Exploring Knowledge Learning in Collaborative Information Seeking Process

Yu Chi
University of Pittsburgh
135 North Bellefield Ave.
Pittsburgh, PA, 15260
yuc73@pitt.edu

Shuguang Han
University of Pittsburgh
135 North Bellefield Ave.
Pittsburgh, PA, 15260
shh69@pitt.edu

Daqing He
University of Pittsburgh
135 North Bellefield Ave.
Pittsburgh, PA, 15260
dah44@pitt.edu

Rui Meng
University of Pittsburgh
135 North Bellefield Ave.
Pittsburgh, PA, 15260
rum20@pitt.edu

ABSTRACT
Knowledge learning is recognized as an important component in people's search process. Existing studies on this topic usually measure the knowledge growth before and after a search. However, there still lacks a fine-grained understanding of users' knowledge change patterns within a search process and users' adoption of different sources for learning. In this on-going project, we are exploring answers to both questions in collaborative information seeking (CIS) since the CIS tasks are usually exploratory, which triggers learning, and involve diverse learning resources such as self-explored search content, partners' search content and explicit communication between them. Through analyzing the data from a controlled laboratory user study with both collaborative and individual information seeking conditions, we demonstrated that users' knowledge keeps growing in both conditions, but they issue significantly more diverse queries in the collaborative condition. Our analysis of users' queries also revealed that the adoption of different learning resources varies at different information seeking stages, and the adoption is influenced by the nature of search tasks too. Finally, we propose several insights for system design to enhance knowledge learning in collaborative information seeking process.

CCS Concepts
• **Information systems → Information retrieval → Users and interactive retrieval → Collaborative search**

Keywords
Searching as learning; collaborative information seeking process; knowledge learning

1. INTRODUCTION
Though gained plenty of attentions recently, examining searching as a learning process is not a new topic in information retrieval. It has long been aware that knowledge acquisition is an important component in information seeking process. As stated in ASK (Anomalous State of Knowledge) model [1], Belkin argued that information seeking is a process to resolve the anomaly between users' current states of knowledge the problem they faced. Marchionini [5] claimed that beyond simple lookup search, people often engage in exploratory search tasks where learning and investigation could play essential roles.

Besides the aforementioned theoretical models, empirical studies also provided substantial evidence that suggests learning to be a very common phenomenon in people's search process [3, 10, 12]. Rieh [7] further identified two roles of learning in a search process – *learning to search* and *searching to learn*, where the former refers to how people learn search experience and expertise while the latter regards learning as a byproduct of search and ends with relevant knowledge increased. This on-going project focuses on the latter role and is interested in studying how users gain domain knowledge and how the knowledge affects follow-up search behaviors such as term selection and search tactics.

Prior related studies in the literature examined users' knowledge learning in both long-term and short-term periods. Vakkari [10] and Wildemuth [12] explored students' learning and searching activities during a course that lasted for several months. Recent studies [2, 3] discovered that knowledge learning can also occur in a single short-term search session. In these studies, how to measure knowledge growth was identified as one crucial challenge, and a commonly-adopted method was to survey user knowledge with questionnaires [13]. However, this approach heavily depends on the effectiveness of the questionnaire and the accuracy of self-reported knowledge levels. Several other studies [3, 15] regarded users' knowledge as a function of users' search behaviors and thus the knowledge change can be implicitly reflected by the changes of users' search behaviors.

To the best of our knowledge, there is little, if any, research investigating how people learn knowledge in collaborative information seeking (CIS) process. We believe this is an important topic for two reasons. Firstly, as Shah [9] pointed out, CIS tasks are usually complex and exploratory in nature. Individual users often possess insufficient knowledge or skills for solving the task. This triggers them to engage in CIS. Through multiple interactions among team members and with a CIS system, users would learn knowledge to address the task via collaboration. Secondly, comparing to an individual search process, users in CIS are provided with richer information sources for their knowledge gain. Beyond learning from one's own search, a user can also directly or indirectly communicate with and learn from the partners. Consequently, it is important to understand people's knowledge learning from different sources in CIS so that better CIS interfaces and systems for enhancing knowledge learning can be designed.

To summarize, in order to investigate the knowledge learning process in CIS, we attempt to study two research questions:

Searching as Learning (SAL), July 21, 2016, Pisa, Italy.

- RQ1: How do users gain knowledge in a collaborative information seeking process, and how does it differ to the knowledge learning in individual search (more in Section §3)?
- RQ2: How do users adopt different learning sources in CIS, and how can different tasks affect the adoption (more in Section §4)?

2. Obtaining Data Collection

2.1 User Study Dataset

To investigate the above questions, we adopted the user study data from previous research by Yue et al. [14]. We choose this dataset for three reasons. Firstly, it includes both individual and collaborative information seeking conditions. This enables us to examine and to compare the knowledge learning in both scenarios, which can help us to answer our RQ1. Secondly, the search system used in [14] provides functionalities for easy accessing to different resources. The CIS system screenshot can be illustrated in Figure 1, which consists of four components – the chat panel (see Area 1), the topic statement panel (see Area 2), the team workspace panel (see Area 3) and the web search panel (see Area 4). The chat panel is always displayed on the left side of the screen in the CIS condition and facilitates the collaborative searchers to directly communicate with each other by sending instant messages. On the remaining right side of the system, participants can switch between the other three panels at any time. Topic statement panel presents the task description. Web search panel consists of a Google search page and a search history list. Besides, participants can view the documents either saved by themselves or by their partners in the team workspace panel. The IIS condition adopted the same system except that the chat panel is hidden, and the workspace is accessible to only one searcher. This system with multiple functions allows us to distinguish their knowledge learning in terms of different sources (i.e., RQ2).

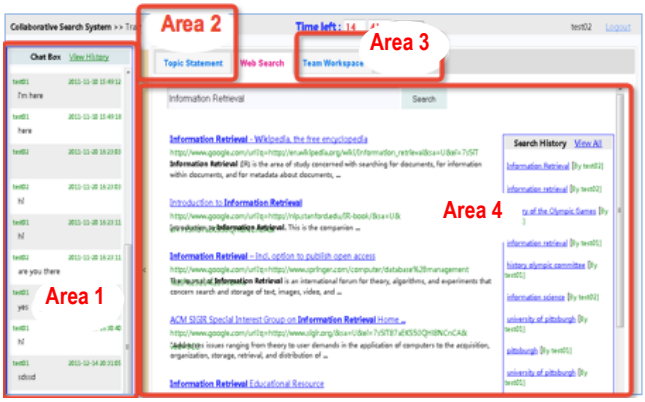

Figure 1: Collaborative information seeking system screenshot

Thirdly, their system logged detailed users' search behaviors for the whole sessions. This rich data helps us probe into each step of their search processes to obtain a fine-grained understanding of users' knowledge learning.

Despite that we borrowed data collection and search tasks from Yue et al. [14], our research focus and research questions are significantly different to theirs. They focused on examining search patterns and using HMM to model such patterns in CIS, whereas we concentrated on knowledge learning in CIS.

In summary, the data collection consists of the search logs of 54 university students. Among them, 18 are individual searchers and 36 participants (18 pairs) for collaborative search. In total, there are complete logs for 108 search sessions, in which 36 sessions (i.e., 18 users × 2 tasks/user) are for individual searches and 72 sessions (i.e., 18 pairs × 2 users/team × 2 tasks/user) are for collaborative searches.

2.2 Task Description

By reusing Yue et al. [14]'s search log data collection, we inherited two search tasks in their study too. The first task is an *information-gathering task (T1)*, where the participants were asked to collect information for a report on the effect of social networking services. This is a recall-oriented task. The second one is a *decision-making task (T2)*, which asked the participants to collect information for planning a trip to Helsinki. This one expects the participants to negotiate with their partners to make joint decisions.

We pay particular attention to different task types for two reasons. One is that topic knowledge change is found to be affected by task type in individual search [6]. But it is unknown if this affection also occurs in collaborative information seeking tasks. Besides, we are curious about whether these two specific tasks designed as different chat-intensive levels would affect the searchers' adoption of learning sources. For example, we expect that people are more likely to communicate and learn from their partners in the decision-making task while they probably gain more knowledge from self-exploration in the information gathering task.

3. KNOWLEDGE LEARNING IN CIS & IIS

3.1 Implicit Measure of Knowledge

Due to the difficulty of direct measuring of knowledge [13], recent studies [3, 15] proposed to utilize implicit behavioral measures such as query complexity to reflect users' knowledge differences. This method is usually based on two assumptions: with the increase of a user's knowledge, she would be likely to either (1) click and view more authoritative websites [11] or (2) use more domain-specific and diverse vocabulary in queries [12, 10]. Previous studies on this topic often defined domain-specific authoritative websites and vocabularies within a specific domain (e.g., medicine, psychology). However, the authoritative websites and vocabularies in an open domain like in our tasks are hard to acquire for lack of existing knowledge resources.

To build proper "authoritative websites" and "domain-specific vocabularies" for open domain tasks, we explore the idea about the *likelihood of discovery* proposed by Shah [8], which was developed to evaluate participant's ability to discover hard-to-find information. We believe that some documents/queries are *easy-to-be-found* among most users while some others require a higher level of users' knowledge to be clicked/issued (thus they are *hard-to-be-found*). A person with more knowledge about a task has higher probabilities to recognize and click those hard-to-be-found webpages and issue more specific queries. Therefore, we link document/query's required knowledge with its *findability*. Specifically, we define click complexity and query complexity to measure the knowledge required to reach the clicked webpages and the queries.

Formally, for each clicked document d_j, its click complexity $C(d_j)$ is calculated by Formula (1), where N is the total number of participants, and n_{dj} denotes the number of participants who clicked d_j. We name $C(d_j)$ as the *click complexity*. Here, we are only interested in the clicked documents that are also relevant, in

which the clicked documents with users' post-task rating as non-relevant are removed.

$$C(d_j) = \log \frac{N}{n_{d_j}} \quad (1)$$

We define query complexity in the same manner except changing n_{dj} to n_{qj} that denotes the number of participants who issued query q_j. Note that we treat two queries as the same if they are exactly matched after stemming and stop word removal. Alternative query complexity measures such as query length (i.e., #terms in a query) were also adopted to evaluate the learning in a search process [3]. However, due to the lack of an enough amount of queries in a user study dataset, we decided to utilize the above metric (as shown in Formula 1) through a pooling of all users' search queries, instead of measuring the number of terms for each unique query.

3.2 Results and Discussion

After computing the query complexity and click complexity for both collaborative and individual conditions, we plot the values over six evenly divided search stages averaged over all sessions, and each stage represents 5 minutes' search during the whole task.

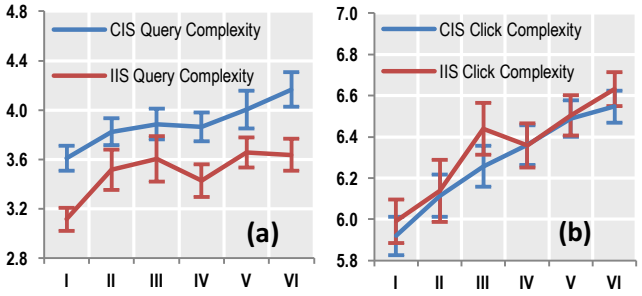

Figure 2: Query (a) and click-through (b) complexity in CIS and IIS over different search stages.

Query Complexity. Figure 2 (a) provides an overall change of query complexity in different search conditions and across different search stages. We can see that user knowledge starts from a relatively low level at the beginning and keeps growing during her seeking process in both individual and collaborative search conditions. Since our data is not normally distributed, we perform Wilcoxon test to examine the significance between different search stages within each condition and Mann-Whitney test to compare significance between CIS and IIS.

We find that the query complexity at the first stage of both conditions is significantly lower than the rest of the following stages, indicating that users indeed searched more specific and unique queries with time goes by.

Comparing between the two conditions, users in CIS issued more complex queries in each stage than the individual searchers. The results show that the query complexity in CIS is significantly higher than IIS in stage I, II, IV, VI. This indicates that CIS which includes partners could provide richer learning sources thus further enables the users to generate more diverse queries. This result triggered us to examine how the users adopt different learning sources to issue queries in CIS in Section §4.

Click Complexity. The results of click complexity are plotted in Figure 2 (b). Similar to the query complexity, click complexity shows an overall trend of increasing over different search stages in both CIS and IIS. However, comparing between CIS and IIS, the statistical test shows no significant difference between the two conditions in all stages, which differs from the results of query complexity. This might indicate that searching collaboratively may help directly on generating diverse or difficult queries, but its impacts on finding and clicking complex documents might be limited. There are several possible explanations of this insufficiency: that collaborative searchers failed to share the knowledge in their clicked documents thus members in a group kept clicking duplicate documents, or that even the collaborative searchers issued more diverse and specific queries, the documents returned were heavily affected by the search system. We would like to explore further on this topic in the future.

4. LEARNING SOURCE IN CIS

This section plans to work on results related to RQ2, which examines the information sources where the users learn their knowledge. Particularly the results presented in Section 3 highlight two motivations for us to study this. Firstly, although the participants in CIS and IIS were given the same set of exploratory tasks in our study, their query complexity was higher in CIS, which indicates a higher learning outcome in CIS. This is most probably due to the richness of the sources involved in the CIS. Secondly, despite the benefits (e.g., higher knowledge gain) of CIS involved, it also requires users to spend more time to communicate and negotiate with each other so that it usually brings higher cognition load. A better understanding of learning sources in CIS may help us design a better user interface that can enhance users' learning process. Additionally, task type is often thought as an important factor in studying information seeking behaviors [6]. We are curious about how the task type can affect users' adoption of learning sources.

4.1 Content Analysis of Learning Sources

Since query complexity is significantly higher in collaborative information seeking condition comparing to individual search, we focused on analyzing search queries and regarded it as the explicit reflection of the knowledge learning trace.

We drew upon content analysis as a methodology to examine the source of each query. This is because the research technique is widely used to reveal meaningful information from the textual content and applicable to our data set which includes plenty of colloquial chat message. Besides, this method with intelligent human judgment allows us to understand the semantic meaning in each piece of information, beyond computing the similarities or matching the exact terms [16] between the current query and the previous actions.

Since existing theory and research literature on learning sources in CIS is limited, conventional content analysis [4], an inductive process, was employed to establish the coding scheme. This method requires the researchers to first immerse themselves in the data to come up with the initial categories. To begin with, 5 teams' search logs (i.e., 20 sessions=5 pairs × 2 users/team × 2 tasks/user) among all the 18 teams were randomly selected. The query was treated as the analysis unit, and for each query, we manually examined its content and all search records before it (including the topic statement, clicks, queries, and chat content) to judge where the user obtained this query and the terms in it. Each query (217 queries in total) was coded by two researchers. At last, four overarching categories emerged, namely learn from self, learn from collaborator, learn from task description and learn from prior knowledge. With this coding scheme, two coders' inter-rater reliability on the 5 teams' logs is acceptable (Cohen's kappa=.66).

We then annotated all the queries (697 queries in total) using the four categories as the codes. The descriptions and examples of the four categories and are:

Learn from Self (LS)*:* The query is generated based on the user's own existing search histories, which include their own previous queries in the session, clicked documents, or search result pages. For example, after clicking document D1 "Negative impact of social networking websites", a result page of query Q1"social networking impact", S4 submitted query Q2 "social networking impact, pros/cons". Therefore, we annotated Q2 as LS.

Learn from Collaborator (LC)*:* Basically, there are two ways for the collaborators to communicate with each other: sending an instant message on the chat panel (Area 1 in Figure 1) or reading related documents shared by the partner in the team workspace (Area 3 in Figure 1). When the query is generated from either of the two, it is treated as LC. LC is a unique learning source in CIS. For instance, S20 shared personal knowledge about "Steubenville sexual assault and social media impact" to her teammate S21 through chat, and S21 started to search for related materials. In such case, we annotated S21's learning source as LC.

Learn from Task description (LT)*:* The user study [14] provided a detailed description for each task, and the topic statement panel (Area 2 in Figure 1) allowed the participants to view the current task description at any time during the search session; thus, participants could learn and select query terms directly from reading the task description.

Learn from Prior knowledge (LP): Participants can bring their own knowledge on the task; therefore, none of the terms in a query appeared in the task description or her and the partners' former search activities. In this case, we mark the query as LP. For example, S31's first query in T2 was Q1 "finland hockey league", while hockey league is not described in the task description nor raised by the partner. Therefore, we annotated S31 generated Q1 with LP. Note, for the queries that are not the first in the log, we carefully examine the records before to distinguish between LP and LS. Only if there is no evidence suggesting that the query is learned from self-search history will it be marked as LP.

4.2 Results and Discussion

The results of users' learning sources through a CIS process in the two tasks are presented respectively in Figure 3. We use area chart to visualize the portion of each learning source change over the whole 30-minute session, and each color denotes one source. The 30-minute search session is evenly divided into four stages based on the time since the data would be too sparse to present the trend if divided into six stages.

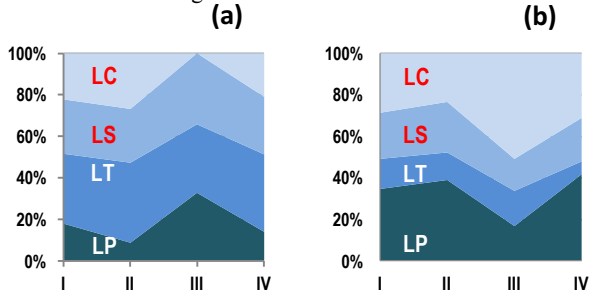

Figure 3: Learning sources across the CIS process in the information-gathering task (a) and decision-making task (b)

4.2.1 Learning in Information-gathering Task

According to Figure 3 (a), both LT and LS are consistently the top learning sources across the whole process. This indicates that users learn from the task description and their own search activities a lot in the information-gathering task. Learning through collaboration (i.e., LC) also plays an important role at the beginning (stages I and II) but not in later stages, particularly in stage III.

We think the results are probably due to the nature of the task, where participants care more about relevance and coverage of search topics presented in the task descriptions (i.e., LT) and consistently learn from their own search (i.e., LS). Under this task, team members tend to exchange their knowledge and conduct labor division at the beginning, and they chat and read each other' documents to check if the task is completed at the end, which may trigger them to issue new queries. That's why LC mainly occurs at the beginning and the end of a search task. Additionally, we find that users do not rely too much on prior knowledge (i.e., LP) at the beginning but start to bring their own knowledge in the third stage which is probably because more understanding of topic relevant document let them recall previous related knowledge.

4.2.2 Learning in Decision-making Task

Comparing to the information-gathering task, the decision-making task in Figure 3 (b) exhibits different patterns for the learning process. LC plays an extremely important role across the whole search process, and even increases at the last two stages. This is consistent with our expectation. Decision-making task requires users to negotiate with their partners and reach a final agreed conclusion so that people frequently interact with each other and obtain information. Notably, they heavily interact with each other in the third and fourth phases of a CIS process because these are the stages that they either need to exchange knowledge or make a decision. In LC, we did not further separate the *learning from chat* with the *learning from partners' search histories* (e.g., query, history), for which we will conduct more fine-grained analysis in the future.

Also, comparing to the information-gathering task, LP is more important in the decision-making task while LS and LT are less important. A possible explanation is that the participants select documents more based on their subjective judgments in this travel plan topic task, rather than only the relevance criteria described in the assigned task descriptions.

5. IMPLICATIONS FOR CIS SYSTEM DESIGN

CIS enables information seekers to solve complex and exploratory search tasks collaboratively. In addition, our study indicated that CIS users issue more diverse queries comparing to IIS, and their knowledge keeps growing during the whole search process. Although with the benefit of gaining more knowledge, CIS is often observed to introduce more cognitive loads [9]. One potential reason is the mismatch of the knowledge states among collaborated team members, particularly for the tasks that intrinsically require the team members to reach an agreement. Existing CIS systems, however, lack sufficient supports to such knowledge learning process. We think that two potential implications can be drawn from above findings for designing a better learning-enhanced CIS system.

5.1 Promoting Knowledge Understanding

In Section §3, we found that an overall trend is that knowledge increases across the whole search process in both CIS and IIS.

Particularly, compared to IIS, CIS users can generate more complex queries. We attribute such difference to the knowledge sharing among team members in CIS. However, we also discovered that there is almost no difference between CIS and IIS in click-through complexity. This might indicate that the current knowledge sharing support through accessing team's workspace and explicit communication is enough to generate proper queries, but is still unable to facilitate a truly understanding of certain knowledge in a clicked document. Therefore, a possible future direction could be exploring more support functions that not only aim to promote knowledge sharing but also knowledge understanding in the clicked documents. For example, proper information visualization techniques can be employed to summarize the knowledge states of the team members and/or the whole groups, or better awareness function can be developed for team members to know whether certain documents have been learned by their team members so that they do not need to visit duplicated documents.

5.2 Task-based Differentiation Support

Results from Section §4 demonstrated that task type can affect users' adoption of different learning sources during the CIS process. However, existing system attempts to differentiate search support for different tasks. We believe that CIS systems should facilitate users to understand and make sense of their shared information needs. For example, in tasks that require intensive communication as T2, the CIS system should assist users to acquire information from their partners especially in the final stage of the search process. Showing team members' behaviors as contextual information in the chat interface might be a helpful approach.

6. Conclusions and Future Work

In recent years, more and more evidence has shown that people learn knowledge in search [3, 10, 12]. While most of the existing studies remain focused on understanding how people learn individually, this paper aims to fill the gap where people learn collaboratively with their partners in a CIS process. Particularly, based on an existing dataset with 54 participants and both collaborative and individual search conditions, we studied how people learn their knowledge individually and collaboratively in search. We find that although user knowledge keeps increasing in both CIS and IIS process, there is a significant difference - CIS users tend to issue more diverse queries than IIS. Further analysis reveals that users in CIS adopt different learning sources, and the adoption also varies in different types of search tasks. Consequently, we propose several potential implications for CIS system to enhance the learning.

We do acknowledge several limitations of this study and plan to explore more in the future. Firstly, more measures of knowledge will be examined as evidence of learning. Secondly, only four types of learning sources were analyzed in this study. For instance, learning from chat content and learning from partners' saved documents are not separated. A more fine-grained analysis should be adopted for deeper understanding.

7. REFERENCES

[1] Belkin, N. (1980). Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, 5,133-143.

[2] Collins-Thompson, K., Rieh, S. Y., Haynes, C. C., & Syed, R. (2016, March). Assessing Learning Outcomes in Web Search: A Comparison of Tasks and Query Strategies. In *the 2016 ACM on Conference on Human Information Interaction and Retrieval* (pp. 163-172).

[3] Eickhoff, C., Teevan, J., White, R., & Dumais, S. (2014, February). Lessons from the journey: A query log analysis of within-session learning. In *the 7th ACM international conference on Web search and data mining* (pp. 223-232).

[4] Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. Qualitative health research, 15(9), 1277-1288.

[5] Marchionini, G. (2006). Exploratory search: from finding to understanding. *Communications of the ACM*, *49*(4), 41-46.

[6] Liu, J., Belkin, N. J., Zhang, X., & Yuan, X. (2013). Examining users' knowledge change in the task completion process. *Information Processing & Management*, *49*(5), 1058-1074.

[7] Rieh, S. Y., Collins-Thompson, K., Hansen, P., & Lee, H. J. (2016). Towards searching as a learning process: A review of current perspectives and future directions. *Journal of Information Science*, *42*(1), 19-34.

[8] Shah, C., & González-Ibáñez, R. (2011). Evaluating the synergic effect of collaboration in information seeking. In *the 34th international ACM SIGIR conference on Research and development in Information Retrieval* (913-922).

[9] Shah, C. (2014). Collaborative information seeking. *Journal of the Association for Information Science and Technology*, 65(2), 215-236.

[10] Vakkari, P., Pennanen, M., & Serola, S. (2003). Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Information processing & management*, *39*(3), 445-463.

[11] White, R. W., Dumais, S. T., & Teevan, J. (2009). Characterizing the influence of domain expertise on web search behavior. In *the Second ACM International Conference on Web Search and Data Mining* (pp. 132-141).

[12] Wildemuth, B. M. (2004). The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology*, *55*(3), 246-258.

[13] Wilson, M. J., & Wilson, M. L. (2013). A comparison of techniques for measuring sensemaking and learning within participant-generated summaries. *Journal of the Association for Information Science and Technology*, *64*(2), 291-306.

[14] Yue, Z., Han, S., & He, D. (2014, February). Modeling search processes using hidden states in collaborative exploratory web search. In *the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 820-830).

[15] Zhang, X., Cole, M., & Belkin, N. (2011, July). Predicting users' domain knowledge from search behaviors. In *the 34th international ACM SIGIR conference on Research and development in Information Retrieval* (pp. 1225-1226).

[16] Yue, Z., Jiang, J., Han, S., & He, D. (2012, October). Where do the query terms come from? an analysis of query reformulation in collaborative web search. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 2595-2598).