Investigation of User Search Behavior While Facing Heterogeneous Search Services

Xin Li, Yiqun Liu*, Rongjie Cai, Shaoping Ma
State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science & Technology, Tsinghua University, Beijing, 100084, China
yiqunliu@tsinghua.edu.cn

ABSTRACT

With Web users' search tasks becoming increasingly complex, a single information source cannot necessarily satisfy their information needs. Searchers may rely on heterogeneous sources to complete their tasks, such as search engines, Community Question Answering (CQA), encyclopedia sites, and crowdsourcing platforms. Previous works focus on interaction behaviors with federated search results, including how to compose a federated Web search result page and what factors affect users' interaction behavior on aggregated search interfaces. However, little is known about which factors are crucial in determining users' search outcomes while facing multiple heterogeneous search services. In this paper, we design a lab-based user study to analyze what explicit and implicit factors affect search outcomes (information gain and user satisfaction) when users have access to heterogeneous information sources. In the study, each participant can access three different kinds of search services: a general search engine (Bing), a general CQA portal (Baidu Knows), and a high-quality CQA portal (Zhihu). Using questionnaires and interaction log data, we extract explicit and implicit signals to analyze how users' search outcomes are correlated with their behaviors on different information sources. Experimental results indicate that users' search experiences on CQA portals (such as users' perceived usefulness and number of result clicks) positively affect search outcome (information gain), while search satisfaction is significantly correlated with some other factors such as users' familiarity, interest and difficulty of the task. Besides, users' search satisfaction can be more accurately predicted by the implicit factors than search outcomes.

Keywords

Search outcome; Heterogeneous information; User study

*Corresponding author

WSDM 2017, February 06-10, 2017, Cambridge, United Kingdom

DOI: http://dx.doi.org/10.1145/3018661.3018673

1. INTRODUCTION

With the abundant information available on the Web, users' search tasks are becoming more and more complex [4, 25, 6]. Besides general search engines, information seekers also rely on CQA portals [12], encyclopedia sites [9] and crowdsourcing platforms [1] to fulfill their information needs. It has been a challenging problem of how to integrate information from heterogeneous sources and uniformly provide an optimal search result list for users. A current solution is to aggregate search results from various heterogeneous sources or verticals into a single search engine result page (SERP) [23, 3].

Though aggregated search techniques have been well studied in existing works on how to compose a federated Web search result page [18], what factors affect users' click-through behavior on aggregated search interfaces [20], and which verticals are relevant for a given query [24], little is known about what explicit and implicit factors affect search outcomes when users have access to heterogeneous search services. A prior study has found that users' learning outcome is closely correlated with their search interaction strategies and perceived learning outcomes when performing learning-related tasks [7]. It has also been found that an individual's source selection for health search task is greatly affected by his/her health literacy and the frequency of using a source [19]. Under vertical circumstances, a user study shows that more complex tasks require significantly more interaction with vertical results [4]. In this paper, we focus on more general search tasks that are not limited to a specific domain. Besides, we aim to investigate the factors that affect users' search outcome and satisfaction when they can freely choose from heterogeneous information sources.

As a preliminary attempt to take up the heterogeneity challenge, we perform a lab-based user study, the framework of which is shown in Figure 1. First, we design tasks on multiple topics and recruit participants with different levels of search background to complete the tasks. Each participant needs to answer a question for each task after searching with heterogeneous information sources. They were also asked to report their perceived satisfaction during the search process. Then, the same domain experts who designed the tasks were hired to evaluate the correctness of their answers. Participants will fill out pre-task and post-task questionnaires as their explicit feedbacks. We also derive features from their search interaction process as implicit factors. Finally, we analyze

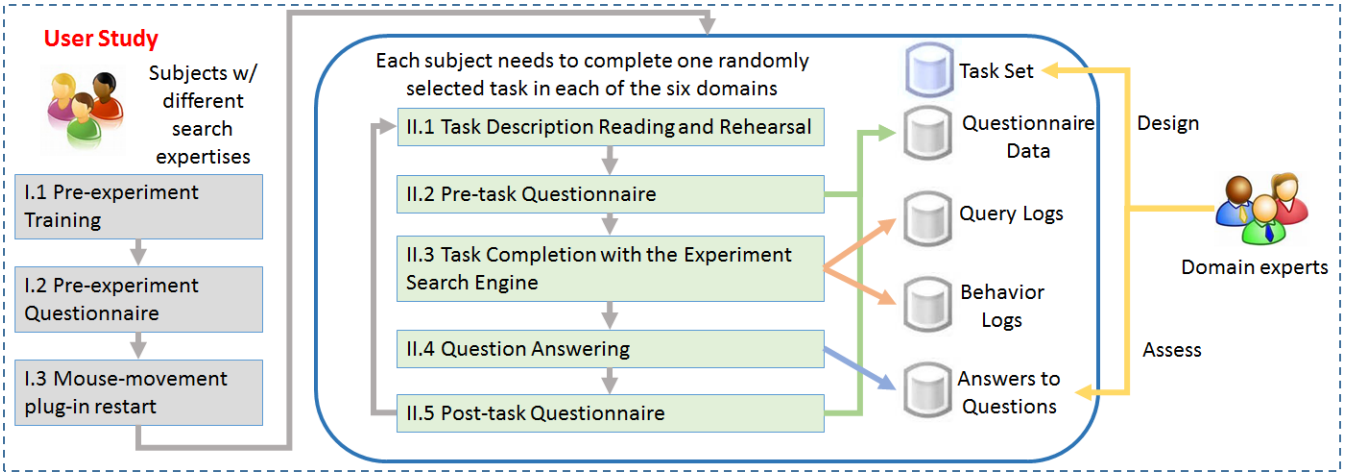

Figure 1: User study framework

the key factors that affect users' search outcomes when accessing heterogeneous information sources. Specifically, this study addresses the following research questions:

- **RQ1**: How does search background affect search outcomes while finishing tasks with heterogeneous search services?
- **RQ2**: What are the key (explicit and implicit) factors that affect a user's search outcome as measured by domain experts?
- **RQ3**: What are the key (explicit and implicit) factors that affect a user's perceived search satisfaction?
- **RQ4**: Can we predict search outcome/perceived search satisfaction with the interaction behaviors collected from users?

Our work has many potential applications in improving the designing of better Web search services and helping search users better fulfill their information needs. For example, study on heterogeneous search services helps us gain more knowledge on users' real search interactions with different information sources and have the potential to improve aggregated search performances. We can also make use of the features derived from users' search interactions to predict their search engine switching behaviors and improve the performance of federated search.

The remainder of this paper is organized as follows. After a discussion of the related work in the next section, we describe in detail the components of our designed user study in the third section. Then, we present the data analysis and search outcome prediction results in the fourth and fifth section, respectively. Finally, the sixth section concludes the paper.

2. RELATED WORK

Search outcome assessment with heterogeneous information access has not attracted much attention so far. However, it is related to several research fields. In this section, we review the most related ones.

2.1 Aggregated Search

Aggregated search is a current solution to handle the heterogeneous information sources and has been widely investigated. Ponnuswami et al. [18] design a machine-learning framework for SERP composition in the presence of multiple relevant verticals. They perform a user study to estimate the pairwise click preference of Web results and vertical ones. Their results show that they can get very high correlation with user engagement clicks by building models that use click preference as judgments. Sushmita et al. [20] investigate factors affecting users' click-through behavior on aggregated search interfaces. They test two aggregated search interfaces: one with results from different sources blended into a single list, and the other with results from each source presented in a separate panel. Their user study results show that the position of search results is only significant for the former interface. Arguello et al. [2] describe a new methodology for evaluating aggregate search result. They derive a reference presentation for a query with the preference judgements on block-pairs. Then, any arbitrary presentation for the query can be evaluated based on its distance to the reference. They present a user study to empirically validate their metic and find that the metric's agreement with the majority preference is in the 67-73% range. Liu et al. [14] use visual saliency to predict users' attention distribution on heterogeneous search components. Our work differs from these existing works in that instead of a federated search result list, we aim to analyze users' search behavior when accessing pages from heterogeneous information sources.

2.2 Search Performance Evaluation

Evaluation is one of the prime concerns in search related studies. Hassan et al. [10] propose to predict search satisfaction with query-based features. They identify a rule indicating dissatisfaction that a similar query issued within a time interval that is short enough implies dissatisfaction. Their results show that a query-based model can indicate satisfaction more accurately than click-based models, and that search success is an incremental process for successful tasks with multiple queries. Wang et al. [21] argue that

Table 1: Search tasks in the user study

Domain	Task Description	Category
Politics	Political scientists have noted that the trend of political polarization during the US presidential election is increasingly evident. What are the reasons behind it?	Intellectual
	In order to achieve their own interests, what kind of strategies do the US interest groups often take?	Intellectual
Economy	What do you think is the cause of the European debt crisis? What are the policy responses of related countries?	Intellectual
	What do you think is the university-based technology transfer rate in China? Is it higher or lower than western countries? What are the reasons behind it?	Intellectual
Medicine	What are the most commonly-used methods for cancer treatment in clinics?	Factual
	What are the potential applications of 3D printing for "Precision Medicine"?	Intellectual
Environment	Why ultraviolet disinfection cannot completely supplant chlorination when disinfecting drinking water?	Intellectual
	What are the characteristics of particulate matter in China? Your answer includes but not limited to: national level, regional level, time-varying and composition.	Intellectual
Physics	What is the gravitational wave, and where does it come from? What is the difference between gravitational wave and electromagnetic wave?	Factual
	There is a fundamental particle named Higgs boson. What is the relation between this particle and the origin of inertial mass?	Factual
Law	Under the Chinese Contract Law, a lessee may not discontinue the lease on a sale of the leased premises. Question: A has leased his own house to B for B's living. During the lease, A sales the house to C. Whether can B claim to C that the lease should be maintained or not?	Intellectual
	Under the law, a property can be sold at the owner's disposal. But under the Partnership Enterprise Law, there are some special rules. Question: A has bought a fishing vessel, now B is in charge of the operation management of the vessel. B deposits part of the profit into A's account regularly. Now A wants to sell the vessel to C. Please answer whether B has the right to fight against A's selling.	Intellectual

searchers' latent action-level satisfaction influences their observed search behaviors and contributes to overall search satisfaction. Therefore, more complete and more accurate predictors of search-task satisfaction can be built by modeling search satisfaction at the action level. Experimental results demonstrate significant value in modeling action-level satisfaction in search-task satisfaction prediction. Liu et al. [13] compare search users' and external assessors' opinions on satisfaction. They find that search users pay more attention to the utility of results while external assessors emphasize on the efforts spent in search sessions. In our work, we evaluate participants' search outcome both by experts' scores on their answers to the tasks and by their feedbacks to the questionnaires.

2.3 Search Expertise

Search expertise plays an important role in determining a searcher's effectiveness and efficiency to gain information on the Web. Boydell and Smyth [5] define search expertise as "the ability to quickly and accurately locate information according to a specific information need". They capture search expertise within a community of like minded searchers by mining the title and snippet texts of results that have been selected by community members in response to their queries. They also build a community-based snippet index, which is used to re-rank the results by boosting the key results that have been frequently selected for similar queries by community members in the past. Moraveji [16] believes that search expertise includes several abilities such as generating appropriate keywords, discerning legitimate from illegitimate pages, reformulating queries, and so on. They deploy a live system to enable the human work that goes into conducting exploratory searches to be efficiently captured and transmitted to other learners [17]. White and Morris [22] try to help all search engine users be more successful in their searches by investigating the interaction logs of advanced search engine users. Experimental results show that there are remarkable differences in the queries, result clicks, post-query browsing behaviors, and search success between advanced and non-advanced searchers.

3. USER STUDY DESIGNATION

In this section, we introduce our user study design using an interactive multi-source search system in a laboratory setting.

3.1 Tasks

To cover a variety of topics, we design tasks on six domains: Politics, Economy, Medicine, Environment, Physics, and Law. For each domain, we recruit a senior graduate student (as "domain expert") with the corresponding major from our university. We request each expert to design two tasks based on the following requirements:

- The task should be non-trival for participants without corresponding domain knowledge so that they have to turn to search services for necessary information to accomplish it.
- The task should be reasonably complex so that the participants cannot complete it with a few simple search interactions.

We set the requirements so that participants all have a similar pre-task knowledge background while finishing the tasks. Through these settings, we aim to investigate user behavior of information integration when performing complex tasks. With the help of the domain experts, we designed 12 tasks in total, two in each of the six domains. A complete list of task descriptions is shown in Table 1. The category of each task listed in the last column is determined according to TREC Session Track[1]. Except for designing tasks, experts are also requested to make scoring criteria for the tasks and assign a score of 0-10 for each

[1]http://ir.cis.udel.edu/sessions/index.html

Table 2: Questionnaire description

Stage	ID	Question text	Scale
Pre-experiment	*general_bing*	How frequent do you use Bing for search?	1=not frequent;...;5=very frequent
	general_baidu	How frequent do you use Baidu Knows for search?	1=not frequent;...;5=very frequent
	general_zhihu	How frequent do you use Zhihu for search?	1=not frequent;...;5=very frequent
	general_skill	How is your skill of collecting information on the Web?	1=not skillful;...;5=very skillful
Pre-task	*pre_knowledge*	How much do you know about the task?	1=not at all;...;5=I know a lot
	pre_interest	Are you interested in the task?	1=not interested;...;5=very interested
	pre_difficulty	How difficult do you think to complete the task?	1=not difficult;...;5=very difficult
Post-task	*post_knowledge*	After searching, how much do you know about the task?	1=not at all;...;5=I know a lot
	post_interest	After searching, are you interested in the task?	1=not interested;...;5=very interested
	post_difficulty	After searching, how difficult do you feel to complete the task?	1=not difficult;...;5=very difficult
	post_bing	How useful is Bing for you to complete the task?	1=not useful;...;5=very useful
	post_baidu	How useful is Baidu Knows for you to complete the task?	1=not useful;...;5=very useful
	post_zhihu	How useful is Zhihu for you to complete the task?	1=not useful;...;5=very useful
	post_satisfaction	How satisfied are you with your search experience?	1=not satisfied;...;5=very satisfied

participant's answer. The scores given by experts are used to measure the search outcome for the task.

3.2 Study Participants

After designing tasks, we need to recruit participants to complete these tasks. On one hand, we wish the participants to have similar background knowledge so that their prior knowledge will not affect the study. On the other hand, we want to investigate the effect of search background on search outcomes while facing heterogenous information sources, so it is necessary to recruit participants with different levels of search background. Therefore, we limit the participants to students from majors different from the six domains in Table 1. We sent a recruiting email to these departments and students can opt in to participate. As a result, a total of 33 participants signed up for the study, including 12 computer science (CS) undergraduate students, 5 CS graduate students, and 16 non-CS students. Each participant needs to complete six tasks containing one randomly selected task in each domain. We balance the number of complete times of the tasks so that each task is presented to 16 or 17 participants.

3.3 Search System and Interface

Instead of a federated search result list, we provide participants with a heterogeneous search environment. We choose three typical information sources that are popular for the participants in our experiment: a general search engine, a general CQA portal, and a specialized high-quality CQA portal. Figure 2 shows the search interface of our study.

Figure 2: Search interface

- **General search engine**: While completing a search task, Web users often turn to a general search engine like Google or Baidu to fulfill their information needs. We choose Bing as the general search engine because it is a popular service provider, and also because its Search API is publicly accessible. To eliminate redundant information, we remove the results of Baidu Knows and Zhihu from Bing result lists (which are not common because these sources do not have close cooperations with Bing search). We reserve the top results for a given query after filtering out redundant results and show the participants 10 results on one SERP.

- **General CQA portal**: General CQA portals like Yahoo! Answers provide a platform for Web users to seek and provide information. Baidu Knows is a widely used CQA portal with a large number of active users. We obtain the top 10 results for a given query and show them to the participants on a single SERP.

- **Specialized High-quality CQA portal**: Zhihu, as a popular Chinese specialized CQA portal, is able to provide questions with detailed and reliable answers that are voted by a large number of users, which is similar to StackExchange or Quora. Again, we obtain the top 10 results for a given query from Zhihu and show them to the participants on a single SERP.

When performing tasks, participants can freely formulate queries and switch platforms during the search process. When participants believe that they have collected enough information for the task in their interaction with the information sources, they can press the "Complete" button to finish the search process and give an answer of 100-200 words according to what they have learned from the search processes. Then the answers will be evaluated and scored by the corresponding experts who design the tasks.

3.4 Questionnaires

To obtain explicit factors of search users, we ask each participant to fill out different questionnaires at three stages of the study. Before performing tasks, they need to provide some general demographic information, including their frequency of using the three information sources for search (denoted as *general_bing*, *general_baidu*, and

general_zhihu) and their self-rated skills of collecting information on the Web (denoted as *general_skill*). After reading the description of each task and before searching, they need to rate their prior knowledge, interest, and difficulty about the task (denoted as *pre_knowledge*, *pre_interest*, and *pre_difficulty*). After completing the search processes and giving answers to the task, they are asked about their perceived knowledge, interest, and difficulty about the task (denoted as *post_knowledge*, *post_interest*, and *post_difficulty*) to see if they have changed their self-belief after searching. Besides, they need to tell their perceived usefulness of the three information sources for them to complete the task (denoted as *post_bing*, *post_baidu*, and *post_zhihu*). Finally, they will assess overall satisfaction of their search experience in the task (denoted as *post_satisfaction*). A complete set of questionnaires are shown in Table 2.

3.5 Logging of Search Interaction

During the participants' search processes, we log their interaction behaviors with the experimental system. Besides query and click behaviors, we also record the dwell time of each page during the search interaction, including SERPs of each information source and landing pages. To identify implicit indicators of search outcome from search interaction, we extract the following features from the search process of each task.

- **Unique query count**: Number of unique queries submitted to each of the three information sources (denoted as *unique_query_count_bing*, *unique_query_count_baidu*, and *unique_query_count_zhihu*).
- **Unique click count**: Number of unique clicks on results from each of the three information sources (denoted as *unique_click_count_bing*, *unique_click_count_baidu*, and *unique_click_count_zhihu*).
- **Total SERP time**: Total dwell time on the SERPs of each of the three information sources (denoted as *total_serp_time_bing*, *total_serp_time_baidu*, and *total_serp_time_zhihu*).
- **Average SERP time per query**: Average dwell time on the SERPs per query of each of the three information sources (denoted as *average_serp_time_bing*, *average_serp_time_baidu*, and *average_serp_time_zhihu*).
- **Total landing time**: Total dwell time on landing pages originated from each of the three information sources (denoted as *total_landing_time_bing*, *total_landing_time_baidu*, and *total_landing_time_zhihu*).
- **Average landing time per query**: Average dwell time on landing pages per query from each of the three information sources (denoted as *average_landing_time_bing*, *average_landing_time_baidu*, and *average_landing_time_zhihu*).
- **Satisfied click count**: Number of satisfied clicks on each of the three information sources (denoted as *sat_click_count_bing*, *sat_click_count_baidu*, and *sat_click_count_zhihu*). Following previous work on search success prediction [8, 11], we define satisfied click as clicks with dwell time on landing pages longer than 30s, and dissatisfied click as clicks on landing pages with dwell time shorted than 10s.
- **Satisfied click ratio**: Ratio between satisfied clicks and all the clicks on each of the three information sources (denoted as *sat_click_ratio_bing*, *sat_click_ratio_baidu*, and *sat_click_ratio_zhihu*).
- **Dissatisfied click count**: Number of dissatisfied clicks on each of the three information sources (denoted as *dsat_click_count_bing*, *dsat_click_count_baidu*, and *dsat_click_count_zhihu*).
- **Dissatisfied click ratio**: Ratio between dissatisfied clicks and all the clicks on each of the three information sources (denoted as *dsat_click_ratio_bing*, *dsat_click_ratio_baidu*, and *dsat_click_ratio_zhihu*).

4. DATA ANALYSIS

Our goal is to investigate the effects of explicit and implicit factors on both search outcome and search satisfaction. Search outcome is evaluated by a domain expert's score on each task. Search satisfaction is directly given by the participant's feedback. We derive explicit factors from participants' feedbacks of questionnaires shown in Table 2 and implicit factors from the search interaction in each task as described in Section 3.5. The main analysis method we adopt is the regression approach to the analysis of variance (ANOVA) [15]. Multiple linear regression attempts to fit a regression line for the response variable using multiple explanatory variables. To evaluate the importance of explanatory variables on fitting the response variable, ANOVA adopts an *F test*. A large *F value* provides evidence against the null hypothesis and indicates high importance of the corresponding explanatory variable on fitting the response variable. Besides, we compute the Pearson Correlation Coefficient (PCC) between search outcomes/satisfaction and the factors to estimate whether they are positively or negatively correlated.

In this section, we present the experimental results regarding our research questions. For RQ1, we evaluate the qualities of the answers given by different groups of participants and compare their outcomes. For RQ2 and RQ3, we aggregate the results of the 198 tasks completed by the 33 participants. We regard the explicit and implicit factors as explanatory variables, and search outcomes or search satisfaction as response variables. We perform ANOVA to test the importance of different factors on fitting. We also show the PCC results to reveal their relations.

4.1 Effect of Search Background on Search Outcome

We recruit participants with different search backgrounds and we aim to investigate their effects on search correctness. Figure 3 compares the scores of the tasks completed by different groups of participants and in different domains. The rightmost part of the figure indicates that the participants with more specialized search backgrounds are able to produce more correct answers. The mean score of CS graduate students is 7.34 (SD=1.96), which is higher than Non-CS students (M=6.63, SD=2.30)

and CS undergraduate students (M=6.63, SD=2.52). CS graduate students achieve higher scores in 4 out of the 6 domains of tasks. Non-CS students do better jobs on Physics tasks, while CS undergraduate students are better at Economy tasks. The results show that though with equal knowledge of the tasks, participants with more specialized search backgrounds are better at formulating queries and choosing search results to complete the tasks, so that they are able to achieve better search outcomes.

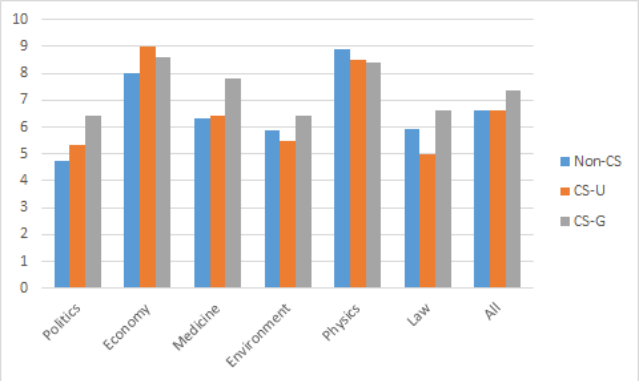

Figure 3: Comparisons of scores achieved by different groups of participants in different taskdomains (Non-CS: non-CS students, CS-U: CS undergraduate students, CS-G: CS graduate students)

4.2 Factor Analysis on Search Outcome

Each answer by a participant is evaluated and assigned a score by the corresponding domain expert to represent the quality of outcome. We analyze the effects of explicit and implicit factors on search outcomes separately. The ANOVA and PCC results are shown in Table 3 and Table 4, respectively.

Table 3: Effects of explicit factors on search outcome (* indicates statistical significance at p<0.1 level, ** indicates statistical significance at p<0.05 level)

Factor	F value	Pr(>F)	PCC
general_bing	0.8079	0.3699	-0.0606
general_baidu	2.8579	0.0926	-0.1057
general_zhihu	**4.2617**	**0.0404***	**0.1583**
general_skill	0.0233	0.8789	-0.0225
pre_knowledge	2.5421	0.1126	-0.0651
pre_interest	2.5210	0.1141	0.0660
pre_difficulty	**8.8329**	**0.0034****	**-0.1642**
post_bing	0.3956	0.5302	0.0518
post_baidu	**8.1941**	**0.0047****	**0.0973**
post_zhihu	**4.9589**	**0.0272***	**0.1023**
post_knowledge	0.0393	0.8430	0.0884
post_interest	0.0005	0.9825	0.0105
post_difficulty	0.6672	0.4151	-0.0238

Table 3 shows that significant explicit factors on fitting search outcomes include *general_zhihu*, *pre_difficulty*, *post_baidu*, and *post_zhihu*. PCC results show that participants' search outcome is positively correlated with *general_zhihu*, *post_baidu*, and *post_zhihu*, while is negatively correlated with *pre_difficulty*. It means that the more frequently a participant uses Zhihu for search, and the more useful he/she perceives Baidu Knows and Zhihu are when performing a task, the higher scores he/she will get on the task. The results indicate the importance of both general and specialized CQA portals when searchers are performing tasks. Another interesting finding is that though participants' outcomes negatively correlate with their expected difficulties before searching, they are not significantly correlated with their perceived difficulties after searching. This means that a user's first sense of a task's difficulty is more precise. As the user searches and gains more knowledge about the task, his/her perceived difficulty of the task will be different. Additionally, from the table we can see that participants' perceived knowledge (both pre-task and post-task) of the tasks is not significantly correlated with their outcomes, which validates of our recruitment policy is promising and all the participants have similar prior knowledge about the tasks.

Table 4: Effects of implicit factors on search outcome (* indicates statistical significance at p<0.1 level, ** indicates statistical significance at p<0.05 level)

Factor	F value	Pr(>F)	PCC
unique_query_count_bing	0.7090	0.4010	0.0581
unique_query_count_baidu	0.5843	0.4457	-0.0524
unique_query_count_zhihu	1.0098	0.3164	0.0746
unique_click_count_bing	0.2414	0.6238	0.0603
unique_click_count_baidu	**4.1511**	**0.0432***	**0.0909**
unique_click_count_zhihu	**5.9446**	**0.0158***	**0.1766**
total_serp_time_bing	0.1952	0.6592	-0.0018
total_serp_time_baidu	0.1512	0.6979	-0.0525
total_serp_time_zhihu	0.8104	0.3693	0.1273
average_serp_time_bing	0.0960	0.7571	-0.0051
average_serp_time_baidu	2.9255	0.0890	-0.0900
average_serp_time_zhihu	0.0015	0.9695	0.1270
total_landing_time_bing	2.2202	0.1381	0.0909
total_landing_time_baidu	**7.1839**	**0.0081****	**0.1996**
total_landing_time_zhihu	1.0735	0.3017	0.0358
average_landing_time_bing	0.5299	0.4677	0.0656
average_landing_time_baidu	0.6274	0.4294	0.1826
average_landing_time_zhihu	0.0013	0.9716	-0.0268
sat_click_count_bing	0.0628	0.8025	0.0683
sat_click_count_baidu	0.2373	0.6268	0.1458
sat_click_count_zhihu	0.2320	0.6307	0.0216
sat_click_ratio_bing	0.0816	0.7755	0.0415
sat_click_ratio_baidu	0.6965	0.4051	0.0866
sat_click_ratio_zhihu	0.0469	0.8289	-0.0085
dsat_click_count_bing	3.1765	0.0765	0.0158
dsat_click_count_baidu	**4.2461**	**0.0409***	**-0.0041**
dsat_click_count_zhihu	0.1299	0.7190	0.1388
dsat_click_ratio_bing	3.7261	0.0553	0.0392
dsat_click_ratio_baidu	0.3829	0.5369	-0.0697
dsat_click_ratio_zhihu	1.2753	0.2604	0.0567

From Table 4 we can see that significant implicit factors on fitting search outcome are *unique_click_count_baidu*, *unique_click_count_zhihu*, *total_landing_time_baidu*, and *dsat_click_count_baidu*. PCC results show that participants' search outcome is positively correlated with *unique_click_count_baidu*, *unique_click_count_zhihu*, and *total_landing_time_baidu*, while is negatively correlated with *dsat_click_count_baidu*. Again, the results reveal the importance of the use of CQA portals on searchers' search outcomes. When participants click on more results from Baidu Knows

and Zhihu, and when they spend more time reading the results from Baidu Knows, they will achieve higher scores for their completed tasks. On the contrary, if a participant is dissatisfied with a larger number of clicked results from Baidu Knows, it is more likely that he/she will give an unsatisfactory answer. Additionally, from the results we can see that significant implicit factors on fitting search outcome do not include general search related factors, which means that general search engines may not be that useful when searchers are performing complex tasks. They can find more specified and detailed solutions to the tasks on CQA portals.

4.3 Factor Analysis on Search Satisfaction

Participants' perceived search satisfaction is directly given from their post-task questionnaires. Similar with search outcome, we analyze the effects of explicit and implicit factors on search satisfaction separately. The ANOVA and PCC results are shown in Table 5 and Table 6, respectively.

Table 5: Effects of explicit factors on search satisfaction (* indicates statistical significance at p<0.1 level, ** indicates statistical significance at p<0.05 level)

Factor	F value	Pr(>F)	PCC
general_bing	2.8687	0.0920	0.0852
general_baidu	0.0004	0.9849	-0.0094
general_zhihu	0.7316	0.3935	-0.0359
general_skill	0.5576	0.4562	0.0671
pre_knowledge	1.1517	0.2846	-0.0538
pre_interest	**5.7664**	**0.0173***	**0.1197**
pre_difficulty	**13.4719**	**0.0003****	**-0.1505**
post_bing	**14.6958**	**0.0002****	**0.2164**
post_baidu	**6.7573**	**0.0101***	**0.0622**
post_zhihu	3.3442	0.0691	-0.0341
post_knowledge	**44.8043**	**0.0000****	**0.3902**
post_interest	**24.4982**	**0.0000****	**0.3329**
post_difficulty	**92.6576**	**0.0000****	**-0.6509**

As shown in Table 5, different from search outcomes, search satisfaction is significantly correlated with a large number of explicit factors, including participants' perceived knowledge, interest and difficulty about the task (both pre-task and post-task), and their perceived usefulness of Bing and Baidu Knows during the search process. When a participant believes that he/she has rich knowledge about the task, or he/she is interested in the task, it is highly likely that he/she will be satisfied with the process of performing the task. On the other hand, if a participant feels that the task is difficult to complete (both before and after searching), he/she will be probably unsatisfied with the task. Interestingly, participants' search satisfaction is significantly positively correlated with their perceived usefulness of Bing during the search process, which is very different with the analysis results of search outcomes. This means that when a user believes that general search engine is useful for him/her to complete the task, though he/she will be satisfied with the whole search process, he/she may not give a correct answer as judged by domain experts to the task in the end. This result indicates that we cannot regard users' subjectively perceived search satisfaction as equivalent with their actual search outcomes (or information gain).

Table 6: Effects of implicit factors on search satisfaction (* indicates statistical significance at p<0.1 level, ** indicates statistical significance at p<0.05 level)

Factor	F value	Pr(>F)	PCC
unique_query_count_bing	3.5129	0.0626	-0.1285
unique_query_count_baidu	**5.2450**	**0.0233***	**-0.1577**
unique_query_count_zhihu	**12.1398**	**0.0006****	**-0.2861**
unique_click_count_bing	0.5107	0.4759	-0.0460
unique_click_count_baidu	0.3219	0.5712	-0.0382
unique_click_count_zhihu	0.6121	0.4351	-0.0355
total_serp_time_bing	0.0931	0.7606	-0.0711
total_serp_time_baidu	2.5515	0.1121	-0.1661
total_serp_time_zhihu	0.2651	0.6073	-0.1859
average_serp_time_bing	0.5049	0.4783	-0.0041
average_serp_time_baidu	0.6680	0.4149	-0.1417
average_serp_time_zhihu	0.1631	0.6868	-0.0994
total_landing_time_bing	0.7793	0.3786	0.0119
total_landing_time_baidu	1.1591	0.2832	0.0441
total_landing_time_zhihu	1.0996	0.2959	0.0325
average_landing_time_bing	1.8260	0.1784	0.0933
average_landing_time_baidu	0.0069	0.9339	0.0342
average_landing_time_zhihu	0.2206	0.6392	0.0130
sat_click_count_bing	0.0973	0.7555	0.0051
sat_click_count_baidu	0.0437	0.8347	0.0057
sat_click_count_zhihu	1.0345	0.3106	0.0499
sat_click_ratio_bing	2.6855	0.1031	0.0869
sat_click_ratio_baidu	0.0957	0.7574	-0.0330
sat_click_ratio_zhihu	3.8445	0.0516	0.0745
dsat_click_count_bing	1.0975	0.2963	-0.0363
dsat_click_count_baidu	0.0437	0.8346	-0.0537
dsat_click_count_zhihu	1.2510	0.2650	-0.0050
dsat_click_ratio_bing	1.5708	0.2118	-0.1185
dsat_click_ratio_baidu	1.7726	0.1849	0.0238
dsat_click_ratio_zhihu	0.6720	0.4135	-0.0038

Table 6 shows that significant implicit factors on fitting search satisfaction only contain *unique_query_count_baidu* and *unique_query_count_zhihu*. PCC results show that the more queries a participant submits to Baidu Knows or Zhihu, the less satisfied he/she will be with the search process. This is reasonable because if a searcher submits multiple queries, it means that the search results of the former queries cannot satisfy his/her information needs and he/she has to reformulate the query and find more search results to complete their tasks, which will degrade their search experience. The results in Table 6 indicate that search satisfaction cannot be fitted well by implicit factors since most of the implicit factors are not significantly correlated with search satisfaction. On the other hand, explicit factors do a better job in fitting search satisfaction.

5. PREDICTION RESULTS

One of the main applications of our user study is to predict users' search outcomes and search satisfaction with the features extracted from their search interactions with heterogeneous information sources. Regarding RQ4, we aim to investigate how well searchers' search outcomes and search satisfaction can be fitted by the implicit factors as described in section 3.5. Since our training set is relatively small, including 198 tasks completed by 33 participants, we adopt linear regression models to reduce the effect of overfitting. We use mean square error (MSE) to evaluate

the prediction performance. For each regression task, we perform 10-fold cross validation and take the average of MSEs as the final result. Our experiments serve as a preliminary attempt to prove that the collected user behavior signals on heterogeneous information sources have impact on the prediction of search outcomes and search satisfaction. The raw data of our experiments is available[2].

5.1 Predicting Search Outcome

We regard the score assigned to each task by the corresponding expert as the target variable and the implicit factors as features. We compare several linear regression models, including Ridge, Lasso, and Elastic Net (implemented with *sklearn*[3]), to fit search outcomes with the implicit factors. The comparison of prediction results is shown in Table 7.

Table 7: Comparison of regression models on predicting search outcome

Model	MSE	Gain
Ridge	4.3495	–
Lasso	4.8967	12.6%
Elastic Net	4.7396	9.0%

We can see from the results that Ridge gives the best performance and we adopt Ridge in the following experiment. To evaluate the importance of features from different sources, we divide the implicit factors into three groups: Bing related features, Baidu Knows related features, and Zhihu related features. We adopt a leave-one-out strategy. Each time we use the whole feature set except one group of features to evaluate the prediction performance. The results are shown in Table 8.

Table 8: Feature importance analysis in predicting search outcome

Feature left out	MSE	Gain
None	4.3495	–
Bing	4.6812	7.6%
Baidu Knows	**4.9409**	**13.6%**
Zhihu	4.6575	7.1%

The results show that leaving out Baidu Knows related features introduces the largest gain in MSE, indicating that users' search interactions with Baidu Knows are the most important in predicting their search correctness. Besides, we note that MSE is relatively large compared to the scale of scores, which means that search outcome cannot be accurately predicted by implicit factors.

5.2 Predicting Search Satisfaction

We regard the search satisfaction answered by each participant as the target variable and the implicit factors as features. The comparison of different linear regression models is shown in Table 9.

Again, Ridge shows the best performance. The importance of different feature groups in predicting search satisfaction using Ridge is shown in Table 10.

[2]http://www.thuir.cn/group/~YQLiu/publications/wsdm2016Li.zip

[3]http://scikit-learn.org/stable/modules/linear_model.html

Table 9: Comparison of regression models on predicting search satisfaction

Model	MSE	Gain
Ridge	0.8022	–
Lasso	0.9140	13.9%
Elastic Net	0.8684	8.3%

Table 10: Feature importance analysis in predicting search satisfaction

Feature left out	MSE	Gain
None	0.8022	–
Bing	0.8425	5.0%
Baidu Knows	0.8587	7.0%
Zhihu	**0.9165**	**14.2%**

We can see from the results that leaving out Zhihu related features introduces the largest gain in MSE. We rely more on the features extracted from users' search process on Zhihu to predict their search satisfaction. Besides, though ANOVA results show that individual implicit factors are not significantly correlated with search satisfaction, the implicit factors as a whole can be effectively used to predict users' search satisfaction.

Both the experiments show that CQA portal related factors are crucial in predicting users' search outcomes, which is consistent with previous results. Additionally, search outcome is an objective evaluation criterion, which is weakly correlated with user behaviors, so that the overall prediction performance is relatively low. On the other hand, search satisfaction is users' subjective evaluation, which can be predicted more accurately.

5.3 Result Discussions

A notable conclusion drawn from the results is that CQA portals play an important role on users' search outcomes when they are performing complex tasks. The results of the effects of both explicit and implicit factors on search outcomes show that the more frequently a searcher uses CQA portals to complete the task, the more likely he/she will give a correct answer. This may guide current search engines to incorporate more results from CQA portals into the search result list when searchers are performing complex tasks.

The data analysis results indicate that users' search satisfaction cannot be equivalent with their outcome. For example, the results of the effects of explicit factors on search satisfaction show that if a searcher's perceived knowledge and interest about the search task is high, it is more likely that he/she will be satisfied with the search process. However, the results in Table 3 indicate that he/she may not get a correct answer to the search task. Therefore, when the optimization goals are different (search outcome or search satisfaction), different strategies should be adopted.

We regard a satisfaction score of 4 or 5 to be satisfied. For 98 out of the 198 tasks, the participants are satisfied. Figure 4 shows the search outcome distribution for the 98 satisfied tasks. We can see that for 40% of the tasks, the participants get lower than or equal to 60% of the full score. The result again implies that searchers' perceived satisfaction of the search process is not reliable to reflect their search outcome.

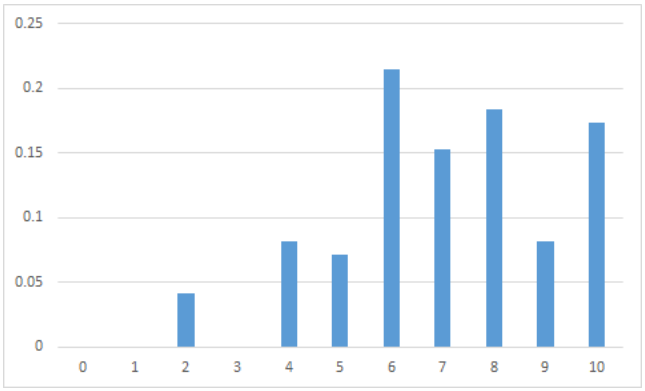

Figure 4: Search outcome distribution for satisfied tasks

6. CONCLUSION

Searchers' information needs are becoming more and more complex and require various sources of information to complete their search task. In this paper, we design a user study to investigate the effects of explicit and implicit factors on users' search outcomes (including information gain and search satisfaction) when they can access heterogeneous information sources. We adopt ANOVA to analyze the significant factors on fitting each search outcome. The results show that CQA portals are crucial in determining searchers' search outcomes when they are performing complex tasks. Besides, we cannot regard search satisfaction as equivalent with search outcome and they can be estimated by different sets of factors. We also adopt linear regression models to predict search outcomes with implicit factors. The results show that users' search satisfaction can be more accurately predicted by the implicit factors than search outcomes.

Acknowledgments

This work was supported by Natural Science Foundation (61622208, 61532011, 61672311) of China and National Key Basic Research Program (2015CB358700).

7. REFERENCES

[1] A. Afuah and C. L. Tucci. Crowdsourcing as a solution to distant search. *Academy of Management Review*, 37(3):355–375, 2012.

[2] J. Arguello, F. Diaz, J. Callan, and B. Carterette. A methodology for evaluating aggregated search results. In *European Conference on Information Retrieval*, pages 141–152. Springer, 2011.

[3] J. Arguello, F. Diaz, J. Callan, and J.-F. Crespo. Sources of evidence for vertical selection. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 315–322. ACM, 2009.

[4] J. Arguello, W.-C. Wu, D. Kelly, and A. Edwards. Task complexity, vertical display and user interaction in aggregated search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 435–444. ACM, 2012.

[5] O. Boydell and B. Smyth. Capturing community search expertise for personalized web search using snippet-indexes. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 277–286. ACM, 2006.

[6] A. Chuklin, K. Zhou, A. Schuth, F. Sietsma, and M. De Rijke. Evaluating intuitiveness of vertical-aware click models. In *Proceedings of the 37th international ACM SIGIR conference on Research and development in information retrieval*, pages 1075–1078. ACM, 2014.

[7] K. Collins-Thompson, S. Y. Rieh, C. C. Haynes, and R. Syed. Assessing learning outcomes in web search: A comparison of tasks and query strategies. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pages 163–172. ACM, 2016.

[8] Q. Guo, D. Lagun, and E. Agichtein. Predicting web search success with fine-grained interaction data. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2050–2054. ACM, 2012.

[9] R. Hahn, C. Bizer, C. Sahnwaldt, C. Herta, S. Robinson, M. Bürgle, H. Düwiger, and U. Scheel. Faceted wikipedia search. In *Business Information Systems*, pages 1–11. Springer, 2010.

[10] A. Hassan, X. Shi, N. Craswell, and B. Ramsey. Beyond clicks: query reformulation as a predictor of search satisfaction. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 2019–2028. ACM, 2013.

[11] J. Jiang, A. Hassan Awadallah, X. Shi, and R. W. White. Understanding and predicting graded search satisfaction. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 57–66. ACM, 2015.

[12] Y. Liu, J. Bian, and E. Agichtein. Predicting information seeker satisfaction in community question answering. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 483–490. ACM, 2008.

[13] Y. Liu, Y. Chen, J. Tang, J. Sun, M. Zhang, S. Ma, and X. Zhu. Different users, different opinions: Predicting search satisfaction with mouse movement information. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 493–502. ACM, 2015.

[14] Y. Liu, Z. Liu, K. Zhou, M. Wang, H. Luan, C. Wang, M. Zhang, and S. Ma. Predicting search user examination with visual saliency. pages 619–628, 2016.

[15] D. C. Montgomery. *Design and analysis of experiments*. John Wiley & Sons, 2008.

[16] N. Moraveji. User interface designs to support the social transfer of web search expertise. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 915–915. ACM, 2010.

[17] N. Moraveji, S. Ahmad, C. Kita, F. Chen, and S. Kamvar. Weblines: Enabling the social transfer of web search expertise using user-generated short-form

timelines. In *Proceedings of the 9th International Computer-Supported Collaborative Learning Conference*, pages 112–119. ISLS, 2011.

[18] A. K. Ponnuswami, K. Pattabiraman, Q. Wu, R. Gilad-Bachrach, and T. Kanungo. On composition of a federated web search result page: using online users to provide pairwise preference for heterogeneous verticals. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 715–724. ACM, 2011.

[19] Y. Sun and Y. Zhang. Individual differences and online health information source selection. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pages 321–324. ACM, 2016.

[20] S. Sushmita, H. Joho, M. Lalmas, and R. Villa. Factors affecting click-through behavior in aggregated search interfaces. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 519–528. ACM, 2010.

[21] H. Wang, Y. Song, M.-W. Chang, X. He, A. Hassan, and R. W. White. Modeling action-level satisfaction for search task satisfaction prediction. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 123–132. ACM, 2014.

[22] R. W. White and D. Morris. 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 in information retrieval*, pages 255–262. ACM, 2007.

[23] K. Zhou, R. Cummins, M. Lalmas, and J. M. Jose. Evaluating aggregated search pages. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 115–124. ACM, 2012.

[24] K. Zhou, R. Cummins, M. Lalmas, and J. M. Jose. Which vertical search engines are relevant? In *Proceedings of the 22nd international conference on World Wide Web*, pages 1557–1568. ACM, 2013.

[25] K. Zhou, T. Sakai, M. Lalmas, Z. Dou, and J. M. Jose. Evaluating heterogeneous information access. In *Proceedings of MUBE workshop*, 2013.