Studying PubMed Usages in the Field for Complex Problem Solving: Implications for Tool Design

Barbara Mirel and Jennifer Steiner Tonks
School of Education, University of Michigan, 610 East University, Ann Arbor, MI 48109. E-mail: {bmirel, jrsteine}@umich.edu

Jean Song
Health Sciences Library, University of Michigan, 1135 Catherine Street, Ann Arbor, MI 48109. E-mail: jeansong@umich.edu

Fan Meng and Weijian Xuan
Psychiatry Department, University of Michigan Medical School, 205 Zina Pitcher Place, Ann Arbor, MI 48109. E-mail: {mengf, wxuan}@umich.edu

Rafiqa Ameziane
Molecular, Cellular, and Developmental Biology Department, University of Michigan, 830 North University, Ann Arbor, MI 48109. E-mail: rafiqaa@umich.edu

Many recent studies on MEDLINE-based information seeking have shed light on scientists' behaviors and associated tool innovations that may improve efficiency and effectiveness. Few, if any, studies, however, examine scientists' problem-solving uses of PubMed in actual contexts of work and corresponding needs for better tool support. Addressing this gap, we conducted a field study of novice scientists (14 upper-level undergraduate majors in molecular biology) as they engaged in a problem-solving activity with PubMed in a laboratory setting. Findings reveal many common stages and patterns of information seeking across users as well as variations, especially variations in cognitive search styles. Based on these findings, we suggest tool improvements that both confirm and qualify many results found in other recent studies. Our findings highlight the need to use results from context-rich studies to inform decisions in tool design about when to offer improved features to users.

Introduction

Recent research and innovations related to PubMed and other MEDLINE information retrieval (IR) systems have expanded our knowledge about scientists' information-seeking behaviors and relevant tool-based support. Little current research, however, specifically examines scientists' information seeking for literature-centric exploratory analysis in actual contexts. Our field study begins to fill this gap. We observed 14 undergraduate majors in molecular, cellular, and developmental biology as they interacted with PubMed for real-world exploratory analysis. This article describes the workflows and tool support that these users demonstrably needed for their PubMed-driven problem solving.

Our field study findings reinforce White and Morris's (2007) contention that "unless [usage log data] are coupled with a qualitative technique [,]. . . it can be difficult to associate interactions with user characteristics" (p. 256). Our findings add to the research that strives to make these associations specifically for PubMed usage. To our knowledge few, if any, field studies specifically on PubMed exist in the research literature. Our results confirm many insights derived from PubMed quantitative studies about users' information-seeking behaviors and tool needs and from qualitative research on similar IR systems. They also add a number of qualifications. For example, our findings confirm that additional conceptual information integrated into retrieved results could expedite getting to relevance. Yet—as a qualification—evidence from our field cases suggests that presentations of this information need to be strategically

Received February 12, 2012; revised June 29, 2012, August 1, 2012; accepted August 1, 2012

 DOI: 10.1002/asi.22796

apportioned and staged or they may inadvertently become counterproductive due to cognitive overload. Based on our findings, we describe and explain users' information-seeking behaviors, reasoning, and needs at various points in problem solving. We characterize their varied cognitive search styles and effects of styles on outcomes. Echoing Bates (1990), our findings emphasize that tool designing needs to be a deliberate process of choosing which features among legitimate contending options best accommodate users (in our case, novice scientists) at specific stages of iterative exploration and knowledge construction.

The rest of the article reviews related work and presents our research questions, methods, and limitations. We then present results on common and variant behaviors and reasoning, and we discuss implications for tool improvements and conclusions, respectively.

Related Work

A synthesis of related work underscores the importance of complementing quantitative findings about PubMed users' information-seeking behaviors with qualitative insights into information seeking in context. At present, most evidence about PubMed users' information seeking comes from quantitative analyses of usage logs. Log data capture sequences of program interactions from which to generalize patterns of behavior, but they are removed from the surrounding details of scientists' actual contexts of work. Based in part on such quantitative findings, many researchers have proposed and/or implemented tool innovations for MEDLINE IR systems to facilitate and enhance scientists' information seeking. As a qualitative complement to this quantitative research, little empirical data exist on scientists' actual uses of PubMed and other MEDLINE IR systems in context for various real-world purposes. Consequently, it is difficult to gauge the extent to which proposed innovations might be effective for actual problem-solving purposes under different conditions and for different types of users. Some qualitative field research characterizes professionals' search and analysis activities in other domains with comparable IR systems to PubMed. Indirectly, these studies suggest many relevant insights for PubMed usages. One other area of related work focuses on demonstrated differences between expert and novice scientists (akin to students) in regard to their respective behaviors and reasoning during exploratory analysis. We now turn to the complementary pictures of information seeking that a synthesis of these areas of related work provides.

Sets of quantitative and qualitative researchers alike have described patterns of information-seeking workflows, but patterns differ according to the two sets of researchers. One reason is that the quantitative and qualitative researchers use different methods for collecting and structuring data into the action segments that define what counts as a pattern. For PubMed, quantitative researchers have analyzed huge volumes of usage logs and have extracted meaningful actions and generalizable patterns of action (Dogan, Murray, Neveol, & Lu, 2009; Herskovic et al., 2007; Lin, DiCuccio, Grigory, & Wilbur, 2008; Lin & Wilbur, 2009; Radlinski, Jurup, & Joachiams, 2008). Actions include, for example, querying (Q), retrieving (R) a MEDLINE article, clicking the Next (N) page of results link, clicking on a Related Article title/link (L), clicking to see more (M) related articles, and using of Advanced features (V) such as Limits to modify views (Lin & Wilbur, 2009). Log researchers also include a catchall category (P) for actions involving Preview/Index, History, Clipboard, Details tabs and LinkOut (Lin & Wilbur, 2009). From coded logs, researchers have abstracted patterns of actions that represent significantly frequent co-occurring strings of two, three, or four actions. Lin and Wilbur (2009), for example, have found that two, three, or four serial clicks for a query, retrieved title, or both combined constitute 75% of all log interactions, making serial queries, retrievals, or combinations of the two "the core of [PubMed] information seeking behavior" (Lin & Wilbur, 2009, p. 499). Other frequent action patterns include several clicks on Related Article titles (indicating chaining behaviors), as well as serial clicks on Next page of results (suggesting difficulty locating a relevant article). Usage logs also reveal patterns in query content. The majority of queries consist of terms for genes, diseases, and/or biological processes (Herkovic et al., 2007).

PubMed log researchers derive patterns from the segments they define in workflows. Typically, they parse actions into segments that are based on durations between milestone program operations, for example, queries, or on some arbitrary time span, for example, 24 hours or 31 days or time elapsed between periods of inactivity (which commonly means 10 to 30 minutes [Dogan et al., 2009; Herskovic et al., 2007; Lin & Wilbur, 2009; Radlinski et al., 2008]). To temper the assumption that successive moves from one query action to another signal a shift to a new topic, some log researchers have calculated semantic distance between query terms (Teevan et al., 2008; Xie, 2000). When semantic distances are close, these researchers put successive queries in the same workflow segment (Herskovic et al., 2007). Calculating semantic distance clearly can account for query refinements, but does not necessarily account for instances in which queries for the same topic use terms that are not obviously semantically related.

Methods for parsing action sequences determine what counts as a pattern, but because segmenting rules are context free, log researchers acknowledge that patterns have limited face value. Qualitative field studies offer an alternative perspective on workflow segments, one grounded in context-of-use. To parse flows of behaviors, field study researchers typically use the same organizing logic that they observe information seekers tacitly enacting in the field. Commonly, information seekers—and by extension field researchers—organize workflow activities by topics implicit in the problem they are exploring (e.g., in our case, different relevant diseases). Information seekers further organize each topic-based segment into stages, which represent tasks they

DOI: 10.1002/asi

use to achieve specific analytical objectives. When descriptions of information-seeking flows are structured in this way, any one topic-based flow of actions can include many different queries, semantically close and not. It may extend for 20 minutes straight or 6 hours intermittently, and periods of inactivity do not necessarily indicate a move to a new topic.

Topics are problem and domain specific, but the stages (objectives) for exploring topics appear to be generalizable. Field study researchers have found that information seekers enact many of the same stages, as follows (Tenopir, Wang, Zhang, Simmons, & Pollard, 2006):

- Start/Select topic/query: "activities [for] the initial search for information" (Ellis, Cox, & Hall, 1993, p. 359).
- Retrieve, explore, browse: "a semi-directed search in an area of potential interest" (Ellis, 1989, p. 179). Users may enact strategies such as chaining ("following chains of citations or other forms of referential connections between material") or filtering "mechanisms . . . to make the information as relevant and as precise as possible" (Ellis, 1989, p. 179; Ellis & Haugan, 1997).
- Extract information, focus, and formulate: "systematic [and deliberate] work [with] a particular source to identify material of interest," for example, close reading, scanning, skimming (Ellis et al., 1993, p. 359).
- Results evaluation: Evaluation of retrieved information against information needs. It may include rereading, assessing relevance, and distinguishing and selecting useful information (Makri, Blandford, & Cox, 2008).
- Analyze and/or synthesize: Interpret relationships meaningful to a problem and goal.
- Verify: "Checks on information and sources for accuracy and errors" and credibility (Ellis et al., 1993, p. 364).
- Record/collect information/present: Composition of "a record, of information resources . . . used, documents or content found. . . . or results of a search" or the "the scribbling and jotting of ideas" (Makri et al., 2008; Makri & Warwick, 2010, pp. 1,749–1,750).

As the stage descriptions imply, the actions that quantitative usage logs can capture predominantly relate to the first two stages, but qualitative studies show that information seekers' interactions with IR and full-text display systems during the last three stages influence their subsequent choices of moves and strategies (Kupulainen & Järvelin, 2010; Toms, 2002). Choices and outcomes of actions in the last three stages, for example, incite information seekers to iterate back to query and retrieve stages, and logs cannot trace these later-stage motivations for query and retrieve patterns. Iteration across stages is an inherent part of information seeking for exploratory analysis (Bates, 1999; Ellis & Haugan, 1997; Kuhlthau, 1999). Users iterate, for example, to reduce uncertainty and to better construct meanings (Andersen, 2006; Blake & Pratt, 2006a; Chowdury, Gibb, & Landoni, 2011) Even experienced IR tool users on average conduct at least two cycles of searching within a topic (Wildemuth, 2006). From Kuhlthau (1993) to the present, researchers of information-seeking behaviors have stressed the importance of providing adequate support in tools and training for users' iterations within and across stages. Clearly, a complete picture of search-and-analysis and associated tool improvements depends on describing all stages of exploring a topic.

Beyond defining and identifying stages and their functional roles in problem-solving, context-rich field studies complement quantitative analyses by revealing diverse factors influencing information seekers' choices for moves and strategies. Factors include domain traits, users' goals, stages of performance, and perceptions of a task and such subjective characteristics as prior knowledge, tool experience, role, and cognitive search styles (Ingwersen & Jävelin, 2005; Leckie, Pettigrew, & Sylvain, 1996; Li & Belkin, 2010). Complementing qualitative research, some usage log studies also show effects of users' level of tool experience on information seeking. For example, White and Morris (2007) distinguished between advanced and nonadvanced tool experience based on the complexity of users' query syntax. They found that nonadvanced users spent more time on a topic, had more and shorter trails, and returned more frequently to previously encountered pages (White & Morris, 2007).

As other factors influencing information seeking, domain knowledge and its effects have long been topics of research in cognitively oriented science studies. Findings related to our research reveal that novice scientists differ from experts in many ways. They typically fail to perceive nuances, tend to focus on irrelevant details, and are less skilled in self-monitoring for understanding and errors (Patel, Kaufman, & Arocha, 2002). Students particularly summarize instead of synthesize information from multiple sources, and they need to be reminded to relate prior knowledge to evidence uncovered in information seeking (Koslowski, Marasia, Chelenze, & Dublin, 2008). Additionally, novices—in our study scientists-in-training—often lapse into confusion even after problem solving has proceeded for a while. This confusion is intrinsic to learning. As they read more, domain novices become more uncertain about how to absorb and organize previously unfamiliar conceptual relationships into their evolving working knowledge (Patel & Kaufman, 2006). In our study, we complement these other quantitative and qualitative studies to analyze users' behaviors with PubMed with an eye on tool design.

Research Questions

Our field study concentrates on students, most of whom were weeks away from earning a bachelor of science degree in molecular, cellular, and developmental biology and who had previously worked as interns in laboratories. Students interacted with PubMed for the open-ended problem of explaining how certain molecular level interactions may combine to influence disease processes. We address the following three research questions:

1. How do these users commonly structure literature-based problem solving workflows and tasks?

DOI: 10.1002/asi

2. Within these structures, what variations occur and why?
3. What do field study findings imply about tool support that would benefit these users?

Qualitative methods were used as a means to fill out current knowledge about aspects of PubMed information-seeking behaviors relevant to tool support (Cresswell & Plano Clark, 2006).

Method

Participants

We observed and gathered think-alouds on 14 upper-level, undergraduate biology majors in a neuroscience laboratory course. These users worked with PubMed for a problem-solving module that was part of the course requirements. The module asked students to find and explain how interactions between genes related to the P2X receptor might influence a complex disease. The students had been studying the P2X receptor structurally and electrochemically all term in the laboratory. The field study occurred in the last month of the winter semester, after 13 weeks (104 hours) of studying the receptor. All but three of the students were graduating at the end of the month; most of them had served as interns or assistants in research laboratories. As research shows, at this point in academic training students are initiated enough into the discipline and scientific learning and reasoning to be called novice scientists (Gehring & Eastman, 2008; McCune & Hounsell, 2005). Additionally, at this point students generally are somewhat practiced in information literacy skills relevant to using PubMed for search and analysis purposes, for example, retrieving relevant articles, identifying criteria for relevance, understanding articles, and identifying information for answering research questions (Gehring & Eastman, 2008; McCune & Hounsell, 2005). The participants in our case all had at least 2 years' prior experience using PubMed to search for and read primary research articles for courses and laboratory jobs. Anecdotally, according to the neuroscience laboratory instructor, none had very much if any experience using PubMed for exploratory analysis.

Participants were a convenience sample. We focused on biology majors in a laboratory course to control some important variables in our selected participants, such as their level of domain expertise, the time pressures under which they conducted the problem solving, their degree of knowledge relevant to the receptor under study, and their reporting purpose. From previous field study research with biomedical domain experts, we knew that these variables are difficult to control when studying experts who define their own investigative problems. Our students all investigated the same problem—jointly with the same laboratory partners with whom they had been working all term—that is, for 13 weeks, 8 hours a week. Beyond these baselines, the students had diverse backgrounds. We refrained from conducting preliminary assessments of background issues, such as students' prior scientific knowledge, comprehension abilities, or skills with PubMed before and after training. Although a limitation to our study, we refrained because this module was part of the students' ongoing coursework, and we sought to disrupt the natural flow of the course as little as possible.

Students gave institutional review board (IRB) consent. During the workflows, partners asked each other questions and gave voice to their thinking processes, making think-alouds more authentic than they are in solo work.

Tasks and Materials

The authors of this article collaborated in designing the inquiry-based module that the students conducted for credit. In it students explored a new area of knowledge for them—the functional roles and interactions of P2X-associated genes in relation to disease phenotypes. The students' search-and-analysis task had two parts and, correspondingly, two deliverables. (See the Supplemental Material for the module assignment and deliverable worksheets.) For the first deliverable, students focused on at least three diseases and recorded explanations about how P2X-related genes influence each disease (see the Supplemental Material for the worksheet). For the second deliverable, students focused on one of the diseases they had recorded in the first deliverable and explained in greater depth the causal, contextual, and conditional relationships and biological events associated with P2X-related genes and the disease. This search-and-analysis is complex and open-ended, not a straightforward fact-reporting situation. It has no one set of "right" methods, no predefined starting and stopping points, and no single right solution (Kumpulainen & Järvelin, 2010). The cumulative nature of the problem-solving module—involving between $1^1/_4$ and 2 hours of literature searching and reading—fits with processes involved in coming to valid explanations over time. As Koslowski, Marasia, Chelenza, and Dublin (2008) note, explanations become increasingly convincing to the scientists who are building the explanations as evidence progressively mounts in an evolving investigation.

Data Collection

Each pair of users shared a workstation with Internet access and university proxies to PubMed (guaranteeing access to full-text articles). Each workstation had screen capture software installed and a high-power microphone connected to capture think-aloud comments. We collected uninterrupted time-stamped, audio-video recordings from each pair marking their activity start to finish. At the end of the class, we also collected the worksheets (deliverables) on which they recorded explanations based on reading.

Procedure

Before the day of the module, the instructor (a co-author) prepared students by discussing in class a review article they had been assigned to read explaining the role of the P2X

DOI: 10.1002/asi

receptor in various complex diseases. The module day was divided into three parts across the 4-hour class period. In the first hour, we prepared students for the information-seeking module. One of the co-authors, an experienced health sciences library informatician, gave students hands-on training with PubMed. She trained them in another tool as well—a term-enrichment tool that they could use to associate MeSH terms to genes. The training included passing out a user guide to which students could refer during the module (presented in the Supplemental Material). The lead co-author then explained the study and gathered signed IRB consent forms from participants. Next, the instructor explained the search-and-analysis module and deliverables. Students were given the next $2^1/_2$ hours to conduct the module with PubMed in pairs in whatever ways they preferred, and they were video- and audio-recorded throughout. None took longer than 2 hours. Students verbally collaborated just as in every other laboratory session, making the think-aloud protocol a very natural experience for them. Students wrote explanations and references for both deliverables on a worksheet, and they moved freely between searching, analyzing, and recording. For the last hour of the class—outside the scope of our study—students planned presentations for the next class period based on their findings from PubMed.

Data Analysis

Recordings of the search-and-analysis activities were transcribed to capture think-alouds as well as PubMed interactions. Working individually at first, three of the co-authors holistically viewed the videos and read the transcripts several times, a qualitative technique for abstracting themes, commonalities, and variations (Strauss & Corbin, 1994). An outside reviewer was brought in as well to do the same holistic reviewing. The investigators met as a group repeatedly to go over findings. We agreed on common ways in which users structured their workflows into segments. We defined these segments as sessions, episodes, stages, and tasks. We used the term *session* to refer to the span of information seeking devoted to producing each deliverable. We called the topic-driven flows of information-seeking within a session *episodes*. We deliberately chose not to use the term *session* to name topic-driven segments even though other studies do so. *Session* is an ambiguous term, used by both qualitative and quantitative researchers to describe segments of actions, but with different criteria for defining them. We used *episode* to avoid ambiguity. We structured the transcripts and video data into sessions, and episodes and within episodes identified common stages across pairs. Additionally, we identified and described variations across pairs of users evident in the data.

Based on findings, we developed codes for the transcript to enable us to find patterns of interactions in users' PubMed stages. A codebook and sample coding sheet are presented in the Supplemental Material. Developing and piloting a coding scheme was a highly iterative process. We coded time-stamped videos for demonstrated program operations—for example, query, retrieve, external link outs, PubMed internal links; and for demonstrated cognitive processes—for example, read, evaluate, record explanations. We also coded the location at which users performed an action—for example, query box, results first page, results second page, abstract, MeSH, related articles, citation, and worksheet. Opening the MeSH term list, for example, was recorded as a PubMed "internal links" action with a MeSH location. External link locations included, for example, Wikipedia. Given conventions in log coding, chaining by citation or by related article was coded as a query action (surrogate query) with, respectively, the citation and related article location.

We analyzed results from coding to abstract patterns of interactions, and we related findings to workflow categories and the commonalities and variations we had initially described. Descriptions evolved accordingly. Many interesting information-seeking dynamics associated with diverse user traits and contextual factors surfaced in our analysis, but they require additional research to do them justice, for example, communications between collaborating partners. To keep our research manageable we had to limit our focus of analysis. Because our aim was to uncover specific demands of problem solving on PubMed users' patterns of information seeking and associated needs for tool support, we concentrated on what turned out to be the most tangibly demonstrated factor influencing problem-solving behaviors and obstacles—cognitive search styles. Thus, we concentrated on user traits (subjective factors) more than sociocontextual factors. We progressively fine-tuned our descriptions of cognitive search styles, their distinguishing traits, and their roles in users' choices of moves and strategies and outcomes.

For meaning-making, we defined and coded worksheets/deliverables for relative depth of explanation through rich feature/discourse analysis (Barton, 2004). We distinguished deeper from more surface explanations by causal structures and linguistic markers in the write-ups and through level and type of detail in content, for example, the inclusion of interdependencies and/or indirect relationships or processes and interlinked/chains of behaviors.

Limitations

As with all qualitative research, ours does not support generalizing beyond our cases. Nonetheless, it provides new evidence about aspects of PubMed information-seeking stages, tasks, and variations that are understudied and need further research. A number of other constraints limit our study. One is that the training, problem solving, and write-ups had to fit within the time frame of a 4-hour class period. Users could not employ alternate means that typify many scientists' actual information-seeking approaches, such as skimming articles to determine which to save and read later. Also, users did not define their own problem or create their own worksheets for reporting, which can affect individuals' directions in reasoning, behaviors, and stages (Blake & Pratt, 2006a). Moreover, we cannot readily distinguish between behaviors that were part of individuals' habitual

DOI: 10.1002/asi

approaches to coursework (i.e., "studenting" behaviors) and behaviors that they would be likely to use in an exploratory task assigned by a laboratory supervisor in a professional context. Nor, as mentioned earlier, can we correlate behaviors and students' prior knowledge or comprehension skills. Another limitation is that we cannot infer with confidence training on PubMed that proved to be beneficial or alternate training designs and content that might have helped our field study users overcome obstacles they experienced. In our findings we recommend tool improvements, but not training. Training is vital, but proposed tool improvements connect best to the data we collected. Additionally, we did not evaluate the quality of the deliverables beyond a description of the relative depth of explanation in users' write-ups. Other quality measures and outcomes of learning and reading—and the cognitive processes and domain knowledge they involve—were outside the scope of our study. Examining collaboration dynamics was also outside the scope of our research and our analytical expertise.

Results

Common Ways of Structuring Information Seeking

Organize workflows into sessions, episodes, stages, and tasks. All the students structured their work by problem-driven goals as users have done in analogous field studies. They broadly divided their work by deliverables into what we call Sessions 1 and 2. Episodes, as mentioned earlier, involved activities devoted to exploring a specific topic/disease. Users engaged in many episodes in Session 1. In Session 2, they drilled down into just one disease (one episode), now in more detail. Users all demonstrated the same objectives in each episode, thereby enacting the same stages.

Figure 1 identifies these stages (adapted from Kulhthau, 1999) with arrows indicating that users iterated a good deal across them. As Figure 1 shows, stages included common sets of tasks, which were a mix of cognitive behaviors (synonymous with cognitive tasks) and program operations. Cognitive tasks included reading results, abstracts, and full texts; applying relevance criteria and judging relevance; transforming textual information into their own words (self-explaining); and validating their interpretations by bringing in prior knowledge or other texts. Operational tasks related to interacting with program features and included entering query statements, linking out, clicking on and viewing MeSH terms, opening related article links, and closing open windows (housecleaning) when too many cluttered the workspace. Figure 1 also shows that users fairly consistently interacted with the same pages to perform the tasks of each stage. PubMed has other pages (e.g., advanced search) that users did not access.

Brief stage descriptions follow, with stages names used in other field studies presented in parentheses.

Query (Start/Select Topic/Query). This stage involved issuing Boolean expressions to retrieve articles—all users in the study preferred Booleans over single keywords—and using related articles or citation chaining, which all but one pair of users did at least once in their workflows. Query also included the split-second judgments users made when results first displayed, and they determined from counts alone whether to stay on the page or requery. Query occurred at several page locations: PubMed home, Results, and Abstract.

Read Results (Retrieve/Explore/Browse). Read Results included reading and skimming lists of results and using

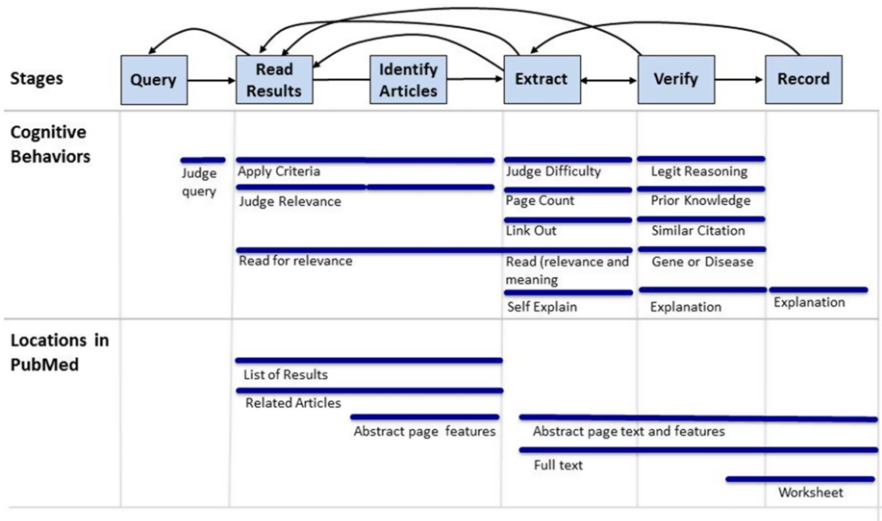

FIG. 1. Model of exploring literature for problem solving. Adapts Kuhlthau's (1999) representation in the Information Search Process (ISP) model to specific processes of novice scientists' uses of PubMed. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

DOI: 10.1002/asi

features on the Abstract page that might help identify if relevance criteria were met—but not the abstract itself. Users judged relevance of titles during this stage by applying the following three criteria: Was there a focus on P2X-related genes or proteins? Was a disease mentioned? Were there cues that the article presented causal explanations?

Identify (Retrieve). We made Identify a separate albeit short stage because it marked the point at which users selected a title and clicked on it to move to the extract stage. We note this stage when discussing time spent on task later (it marks the move from reading results to finding one of interest). We do not include it later in discussing sample information-seeking trails. For trails, this stage is implicit in moving to Extract.

Extract (Extract Information, Evaluate, Analyze, Synthesize). In Extract and the next two stages, users read texts—abstracts, full-text articles, outside information, or some combination of the three. Extract combined reading for relevance, understanding, interpreting, evaluating, and synthesizing. These modes of analytical reading and meaning making were dynamically intertwined, making it necessary for us to combine all of them in one stage. During this stage, users applied the same three relevance criteria they did to the results lists and added a fourth criterion—level of difficulty. They abandoned articles that were hard to understand. After users deemed texts relevant, they read for understanding and explanatory meaning.

Verify (Verify). Users verified in a number of ways: confirming judgments of relevance, reviewing content for accuracy and comprehension, validating interpretations with additional evidence, verifying spelling, and matching retrieved titles with citations when engaged in citation chaining.

Record (Record). Users composed two deliverables in which they recorded explanations of molecular-level influences on a disease. They cited references in each deliverable. Writing during Record threw all users back to reading the texts, realizing that they had to still better understand biological relationships, events, and outcomes.

The overall time that each pair of users devoted to episodes and stages in Session 1 hovered at an hour (see Table 1). Time spent in Session 2 varied based on how well each pair initially understood the disease they chose to investigate further. Users who spent longer in Session 2 (e.g., Pairs 4 and 6) had read and written shallowly about all diseases for the first deliverable and now read longer and more closely to better understand potential mechanisms and the role P2X played for the targeted disease. Pairs more markedly varied in time spent in terms of apportioning their time across stages in any one episode, described in detail in a later section on variations.

Patterns: Iterations and interdependence between stages. The students conducted the same low-level actions and series of actions as usage log researchers have identified. The field study results, however, reveal various functional roles these actions played in actual problem solving. As in usage logs, repeated query and retrieval actions were prominent, but field study findings suggest different implications about this pattern. Repeated Qs, Rs, and mixed Qs and Rs were not necessarily due to users' difficulty in matching query terms to relevance criteria. As detailed later, many pairs of participants frequently iterated back to query and retrieval actions, seemingly because other cues for relevance and content that "spoke" to them were not available elsewhere.

TABLE 1. Time spent in each session.

Pair ID	Session 1 (minutes)	Session 2 (minutes)
1	67	29
2	54	20
3	69	20
4	70	55
5	52	22
6	56	50
7	44	29

Action trails in Figures 2 and 3 illustrate some of these iterating actions for want of relevance cues elsewhere. As diagrammed in Figure 2, one pair of users queried, clicked on a result to read its abstract, and then moved to its full text (noted in blue). Then in actions that would fall outside the bounds of log capturing, these users proceeded to read the full text and click on a citation (noted in green). They next read the full text of the cited article and subsequently cross-referenced information from both full texts. At this point, the users were not sure if either of these texts was relevant in terms of discussing a disease potentially influenced by P2X-related genes. They, therefore, went back into what would be "usage log radar" and reopened the Results page. In logs, the capture at this point would be query-retrieval-retrieval (QRR), but in fact the trail so far was query—retrieval-read the full text—chain by citation in the retrieved result full text—read the cited full text article—reread the full text of retrieved result—retrieval.

The users (in Figure 2) continued their flow of work. Now they clicked and opened the abstract of the cited article. They wanted to view other information on this page for background. Next in a long series of subsequent moves (once more outside log capture range), the users read and verified information from both articles. They moved between full texts by clicking open windows of the full texts on their desktop. Ultimately, they synthesized and recorded explanations using both articles.

In what would be a log-recorded QRR pattern, this instance of PubMed use-in-context shows that users' repeated retrieve actions functioned to reduce users' uncertainty about a potentially interesting and interrelated pair of articles. Unable to glean relevance cues quickly from full

DOI: 10.1002/asi

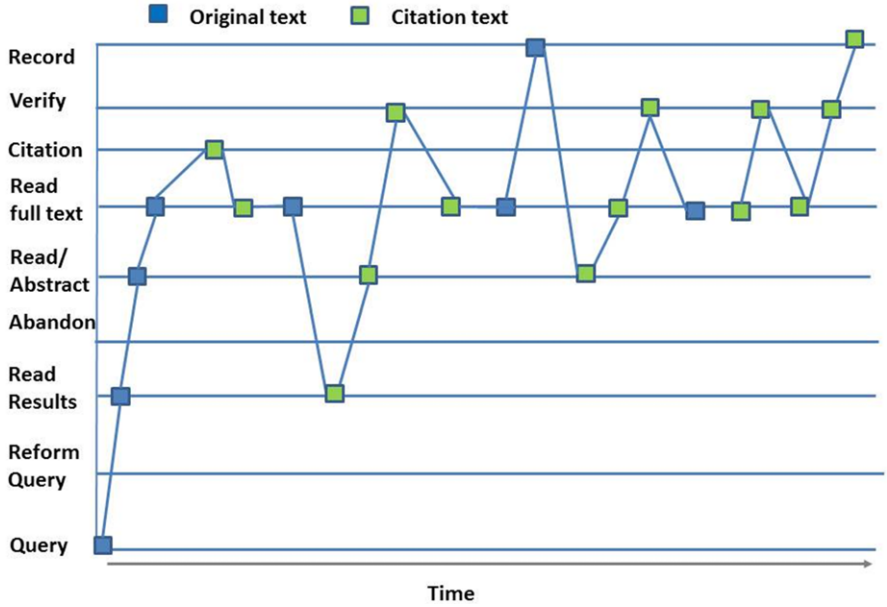

FIG. 2. One pair of users' flow of tasks. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

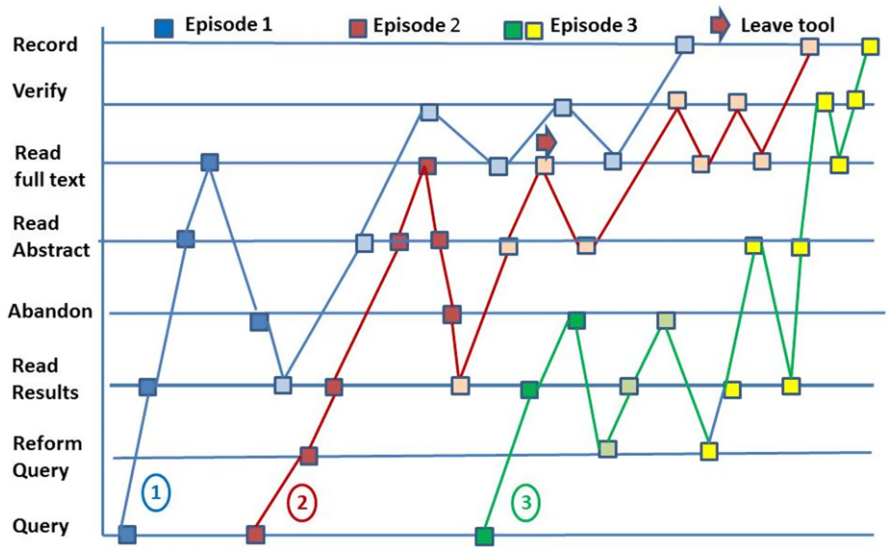

FIG. 3. Another pair of users' flow for three episodes. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

texts and lacking information elsewhere to resolve their uncertainty about relevance, users in this QRR instance returned to results to find something more about the articles.

The action trails as in illustrated in Figure 3 tell a somewhat different story. This figure shows three episodes that a different pair of users from those in Figure 2 conducted to investigate, respectively, three diseases. In all three episodes, the users, like those in described in Figure 2, regularly returned to the same results list to select a new potentially relevant article. Yet the users in Figure 3 had a more generalized and larger results list (193 hits from the broad query "P2X AND disease") from Figure 2's 58 hits from the narrower query, "P2X AND schizophrenia." The users in Figure 3 repeatedly returned to results each time they abandoned an article they had selected either as not relevant or too difficult. For ground truth, moves in Figures 2 and 3 should not be represented the same way. The differing contexts for returns to results (retrieval moves) signal distinct user motivations and likely different needs for improved tool support.

Field study data also show (see Figure 4) that reading (extract in Figure 4) and recording were core information-seeking activities along with querying and retrieving (read results in Figure 4). In fact, it was mostly during reading

DOI: 10.1002/asi

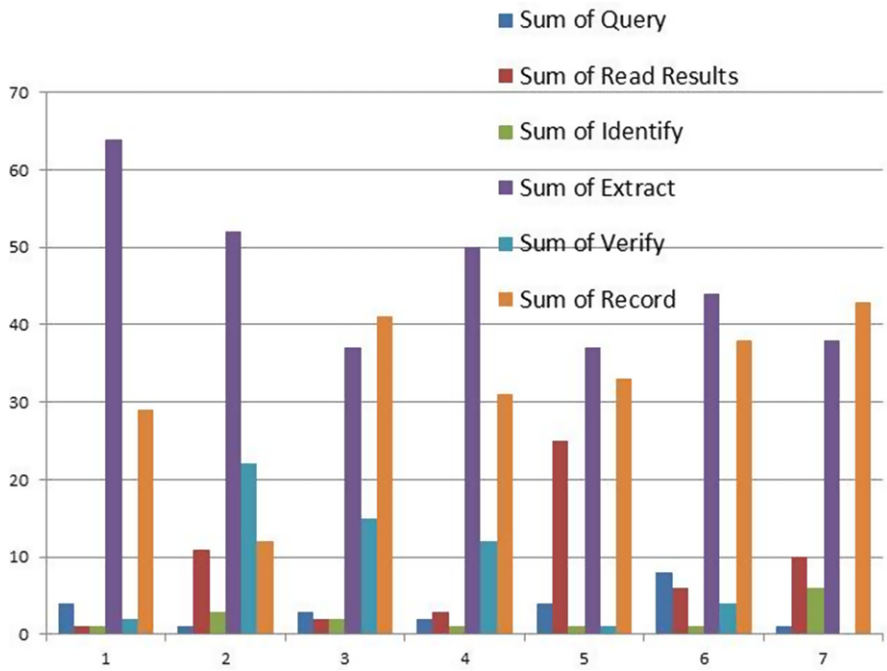

FIG. 4. Comparison across pairs of the proportion of time each spent per stage during an episode. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

and recording that information seekers experienced uncertainties either in disambiguating relevance criteria or in turning their knowledge into words. As a result—and often for lack of support elsewhere—users returned repeatedly to the stages of Query and Read Results, and then again back to Extract and Verify. The ways in which users iterated seemed to vary by users' cognitive styles of searching, described below.

Shared approaches and tasks within stages. Users enacted many cognitive tasks in common during each episode. They also encountered many of the same difficulties. We describe these tasks and difficulties structured by analytical objectives.

Experimenting during a prelude. Across all pairs, users started Session 1 spending 10 to 20 minutes experimenting with possible investigative approaches—that is, a prelude to investigation. During this time—and only then—users all employed the count of retrieved titles as a criterion for either abandoning or staying with a results list. In one query case, a pair of users accepted an autocompletion suggestion of "P2X receptors" for their query and consequently retrieved 1013 titles. The count was daunting, and they immediately abandoned the list and requeried, now adding a disease as well. During the Prelude, users typically issued many queries to experiment with terms and result counts. They did not seriously look for or open and read long articles from the result lists. Early experimenting influenced later actions. For example, the users who had accepted autocompletion thereafter explicitly shunned it.

Formulating Queries. Once users started to write queries in earnest and began the first episode, they encountered some cognitive and operational problems. Cognitively, null results occurred from queries in which a gene in the query was not associated with the disease users had also included in the query, for example, "P2X2 AND schizophrenia." Users' query term specificity was not backed up with biological knowledge about the gene.

Operationally, users often were unsure about the spelling of disease terms and linked out for help to Google and other websites. PubMed's autocompletion helped one pair of users to spell "epithelial," but they did not accept the autocompletion (avoiding its automatic launch to search because they had more terms to enter). After closing autocompletion they forgot the spelling and (inefficiently) had to refer once more to autocompletion. To check the accuracy of gene names, users linked to Entrez Gene. Link-outs added time and the risk of getting lost on return once numerous windows or tabs were open. Getting null results from a query also had an operational dimension. As in usage logs, null results occurred, at times due to incorrectly formatted queries with bibliographic information. Most often, bibliographic querying was part of a pair of users' chaining strategy. Recovering from bibliographic formatting errors was time consuming—an irony because chaining is a means to bypass inefficiencies caused by less-than-useful query terms. For example, upon experiencing problems with citation formatting users typically returned to source articles to verify the bibliographic information. They then copied and pasted the citation in the query box, shortened the string to just authors or titles, and painstakingly removed commas

DOI: 10.1002/asi

and periods. None of the users turned to advanced PubMed features for bibliographic queries covered in the training. Users' observed behaviors and think-alouds suggest that this feature did not match the quick copy-and-paste technique that seemed to be habitual to all users. Several pairs of users avoided the inaccuracies or faulty formatting that came from alternately copying and pasting citations into the query box by directly clicking reference hyperlinks in full texts. The downside to clicking on references was that each reference opened in a new tab. When too many tabs were open users got lost trying to find and reread an earlier article. In terms of query content and in accord with usage log findings, all users' queries in our study consisted of a disease name, the Boolean AND operator, and a gene name—typically with a P2X root instead of less obvious genes such as P2Y or IL-1 beta.

Getting to and judging relevance. Two striking features characterized all users' processes of getting to relevance. First, they read titles almost to the exclusion of any other relevance cues. Second, their initial certainty about relevance criteria soon gave way early in the first or second episode to confusion about what counted as a gene, disease, or explanation—the three criteria for relevance. Both of these experiences made getting to relevance time-consuming interactions.

First, by attending primarily to titles, users rarely if ever noted or strategically used authors, journal titles, and dates. They often pondered long over titles, and their limited domain knowledge seemed to constrain inferences about relevance they could draw from the titles. For example, they mistakenly eliminated articles with non-P2X derivative genes that were relevant. Users often commented that they wished they could find relevant titles more quickly and easily. Users who primarily chained by citation or intermittently clicked related articles also disregarded almost any other cues, but titles. Consequently, they often chose and successively read articles by the same authors. None of these users deliberately skipped over authors whose work they had already read as a means for gaining varied perspectives on the same topic. As a result, one pair commented, "I feel like we keep reading the same articles."

The only other relevance cue that a majority of users sought was MeSH terms. The MeSH feature and the additional Gene2Mesh tool were part of the preliminary PubMed training, and users hoped that MeSH terms would tell them if an article contained a relevant gene. Except for one case, however, MeSH terms did not help the users, and none interacted with Gene2MeSH. In the one case, the MeSH term "genetic" signaled to the pair of users that the article must involve a gene. No users inferred conceptual relationships from the list of MeSH terms combined, perhaps reflecting their novice traits as scientists.

As the second prominent feature of getting to relevance, users all became less certain than they originally were about what counted as a gene or disease. For example, they questioned whether the terms P2X and P2X7 were receptors or genes. They also questioned whether some title terms were actually diseases, such as "analgesic involvement" ("No, that's therapeutic") or "chemoprevention" ("No, that's a treatment"). As time progressed and as users read more abstracts and full texts, their uncertainty about relevance and relevance criteria actually increased rather than diminished. That is, the more users read, the less sure they became about what constituted a gene, disease, or explanation. Increased uncertainty is a novice scientist trait. To deal with the rising uncertainty, our users spent less time examining titles and more time reading texts to disambiguate relevance criteria. Texts were more likely to provide the conceptual knowledge required for reducing uncertainty. These trends and effects of uncertainty cannot be captured in usage log data.

Making the transition to full texts. Users' transitions from a PubMed-displayed abstract to its full text were at times disruptive. The full text was displayed by a different system. One pair of users who clicked on a link to display a full-text proceedings article, for example, instead got a display of the table of contents of the conference proceedings. Another pair retrieved a journal homepage, not the intended article. Both pairs of users spent relatively long getting to their intended articles (e.g., 3 minutes instead of the 3 seconds they expected).

Reading texts to verify/judge relevance. As mentioned, across users a good proportion of text-reading time went toward disambiguating what counted as a gene, disease, and explanation. For cues from abstracts and full texts, all users at some point skim-read texts and often employed "Find" features to locate gene and disease names. "Find" was effective unless an article used an unfamiliar gene synonym, in which case it took users several minutes and link-outs to figure out that the article used a synonym. Even in full-text articles it took users a good amount of time to sort out ambiguities about relevance criteria. As mentioned, a majority of them puzzled over whether P2X referred to a receptor, a gene, or a protein; for goal clarity, they wondered whether these were distinctions that mattered for the purpose of the deliverable. Some users changed their minds over time about what legitimately counted as a disease. One pair noted early on, for example, "This [inflammation] is sort of a disease; but it's not real. Let's find something with an actual name." Later, after reading more, the pair reversed and rationalized: "I'm not sure P2X causes anything. It seems like it changes in all these . . . like it's a marker for everything . . . P2X is in everything. Therefore, this broad disease is okay."

Uncertainty about relevance cues reared its head even during the Record stage. For example, one pair interrupted their recording of an explanation about P2X and bladder hyperactivity to re-read the text and then had second thoughts about whether the article actually met the disease criterion. Reading the sentence "C-fibers that express P2X3 and P2X2/3 receptors mediate at least in part acetic

DOI: 10.1002/asi

acid-induced bladder hyperactivity-mediate" one user said to the other:

> User 1: This is acetic-acid induced bladder activity. So it isn't a disease. It's something that is being induced.
> User 2: It is a disease. It's bladder hyperactivity.
> User 1: But it's being induced.
> User2: So? So like fail-of-your-immune-system-induced AIDS, that's not a disease? I don't think so.

Reaching no overt resolution, the pair tacitly accepted the article as relevant and continued with the worksheet.

Determining whether an article met the explanation criterion also threw users into a quandary. For example, one pair discussed whether they could legitimately consider a cause and effect statement to be an explanation because the statement used the word *may*. ("It's not an explanation. It's not strong enough.") Other users wondered if it was an explanation when an article discussed how molecular interactions functioned biologically but did not overtly discuss how the functions influenced disease processes. For example, reading about the P2X role in up-regulation, a pair of users said to each other:

> User 1: Is that how P2X causes it [pancreatic cancer]?
> User 2: I don't know if they know how it causes it. They just say that it's increased.

Reading texts to make meaning. As the preceding examples about relevance judgments suggest, users often slid almost imperceptibly from judging relevance to interpreting functional, causal, and conditional molecular interactions. As we detail more fully below, variations in cognitive search styles seemed to go hand-in-hand with students' approaches to making meaning and to achieving in-depth explanations (though the direction of influence is not clear from the data). In terms of commonalities in reading for meaning, however, all of the students employed the same approach to get themselves oriented in an article. They scrolled the text, noted the page count, and sought a section to use as a starting point. The majority started with the Discussion. Two pairs first read the Introduction and then moved to the Discussion. Many users who read reviews also sought to start with a Discussion section and were surprised, dismayed, and disoriented when they could not find one. Even when these users finally gathered that their article did not follow primary research conventions for section headings, it is not clear if they recognized that they were reading a review. None articulated this recognition. For these readers, getting acclimated to review section headings—which were typically thematic—took time. They were not immediately able to find an optimally relevant section, uncertain about which theme came closest to their goal.

After getting oriented, another commonality across users was to iteratively extract, verify, and record—at times returning to retrieved results or doing more chaining, as well. Users all accurately realized that P2X rarely was a direct cause of a possible disease and that they would have to "look for indirect relationships." They iterated across stages to better understand indirect relationships and to obtain evidence to support explanations. Finally, while reading and verifying for meaning, all users at least once linked out to external information to acquire background knowledge relevant for understanding the article.

Staying oriented. Given that users iterated across workflow stages and frequently linked out and back, they had to struggle, at times, to stay oriented for coherence in thought. In one case of linking out, users had eight open tabs of abstracts and texts, and it took them a while to find the text they wanted upon return. Additionally, to extract information and to construct and validate knowledge field users commonly backtracked, reread, and explored new or tangential content. They linked out and back and needed to find where they left off. They strived to remember sources of relevant knowledge and passages of interest in sources. In the field study, when users got lost at various points it diminished efficiency and interrupted their analytical coherence.

Variations across users and implications for task performance and outcomes. **Variations in time on task.** In addition to similarities across users' workflow structures and episode-based cognitive tasks variations occurred—especially in how users ordered and combined certain tasks and stages. The prevalence of some task actions over others varied as did the time users spent on particular tasks and stages. For example, one pair read more and jointly discussed texts more thoroughly than any other users (Pair 1 in Table 2). Extensive time reading caused them to conduct fewer episodes (e.g., four) and write about fewer diseases (e.g., three) in their worksheets than users who did not read closely. In fact, these close readers explicitly worried about "lagging behind." By contrast, another pair of users who tended to rapidly scan texts and were explicitly proud of their speed moved through more episodes (eight) than other users, recorded more diseases (e.g., five), and included more references for each explained disease (at least two per disease). Still one other pair spent 50% of their extract (text reading) time in link-outs acquiring background knowledge. They read more reviews than original research articles and read and used more abstracts than articles for their written explanations. In these examples, the first pair of users developed deep explanations in their deliverables for Session 1. The last two wrote surface explanations.

Table 2 shows the amount of time each user pair spent during a successful episode in Session 1. We define a successful episode as one that resulted in a recorded explanation on the worksheet. As Table 2 shows, across the stages of users' respective sample episodes all pairs invested the greatest proportions of time in extracting (reading and verifying texts) and recording. Given the requirements of the course module, this apportioning is not surprising. But it underscores how end-to-end workflows in context complement what usage logs can capture.

DOI: 10.1002/asi

TABLE 2. Time spent in a successful episode for each pair and apportioned time per stage.

Pair ID	Query (%)	Read results (%)	Identify (%)	Extract (%)	Verify (%)	Record (%)	Raw time (minutes)
1	4	1	1	64	2	29	20
2	1	11	3	52	22	12	15
3	3	2	2	37	15	41	14
4	2	3	1	50	12	31	9
5	4	25	1	37	1	33	19
6	8	6	1	44	4	38	9
7	1	10	6	38	0	43	7

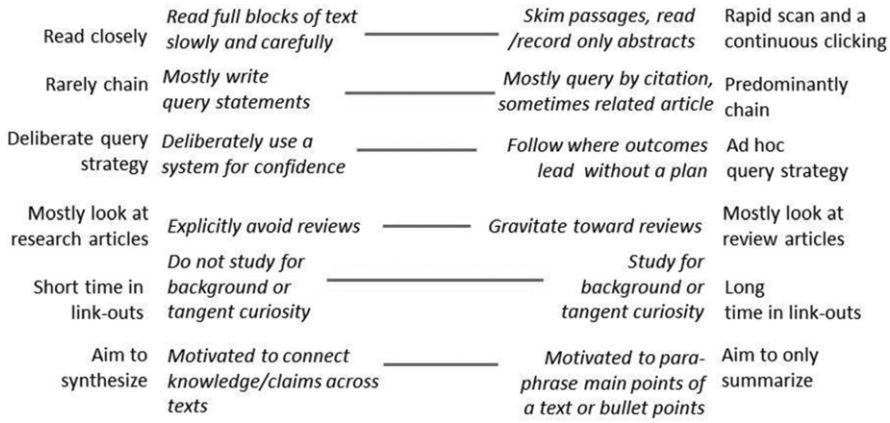

FIG. 5. Template of six dimensions composing cognitive style.

Other than such broad patterns, each user pair in Table 2 had their own story. For example, Pair 1 spent 4% of their episode querying as did Pair 5, but for different reasons. Pair 1 issued only one query, but it took time because of misspellings. But the pair did not have any difficulty in knowing what they wanted to query on because they based their disease term on an authoritative source—namely, on Science Direct facets, explained later; and they identified a good result efficiently. This pair only read one article in their episode, and never spent time returning to the results list, hence their high proportion of time in Extract. That is, they apportioned their attention and energy to productive ends. Pair 5, by contrast, issued four queries and took an ad hoc approach to query formulation. They returned to the query stage frequently to issue new queries on the same disease (cancer), often entering terms that caught their eye while reading retrieved texts (e.g., "immunology" and "apoptosis"). Not advanced in query statement construction, they wrote, for example, "P2X AND immunology" rather than "P2X AND cancer AND immunology." They also spent a long time reading results lists, largely to see what, in fact, they had queried for. Toward the end of the episode they split their time and iterations almost evenly between extract and record, as Pair 1 did, but again for different reasons. Pair 5'ss ad hoc querying led to a lack of focus in their cumulative construction of knowledge. Once they started writing they had to focus better. At this point, they interleaved extract and record and turned back to texts regularly to fill in the gaps in what they were writing.

These and other pairs' apportioned times in exploring a topic suggest that complex relationships among what is likely an array of factors affected users' information-seeking directions and outcomes in ways that are not yet studied or understood for PubMed usage. For our cases, we found that we could derive few patterns from the coded transcripts from which to make claims about task durations or priorities because each pair of users seemed to have a different cognitive style of search that shaped task performance in different ways.

Variations in cognitive style. Field study results reveal aspects of PubMed users' cognitive styles of search that have not been examined before, and we found that these styles could be described by six dimensions. Every pair of users had a somewhat different profile on these dimensions. In Figure 5, we present a template of these dimensions, represented as continua because except for one pair (Pair 1), no users embodied pure instances (end points) on every dimension. For example, many users read both primary research articles and reviews but varied in which type they read more.

We present a filled-out cognitive style template for each pair of users in the Supplemental Material. These detailed

DOI: 10.1002/asi

profiles are reference material because they were too idiosyncratic to afford reliable inferences or correlations. Evidence suggests associations between cognitive style and users' moves, strategies, and outcomes, but additional research—and larger groups of participants—are needed to explain the interrelated factors and effects conclusively. Descriptively, variations in cognitive styles often led users in different directions, and we describe two such examples below.

1. *Example 1.* Variations in styles for formulating queries and interacting with results: For these activities, users' styles ranged from systematic to ad hoc. Systematically, one pair of users queried by disease terms they had copied from disease facets in Science Direct, a full-text delivery system that provides facets for articles it displays. The users were confident in the disease terms they took from Science Direct because it was an authoritative source. They skimmed results looking for other relevance criteria (genes and explanations) and got to relevant items quickly. Another pair of users enacted a different systematic style by chaining by citations. They eventually used reference lists within articles as proxies for result lists. Similarity-based chaining made them confident in the likely relevance of each retrieved article. By contrast, ad hoc styles meandered more. In one case, a pair queried broadly ("P2X AND disease") and spent the next 38 minutes chipping away somewhat randomly at the retrieved list of 193 results looking for disease, gene, and explanation cues for relevance. They opened and closed many irrelevant articles, each time returning to the original list still uncertain about what to look for in titles as indicators of relevance. As mentioned earlier, a different pair issued the same broad query and then in ad hoc fashion entered biological process terms encountered while reading relevant texts. Yet another pair Googled "-itis" to find random disease names to enter.
2. *Example 2.* Variations in processes related to reading-for-meaning: To construct meaning from texts, some users consistently read closely while others rapidly scanned. For one pair of users, power browsing and speed seemed to be ends in themselves. These reading-style variations affected the knowledge and meaning users constructed from texts. In the power browsing case, the users read the Discussion (quickly) backwards—starting at the end and stopping when they believed they had acquired enough information. Yet they stopped prematurely and missed a paragraph that qualified the claims they had read in parts of the paragraph they had covered, which affected the accuracy of their deliverable. Another pair who tended toward rapid scanning spent a good deal of time skimming conceptual information in link-outs. In one instance, they spent almost 4 minutes reading external information about cardiomyopic calsequestrin—longer than they ever spent consecutively reading an abstract or full text. Cognitive search style and prior domain knowledge were likely entangled in instances such as this one. Users' depths of explanation in the deliverable were associated with cognitive search style—and likely tied to prior knowledge, as well. Users who consistently read closely throughout Sessions 1 and 2 produced the deepest explanations in both deliverables. Table 3 shows the numbers of full texts each pair read and the number of diseases they recorded in their deliverables. Numbers of diseases and full texts for many pairs were fairly close (e.g., four to five diseases for all but one pair, and four to five texts for all but two pairs). But numbers alone were not ample signs of explanatory depth in deliverables.

TABLE 3. Diseases covered and number of articles read by each pair.

Pair ID	# Diseases	# Full texts
1	3	4
2	4	4
3	5	7
4	4	5
5	4	5
6	5	11[a]
7	4	5

[a]This pair read only abstracts (eight of them) for Deliverable 1 and three full texts for Deliverable 2.

Variations in depth of explanation in the deliverables. Results from analyzing the relative depth of explanation in student deliverables showed that some clear distinctions separated deeper from more surface explanations, as follows:

- A deep explanation captured interdependencies or cascades of processes, and specifically identified the role of a P2X gene in them. A template description of such an explanation is *During biological event or process V, Gene A changes and has Effect E in the cell. Effect E influences Gene B, and all these interrelated processes influence aspect D of the disease.* To compose the following example of a deep explanation, the users spent a good deal of reading time and conversation figuring out the direction of influence:
  Example: "Based on the findings in this paper there appears to be an increase in P2X7 receptor during acute or chronic phases of temporal lobe epilepsy. This overactivity causes excitotoxicity and hyperexcitability. The cells try to compensate by decreasing expression of P2X4. The resulting cell death from the seizure may be due to P2X7 over-expression."
- Surface explanations typically included more than two interacting entities, processes, or both as influences on a disease, but not as many factors as deep explanations. It was common for users to record this string of interactions without tying them to disease processes. Templates for surface explanations include *During process or event V, Effect E occurs to Gene A (in Location L). Gene A becomes involved in process or event W [which is related to the disease].* Two examples show a range of this level of explanation:
  Example A: "P2X7 expression is increased during epileptogenesis. This upregulation of the receptor occurs at glial cells and glutamatergic nerve terminals. This suggests that this receptor may be involved with inflammatory processes during the onset of epilepsy."
  Example B: "Alzheimers results from an upregulation in P2X receptors, which induces an inflammatory response."

Close reading during Session 1 seemed to prepare users to explain molecular interactions in the first deliverable in

DOI: 10.1002/asi

complex ways—for example, as serial or cascading events. A heavy dependence on reading reviews instead of primary research articles seemed to lead to shallow explanations.

The second deliverable explicitly called for turning an explanation that users had derived during Session 1 into a deeper explanation. During Session 2, users investigated the following two issues: (a) how factors affecting P2X were involved in events and outcomes associated with susceptibility to the targeted disease and (b) how the factors were activated or inhibited in the first place. For some users, the demands of this interpretive problem solving called for "so much thinking it hurts my head." Even a pair who interpreted, inferred, and synthesized most deeply seemed to regard this thinking as unfamiliar. They synthesized content for complex causality across three articles and expressed it in their own words. Yet they called this process "cutting and pasting." Their deep explanation follows.

> A genetic predisposition or a part injury can cause a person to be more susceptible to seizures which causes an increase in P2X7 expression in neuronal cells like astrocytes. An increase in P2X7 makes the cell more susceptible to extracellular ATP. Greater than normal excitation of P2X7 receptors leads to an increased Ca^{2+} influence than normal [sic] and this causes cell death. Cell death will release ATP and calcium. This released ATP and calcium can activate and pass through other neuronal cells with upregulated P2X7 and contribute to killing them and eventually cause lesions or plaques in this area.

Other pairs did not attain this depth of explanation, but all of them at least somewhat deepened their recorded explanations from Session 1 by writing about more factors and interactions. Table 4 shows a rating of each pair's explanations in the two deliverables.

In summary, our field study findings reveal that complex, multivariate relationships shaped and directed students' intentions, choices about next moves, iterations, adaptations to unanticipated outcomes, and content in deliverables. Results confirm much that has been found in current IR research about information-seeking behaviors and add new dimensions and qualifications. They offer a more complete picture of this particular PubMed exploratory workflow in its naturalistic setting. Contextually grounded, our results reveal the functional roles that users' tasks served in working toward overall and stage-specific goals. If tools are to accommodate users' needs it is these functional roles that tools need to target. Functional tasks are best represented as an action *and* its outcome in relation to a cognitive objective (Jansen, Booth, & Smith, 2009). Our results contribute to this representation.

TABLE 4. Level of explanatory depth in deliverable explanations by each pair.

Pair ID	Deliverable 1	Deliverable 2
1	Deep	Deep
2	Surface	Deepened
3	Surface	Deepened
4	Surface	Deepened
5	Surface	Surface
6	Surface	Deepened
7	Medium	Deep

Discussion: Implications for Tool Support and Their Contributions to Current Research

In this section, we examine the implications of field study findings for improvements in tool support and examine ways in which findings confirm, complement, and qualify improvements already proposed in the research literature. Based on our results, we discuss implications for tool improvements for obstacles in (a) formulating queries, (b) getting to relevance in retrieved results, and (c) reading for meaning with a coherent train of thought. In discussing each area in light of our cases we first suggest relevant operation-level improvements and then propose improved support for the higher order reasoning and behaviors. As an advance caveat, some PubMed features already address obstacles that users encountered, but users did not access them. Our findings, reinforced by other studies, suggest that sometimes PubMed features may not have jived with users' habitual and preferred approaches (e.g., copying and pasting citations). At other times, the features may rest on search system/information science notations unfamiliar to the students. Students may have chosen to forgo learning these conventions in favor of devoting their mental energies toward gaining and organizing relevant new knowledge for problem solving.

Formulating Queries: Obstacles and Proposed Improvements

Program operation level obstacles and improvements. As in usage logs, field users ran into difficulties in spelling and bibliographic formatting (Lin & Wilbur, 2009). Additionally, they spent a good deal of cumulative time trying to decide if entering a P2X-prefix term would retrieve articles related to a gene, receptor, or both—and wondering how to distinguish between the two in the retrieved titles. For misspellings, current research proposes autospelling corrections; for problems with bibliographic formatting it proposes citation builders; for term ambiguity it suggests autocompleted terms accompanied by their conceptual categories (e.g., gene, disease, process; Eaton, 2006; Herskovic et al., 2007, Ramampiearo & Li, 2011; Wilbur, Kim, & Xie, 2007). For our users, autocorrection to spelling would have increased querying efficiency as long as the program had given students control over when to launch the query. For citation builders, our results add the qualify that this proposed improvement should let users copy and paste without the need to reformat. Finally, for conceptual categories within autocompletion our results indicate that this feature indeed may have been an efficient way to provide the authoritative

DOI: 10.1002/asi

categorization that our users wanted and looked for (e.g., in MeSH terms and Science Direct facets).

Obstacles and improvements for higher-order reasoning and behaviors. Users' higher-order obstacles in formulated queries pertained to writing Boolean expressions, matching query terms to problems at various stages of episodes, and chaining by citation. When usage log researchers note obstacles in Boolean expression writing they note that five or more terms commonly leads to null results (Lin & Wilbur, 2009). This was not our students' problem. Rather, null results largely came from two- or three-term expressions that contained mutually exclusive terms. Current solutions for problems tied to writing Boolean expressions—both in the research literature and implemented in some MEDLINE IR tools, including PubMed—commonly take the form of Boolean expression builders (Ding, Hughes, Berleant Fulmer, & Wurtele, 2006; National Library of Medicine, 2011). Based on our users' experiences, Boolean expression builders may be more useful later in an episode after students had filled out their spotty domain knowledge. Early in an episode, students did not seem to want to do more than write simple two-term, AND queries. This is not to say that they, at times, might not have benefited from writing more complicated queries, but to write more complicated queries they needed more domain knowledge. Compound expressions—when our users wrote them—only came later in an episode, after the students knew more from their reading about a disease and P2X associations. Boolean expression builders help with syntax not semantics. Support for acquiring the underlying conceptual knowledge that students needed for writing optimally productive query statements may have been better addressed through features that let users filter by conceptual facets or clusters (discussed later). After gaining a certain level of relational knowledge about terms, users may then benefit from Boolean expression builders to go beyond simple queries. At present, however, little is known about when and if—at a certain level of conceptual understanding—users may prefer complex Boolean statements to faceted searching or vice versa.

A second higher-order obstacle to getting to relevance in results was users' uncertainty about whether their query matched their problem, as reflected in one pair's Googling for "–itis." Current research proposes query expansion features as a solution (Dogan et al, 2009; Matos, Arrais, Maia-Rodrigues, & Oliveira, 2010). Our findings suggest that query expansion like Boolean expression builders would be most helpful later in a workflow, after users familiarize themselves with relevant genes, diseases, concepts, and biological events through reading. In other studies, nonexperts in a domain often chose inappropriate suggestions from query expansion features and had trouble distinguishing good from poor terms (Xu & Croft, 2000). Query expansion and the operational support provided by autocompletion's term tags could perhaps work together incrementally. Our students at first only sought to categorize terms for clarity in relevance criteria, and autocompletion could have helped. Query expansion—introduced later as a more comprehensive supplement—might have helped them construct knowledge beyond what might immediately occur to them.

Finally, in chaining by citation—an often lauded surrogate for formulating queries—unanticipated inefficiencies occurred (Bates, 1999; Lin & Wilbur, 2009). Chaining did not preclude tendencies to disregard cues other than titles, to repeatedly read articles by the same authors, or to grapple with ambiguous relevance criteria. Solutions for these issues pertain to improvements proposed in the next section on getting to relevance.

Getting to Relevance: Obstacles and Proposed Improvements

Program operation level obstacles and improvements. In usage log studies and in our research, users spent a long time between issuing a query and clicking a title and between skimming abstracts or articles and actually finding a text that was relevant (Lin & Wilbur, 2009). At a program operation level, the research literature proposes two tool improvements to expedite navigating to relevant items that we also strongly advocate for the field users. They are (a) to provide many options for ordering result displays beyond chronology, especially the option of relevance ranking, and (b) to provide mechanisms for filtering results to get quickly to indicators of relevance, for example, through facets, clusters, tables, or tabs (Hearst, 1999; Lu, 2011; Muin, Fontelo, Lie, & Ackerman, 2005; Ramampiearo & Li, 2011; Sarkar, Schenk, Miller & Norton, 2009; Tang, 2006; Xuan et al., 2010). PubMed's chronological ordering of results gave quick access to the most recent research first, but its lack of relevance ranking was an obstacle. Students likely could have benefited from results ordered by relevance as an effortless means to enrich their impoverished strategy of relying almost exclusively on titles for relevance cues. For obstacles in getting to relevance, our users operational (navigational) and higher order reasoning needs were tightly coupled; we now discuss the latter.

Obstacles and improvements for higher-order reasoning and behaviors. Result ranking and filtering through predefined categories or clusters are common solutions for helping users navigate and assimilate conceptual knowledge as they skim listed or grouped results. Unfortunately, facets and clusters—common features in other search systems—have scarcely been implemented or user-tested in context for MEDLINE IR systems and not at all for iterative problem solving by domain novices (Leroy, Xu, Chung, Eggers, & Chen, 2007; Wilson, Schraefel, & White, 2009). Some MEDLINE-related studies have shown that users preferred categorical presentations of results to clusters, but it has yet to be determined what category names and aggregations align best with users' domain knowledge and problem-based interests vis-á-vis MEDLINE literature and what type and amount of user control is optimally useful (Hearst, 1999; Lin

DOI: 10.1002/asi

et al., 2008; Pratt & Fagan, 2000; Quinones, Su, Marshall, Eggers, & Chen, 2007).

Toward the same end of supporting users' higher-order needs for judging relevance from result displays, current research proposes presenting extra conceptual information with results, for example, information on genes, diseases, gene ontology terms, and/or pathways discussed in the article and visualized conceptual similarities among related articles and between social media metadata (Kilicoglu et al., 2008; Lin et al., 2008; Lu, 2011). Extra information also includes similarities between articles based on indirect as well as direct associations between genes in the articles (Pratt & Yetisgen-Yildiz, 2003; Smalheiser, Torvik, & Zhou, 2009). Our field results suggest potential merits in this solution, but with several qualifications. As with Boolean-expression builders and query expansion, the value of extra conceptual information for domain novices ultimately turns on when in the workflow it is given—that is, at the point when users are ready for particular information—and how much of it to give at once. Field study results suggest that extra conceptual information may need to be distributed across more displays than screens of results and that information might need to be redundant. As action trails showed, students relied heavily on result lists because they did not readily find knowledge or relevance cues that "spoke" to them elsewhere. Giving such users, more information from the start is no guarantee that it will capture their attention. It could be that field users did not pay attention to other metadata (dates, authors, journals) because they were not familiar enough with the subspecialty to know the significance or relevance that cues connoted. As results show, they did not seem to have the background to draw meanings from sets of MeSH terms combined. Other field studies have shown that users who are not domain experts often avoid additional information, and when they use it they do not necessarily make better decisions and or correct ones (Jansen & Rieh, 2010).

Our results suggest two issues to consider in tool design in regard to when, what, and how much extra information to provide. The first consideration is domain novices' tendency to decline in analysis performance after reading texts due to the cognitive demands of organizing previously unfamiliar knowledge into meanings for problem solving (Patel & Kaufman, 2006). Field study users experienced this downturn and worked through it by reading texts more closely. The second consideration is cognitive style. Users who rapidly scanned pursued breadth over depth; for them being presented with extra conceptual information may have inadvertently promoted their tendency to skim the surface now of yet more information. They would have achieved more coverage quickly, but not necessarily more depth. It is a challenge to provide additional conceptual information that is strategically attuned to users' cumulative learning and knowledge at various episodes and stages. Based on information our field users sought, we suggest limiting additional conceptual information early in an episode to genes (and synonyms), diseases and overrepresented GO terms (biological processes and molecular functions). Later in the workflow, after having acquired more understanding of the problem space, users may be ready to use other additional information, including social metadata, co-occurring concepts, and causal or indirect relationships. Once our users realized that P2X was not a direct cause in disease influences, they were ready to learn about indirect relationships.

Reading for Meaning: Obstacles and Proposed Improvements

Program operation level obstacles and improvements. Field study results underscore that read/extract and record stages played pivotal roles in users efforts to clarify and judge relevance. Read and record also were central to filling out causal relationships. Usage log research has shown that users spent relatively short amounts of time reading full texts, for example, 18% of a total session. In our cases, students spent far longer (Herskovic et al., 2007). Unfortunately, the current IR literature focuses less on tool improvements for meaning-making through iterative reading, recording, and verifying than it does on getting to relevant results.

In the relatively scant research that addresses IR support for meaning making, researchers have noted that one obstacle is disrupted trains of thought when users move between a query-and-retrieval system and other systems, such as full text display systems or link-outs (Blake & Pratt, 2006b; Nicholas, Huntington, Jamali, & Watkinson, 2006). Researchers believe that disruptions occur because users perceive and have difficulty dealing with inconsistences between site structures or with discontinuities between conceptual cues in the two different systems Users in our study experienced such disruptions, for example, getting a conference proceedings table of contents or getting lost after returning from a link-out to eight open tabs of information. For operation-level support, tool improvements could include features for annotating texts to facilitate recall, interactive history lists, and a means for excerpting textual elements of interest and tying them (hyperlinking) to their sources (Bishop, 1999). Also, search tools and text display systems could give the impression of integration by repeating the same conceptual information across systems, for example, relevant conceptual facets from the gene ontology and ontologies related to pathways and diseases (Blake & Pratt, 2006; Makri et al., 2008; Sandusky & Tenopir, 2007). As Komlodi and Soergel note (2002), only a handful of current IR researchers highlight the need to support users in staying oriented.

Obstacles for improvement for higher-order reasoning and cognition. At higher cognitive levels, the textual components of full texts play multiple roles in meaning-making (Bishop, 1999). As found in other IR research, our users read nonlinearly and depended on section headings to find information that could help in constructing relevant meaning. To our knowledge, little current research addresses tool improvements for such obstacles as incomplete "backwards" reading during power browsing or difficulties in matching thematic review sections to problem goals.

DOI: 10.1002/asi

Additionally, students needed and recognized a need for preunderstanding knowledge, which is a vital dimension of productive searches by students and domain novices (Koslowski et al., 2008). Students linked out for this knowledge and gravitated toward review articles. Some of them, however, did not transition from acquiring preunderstanding knowledge to actively constructing interpretive knowledge. They preferred to spend time reading Wikipedia or review articles rather than primary research articles—despite being experienced in the latter from prior courses and jobs. As Koslowski et al. (2008) argue, novice scientists often need overt cues to prompt them to gain preunderstanding knowledge from different, more appropriate sources. To our knowledge, current proposed improvements do not address this need.

Improvements proposed in current research focus mostly on presenting information derived from Natural Language Processing about biological events or claims in an article, perhaps as advanced organizers in the full-text displays (PubMed Assistant, http://metnet.vrac.iastate.edu/browser) (Agarawal & Yu, 2009; Blake, 2010; Cohen et al., 2008; Lin et al., 2009). As with other improvements, it is hard to assess definitively the helpfulness of such information since our users did not have it. The issue—as with query expansion and additional conceptual information in result displays—is for designers to better understand *when* and *what* information is right for domain novices' needs and what overreaches their abilities to process meanings at certain points in information seeking. More needs to be known about the trigger information that can help domain novices construct explanations of disease mechanisms and—referring back to Jansen et al. (2009)—the functional tasks domain novices apply.

Conclusions

Our study reinforces many results found in PubMed usage log research, in field studies of other domains, and in usability assessments of similar IR systems. For example, our study confirms that query and retrieve are core activities; that inefficiencies often occur that could be reduced through tool improvements; that users perform stages and many constituent cognitive tasks in common; and that they iterate a great deal between stages, often due to uncertainties. Our study adds to current research about commonalities among users as well, demonstrating that reading, recording, and verifying are just as central (cognitively) as query and retrieve for getting to and judging relevance. Amid commonalities, the contextual data of our study also suggest that data in logs demonstrating frequent and long times in query and retrieve actions may actually reflect technology effects more than generic processes of knowledge construction. In our cases, users returned with great frequency to query and retrieve actions because cues for relevance that "spoke" to them were not adequate or adequately distributed across screens.

Our results also contribute a better understanding of variations among users that affect workflow moves, strategies, and outcomes. We characterize six indicators of cognitive search styles and describe their effects. Our evidence also suggests that having more (or less) domain-specific background knowledge strongly affected users' moves, strategies, and outcomes; however, we do not have firm indicators or measures of our users' prior knowledge. Our PubMed uses-in-context also reveal that action trails that would receive the same coding in log analysis varied in motivations, in actions outside of "log radar" and, importantly, in users' needs for tool support.

Operationally and cognitively, the field study users experienced obstacles that interfered with information-seeking efficiency and effectiveness. Tool improvements could have helped. Some improvements proposed in the current IR research literature seem applicable to our users. Others need to be qualified. Yet other areas, such as support for reading for meaning, need to be more fully researched. Qualifications that our study suggest largely center on the need to attune choices of features and functions to users' cognitive needs at specific stages of their cumulative workflow to direct when to give information, what and how much information to offer, how to categorize it, how to choose between or combine features when they address similar needs, and what control to give users. Our findings on undergraduate biology majors underscore the challenge of these qualifications, for example, attuning features and calibrating presentations to the help students needed for figuring out what counts as relevant; what constitutes a cause for their recorded explanations, especially at a complex level; and when rapid scanning—"high efficiency"—may run counter to effective understanding. Given the wide range of tool improvements that are proposed in current research or implemented in recent innovations, technical feasibility is not the bottleneck at present. Technological solutions are proceeding at a faster pace than our understanding of how to best implement them. To reach this understanding more field studies, quantitative analyses, and mixed methods research are needed, and they need to build on each other with a focus on functional tasks in exploratory information seeking.

Acknowledgments

This research was funded by two NIH grants: 1 RO1LM009812-01A2 and U54 DA021519.

References

Agarwal S., & Yu, H. (2009). Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. Bioinformatics, 25, 3174–3180.

Anderson, T.D. (2006). Uncertainty in action: Observing information seeking within the creative process of scholarly research. Information Research, 12. Retrieved from http://informationr.net/ir/12-1/paper283.html

Barton, E. (2004). Linguistic discourse analysis: how the language in texts works. In C. Bazerman (Ed.), What writing does and how it does it: An introduction to analyzing texts and textual practices (pp. 57–82). Mahwah, NJ: Erlbaum.

DOI: 10.1002/asi

Bates, M. (1990). Where should the person stop and the information search interface start? Information Processing & Management, 26(5), 575–591.
Bates, M.J. (1999). The invisible substrate of information science. Journal of the American Society for Information Science, 50, 1043–1050.
Bishop, A.P. (1999). Document structure and digital libraries: How researchers mobilize information in journal articles. Information Processing and Management, 35, 255–279.
Blake, C. (2010). Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. Journal of Biomedical Informatics, 43, 173–189.
Blake, C., & Pratt, W. (2006a). Collaborative information synthesis I: A model of information behaviors of scientists in medicine and public health. Journal of the American Society for Information Science and Technology, 57, 1740–1749.
Blake, C., & Pratt, W. (2006b). Collaborative information synthesis II: Recommendations for information systems to support synthesis activities. Journal of the American Society for Information Science and Technology, 57, 1888–1895.
Chowdury, S., Gibb, F., & Landoni, M. (2011). Uncertainty in information seeking and retrieval: A study in an academic environment. Information Processing and Management, 47, 157–175.
Cohen, K.B., Vespoor, K., Johnson, H.L., Roeder, C., Ogren, P.V., Baumgartner, W.A., Jr., . . . Hunter, L. (2008). High-precision biological event extraction: Effects of system and data. Computational Intelligence, 27, 681–701.
Cresswell, J., & Plano Clark, V. (2006). Designing and conducting mixed methods research. Thousand Oaks, CA: Sage.
Ding, J., Hughes, L., Berleant, D., Fulmer, A., & Wurtele, E. (2006). PubMed Assistant: A biologist-friendly interface for enhanded PubMed search. Bioinformatics, 22, 378–380.
Dogan, R., Murray, G. C., Neveol, A., & Lu, Z. (2009). Understanding PubMed user search behavior through log analysis. Database, 2009, bap018.
Eaton, A. (2006). HubMed:Aweb-based biomedical literature search interface. Nucleic Acids Research, 34, W745–W747.
Ellis, D. (1989). A behavioural approach to information retrieval system design. Journal of Documentation 45, 171–212.
Ellis, D., & Haugan, M. (1997). Modeling the information seeking patterns of engineers and research scientists in an industrial environment. Journal of Documentation, 53, 384–403.
Ellis, D., Cox, D., & Hall, K. (1993). A comparison of the information seeking patterns of researchers in the physical and social sciences. Journal of Documentation, 49, 356–369.
Gehring, K., & Eastman, D. (2008). Information fluency for undergraduate biology majors: Applications of inquiry based learning in a developmental biology course. CBE–Life Sciences Education, 7, 54–63.
Hearst, M. (1999). The use of categories and clusters for organizing retrieval results. In T. Strzalkowski (Ed.), Natural language information retrieval (pp. 333–373). Dordrecht, The Netherlands: Kluwer.
Herskovic, J., Tanaka, L., Hersh, W., & Berstam, E.V. (2007). A day in the life of PubMed: Analysis of a typical day's query log. Journal of the American Medical Informatics Association, 14, 212–220.
Ingwersen, P., & Järvelin, K. (2005). The turn: Integration if information seeking and retrieval in context. Heidelberg, Germany: Springer.
Jansen, B.J., & Rieh, S.Y. (2010). The seventeen theoretical constructs of information searching and information retrieval. Journal of the American Society for Information Science and Technology, 61, 1517–1534.
Jansen, R., Booth, D., & Smith, B. (2009). Using the taxonomy of cognitive learning to model online searching. Information Processing and Management, 45, 643–663.
Kilicoglu, H., Fiszman, M., Rodriguez, A., Shin, D., Ripple, A.M., & Rindflesch, T.C. (2008). Semantic MEDLINE: A web application for managing the results of PubMed Searches. In Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008) (pp. 69–76). Turku, Finland: Turku Centre for Computer Science.
Komlodi, A., & Soergel, D. (2002). Attorneys interacting with legal information systems: Tools for mental model building and task integration. Proceedings of the American Society for Information Science and Technology, 39, 152–163.
Koslowski, B., Marasia, J., Chelenza, M., & Dublin, R. (2008). Information becomes evidence when an explanation can incorporate it into a causal framework. Cognitive Development, 23, 472–487.
Kuhlthau, C.C. (1993). Seeking meaning. A process approach to library and information services. Norwood, NJ: Ablex.
Kuhlthau, C.C. (1999). Accommodating the user's information search process: Challenges for information retrieval system designers. Bulletin of the American Society for Information Science, 25, 12–16.
Kumpulainen, S., & Järvelin, K. (2010). Information integration in molecular medicine: Integrated use of multiple channels. In Proceedings of the Third Symposium on Information Interaction in Context (IIIX '10) (pp. 95–104). New York: ACM Press.
Leckie, G.J., Pettigrew, K.E., & Sylvain, C. (1996). Modeling the information seeking of professionals: A general model derived from reseach on engineers, health care professions, and lawyers. Library Quarterly, 66, 161–193.
Leroy, G., Xu, J., Chung, W., Eggers, S., & Chen, H. (2007). An end user evaluation of query formulation and results review tools in three medical meta-search engines. International Journal of Medical Informatics, 76, 780–789.
Li, Y., & Belkin, N.J. (2010). An exploration of the relationships between work task and interactive information search behavior. Journal of the American Society for Information Science, 61, 1771–1789.
Lin, J., DiCuccio, M., Grigorya, V., & Wilbur W.J. (2008). Navigating information spaces: A case study of related article search in PubMed. Information Processing and Management, 44, 1771–1783.
Lin, J., & Wilbur W.J. (2009). Modeling actions of PubMed users with n-gram language models. Information Retrieval, 12, 487–503.
Lu, A. (2011). PubMed and beyond: A survey of web tools for searching biomedical literature. Database, 2011, baq036.
Makri, S., & Warwick, C. (2010) Information for inspiration: Understanding architects' information seeking and use behaviors to inform design. Journal of the American Society for Information Science and Technology, 61, 1745–1770.
Makri, S., Blandford, A., & Cox, A. (2008). Investigations the information-seeking behavior of academic lawyers: From Ellis's model to design. Information Processing and Management, 44, 613–634.
Matos, S., Arrais, J.P., Maia-Rodrigues, J., & Oliveira, J.L. (2010). Concept-based query expansion for retrieving gene related publications from MEDLINE. BMC Bioinformatics, 11, 212.
McCune, V., & Hounsell, D. (2005). The development of students' ways of thinking and practicing in threefinal-year biology courses. Higher Education, 49, 255–289.
Muin, M., Fontelo, P, Liu, F., & Acklerman, M. (2005). SLIM: An alternative web interface for MEDLINE/PubMed searches–a preliminary study. BMC Bioinformatics and Decision Making, 5, 37–46.
Nicholas, D., Huntington, P., Jamali, H.R., & Watkinson, A. (2006). The information seeking behavior of the users of digital scholarly journals. Information Processing & Management, 42, 1345–1365.
National Institutes of Health. (2011). National Library of Medicine (NLM) Technical Bulletin. PubMed advanced search page updated. Retrieved from http://www.nlm.nih.gov/pubs/techbull/nd11/nd11_pm_advanced_search.html
Patel, V., & Kaufman, D. (2006). Cognitive science and biomedical informatics. In E. Shortliffe & J.J. Cimino (Eds.), Biomedical informatics: Computer applications in health care and biomedicine (3rd ed.). New York: Springer-Verlag.
Patel, V., Kaufman, D., & Arocha, J. (2002). Emerging paradigms of cognition in medical decision-making. Journal of Biomedical Informatics, 35, 52–75.
Pratt, W., & Fagan, L. (2000). The usefulness of dynamically categorizing search results. Journal of American Medical Informatics Association, 7, 605–617
Pratt, W., & Yetisgen-Yildiz, M. (2003). LitLinker: Capturing connections across the biomedical literature. In Proceedings of the International

DOI: 10.1002/asi

Conference on Knowledge Capture (pp. 105–112). New York: ACM Press.

Quinones, K., Su, H., Marshall, B., Eggers, S., & Chen, H. (2007). User-centered evaluation of Arizona BioPathway: An information extraction, integraton, and visualization system. IEEE Transactions on Information Technology in Biomedicine, 11, 527–536.

Radlinski, R., Jurup, M., & Joachims, T. (2008). How does clickthrough data reflect information retrieval quality? In Proceedings of the 17th Annual ACM Conference on Information and Knowledge Management (CIKM'08) (pp. 73–82). New York: ACM Press.

Ramampiaro, H., & Li, C (2011). Supporting biomedical information retrieval: The BioTracer approach. In Proceedings of the Transactions on Large Scale Data and Knowledge Centered Systems (pp. 73–94). Heidelberg, Germany: Springer-Verlag.

Sandusky, R., & Tenopir, C. (2007). Finding and using journal article components: Impacts of disaggregation on teaching and research practices. Journal of the American Society for Information Science and Technology, 59, 970–982.

Sarkar, I., Schenk, R., Miller, H., & Norton, C. (2009). LigerCat: Using "MeSH clouds" from journal, article, or gene citations to facilitate the identification of relevant biomedical literature. American Medical Informatics Association Annual Symposium Proceedings, 2009, 563–567.

Smalheiser, N.R., Torvik, V.I., & Zhou, W. (2009). Arrowsmith two-node search interface: A tutorial on finding meaningful links between two disparate sets of articles in MEDLINE. Computer Methods and Programs in Biomedicine 94, 190–197. Retrieved from http://arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/start.cgi

Strauss, A., & Corbin, J. (1994) Grounded theory methodology: An overview. In N. Denzin & Y.S. Lincoln (Eds.), Handbook of qualitative research (pp. 273–285). Thousand Oaks, CA: Sage.

Tang, M-C. (2006). Browsing and searching in a faceted information space: A naturalistic study of PubMed users' interaction with a display tool. Journal of the American Society for Information Science and Technology 58, 1998–2006.

Teevan, J., Dumais, S., & Liebling, D. (2008). To personalize or not to personalize: Modeling queries with variation in user intent. Proceedings of the 31st Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '08) (pp. 163–170). New York: ACM.

Tenopir, C., Wang, P., Zhang, Y., Simmons, B., & Pollard R. (2006). Academic users' interactions with ScienceDirect in search tasks: Affective and cognitive behaviors. Information Processing and Management, 44, 105–121.

Toms, E.G. (2002). Information interaction: Providing a framework for information architecture. Journal of the American Society for Information Science and Technology, 53, 858–862.

White, R., & Morris, D. (2007). Investigating the querying and browsing behavior of advanced search engine users. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR '07) (pp. 255–262) New York: ACM Press.

Wilbur, W.J., Kim, W., & Xie, N. (2007). Spelling correction in the PubMed search engine. Information Retrieval Boston, 9(5), 543–564.

Wildemuth, B. (2006). Evidence-based practice in search interface design. Journal of the American Society for Information Science and Technology, 57, 825–828.

Wilson, M., Schraefel, M.C., & White, R. (2009). Evaluating advanced search interfaces using established information-seeking methods. Journal of the American Society for Information Science and Technology, 60, 1407–1422.

Xie, H.I. (2000). Shifts in interactive intentions and information-seeking strategies on interactive information retrieval. Journal of the American Society of Information Science, 51, 841–857.

Xu, J., & Croft, W.B. (2000). Improving the effectiveness of information retrieval with local context analysis. ACM Transactions on Information Systems, 18, 70–112.

Xuan, W., Dai, M., Buckner, J., Mirel, B., Song, J., Athey, B., . . . Meng, F. (2010). Cross-domain neurobiology data integration and exploration. BMC Genomics, 11(Suppl 3), S6.

DOI: 10.1002/asi