Query Priming for Promoting Critical Thinking in Web Search

Yusuke Yamamoto
Shizuoka University
Hamamatsu, Japan
yamamoto@inf.shizuoka.ac.jp

Takehiro Yamamoto
Kyoto University
Kyoto, Japan
tyamamot@dl.kuis.kyoto-u.ac.jp

ABSTRACT

We propose *query priming* to activate careful user information seeking in web searches. Query priming employs query auto-completion (QAC) and query suggestion (QS) to present search terms that stimulate critical thinking and encourages careful information seeking and decision making.

We conducted an online user study using a crowdsourcing service. Analysis of search behavior logs and questionnaire responses confirmed the following. (1) With query priming, participants issued more queries and (re-)visited search engine result pages more frequently. (2) Query priming promoted webpage selection targeted at evidence-based decision making. (3) The query priming effect varied relative to participant educational background.

This study contributes to search interaction design to enhance user engagement in critical thinking in web searches.

CCS CONCEPTS

• **Information systems** → *Search interfaces*; • **Human-centered computing** → *User studies*;

KEYWORDS

Web search, critical thinking, priming effect, human factor

ACM Reference Format:
Yusuke Yamamoto and Takehiro Yamamoto. 2018. Query Priming for Promoting Critical Thinking in Web Search. In *CHIIR '18: 2018 Conference on Human Information Interaction & Retrieval, March 11–15, 2018, New Brunswick, NJ, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

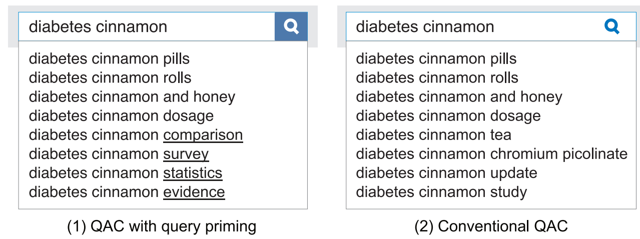


Figure 1: Comparison of QAC with query priming and conventional QAC (prime terms are underlined in this paper so that readers can identify them as manipulated)

1 INTRODUCTION

The web is an important source of information; however, lack of credibility, e.g., fake news disseminated on social networks, is a serious problem. Nevertheless, studies have found that people frequently do not consider the of web information [16, 17]. Therefore, creating information access environments that help people obtain credible information and make effective decisions is important.

Various search support systems have been proposed, such as evidence search systems [13], dispute suggestion systems [7, 22], and systems that visualize credibility-related scores [23]. However, such systems are only useful if users understand their purpose and appropriate use. We propose *query priming* whereby terms that promote careful information seeking and effective decision making are suggested. The proposed query priming is based on the priming effect in cognitive science. In cognitive science, the priming effect refers to a stimulus (hereafter *prime*) that activates a mental concept and influences subsequent behaviors [11]. A well-known example of the priming effect is the Florida effect experiment [3]. The experiment involved two groups of students. One group was asked to construct short sentences using words associated with older people, e.g., *worried*, *Florida*, *old*, *gray*, and *wrinkle*. Then, the researchers measured the students' walking speeds, and they found that the walking speed of students who constructed sentences using words associated with older people was statistically faster than that of those who did not consider such words.

Critical thinking is important when evaluating the quality and credibility of information [2]. We design *prime terms* to evoke critical thinking, careful information seeking, and effective decision making. We propose a search user interface (SUI) that displays prime terms in query auto-completion (QAC) or query suggestion (QS) terms. Figure 1 compares QAC with query priming results to conventional QAC results. As shown, we selected familiar prime terms as complementary information for search tasks and integrated them for query priming in SUIs.

We conducted a user study to determine the effect of query priming relative to promotion of critical thinking in web search processes. Our primary contributions are as follows.

- We consider a new query priming concept and propose an SUI that employs query priming.
- Through a user study and analysis, we show that query priming activates search/browsing behaviors that promote webpage selection for evidence-based decision making.
- We found that the effect of query priming varies relative to the educational background of search users.

2 RELATED WORK

2.1 Supporting credibility judgment on the web

Several studies have proposed algorithms to measure the correctness/credibility of web information, such as tweets, claims, and

CHIIR '18, March 11–15, 2018, New Brunswick, NJ, USA
2018. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

webpages. Pasternack et al. proposed an algorithm to measure information credibility by aggregating multiple sources that support or contradict given information [19]. Dong et al. developed a method to evaluate web page information credibility assuming that credible webpages have few false facts or claims[6]. Castillo et al. proposed a method to automatically assess the credibility level of news propagated through Twitter by analyzing tweets and reposts about news [5].

In addition, some studies focused on helping users assess credibility using qualitative information. Leong et al. developed an algorithm to retrieve evidence information from the web to help users verify the credibility of suspicious statements [13]. DISPUTE FINDER, developed by Ennals et al., highlights suspicious sentences in browsed webpages [7].

2.2 Searcher attitude and bias for careful information seeking on the web

It is known there is a great deal of unverified information on the web; however, several studies have reported that many people accept such information without consideration. For example, Nakamura et al. reported that more than 50% of people inherently perceive webpages retrieved by search engines as somewhat credible [17]. Morris et al. claimed that many people trust information from social network services more than search engine results, even though false information is often spread on social networks [16].

Even if people are aware of suspicious information, they often misjudge its credibility due to the use of incorrect heuristics, i.e., cognitive bias [11]. Leong et al. revealed the existence of domain bias whereby searchers believe that relevant webpages are authorized by specific domains [10]. White et al. studied the relation between beliefs about search topics and search behaviors [20]. They suggested that if a searcher's belief in a search topic is strong, it is difficult to shift the belief after search and browsing.

2.3 Enhancing and Activating effective user search behavior

Some researchers have developed methods to activate and enhance users' search activities. Harvery et al. revealed that providing examples of high-quality queries can help users formulate more effective queries [9]. Search Dashboard, proposed by Bateman et al., provides a UI that reflects search behaviors and summarizes search histories [4]. Their experiments indicated that Search Dashboard can help users modify their search behavior to improve search performance. Note that these studies focused on explicit feedback to enhance user search skills.

Some researchers have studied methods to change user behavior implicitly to provide better search experiences. Yamamoto et al. suggested that users spend more time viewing search engine result pages (SERP) and webpages if disputed topics are highlighted in web search [22]. Agapie et al. proposed an SUI that places a halo around a query box if a user inputs long queries. They reported that their proposed interface encourages users to input longer queries for better search results [1]. The query priming proposed in this paper is a type of implicit approach to encourage careful search experiences.

3 QUERY PRIMING DESIGN

3.1 Critical Thinking in Web Search Processes

The purpose of query priming is to encourage users to adopt critical thinking when seeking information on the web. Ennis defined critical thinking as logical and reflective thinking to determine what to believe or do [8]. Ennis also claimed that ideal critical thinkers are disposed to: seek reasons, consider the total situation, look for alternatives, and use critical thinking, e.g., deductive reasoning. We expect that, if search users are critical thinkers, to obtain correct and credible information from the web during web search processes, they will behave in the same manner which the information literacy researchers or librarians think is important [14]. Their search behaviors are expected to include: (1) spending more time searching, (2) issuing more queries to obtain appropriate information for decision making, (3) browsing more webpages for comparison, (4) checking evidence to support webpage content, such as the expertise of webpage authors, existence of valid references, and the freshness of webpages, and (5) collecting more evidence to support solid decisions. We designed an SUI with query priming to activate the above behaviors in web search processes.

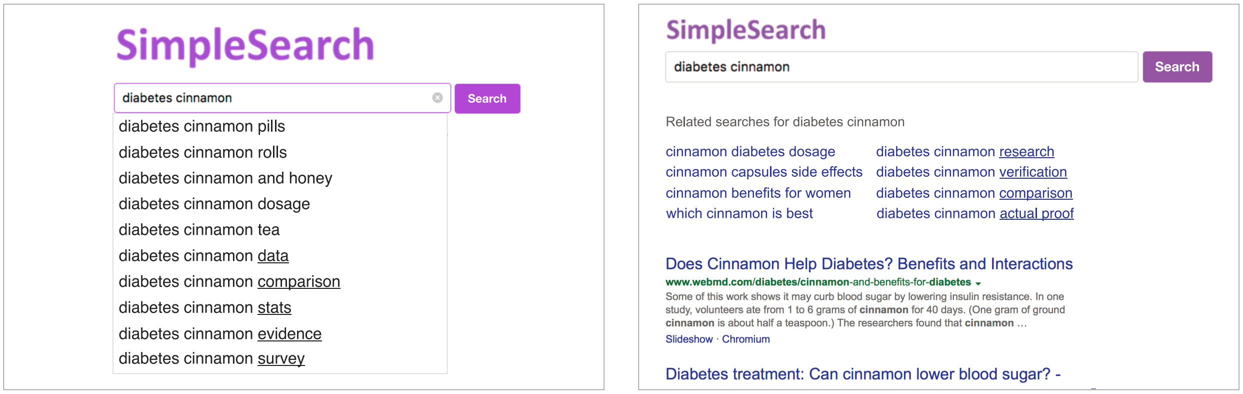
3.2 Designing Priming Effect in Web Search

The query priming is based on the ideomotor effect, which is a type of priming effect [3]. Here, the basic idea is that search engines present users with terms that encourage careful information seeking during or after issuing queries. To encourage careful information seeking without reducing search engine usability, we consider that query priming terms (hereafter *prime terms*) should satisfy several requirements.

First, for the priming effect to trigger a behavior change, prime terms should evoke critical thinking. Second, the prime terms displayed as complimentary information should not be unusual or unfamiliar. Even if the displayed prime terms are effective, if they are irrelevant to a given search task, they will not encourage a user to proceed with a web search. Third, the prime terms should trigger the priming effect in a variety of users. If possible, users should have several opportunities to look at the terms.

We employ the following approach to achieve the above requirements. To design effective prime terms, we focus on the following dispositions related to critical thinking: *awareness for logical thinking*, *inquiry-mindset*, *objectiveness*, and *evidence-based judgment*. We collect terms associated with these four dispositions as prime terms. Furthermore, we collect prime terms that search users capable of critical thinking are likely to use as search queries to guarantee the usability of search engines. We also choose query-independent terms as prime terms such that query priming can be applied to any search topic. To expose search users to the prime terms, we focus on QAC and QS to display prime terms. QAC [15] and QS [12] are very popular search engine features that provide search assistance information displayed close to query boxes. In a web search, a user can issue or modify queries in query boxes. Therefore, if a search engine displays prime terms using QAC/QS functions, search users are expected to consider such terms.

Screenshots of QAC and QS with query priming are shown in Figure 2. We include the prime terms in the QAC/QS term list. This approach promotes careful information seeking on the web.



(1) Query priming at a top page (query auto-completion) (2) Query priming at a SERP (query suggestion)

Figure 2: Screenshot of (1) QAC with query priming and (2) QS with query priming (prime terms are underlined in this paper so that readers can identify them as manipulated)

3.3 Research Questions

We expect that, if query priming activates critical thinking, user search/browsing behaviors would change both superficially and substantially. Furthermore, if the query priming effect is sustained after using an SUI that includes priming, it is possible that such SUIs can be applied to train critical thinking. However, it is unlikely that query priming will influence all users. Several studies have reported that university-level learning activities foster critical thinking abilities [18]. Thus, we assume that the effect of query priming will depend on pre-existing critical thinking abilities.

To explore the effects of query priming, we consider the following research questions.

- **RQ1:** Does query priming affect users' search and browsing behaviors, such as task completion duration, query issuing count, and page view count?
- **RQ2:** Does query priming change users' perspectives relative to searching and reading webpages to support decision making?
- **RQ3:** After using SUIs with query priming and if query priming affects users, are such effects sustained when SUIs without query priming are used?
- **RQ4:** Do the effects of query priming vary relative to educational background, where university education is assumed to represent critical thinking abilities?

4 PRIME TERM COLLECTION

We used a crowdsourcing service to collect and evaluate prime terms suggested in QAC and QS. In this section, we describe the procedure used to obtain the prime terms.

4.1 Collection of prime term candidates

We adopted two approaches to collect prime term candidates to stimulate critical thinking.

Table 1: Search topics to collect prime term candidates

Search topic
moving company contract, internet service contract, cancer treatment, dieting method, English learning method, overseas study option, buy microwave oven, buy TV, stock trading, apply for credit card

First, we collected terms which people imagined from critical thinkers using Lancers.jp[1], a Japanese crowdsourcing service. We randomly displayed one of the four dispositions described in Section 3.2 to the crowdsourced workers. Then, the workers were asked to write three keywords that describe typical attributes or behaviors that people with the shown disposition recalled. Using this approach, we collected 401 prime term candidates from 74 workers.

Second, we collected query terms that critical thinkers were likely to use when searching for a given topic. In this task, we asked crowd workers to assume that a searcher with one of the four dispositions was searching for each topic listed in Table 1. Then, the workers were asked to write three query terms that the fictional searcher was likely to use together with the search topic using the *AND* operator. Using this approach, we collected 317 prime term candidates from 54 workers.

4.2 Evaluation of prime term candidates

Then, we evaluated the collected prime term candidates.

[1] Lancers.jp: http://www.lancers.jp

In this evaluation, we asked crowd workers to imagine that a searcher issued a query containing one of the prime term candidates. Then, the workers evaluated the extent to which they considered the fictional searchers to be critical thinkers. We considered the degree of each evaluation as an association level of each prime term candidate with the critical thinking dispositions.

For each of the four critical thinking dispositions, we targeted the top 20 frequent prime term candidates for the evaluation task. The evaluation task was conducted as follows.

(1) We randomly allocated one of the four dispositions to each worker (here, disposition is denoted d).
(2) For each of the prime term candidates collected by d in Section 4.1, we performed the following (here, prime term candidate is denoted p).
  (a) We asked crowd workers to imagine that a searcher issued a query containing p.
  (b) The workers evaluated the extent to which they felt the searcher would have disposition d using a five-point Likert scale.

The following is an example task description.

> *Suppose that a searcher issued the query "<topic> AND comparison" in a web search. Do you feel that the searcher had an objective perspective? Please evaluate the extent to which you feel the searcher had an objective perspective using a five-point scale where -2 is never feel and +2 is feel very intensely. Note that the searcher often includes the term "comparison" in queries even when searching other topics.*

In this evaluation task, we allocated 50 workers to each of the four dispositions. Table 2 shows the top five highly-evaluated prime term candidates for each disposition. As prime terms for query priming, we used the 10 terms with the top three rankings that did not have similar terms in this table.

5 METHOD

This section describes the experimental design and the procedure employed in the user study.

5.1 Participants

We recruited 200 participants through Lancers.jp.We excluded 82 participants from the analysis because they did not complete all tasks or they unintentionally used other search engines, such as like Google. Thus, we used the data from 118 participants (55 were university educated).

The participants were asked to report how familiar they were with search engines using a five-point Likert scale (-2: completely unfamiliar; +2: completely familiar). We confirmed that most participants were familiar with search engines (*mean*: 1.36, *SD*: 0.73).

Each participant received approximately $4 for their time.

5.2 Tasks

Each participant performed 10 search tasks. As shown in Table 3, we prepared two types of questions for each task, i.e., *open question tasks* and *closed question tasks*. For each open question, the participants were asked to search for answer candidates and to provide one of the candidates as an answer. For each closed questions, they provided a *yes* or *no* response. The task categories included various topics, such as invention, science, and medicine. Note that several answer candidates could be found for each question.

In each task, the participants were asked to use our experimental system to search for answers and find evidential webpages.

Before starting each task, we asked the participants how familiar they were with the task question (five-point Likert scale; -2: completely unfamiliar; +2: completely familiar). On average, the number of the unfamiliar questions (i.e., less than 0) was 8.08 out of 10 (SD = 1.58). The mean task familiarity for the 10 tasks was -1.10 on average (SD = 0.50). This result confirms that most participants did not know the answers or were not confident of the answers prior to participating in the user study.

5.3 Design and procedure

In this experiment, we adopted a 2x2 between-subjects design to examine the effects of two factors, i.e., educational background and UI condition. The educational background factor had two levels, i.e., *university educated* and *not university educated*. The UI condition factor also had two levels. The first level was *control UI*, which provided conventional QAC/QS functions in a web search, and the second was *priming UI*, which displayed several query primes in QAC/QS expressions. Note that we provide a more detailed explanation of the UI condition in Section 5.4.

We allocated the participants to each UI condition randomly. Consequently, the participants were categorized into the four groups shown in Table 4. After the participants agreed to the consent form on the crowdsourcing website, they moved to our experimental website for the user study. The user study comprised four phrases.

The first was a practice phase. Here, we presented one sample search task to allow participants to become familiar with similar tasks and the search system.

The second phase was the *intervention phase*. In the intervention phase, participants were asked to search for answers to eight of the 10 tasks using the experimental search system. At the beginning of each search task, we presented the following description to introduce the task.

> *Does cinnamon help improve diabetes? Click the "Start search" button and search for an answer using our search system. When you find a satisfactory answer, come back to this webpage and report the answer and decisive evidence URLs (webpages). Here, multiple URLs are acceptable.*

Note that the search process was not time limited. Once a participant found a satisfactory answer for each task, they reported it and the URLs of the webpages that supported the answer (i.e., evidence webpages). In this phase, four tasks were selected randomly for each participant from the open and closed tasks. Note that the task order was randomized in the intervention phase, respectively.

The third phase, i.e., the *plain phase* examined the persistence of the query priming effect. Here, participants searched for answers to two of the 10 tasks (one open and one closed) using a search system wherein QAC/QS functions were disabled. Note that the two tasks did not overlap the eight tasks in the intervention phase and the task order was randomized. Before beginning each task, we

Table 2: Examples prime term candidates associated with critical thinkers (numbers in parentheses are the mean crowd workers ratings). Underlined terms were used for query priming in the user study.

Disposition type	Top five prime term candidates				
Awareness for logical thinking	principle (1.37)	evidence (1.22)	mechanism (1.22)	process (1.20)	proof (1.18)
Inquiry mind	survey (1.37)	research (1.33)	validation (1.31)	pursuit (1.16)	comparison (0.96)
Objectiveness	comparison (1.27)	stats (1.24)	analysis (0.98)	difference (0.75)	reputation (0.69)
evidence-based judgment	evidence (1.74)	actual proof (1.74)	data (1.56)	promise (1.52)	proof (1.48)

Table 3: Search task questions

Type	Question
Open	Who invented the light bulb?
	Who invented the telescope?
	Who invented the steam engine?
	What causes global warming?
	Why did dinosaurs become extinct?
Closed	Does cinnamon help improve diabetes?
	Does vitamin C help prevent pneumonia?
	Does ginkgo leaf help improve tinnitus symptoms?
	Can cocoa decrease blood pressure?
	Does garlic help improve and prevent a common cold?

Table 4: Participants allocation

	Educational background	
UI condition	University educated	Not university educated
control	29	31
priming	26	32

presented an introductory description similar to that presented in the intervention phase.

At the end of the experiment, participants were asked to complete a questionnaire (fourth phase). The questionnaire consisted of two sets of questions, i.e., (1) questions about system usefulness and (2) questions about perspectives relative to decision making during the search tasks.

We asked the participants to evaluate system usefulness relative to the following three metrics using a five-point Likert scale (-2: completely useless; +2: completely useful).

QAC/QS to modify input query represents how useful QAC/QS information was to modify the participants' input queries during search tasks.

QAC/QS to consider decision making perspectives represents how useful QAC/QS information was to consider viewpoints for evaluating webpages and answer candidates.

Displayed search result list represents how satisfactory the search results were relative to finding task answers and evidence.

To survey decision-making perspectives, we asked the participants to evaluate the extent to which they considered the following during the search tasks (five-point Likert scale: -2: very little; +2: very much).

Content completeness is the quantity of information in webpages.

Content freshness is the recency of the information in webpages.

Content objectivity is the extent to which participants felt that the content was objective or unbiased.

Content typicality represents how many webpages provided similar information.

Social reputation is the extent to which a webpage had a positive social reputation.

Content author represents who created the webpage content.

5.4 Experimental search system

For the user study, we developed a simple web search system with a UI similar to commonly used search engines, such as Google and Yahoo. The system had a top page and SERPs.

When participants input a query into a search box, the system displayed several query candidates below the search box using the QAC function on both the top and SERPs. If the UI condition was control, the system displayed a maximum of 10 query candidates in the QAC using the input query and the Google Suggest API[2]. If the UI condition was priming, the system replaced half of the Google Suggest queries with queries in which the prime terms in Table 2 were appended to a participant's input query term (Figure 2-(1)).

On SERPs, the system used the Bing Web Search API[3] to provide a list of search results for an given input query. The number of displayed results was set to 150, and each search result comprised a title, a snippet, and a URL. When participants clicked a title, the system opened a webpage in a different browser tab. In addition, the system displayed several queries related to the input query above and below the search results list using the QS function. If the UI condition was control, the system showed a maximum of eight related queries using the Bing Web Search API. If the UI condition was priming, the system replaced half of the Bing-related queries with query primes similar to the QAC function (Figure 2-(2)).

6 RESULTS

We obtained data from 1180 search tasks, and we analyzed the effects of the UI condition and participants' educational backgrounds on participants' search/browsing behaviors and decision making.

[2] http://suggestqueries.google.com/complete/search
[3] https://azure.microsoft.com/services/cognitive-services/bing-web-search-api/

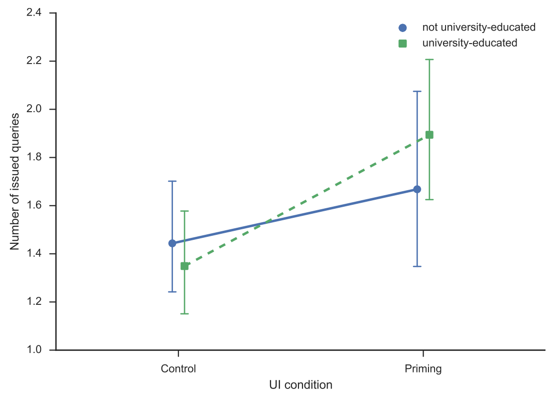


Figure 3: Number of issued queries in each task during the intervention phase divided by educational background and UI condition (error bar means 95% confidence interval)

This analysis was performed for both the intervention and plain phases. Moreover, we studied the effects on participants' search attitudes through an analysis of the questionnaire data.

We employed a non-parametric factorial ANOVA using aligned rank transformation [21] because we did not confirm the normality of the obtained data. All analyses were conducted using the *ARTool* statistical package[4].

6.1 Task performance

To answer RQ1, we first examined the *task completion duration* of each search task, which was measured by monitoring the time required to read the introductory task description and the time required to post an answer and its evidence URLs.

We expected that query priming would make the participants perform search tasks more slowly if the priming promoted critical thinking in the web search process. However, as shown in Tables 5 and 6, we did not find significant effects due to the UI condition and educational background in both the intervention and plain phases.

6.2 Search and browsing Behavior

To answer RQ1, RQ3, and RQ4, we analyzed participants' search and browsing behaviors during each task by examining how carefully they searched for answer candidates.

Query issue count is the number of queries participants issued in each task.

SERP visit count is the number of SERPs participants visited in each task.

Page visit count w/o SERP is the number of webpages participants visited in each task (excluding SERPs).

We found significant effects due to the UI condition and interaction between the UI condition and educational background relative to query issue count in the intervention phase (UI condition: $F_{(1,114)} = 15.8$, $p < 0.001$, partial $\eta^2 = 1.22\text{e-}1$; interaction: $F_{(1,114)} = 10.2$, $p < 0.01$, partial $\eta^2 = 8.19\text{e-}2$). As shown in Figure 5, regardless of whether the participants were university educated, those using the priming UI issued more queries on average than those using the control UI (university-educated group's *mean*: 1.35 vs. 1.89 in Table 5; not university-educated group's *mean*: 1.44 vs. 1.67 in Table 5). For the interaction between the UI condition and educational background, we conducted a post-hoc simple effect analyses of the query issue count running a Mann-Whitney's U test. The simple effect analyses revealed that the query priming UI for university-educated participants significantly elicited more frequent query issuing than the control UI for ones ($Z = -3.49$, $p < 0.001$, $r = 0.47$). This statistical result indicates that query priming led participants with university education to issue more queries than those without university education in the intervention phase, which displays the primes.

For the plain phase, Table 6 shows that participants using the priming UI issued more queries on average than those using the control UI (university-educated group's *mean*: 1.24 vs. 1.88; not university-educated group's *mean*: 1.31 vs. 1.45). Moreover, the query issue count was affected by educational background and the UI condition (educational background: $F_{(1,114)} = 4.71$, $p < 0.05$, partial $\eta^2 = 4.57\text{e-}2$; UI condition: $F_{(1,114)} = 5.46$, $p < 0.05$, partial $\eta^2 = 3.97\text{e-}2$). Note that we did not find statistical significance relative to the interaction between the two factors. These results indicate that, if participants used the priming UI in the intervention phase, even after no longer using the UI, they issued more queries than those who used the control UI.

With the SERP visit count, we found significant effects due to the UI condition in the intervention phase ($F_{(1,114)} = 4.56$, $p < 0.05$, partial $\eta^2 = 3.84\text{e-}2$). As shown in Table 5, participants using the priming UI visited more SERPs than those using the control UI regardless of educational background (university-educated group's *mean*: 4.72 vs. 6.13; not university-educated group's *mean*: 4.14 vs. 4.90). These results suggest that the query priming UI led the participants to (re-)check a list of search results more frequently after issuing queries or browsing webpages than the control UI.

This trend was confirmed in the plain phase ($F_{(1,114)} = 4.96$, $p < 0.05$, partial $\eta^2 = 4.17\text{e-}2$) (university-educated group's *mean*: 3.60 vs. 4.89 in Table 6; not university-educated group's *mean*: 4.30 vs. 5.98 in Table 6). These results suggest that, compared to the control UI, the priming UI led the participants to visit SERPs more frequently even after use of the UI ends.

For the visit count for webpages except SERPs, the participants did not exhibit any difference due to educational background and UI condition, both in the intervention and plain phases.

We can summarize the above results by stating that the priming UI had greater impact on search behavior with the search engines than the control UI. On the other hand, there was no difference between the two UI conditions' effects on browsing behavior outside the search engines.

6.3 Posted evidence

To answer RQ2, RQ3, and RQ4, we examined URLs posted by participants as evidence to support their task answers. We manually checked 309 unique posted URLs (webpages) relative to the following metrics, thereby examining the effects due to query priming on evidence-based decision making.

Number of evidence URLs is how many URLs participants posted as evidence for their task answer in each task.

[4] http://depts.washington.edu/madlab/proj/art/

Table 5: Participant behaviors and submitted answers during each task in the *intervention* phase broken down by educational background (EB) and UI condition (UI) (*: significance level at 0.001, **: 0.01, *: 0.05, and ·: 0.1)**

Metric		University educated		Not university educated		p-value		
		Control	Priming	Control	Priming	EB	UI	Interaction
Task performance	Task completion duration (s)	226.8	239.7	183.4	210.2	0.48	0.48	0.54
Search/browsing behavior	Query issue count	1.35	1.89	1.44	1.67	·	***	**
	SERP visit count	4.72	6.13	4.14	4.90	0.19	*	0.30
	Page visit count w/o SERP	6.06	8.13	5.90	5.97	0.36	0.28	0.40
Submitted evidence	Number of evidence URLs	1.52	1.90	1.47	1.40	*	0.19	*
	Reference validation (%)	49.6	61.1	48.4	53.9	0.41	*	0.41
	Author existence validation (%)	45.7	50.4	33.0	36.3	***	0.12	0.90
	Author expertise validation (%)	20.2	22.7	29.3	35.1	***	0.24	1.00
	TLD validation (%)	8.6	11.1	4.0	3.5	***	0.62	0.30

Table 6: Participant behaviors and submitted answers during each task in the *plain* phase broken down by educational background (EB) and UI condition (UI) (*: Significance level at 0.001, **: 0.01, *: 0.05, and ·: 0.1)**

Metric		University educated		Not university educated		p-value		
		Control	Priming	Control	Priming	EB	UI	Interaction
Task performance	Task completion duration (s)	171.8	228.1	163.8	200.6	0.66	0.12	0.69
Search/browsing behavior	Query issue count	1.24	1.88	1.31	1.45	*	*	·
	SERP visit count	4.30	5.98	3.60	4.89	0.11	*	0.35
	Page visit count w/o SERP	4.95	8.17	5.37	6.70	0.54	0.13	0.37
Submitted evidence	Number of evidence URLs	1.55	1.98	1.50	1.44	*	*	*
	Reference validation (%)	48.3	63.5	53.2	62.5	0.77	0.13	0.54
	Author existence validation (%)	41.4	46.2	43.5	45.3	0.89	0.60	0.79
	Author expertise validation (%)	19.0	32.7	25.8	37.5	0.51	*	0.89
	TLD validation (%)	3.4	15.4	8.1	7.8	0.17	*	·

Author expertise validation is the ratio of tasks where participants posted evidence URLs containing descriptions to confirm if the content authors had expertise in the task topic[5].

Author existence validation is the ratio of tasks where participants posted evidence URLs containing descriptions to confirm if the content authors actually exist[6].

TLD validation is the ratio of tasks where participants posted evidence URLs with credible top-level domains (TLD) suggesting governmental organization or academic organizations[7].

Reference validation is the ratio of tasks where participants posted evidence URLs containing authorized references to support the task answers[8].

For the number of evidence URLs in the intervention phase, we found significant differences due to educational background and the interaction between educational background and UI condition (educational background: $F_{(1,114)} = 5.45$, $p < 0.05$, partial $\eta^2 = 4.57\text{e-}2$; interaction: $F_{(1,114)} = 4.04$, $p < 0.05$, partial $\eta^2 = 3.42\text{e-}2$). As shown in Figure 4, if participants were university educated, participants using the priming UI posted more evidence URLs than those using the control UI (*mean*: 1.90 vs. 1.52 in Table 5). On the other hand, if participants were not university educated, those using the priming UI gave as many URLs as those using the control UI (*mean*: 1.40 vs. 1.47 in Table 5). The simple effect analyses with a Mann-Whitney's U test indicated that the UI condition factor for university-educated participants approached significance ($Z = -1.68, p = 0.094 < .1, r = 0.23$). These statistical results suggest that the query priming promoted participants with university education to find more evidence URLs, although the effect was not so statistically significant. On the other hand, we did not confirm that priming would affect those without university education.

This trend was observed for the number of evidence URLs in the plain phase. In the plain phase, the number of evidence URLs was affected by educational background, the UI condition, and the interaction between educational background and UI condition (educational background: $F_{(1,114)} = 4.16, p < 0.05$, partial $\eta^2 = 3.52\text{e-}2$;

[5]If a content author displayed governmental certificate (medical doctor and lawyer) on their webpage or if the author affiliation was shown as a governmental organization, academic organization, or publicly-listed companies on the webpage, we considered the author demonstrated expertise.
[6]If a webpage showed both a content author's real name and a legitimate affiliation, we considered the author to exist.
[7]We set the following TLDs as credible: *go.jp* (governmental originations in Japan), *.gov*, *ac.jp* (universities in Japan), and *.edu*.
[8]We defined that authorized references were articles published by academic associations or governmental organizations.

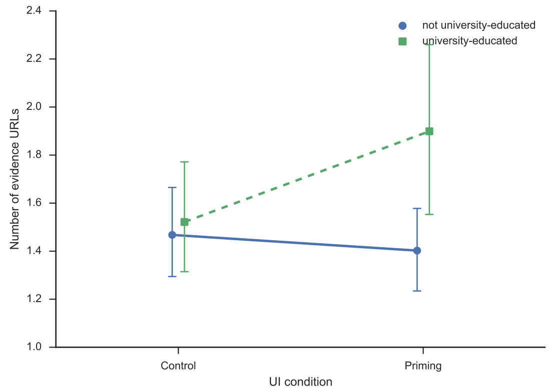


Figure 4: Number of URLs posted as answer evidence in each task during the intervention phase divided by educational background and UI condition (error bar means confidence interval at 95% level).

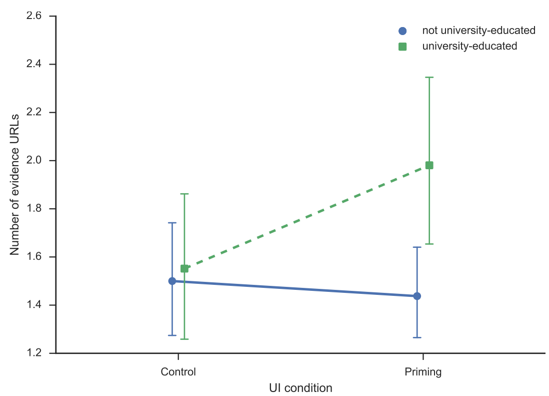


Figure 5: Number of URLs posted as answer evidence in each task during the intervention phase divided by educational background and UI condition (error bar means confidence interval at 95% level).

UI condition: $F_{(1,114)} = 3.99, p < 0.05$, partial $\eta^2 = 3.38\text{e-}2$; interaction: $F_{(1,114)} = 5.33, p < 0.05$, partial $\eta^2 = 4.47\text{e-}2$). Figure 5 shows that the number of posted evidence URLs when using the priming UI was greater than when using the control UI for only university-educated participants (university-educated group's *mean*: 1.98 vs. 1.55; not-university-educated group's *mean*: 1.44 vs. 1.50). The simple effect analyses with a Mann-Whitney's U test revealed that the UI condition factor for university-educated participants was close to being statistically significant ($Z = -1.93$, $p = 0.054 < .1$, $r = 0.26$). These results suggest that the priming UI can promote participants to find and post more evidential webpages when using the UI, even after use of the UI ends.

Note that reference validation in the intervention phase was affected by only the UI condition ($F_{(1,114)} = 4.08$, $p < 0.05$, partial $\eta^2 = 3.46\text{e-}2$). Table 5 suggests that, if participants without university education used the priming UI, they posted evidential webpages with validated reference for 5.5% more tasks on average than those using the control UI (Table 5). Moreover, we found that, if participants with university education used the priming UI, they posted reference-validated webpages as evidence for 11.5% more tasks on average than those using the control UI (Table 5). On the other hand, in the test phase, we did not find a statistical effect due to either educational background or the UI condition relative to the reference validation metric.

With the author expertise validation and TLD validation metrics, we found that the UI condition did not affect these metrics in the intervention phase; however, educational background did affect them (author expertise validation: $F_{(1,114)} = 12.4$, $p < 0.001$, partial $\eta^2 = 9.82\text{e-}2$; TLD validation: $F_{(1,114)} = 11.1$, $p < 0.001$, partial $\eta^2 = 8.87\text{e-}2$). Note that, from these results, we cannot confirm that while participants were using the priming UI, the function difference affected participant attention to author expertise and the TLD of the posted webpages. On the other hand, in the plain phase, a significant effect by the UI condition was observed (author expertise validation: $F_{(1,114)} = 4.16$, $p < 0.05$, partial $\eta^2 = 3.52\text{e-}2$; TLD validation: $F_{(1,114)} = 4.61$, $p < 0.05$, partial $\eta^2 = 3.89\text{e-}2$). The results suggest that, even after the priming UI was no longer used, it affected participant attention to webpage TLD and author expertise, although the effect did not appear while the UI was used.

For the author existing validation metric, we did not observe an effect due to the UI condition in either the intervention or plain phases.

6.4 Questionnaire

To perform a qualitative analysis, we examined the participants' answers to the questionnaire described in Section 5.3. Table 7 shows the average scores for system usefulness and viewpoint in decision making metric (-2: completely disagree; +2: completely agree).

As for the usefulness of QAC/QS for input query modification and the displayed search results, there was no significant effect due to educational background and UI condition, which suggests that participants did not complain about the query priming UI compared to the UI condition. Moreover, we found that the UI condition had significant effect on the usefulness of QAC/QS relative to considering the decision-making perspective ($F_{(1,114)} = 7.88$, $p < 0.01$, partial $\eta^2 = 6.47\text{e-}2$). Table 7 shows that, regardless of educational background, participants gave higher scores on this usefulness on average (university-educated group's *mean*: 0.68 vs. 0.94; not university-educated group's *mean*: 0.72 vs. 0.88). From this result, we consider that participants thought that the query candidates provided by the priming UI terms were more useful when investigating appropriate information to answer the task questions than those provided by the conventional QAC/QS UI.

Relative to the decision-making perspective, only the score of the content author viewpoint exhibited a significant difference relative to educational background and interaction between educational background and the UI condition (educational background: $F_{(1,114)} = 9.93$, $p < 0.01$, partial $\eta^2 = 8.01\text{e-}2$; the interaction: $F_{(1,114)} = 5.63$, $p < 0.05$, partial $\eta^2 = 4.70\text{e-}2$). Figure 6 shows that, if participants were university educated, those using the priming UI gave higher scores than those using the control UI (*mean*: 1.11 vs. 0.34 in Table 7). On the other hand, Figure 6 also indicates that,

Table 7: System usefulness and viewpoint for decision making in each task broken down by educational background (EB) and query completion/suggestion type (UI) (*: significance level at 0.001, **: 0.01, *: 0.05, and ·: 0.1).**

		University educated		Not university educated		*p*-value		
	Metric	Control	Priming	Control	Priming	EB	UI	Interaction
System usefulness	QAC/QS to modify input query	1.03	1.08	0.87	1.00	0.61	0.48	0.65
	QAC/QS to consider decision making viewpoint	0.72	0.88	0.68	0.94	0.61	**	0.92
	Displayed search result list	0.93	0.88	1.06	1.16	0.25	0.85	0.73
Viewpoint for decision making	Content completeness	0.90	1.15	0.84	0.88	0.59	0.49	0.91
	Content freshness	0.24	0.04	0.35	-0.19	0.81	·	0.41
	Content objectivity	1.34	1.35	1.03	1.03	·	·	0.74
	Content typicality	1.00	1.15	1.00	1.00	0.56	0.65	0.52
	Social reputation	-0.24	-0.27	-0.29	-0.41	0.93	0.87	0.86
	Content author	0.34	1.11	0.16	-0.25	**	0.84	*

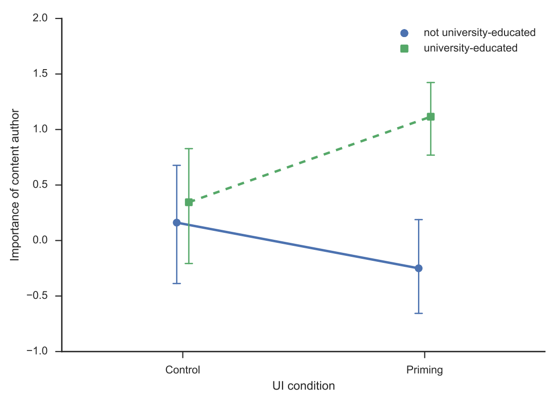


Figure 6: Extent to which participants considered content author in the search task divided by educational background and UI condition (error bar means 95% confidence interval)

if the participants did not have a university education, those using the priming UI gave lower scores than those using the control UI (*mean* -0.25 vs. 0.16 in Table 7). The simple effect analyses with a Mann-Whitney's U test indicated that the UI condition factor for university-educated participants was a certain trend toward significance ($Z = -1.78, p = 0.074 < .1, r = 0.24$). On the other hand, we found that there was no significant influence of the UI condition factor for not-university-educated partisipants. From these results, we consider that there was a trend that the query priming made participants with university education more careful about the content author, while the priming made participants without university education less careful.

7 DISCUSSION

In this section, we discuss answers to RQs 1-5 (Section 1) relative to the results of the user study.

For RQ1 and RQ4, we examined the task performance and search behaviors in a comparison between participants with and without university education. The task performance results demonstrate that both education background and the UI condition did not affect task completion duration (Section 6.1). In addition, the browsing behavior results indicate that query priming did not increase the page visit count (Section 6.2). If query priming promoted careful information seeking, we expected that participants using the priming UI would spend more time in their search tasks and visit more webpages in order to support their decision; however, the experimental results did not support this expectation.

On the other hand, the search behavior results indicate that the priming UI increased the frequency of issuing queries and (re-)visiting SERPs (Section 6.2). In particular, query priming had greater impact on the frequency of issuing queries for university-educated participants. The questionnaire results indicate that participants that used the priming UI had no complaints about the QAC/QS functions compared to those using the control UI. This questionnaire result denies that the participants using the priming UI issued more queries due to low QAC/QS performance (Section 6.4). From these results, we consider that query priming promoted frequent query issuance and SERP (re-)visits to carefully select which webpages to browse in limited time.

For RQ2 and RQ4, we examined the submitted evidence URLs and the questionnaire relative to the decision-making perspective. The questionnaire results indicate that query priming influenced the content author perspective in decision making (Section 6.4). As shown in Figure 6, if participants used the priming UI, those with university education felt more aware of the content author, while those without university education felt less aware.

However, behavioral results relative to evidence submission tell a different story. As shown in Table 5, the priming UI had no effect on content author factors, such as author existence validation, author expertise validation, and TLD validation. On the other hand, Table 5 indicates that more participants who used the priming UI submitted webpages containing valid references than those using the control UI. Moreover, Figure 4 reveals that university-educated participants using the priming UI submitted more evidential webpages than those using the control UI. From these results, we guess that query priming prompted university-educated participants to seek multiple evidence sources with authorized references even though they were not aware of that.

From the above discussion, we confirm that (1) using the query priming UI changed participant search behavior towards careful information seeking and (2) the effect positively appeared for participants with university education.

For RQ3, we do not insist from the experimental results that the query priming effect was sustained after use of the priming UI ended. Relative to the query issue count and SERP visit count, we observed that the priming UI affected participants during and after the use of the UI (Tables 5 and 6, and Figures 4 and 5). In addition, we observed persistence of the query priming effect on the number of posted evidence URLs for only university-educated participants (Figures 4 and 5). However, the effect on reference validation for the submitted evidence URLs was lost after disabling the QAC/QS functions. Furthermore, for author expertise validation and TLD validation on submitted evidence URLs, although a significant difference between the two UI conditions was observed after the functions were disabled, we did not observe a significant effect of the priming UI when the QAC/QS functions were enabled (Tables 5 and 6). Intuitively, we hypothesized that, if query priming promoted the above two points, the effect would at least appear when the QAC/QS functions were enabled. However, the observed results do not support this hypothesis. Here, one possible interpretation is that the query priming effect would appear with delay. In our user study, participants performed only two tasks in the plain phase compared to five tasks in the intervention phase. Therefore, it is possible that we did not obtain stable results relative to the submitted evidence. In future, we must design and perform a more suitable user study to validate the persistence of the query priming effect. In addition, a laboratory user study is required to examine the effect in detail because we had troubles to prevent some participants from unintentional use of the experimental systems.

8 CONCLUSION

In this paper, we have proposed a new *query priming* concept for QS and QAC towards careful information seeking. A search user interface with query priming presents terms that evoke critical thinking, thereby encouraging users to search for webpages carefully. The results of a user study using a crowdsourcing service indicated that proposed query priming concept changed participants' search behaviors and promoted the issuance of more queries and more frequent visits to SERPs. Furthermore, query priming affected the selection of webpages relative to evidence-based decision making. We found that, among university-educated participants, those using the priming UI collected more evidence with valid references.

Although the user study indicated that the priming UI had some effects while enabled, it did not provide sufficient support relative to the expectation that the query priming effect would be sustained even when the priming UI was not used. In future, we plan to conduct a laboratory study to carefully validate the persistence of the query priming effect.

Most current search tools focus on matching webpages to user information requirements. We believe that our work can support the design of search interactions to encourage careful information seeking and decision making on the web rather than solely relying on machine prediction and judgment.

ACKNOWLEDGMENTS

This work was supported in part by Grants-in-Aid for Scientific Research (17K17832, 16H01756, 16K16156) from MEXT of Japan.

REFERENCES

[1] Elena Agapie, Gene Golovchinsky, and Pernilla Qvarfordt. 2013. Leading People to Longer Queries. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2013)*. ACM, New York, NY, USA, 3019–3022.

[2] American Library Association (ALA). 2000. *Information literacy competency standards for higher education.*

[3] John A Bargh, Mark Chen, and Lara Burrows. 1996. Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of personality and social psychology* 71, 2 (1996), 230–244.

[4] Scott Bateman, Jaime Teevan, and Ryen W. White. 2012. The Search Dashboard: How Reflection and Comparison Impact Search Behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2012)*. ACM, New York, NY, USA, 1785–1794.

[5] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information Credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web (WWW 2011)*. 675–684.

[6] Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. 2015. Knowledge-based Trust: Estimating the Trustworthiness of Web Sources. *Proceedings of the VLDB Endowment* 8, 9 (2015), 938–949.

[7] Rob Ennals, Beth Trushkowsky, and John Mark Agosta. 2010. Highlighting Disputed Claims on the Web. In *Proceedings of the 19th International Conference on World Wide Web (WWW 2010)*. 341–350.

[8] Robert H. Ennis. 1987. A taxonomy of critical thinking dispositions and abilities. In *Series of books in psychology. Teaching thinking skills: Theory and practice*, J. B. Baron and R. J. Sternberg (Eds.). W H Freeman/Times Books/ Henry Holt & Co, New York, 9–26.

[9] Morgan Harvey, Claudia Hauff, and David Elsweiler. 2015. Learning by Example: Training Users with High-quality Query Suggestions. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2015)*. ACM, 133–142.

[10] Samuel Ieong, Nina Mishra, Eldar Sadikov, and Li Zhang. 2012. Domain Bias in Web Search. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM 2012)*. ACM, 413–422.

[11] Daniel Kahneman. 2011. *Thinking, fast and slow.* Macmillan.

[12] Diane Kelly, Amber Cushing, Maureen Dostert, Xi Niu, and Karl Gyllstrom. 2010. Effects of popularity and quality on the usage of query suggestions during information search. In *Proceedings of the 28th SIGCHI Conference on Human Factors in Computing Systems (CHI 2010)*. ACM, 45–54.

[13] Chee Wee Leong and Silviu Cucerzan. 2012. Supporting Factual Statements with Evidence from the Web. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM 2012)*. ACM, 1153–1162.

[14] Marc Meola. 2004. Chucking the checklist: A contextual approach to teaching undergraduates Web-site evaluation. *portal: Libraries and the Academy* 4, 3 (2004), 331–344.

[15] Bhaskar Mitra, Milad Shokouhi, Filip Radlinski, and Katja Hofmann. 2014. On user interactions with query auto-completion. In *Proceedings of the 37th international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2014)*. ACM, 1055–1058.

[16] Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich. 2010. What Do People Ask Their Social Networks, and Why?: A Survey Study of Status Message Q&A Behavior. In *Proceedings of the the 28th ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2010)*. ACM, 1739–1748.

[17] Satoshi Nakamura, Shinji Konishi, Adam Jatowt, Hiroaki Ohshima, Hiroyuki Kondo, Taro Tezuka, Satoshi Oyama, and Katsumi Tanaka. 2007. Trustworthiness Analysis of Web Search Results. In *Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2007)*. Springer, 38–49.

[18] Ernest T. Pascarella. 1989. The development of critical thinking: Does college make a difference? *Journal of College Student Development* 30, 1 (1989), 19–26.

[19] Jeff Pasternack and Dan Roth. 2013. Latent Credibility Analysis. In *Proceedings of the 22nd International Conference on World Wide Web (WWW 2013)*. 1009–1020.

[20] Ryen White. 2013. Beliefs and Biases in Web Search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2013)*. ACM, 3–12.

[21] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2011)*. ACM, 143–146.

[22] Yusuke Yamamoto and Satoshi Shimada. 2016. Can Disputed Topic Suggestion Enhance User Consideration of Information Credibility in Web Search?. In *Proceedings of the 27th ACM Conference on Hypertext and Social Media (HT 2016)*. ACM, 169–177.

[23] Yusuke Yamamoto and Katsumi Tanaka. 2011. Enhancing Credibility Judgment of Web Search Results. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2011)*. ACM, 1235–1244.