8.

From the results shown in Table 8, we could observe that the accuracy of identifying searchers' proficiencies with the quality of searchers' task outcomes had reached 87.72%, which outperformed all the other setups. However, identifying searchers' proficiencies with the quality of searchers' task outcomes was significantly worse ( $Z = -3.197$ ,  $p < .001$ ) than the accuracy of identifying searchers' proficiencies with searchers' ability to transform information (shown in Table 3). The advantage in accuracy implies that the ability to transform information is a feasible indicator of searchers' proficiencies. Moreover, by quantitatively measuring searchers' ability to transform information, our proposed model could precisely identify searchers' proficiencies.

## VI. CONCLUSION AND FUTURE WORKS

This paper proposed a unified probabilistic model to reflect the ternary relation between the ability to transform information, the information collected, and the probability of successfully building task outcomes. Based on the IRT, the proposed model parameterized the probability of successfully building task outcomes by the ability to transform information and the effect of information collected modeled as search behaviors, including QS, DS, and IT. We designed heuristic methods to estimate the parameters and hyperparameters for our proposed. We evaluate the proposed model's performance in identifying searchers with high and low programming proficiency under different setups. Experimental results showed that the proposed model could properly estimate the ability to transform information. Our proposed model contributed to the formal understanding of how people search to build task outcomes.

The time complexity of the parameter and hyperparameter estimation methods proposed by us in Section IV-B was  $O(n^2)$ , which may lead to performance issues when applied in large-scale applications. In large-scale applications, the parameter and hyperparameter in our proposed model could be provided by domain experts. Based on the experimental results in Section V-F and Section V-G, our model achieves high accuracy and low time consumption when using parameters and hyper-parameters given by domain experts. Meanwhile, for future works, our parameter and hyperparameter estimation methods could possibly be improved by considering intelligent optimization approaches.

We choose the one parameter logistic (1PL) IRT model as the basis of our improved model. Our choice implies that we make two assumptions: 1) It is impossible for a searcher to transform the information into task outcome by guessing. 2) The discrimination between transforming information into different outcomes is the same. Although the high accuracy in our experiment shows that our assumptions hold in most cases, taking discrimination and searchers' probability of building an outcome by guessing into consideration may lead to a more accurate model.

In our proposed model, we only considered the ability to transform information and ignored other types of abilities, such as generating high-quality queries or selecting related documents. The experimental results showed that simple combinations of various abilities into vectors might be meaningless. We argue to study further why such combinations led to no improvements. Such studies may reveal possibilities to understand how different abilities are related and propose a more comprehensive understanding of how people search to build outcomes.

Due to the high cost of collecting and analyzing data on people searching to build outcomes, the scale of experiments in the domain of Searching as Learning is usually limited [46], [47]. This limitation affects the practical use of studies in the domain of Searching as Learning [48]. In order to facilitate future studies in the domain of Searching

as Learning, we have open-sourced our experimental system at <https://github.com/zhangyin-github/thesallab.allogger>. The system provides a possible solution to the collection of search behaviors including query formulation, document selection and information transformation. The system also enables the collection of task outcomes. We hope the system could advance Searching as Learning research and applications.

Our model may be used as a framework to seek new possible ways to improve session-based search engines, including providing a user with information that could be more likely transferred, given the user's ability to transfer information. The high accuracy of our model in identifying proficient searchers could also suggest new ways to personalize search engines. Our model also suggests a new perspective to understand search behavior as changes of difficulties. Such a perspective may be adopted in reinforcement learning based search methods to develop new dynamic search models.

## REFERENCES

- [1] K. Byström and K. Järvelin, "Task complexity affects information seeking and use," *Inf. Process. Manage.*, vol. 31, no. 2, pp. 191–213, Mar. 1995.
- [2] P. Vakkari and S. Huuskonen, "Search effort degrades search output but improves task outcome," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 63, no. 4, pp. 657–670, Jan. 2012.
- [3] Y. H. Liu, A. Arnold, and G. Dupont, "Evaluation of conversational agents for aerospace domain," Presented at the Joint Conf. Inf. Retr. Communities Eur. (CIRCLE), 2020.
- [4] W. Pian, J. Chi, and F. Ma, "The causes, impacts and countermeasures of COVID-19 'infodemic': A systematic review using narrative synthesis," *Inf. Process. Manage.*, vol. 58, no. 6, Nov. 2021, Art. no. 102713.
- [5] P. Vakkari, "Searching as learning: A systematization based on literature," *J. Inf. Sci.*, vol. 42, no. 1, pp. 7–18, Jan. 2016.
- [6] K. Mercer, K. Weaver, and A. Stables-Kennedy, "Understanding undergraduate engineering student information access and needs: Results from a scoping review," Presented at the ASEE Annu. Conf. Expo., Tampa, FL, USA, Jun. 2019.
- [7] Y. Chi, "Health consumers' knowledge learning in online health information seeking," Ph.D. dissertation, School Comput. Inf., Univ. Pittsburgh, Pittsburgh, PA, USA, 2021.
- [8] P. Vakkari, M. Völske, M. Potthast, M. Hagen, and B. Stein, "Predicting essay quality from search and writing behavior," *J. Assoc. Inf. Sci. Technol.*, vol. 72, no. 7, pp. 839–852, Jan. 2021.
- [9] W. Hersh, *Information Retrieval: A Health and Biomedical Perspective*. New York, NY, USA: Springer, 2008.
- [10] B. M. Wildemuth, R. de Blik, C. P. Friedman, and D. D. File, "Medical students' personal knowledge, searching proficiency, and database use in problem solving," *J. Amer. Soc. Inf. Sci.*, vol. 46, no. 8, pp. 590–607, Sep. 1995.
- [11] M. Bron, J. van Gorp, F. Nack, M. de Rijke, A. Vishneuski, and S. de Leeuw, "A subjunctive exploratory search interface to support media studies researchers," Presented at the 35th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., Aug. 2012.
- [12] K. Collins-Thompson, S. Y. Rieh, C. C. Haynes, and R. Syed, "Assessing learning outcomes in web search: A comparison of tasks and query strategies," Presented at the ACM Conf. Hum. Inf. Interact. Retr., New York, NY, USA, Mar. 2016.
- [13] J. R. Hill, "A conceptual framework for understanding information seeking in open-ended information systems," *Educ. Technol. Res. Develop.*, vol. 47, no. 1, pp. 5–27, Mar. 1999.
- [14] R. W. White, S. T. Dumais, and J. Teevan, "Characterizing the influence of domain expertise on web search behavior," Presented at the 2nd ACM Int. Conf. Web Search Data Mining, Barcelona, Spain, Feb. 2009.
- [15] J. Frerejean, J. L. H. van Strien, P. A. Kirschner, and S. Brand-Gruwel, "Effects of a modelling example for teaching information problem solving skills," *J. Comput. Assist. Learn.*, vol. 34, no. 6, pp. 688–700, Dec. 2018.
- [16] J. R. Hill and M. J. Hannafin, "Teaching and learning in digital environments: The resurgence of resource-based learning," *Educ. Technol. Res. Develop.*, vol. 49, no. 3, pp. 37–52, Sep. 2001.
- [17] N. Y. Yen, Q. Zhao, Y. Liu, and J. C. Tsai, "An intelligent state machine towards task-oriented search support," Presented at the IEEE Int. Conf. Cybern. (CYBCO), Lausanne, Switzerland, Jun. 2013.
- [18] D. Garigliotti and K. Balog, "Generating query suggestions to support task-based search," Presented at the 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., Tokyo, Japan, Aug. 2017.
- [19] S. Ghosh, M. Rath, and C. Shah, "Searching as learning: Exploring search behavior and learning outcomes in learning-related tasks," Presented at the Conf. Hum. Inf. Interact. Retr., New York, NY, USA, Mar. 2018.
- [20] M. Tibau, S. Siqueira, and B. P. Nunes, "A comparison between entity-centric knowledge base and knowledge graph to represent semantic relationships for searching as learning situations," Presented at the VIII Congresso Brasileiro de Informática na Educação, Brasília, Brasil, Nov. 2019.
- [21] Z. Tang and G. H. Yang, "A re-classification of information seeking tasks and their computational solutions," 2019, *arXiv:1909.12425*.
- [22] I. Hsieh-Yee, "Research on web search behavior," *Library Inf. Sci. Res.*, vol. 23, no. 2, pp. 167–185, Jun. 2001.
- [23] C. Hölscher and G. Strube, "Web search behavior of internet experts and newbies," *Comput. Netw.*, vol. 33, nos. 1–6, pp. 337–346, 2000.
- [24] S. Brand-Gruwel, I. Wopereis, and Y. Vermetten, "Information problem solving by experts and novices: Analysis of a complex cognitive skill," *Comput. Hum. Behav.*, vol. 21, no. 3, pp. 487–508, May 2005.
- [25] R. Kalyani and U. Gadiraju, "Understanding user search behavior across varying cognitive levels," Presented at the 30th ACM Conf. Hypertext Social Media, Hof, Germany, Sep. 2019.
- [26] R. K. Hambleton and R. W. Jones, "An NCME instructional module on: Comparison of classical test theory and item response theory and their applications to test development," *Educ. Meas., Issues Pract.*, vol. 12, no. 3, pp. 38–47, Oct. 2005.
- [27] J. P. Fox, *Bayesian Item Response Modeling: Theory and Applications*. New York, NY, USA: Springer, 2010.
- [28] A. J. Heidenberg, "Reducing pretest and gain score correlation using item response theory domain scores," Ph.D. dissertation, Dept. Educ. Policy Stud., Georgia State Univ., Atlanta, GA, USA, 2000.
- [29] W. M. Yen, "The choice of scale for educational measurement: An IRT perspective," *J. Educ. Meas.*, vol. 23, no. 4, pp. 299–325, Dec. 1986.
- [30] M. A. Kim-O and S. E. Embretson, "Item response theory and its application to measurement in behavioral medicine," in *Handbook of Behavioral Medicine*, A. Steptoe, Eds. New York, NY, USA: Springer, 2010, pp. 113–123.
- [31] S. P. Reise and N. G. Waller, "Item response theory and clinical measurement," *Annu. Rev. Clin. Psychol.*, vol. 1, no. 5, pp. 27–48, 2009.
- [32] C.-H. Leng, H.-Y. Huang, and G. Yao, "A social desirability item response theory model: Retrieve–deceive–transfer," *Psychometrika*, vol. 85, no. 1, pp. 56–74, Mar. 2020.
- [33] M. Yarandi, H. Jahankhani, M. Dastbaz, and A. R. Tawil, "Personalised mobile learning system based on item response theory," Presented at the 6th Annu. Conf. School Comput., Inf. Technol. Eng., London, U.K., Jan. 2011.
- [34] G. Marchionini, "Exploratory search: From finding to understanding," *Commun. ACM*, vol. 49, no. 4, pp. 41–46, Apr. 2006.
- [35] C. C. Kuhlthau, J. Heinström, and R. J. Todd, "The 'information search process' revisited: Is the model still useful," *Inf. Res.*, vol. 13, no. 4, pp. 4–13 2008.
- [36] R. Rummer, J. Schweppe, and A. Schwede, "Open-book versus closed-book tests in university classes: A field experiment," *Frontiers Psychol.*, vol. 10, p. 463, Mar. 2019.
- [37] K. Järvelin, P. Vakkari, P. Arvola, F. Baskaya, A. Järvelin, J. Kekäläinen, H. Keskestalo, S. Kumpulainen, M. Saastamoinen, R. Savolainen, and E. Sormunen, "Task-based information interaction evaluation: The viewpoint of program theory," *ACM Trans. Inf. Syst.*, vol. 33, no. 1, pp. 1–30, Feb. 2015.
- [38] R. Raieli, *Multimedia Information Retrieval: Theory and Techniques*. Amsterdam, The Netherlands: Elsevier, 2008.
- [39] M. E. Falagas, E. I. Pitsouni, G. A. Malietzis, and G. Pappas, "Comparison of PubMed, scopus, web of science, and Google scholar: Strengths and weaknesses," *FASEB J.*, vol. 22, no. 2, pp. 338–342, Feb. 2008.
- [40] P. Li, B. Zhang, and Y. Zhang, "Extracting searching as learning tasks based on IBRT approach," *Appl. Sci.*, vol. 12, no. 12, p. 5879, Jun. 2022.

- [41] J. Mao, Y. Liu, N. Kando, M. Zhang, and S. Ma, "How does domain expertise affect users' search interaction and outcome in exploratory search?" *ACM Trans. Inf. Syst.*, vol. 36, no. 4, pp. 1–30, Oct. 2018.
- [42] H. O'Brien, A. Cole, A. Kampen, and K. Brennan, "The effects of domain and search expertise on learning outcomes in digital library use," Presented at the ACM SIGIR Conf. Hum. Inf. Interact. Retr., Mar. 2022.
- [43] R. Sharifpour, M. Wu, and X. Zhang, "Large-scale analysis of query logs to profile users for dataset search," *J. Documentation*, vol. 79, no. 1, pp. 66–85, Jan. 2023.
- [44] S. Scholten and S. B. Moon, "An experimental study on the effect of domain expertise on the consistency of relevance judgements," *J. Inf. Manage. Soc.*, vol. 38, no. 3, pp. 1–22, 2021.
- [45] R. Sobha, "The effects of domain expertise on a user's conversational search," Tech. Rep., 2020.
- [46] X. Song, C. Liu, and Y. Zhang, "Chinese college students' source selection and use in searching for health-related information online," *Inf. Process. Manage.*, vol. 58, no. 3, May 2021, Art. no. 102489.
- [47] J. E. Hinostrroza, A. Ibieta, and C. Labbé, "Using information problem-solving activities to teach: An exploratory study designed to improve teacher competencies related to internet use in the classroom," *Technol. Pedagogy Educ.*, vol. 30, no. 2, pp. 235–255, Mar. 2021.
- [48] I.-C. Wu, B.-X. Huang, and P. Vakkari, "Learning outcomes during information search in digital archives," *Proc. Assoc. Inf. Sci. Technol.*, vol. 58, no. 1, pp. 329–380, 2021.



**YIN ZHANG** was born in Shenyang, Liaoning, China, in 1985. He received the B.S. degree in computer science and technology and the Ph.D. degree in computer application technology from Northeastern University, Shenyang, in 2006 and 2012, respectively. He is currently an Associate Professor with Software College, Northeastern University. His research interests include searching as learning, software engineering, and channel coding theory.



**BIN ZHANG** received the B.S. degree in computer software from Xi'an Jiaotong University, Xi'an, China, in 1984, and the M.S. degree in computer application technology and the Ph.D. degree in computer software and theory from Northeastern University, Shenyang, China, in 1989 and 1997, respectively. He is currently a Professor with Software College, Northeastern University. His research interests include edge computing and service computing.



**YULI ZHAO** received the B.S. and M.S. degrees in software engineering and the Ph.D. degree in communication and information systems from Northeastern University, Shenyang, China, in 2007, 2009, and 2013, respectively. She is currently an Associate Professor with Software College, Northeastern University, Shenyang, China. Her research interests include channel coding theory and software engineering.



**YUYANG BAI** is currently pursuing the M.S. degree in computer science and engineering with Northeastern University, Shenyang, China. In 2022, he was qualified to continue his Ph.D. study at Northeastern University. His research interests include searching as learning, exploratory search, and the relationship between search process and searchers' ability.



**PERTTI VAKKARI** is currently a Professor Emeritus in information studies with the Faculty of Information Technology and Communication, Tampere University, Finland, where his work is focused on information seeking, task-based information searching, fiction searching, evaluation of interactive information retrieval, characteristics of research in information and library science, and outcomes of public libraries. His publications, too numerous to list, include, "Result List Actions in Fiction Search," published in the Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries, for which he won the Vannevar Bush Award for the Best Paper, in 2015.

He was a Board Member of the Nordic Research School for Library and Information Science, a member of the Standing Program Committee of the International Conference on Information Seeking in Context (ISIC), and a member of the editorial boards of *Information Processing and Management* and *Journal of Documentation*. He received the ASIS&T Research in Information Science Award in 2020. He has served as the Chairperson for the Nordic Information Studies Research Education Network.

...