Can Disputed Topic Suggestion Enhance User Consideration of Information Credibility in Web Search?

Yusuke Yamamoto
Kyoto University
Yoshida-honmachi, Sakyo, Kyoto, Japan
yamamoto@gsm.kyoto-u.ac.jp

Satoshi Shimada
Kyoto University
Yoshida-honmachi, Sakyo, Kyoto, Japan
shimada@gsm.kyoto-u.ac.jp

ABSTRACT

During web search and browsing, people often accept misinformation due to their inattention to information credibility and biases. To obtain correct web information and support effective decision making, it is important to enhance searcher credibility assessment and develop algorithms to detect suspicious information. In this paper, we investigate how credibility alarms for web search results affect searcher behavior and decision making in information access systems. This study focuses on disputed topic suggestion as a credibility alarm approach. We conducted an online user study in which 92 participants performed a search task for health information. Through log analysis and user surveys, we confirmed the following. (1) Disputed topic suggestion in a search results list makes participants spend more time browsing pages than ordinary search conditions, thereby promoting careful information seeking. (2) Disputed topic suggestion during web browsing does not change participant behaviors but works as complementary information. This study contributes to system designs to enhance user engagement in critical and careful information seeking.

CCS Concepts

•**Information systems** → **Search interfaces;** *Web searching and information discovery;* •**Human-centered computing** → **Empirical studies in HCI;**

Keywords

Web search; careful information seeking; credibility; decision making; behavior analysis

1. INTRODUCTION

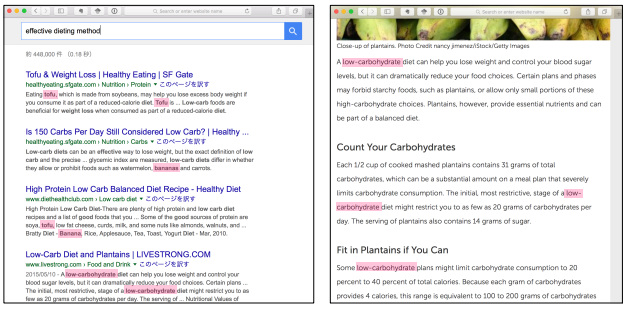
Currently, people frequently rely on web pages to acquire various types of information. Such information varies from lightweight reading to significant information that can affect their lives. However, if people do not consider information credibility, they can be misled easily. Several studies have reported that people often trust web information without considering its credibility. Nakamura et al. reported that more than 50% of people perceive web pages retrieved by search engines as credible [17]. Lindgaard et al. claimed that people often trust visually appealing web pages based on initial impressions [13]. Morris et al. stated that many people trust information from social network services more than search engine results [16], even though false information is often spread on social networks [14].

Unlike information published or broadcast by conventional media, information in web pages is rarely verified before publication. The Pew Research Center reports that more than 70% of American Internet users searched for health information in the period 2011-2012[1]. However, according to Sillence et al., less than 50% of medical websites have been authenticated by medical experts [22]. There is a great deal of unverified information on the web, and many people accept such information without careful consideration. Therefore, it is important to create an information access environment where people can obtain credible information to support effective decision making. In this paper, we study one possible approach to support the assessment of information credibility.

Various types of support systems have been proposed, including evidence information search [11], disputed sentence suggestion [7, 26], and systems that provide scores according to credibility criteria [27, 21]. These systems are helpful if people can use them to support decision making. However, people often search for, interpret, and favor information that confirms their pre-existing beliefs and biases [10]. Therefore, some people accept information as credible when systems provide complementary information conforming to their pre-existing beliefs. In addition, people often do not feel the need for support systems because they assume the information on the web is credible [17], and others believe that they can identify correct information effectively [25]. Thus, we must consider how to enhance "careful" information seeking. Unless people are careful, they will not read web pages critically or compare multiple information sources, and they may accept suspicious information as credible information. Developing careful information seeking enables people to use the above mentioned support systems more appropriately and effectively in order to enhance decision making.

In this paper, we study the relationship between credibility warnings and user behavior. This relationship has implications for the design to enhance user engagement in careful information seeking. In the field of communication and persuasion, researchers have discussed threat appeal as a possible approach to change people's attitudes [20]. Our study focuses on the suggestion of disputed topics as a type of threat appeal. Here, disputed topics are those that some people claim are suspicious, regardless of whether the topics are actually true or not. In this study, we focus on two situations to

[1]PewResearchCenter's Health Fact Sheet, http://www.pewinternet.org/fact-sheets/health-fact-sheet/



(a) Disputed topic suggestion in scanning a list of Web search results.

(b) Disputed topic suggestion in browsing Web pages

Figure 1: Examples of disputed topic suggestion during search and browsing

suggest disputed topics: *scanning a list of web search results* and *browsing web pages in the list* (Figure 1). We conducted an online user study where participants search for health information using our experimental system. Based on the results, we examine how disputed topic suggestion during web search influences user search behavior and attitude under both conditions.

Our primary contributions are as follows.

- The effect of disputed topic suggestion depends on suggestion timing even though the same disputed topic is highlighted.
- Disputed topic suggestion while scanning a list of search results has significant influence on user attitude toward information seeking, leading them to slow search and browsing.
- When users see disputed topics when browsing web pages, they often use them as complementary information to support decision making. The suggestions do not affect user search behaviors significantly.

2. RELATED WORK

2.1 Measuring correctness and credibility of web information

One possible approach against suspicious web information is measuring the correctness or credibility of the given information. Several studies have presented algorithms to measure the correctness probability of information. Galland and Pasternack et al. developed algorithms to measure the credibility of information by aggregating multiple sources that support or contradict the given information [8, 18]. Dong et al. described a method to evaluate web page information credibility assuming that credible web pages have few false facts or claims [6]. Although these algorithms appear hopeful, they only work well for domains where prior knowledge is available.

Rather than measuring correctness, some researchers have developed methods to measure the credibility of web information from specific sources, such as Wikipedia, Twitter, claims, and web pages. Adler et al. proposed an algorithm to measure the credibility of sentences in Wikipedia articles under the hypothesis that sentences are credible if they have not been edited [1]. Castillo et al. studied a method to automatically judge the credibility level of news propagated through Twitter by analyzing tweets and re-posts about news [3]. These algorithms focus on various criteria that comprise credibility, such as objectivity, authority, and freshness [15]. However, credible information is not always correct and vice versa. Measuring credibility is inappropriate for misinformation detection, although the calculated scores can help people assess the credibility of web information.

2.2 Supporting user credibility judgment on the web

Some studies have focused on helping users judge credibility. Suh et al. introduced WIKIDASHBOARD, a system to visualize edit histories in Wikipedia articles [24]. Pirolli et al. examined how WIKIDASHBOARD affects user credibility assessment of Wikipedia articles [19]. Leong et al. developed an algorithm to retrieve evidence information from the web so that users can verify the credibility of suspicious statements [11]. Some researchers have proposed prototype systems that visualize scores of web search results according to various credibility criteria [21, 27]. Measuring the credibility of web information can be useful if users are willing to filter out non-credible web information according to their own credibility criteria. Regarding the credibility of multimedia data, Diakopoulos et al. introduced a system to support credibility judgment of video content by visualizing user evaluations and comments about the content [5].

Some studies have focused on dispute suggestion to notify which information users should consider. Our study extends this approach. DISPUTE FINDER developed by Ennals et al. highlights suspicious sentences in actively browsed web pages [7]. Yamamoto proposed suspicious sentence suggestion as a new type of query suggestion in web search engines [26]. These studies conducted qualitative user studies to evaluate the usefulness of the proposed systems. However, little attention has been paid to how such systems actually influence user attitudes and behaviors. In this paper, we study the effects of suspicious information suggestion on user search attitudes and behaviors.

2.3 Bias in web search and browsing

If all people were strongly motivated toward careful and critical information seeking on the web, the previously described measurements and support systems would be useful. However, few people are conscious of information credibility in web search and browsing. As stated in the previous section, Nakamura et al. reported that many search engine users inherently perceive web pages in search result lists as somewhat credible [17].

Furthermore, even if people are aware of suspicious information, they often make mistakes in credibility judgment owing to the use of incorrect heuristics, which is known as cognitive bias [10]. Some types of cognitive bias have been found in the information retrieval field. Leong et al. revealed the existence of domain bias whereby searchers often believe that relevant web pages are authorized by particular domains [9]. Clarke et al. analyzed the clickthrough logs of a commercial web search engine. They indicated that searchers are more likely to select search results that present more readable snippets [4]. White et al. studied the relation between pre-existing beliefs about search topics and search behaviors [25]. They showed that, if searchers have weak belief in search topics, they are likely to change their belief after performing a search. They also indicated that, if searcher belief in the search topic is strong, it is difficult to shift their belief after search and browsing.

These studies suggest that focusing on credibility and avoiding cognitive biases are important issues to prevent people from consuming incorrect information.

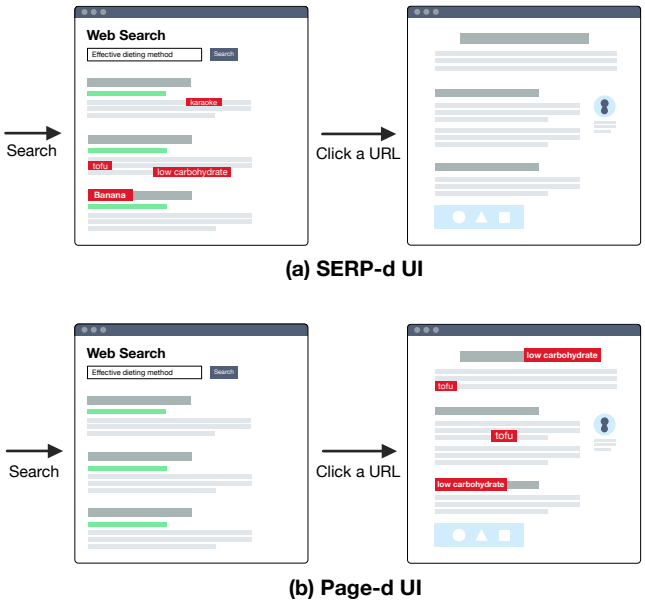


Figure 2: Overviews of (a) SERP-d UI behavior and (b) Page-d UI behavior (each UI highlights disputed topics in red)

3. METHODS

We conducted an online study with 92 participants to understand how different disputed topic suggestion styles affect search behaviors. In the study, we asked participants to search the web for effective treatment or prevention of several health problems using three search user interfaces (UIs). Our study was conducted to address the following questions.

- Does disputed topic suggestion focus searcher attention on suspicious topics on the web?
- Which disputed topic suggestion style encourages searchers to scrutinize web search results in order to obtain the most credible information?
- How do searchers use suggested disputed topic information for their final decision making in search tasks?

The primary objective of this study is to determine whether disputed topic suggestion encourages web searchers to engage in careful information seeking. Therefore, we do not focus on improving search task efficiency and preciseness.

3.1 Conditions

This study adopted a one-way factorial design with three search UI conditions: **Control**, **SERP-d**, and **Page-d**. In the Control UI, our experimental search system simply returns a list of search results for a given query. Each search result contains a title, a snippet (content summary), and a URL, similar to conventional search engines. The SERP-d UI highlights disputed topics only in search engine result pages (SERP). If disputed topics about a given query exist, disputed topics appearing in titles and snippets on SERPs are highlighted in red, as shown in Figure 2(a). In the Page-d UI, disputed topics are highlighted while web pages are being browsed (Figure 2(b)). Note that participants using Page-d cannot notice which topics are considered disputed before visiting the web pages.

Table 1: Search topics and disputed topic examples

Search topic	Disputed topic example
Cancer	Avastin, BCG[2], MTX[3], macrobiotic, carbohydrate, hormone therapy, vitamin C
Dieting (weight loss)	karaoke, banana, water, soda, half-body bathing, low-carbohydrate, tofu
Acne	acerola, olive, steroid, yogurt, Vaseline, water face-wash, laser
Hangover	turmeric, caffeine, sauna, aquarius, cyrenidae, honey, salt plums
Atopy	BCG, eel, chocolate, bread, baby oil, Grifola frondosa, antihistamine
Depression	SNRI[4], SSRI[5], Inderal, screening, alcohol, exercise, mental therapy
Constipation	konjac, sesame, prune, lactobacillales, bok choy laxative, enzyme diet
Pollen allergy	protein, smoking, alcohol, margarine, Celestamine, roe, Chinese herb
Asthma	β2-adrenergic agonists, alcohol, APAP[6], Meptin, smoking, sugar, milk

3.2 Materials

3.2.1 *Topics for search tasks*

We selected nine search topics about health. The topics are listed in Table 1. These search topics are popular in Japan; however, many web pages about the topics contain suspicious approaches to cures and prevention (e.g., home remedies and urban legends).

3.2.2 *Disputed topics*

We define a disputed topic as having negative comments on some web pages. To collect disputed topics related to the chosen search topics, we used WISDOM X, a semantics-oriented web mining service by National Institute of Information and Communications Technology, Japan[7]. WISDOM X provides a question answering function that provides answer candidates using natural language questions. WISDOM X applies natural language processing techniques to indexes of 100 million Japanese web pages to extract a list of answer candidates (e.g., *What is effective for weight loss?* $\rightarrow$ *low-carb diet*) [2].

To obtain disputed topic candidates for each search topic shown in Table 1, we issued the query "What is not effective for {*search topic*}" into WISDOM X. Then, we filtered out unknown or meaningless topic candidates from the gathered candidates manually. Examples of the obtained disputed topics are given in Table 1.

3.2.3 *Web page collection for search tasks*

Web pages displayed during each task were gathered using the Microsoft Bing Search API[8] before the online study. We gathered two types of web pages: (1) pages containing disputed topics

[2]BCG: Bacillus Calmette Guerin
[3]MTX: Methotrexate
[4]SNRI: Serotonin & Norepinephrine Reuptake Inhibitors
[5]SSRI: Selective Serotonin Reuptake Inhibitors
[6]APAP: parracetamol
[7]WISDOM X: http://wisdom-nict.jp (in Japanese)
[8]Microsoft Bing Search API:
https://datamarket.azure.com/dataset/bing/search

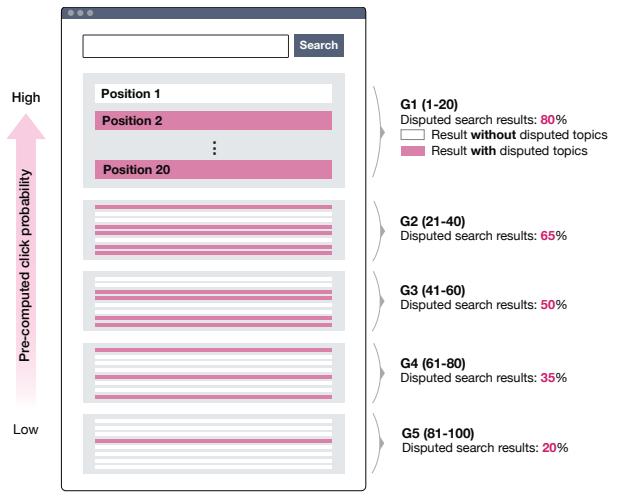


Figure 3: Manipulation of search result listing. Disputed and non-disputed pages are allocated to red and white slots, respectively. web pages with higher click probability appear at higher ranks.

(*disputed web pages*) and (2) pages not containing disputed topics (*non-disputed web pages*). To collect 100 disputed web pages for search topic q and disputed topic d, we issued the query "*effective* $\wedge$ $q \wedge d$" to the Bing API. Hundred non-disputed web pages were collected with the query "*effective* $\wedge$ q." We checked the relevance of the obtained search results for the given query manually and stored the titles, snippets, URLs, and raw HTML of the relevant results as our test set.

3.2.4 *Listing search results*

For any UI condition, immediately after beginning search tasks, our experimental system presented a controlled list of search results for a pre-defined initial query. As is well known, most users look at the first 10 search results only [23]. However, the number of displayed search results was fixed to 100 per query in this experiment. This manipulation was intended to enable participants to investigate as many web pages as possible for obtaining credible information. Each search result comprised a title, a snippet, and a URL. The same list of search results was displayed for the same search topic under each UI condition. Note that the participants were not permitted to change the initial query. They selected and browsed web pages from the fixed list. Furthermore, link navigation was disabled in order to track participant behavior.

Note that we selected search results to display and adjusted the rank position of each result prior to conducting the user study. This was done to avoid cases in which (1) few disputed topics in SERPs would attract participant attention with the SERP-d UI and (2) participants using the Page-d UI would visit few pages with disputed topics.

For case 1, we manipulated lists of search results such that more search results whose titles or snippets contained disputed topics would appear at higher ranks. As shown in Figure 3, the list of search results for each search topic was divided into five groups (G1, G2, G3, G4, and G5). In each group, 80%, 65%, 50%, 35%, and 20% of the search results had disputed topics in their title or snippet, respectively. We randomized the positions of the slots where disputed web pages and *non-disputed web pages* were allocated within each group.

For case 2, we computed the clickthrough probability of the *disputed web pages* and *non-disputed web pages*. We then positioned the pages into the slots for *disputed web pages* (red slots in Fig 3) and *non-disputed web pages* (white slots in Fig 3) in order of clickthrough probability. For this clickthrough probability pre-computation, we used Lancers.jp[9], a crowd sourcing service in Japan. We asked workers on the crowdsourcing service to scan lists of our collected web pages about the search topics and find the web pages to satisfy information needs behind the search topics. The order of the web pages in the list was randomized for each worker. In this task, disputed topics were not highlighted (i.e., the Control UI). The workers' clickthrough logs were stored during their tasks. More than 100 workers were allocated to a single search topic. A total of 1012 workers participated in this preliminary investigation task. Finally, we computed the clickthrough probability of the collected web pages using the obtained logs.

3.3 Participants

A total of 121 participants were recruited via Lancers.jp. None participated in the preliminary investigation tasks. All participants reported that they were familiar with searching and browsing the web. They were randomly assigned to one of 27 groups (3 UIs x 9 topics) per task. Note that the maximum number of tasks each participant could perform was limited to nine. The participants received 50 Japanese yen (around 50 cents) for each response as compensation.

In this study, we collected 507 responses from the 121 participants. We omitted 27 responses for which the participants did not complete the tasks and 180 responses for which the participants' behavior data was not monitored due to browser Javascript settings. Finally, we used 310 valid responses from 92 participants.

3.4 Procedure

The online study consisted of four parts: (1) registration, (2) task introduction, (3) search task, and (4) questionnaire.

First, participant candidates checked a recruiting message on Lancers.jp. The recruiting message stated that the study's objective was to survey what people considered effective for health. The message also showed that the required time per task would be between 5 and 10 minutes. Once participant candidates agreed to participate in the study, they enrolled and proceeded to our study's website.

When the participants visited the website, we randomly allocated them a search topic and a UI condition. The participants were asked to read a brief introduction to the task procedure. Furthermore, for the participants allocated the SERP-d UI or Page-d UI, the website described that disputed topics would be highlighted during the search task using an example about an athlete's foot remedy. The description explained that when some web pages suggested that a topic was ineffective for a target health problem, the system would highlight the topic in red. After the brief introduction, the participants were presented with the following task scenario.

> *You have troubles with your health and want to lose weight. You are about to search the web for effective weight loss diet methods. Please start a web search from the following link, and then find and report effective methods.*

After reading this scenario, the participants clicked a link to start the web search. Once moving to the web search page, a list of

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

Table 2: Mean and standard error of the mean (SEM) of pageviews for the three UIs

UI condition	Control	Page-d	SERP-d
Mean pageviews	5.24	5.55	4.64
(SEM)	(0.634)	(0.537)	(0.547)

100 search results was presented, similar to conventional search engines. The participants were asked to browse the web pages in the list without time limitation. When they found their answer (effective method for the target health problem), they finished the search and reported the answer on the study's website. If they did not find satisfactory answers, they were allowed to report "nothing effective." During the search tasks, the participants' clickthroughs and dwell times on web pages were stored using Javascript.

Finally, the participants using the SERP-d UI or Page-d UI were asked to fill out questionnaire about the usefulness and useful aspects of the UIs. The survey about usefulness was answered on a 5-point Likert scale (1 = never useful, 3 = neutral, 5 = very useful). For the useful aspect survey, the participants were asked to make multiple selections from the following list.

- **Attention calling for credibility-aware search**: to become more conscious of information credibility during web search
- **Notification of information to be checked**: to find which disputed topics should be examined in detail
- **Suspicious topic filtering**: to filter out suspicious topics
- **Others**: unlisted useful aspects
- **Nothing**: no useful aspects

After the search task and survey, a unique confirmation code was issued to the participants. The participants copied and pasted the code to Lancers.jp, thereby allowing us to pay task rewards.

4. RESULTS

Here, we present our results for the behavior analysis and post questionnaire. Our analytical results show that SERP-d had a more significant impact on participant behaviors in search tasks than Page-d.

For statistical testing, we utilized a one-way analysis of variance (ANOVA) with the three UI conditions as factors. For post-hoc pairwise comparisons, we used the Tukey–Kramer test. The analysis of the post questionnaire used Pearson's χ^2 test.

4.1 Pageviews

To evaluate whether Page-d and SERP-d encouraged participants to check multiple information sources for careful decision making, we examined the number of web pages (pageviews) that participants viewed in the three UIs using clickthrough logs. The results are given in Table 2. Our ANOVA results revealed no statistical difference between the mean pageviews per task for the participants using Control (5.24, SEM = 0.634), SERP-d (4.64, SEM = 0.547), and Page-d (5.55, SEM = 0.537) ($F(2, 307)$=0.638, p = 0.529, η^2 = 4.14-3).

To determine how Page-d and SERP-d influenced participant page selection, we measured the mean clickthrough ratio of disputed web search results per task with the three UIs. The clickthrough ratio means how much percentage of web search results which a participant clicked during his/her task contained disputed

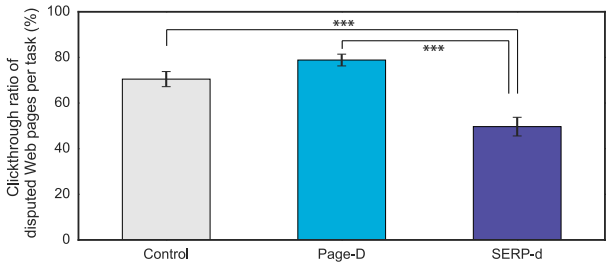


Figure 4: Clickthrough ratio of disputed web pages per task for the three UIs (significant differences are shown by lines)

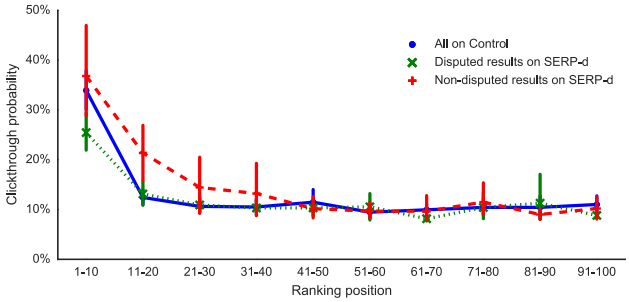


Figure 5: Clickthrough probability for disputed/non-disputed web search results in each ranking position group

topics in the SERPs. Although Control and Page-d did not highlight disputed topics in the SERPs, we checked and computed how often the participants using these UIs clicked disputed web search results that SERP-d highlighted in the SERPs. There was a significant difference between the three UIs for the clickthrough ratio of disputed web pages ($F(2, 307)$=19.8, $p^{**} < .01$, η^2 = 0.114). Figure 4 shows that our post-hoc pairwise comparison analyses reported that the mean clickthrough ratio of SERP-d was much less than that of Control (49.7% < 70.5%; $p^{**} < .01$) and Page-d (49.7% < 78.9%; $p^{**} < .01$). This indicates that the disputed topic highlighting in SERPs helped participants avoid disputed web pages before visiting them.

We also analyzed the clickthrough probability for web search results in each ranking position by comparing SERP-d and Control. Here, the clickthrough probability of a web search result means the percentage of participants that clicked the given search result. As seen in Figure 5, when the participants used SERP-d, non-disputed web pages were clicked in SERPs with higher probability than disputed web pages in the 1–40 rank positions. In particular, for the 1–20 positions, the clickthrough probability for non-disputed web pages was more than 10% higher than that of disputed web pages. These results support the finding about clickthrough ratio per task, indicating that participants clicked non-disputed web search results more frequently than disputed results. Furthermore, as shown in Figure 5, in contrast to Control, the clickthrough probability for non-disputed web pages with SERP-d decreased gradually to lower-ranked results. This suggests that participants using SERP-d tried to check the middle-positioned search results as well as the high-positioned group, because the middle-positioned one had fewer disputed search results than the high-positioned one. On

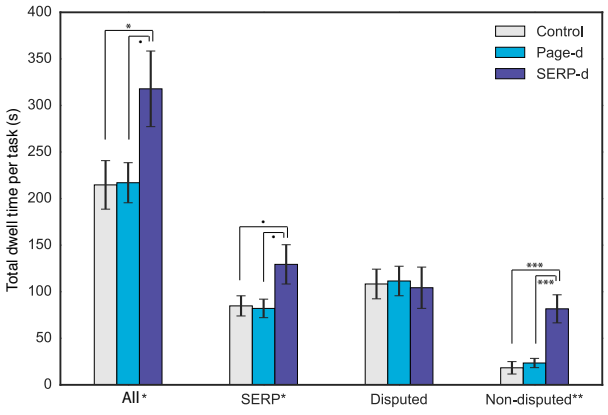


Figure 6: Total dwell time per task for all pages, SERPs, disputed pages, and non-disputed pages with the three UIs (significant differences are indicated by lines (·=p < 0.1, ∗=p < .05, ∗∗=p<.01))

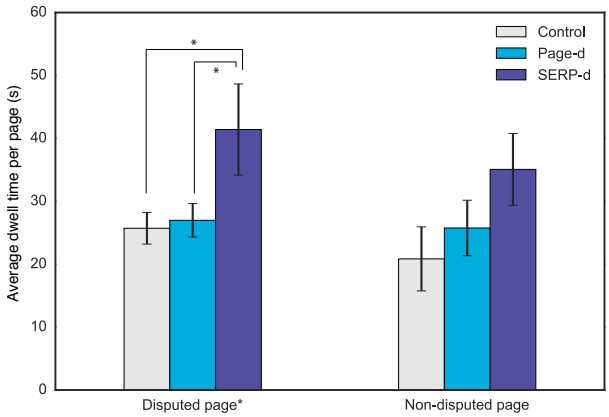


Figure 7: Average dwell times per disputed and non-disputed web pages for the three UIs

the other hand, participants using Control often concentrated on browsing high-ranked web search results.

4.2 Dwell time

To observe how Page-d and SERP-d enhanced participant engagement in careful reading of web information, we examined dwell time while viewing SERPs and web pages with the three UIs. Figure 6 shows the mean total dwell times per task for viewing SERPs, disputed web pages, non-disputed web pages, and all pages. The total dwell times for disputed/non-disputed web pages for Control were measured by checking how long a participant stayed on web pages where disputed topics would/would not have been highlighted when using SERP-d or Page-d. In terms of total dwell time for all web pages (session time), there was significant statistical difference between the three UIs ($F(2, 307)$=3.75, p^* < .05, η^2 = 2.38e-2). Furthermore, as shown in Figure 6, the session time for SERP-d was significantly longer than that for Control (317.8s > 214.7s; p^* < .05). A marginal trend toward significance was found between SERP-d and Page-d (317.8s > 217.0s $p^{\cdot}$=.056 < 0.1). In terms of total dwell time on SERPs, the time for SERP-d was longer than that of the other UIs. ANOVA results revealed a significant difference between the three UIs ($F(2, 307)$=3.27, p^* < .05, η^2 = 2.09e-2). The post-hoc pairwise comparisons suggested marginal trends toward significance between SERP-d and Control and between SERP-d and Page-d (129.4s > 84.8 s; p=.081; 129.4 s > 82.1 s; p=.056, respectively). This indicates that participants using SERP-d might tend to take more time to scan a list of search results than with the other UIs.

With respect to total dwell time on disputed web pages, there was no statistical difference between the three UIs (Control = 117.2 s, SERP-d = 104.3 s, and Page-d = 111.5 s; p = .961; η^2 = 2.60e-4). On the other hand, in terms of dwell time on non-disputed web pages, there was statistical significance between the three UIs ($F(2, 307)$=13.0, p^{**} < .01, η^2 = 7.78e-2). The post-hoc pairwise comparisons indicated that participants using SERP-d viewed non-disputed web pages longer in total than participants using Control (81.6 s > 18.2 s, p^{***} = 8.89e-5) and Page-d (81.6 s > 23.4 s, p^{***} = 1.83e-5). When the findings for dwell time on SERPs were combined, the results indicate that SERP-d made participants stay in SERPs and non-disputed web pages longer than Control and Page-d.

Furthermore, we calculated the average dwell times per disputed and non-disputed web pages. In this calculation, we did not consider participants who did not visit disputed/non-disputed web pages. For the average dwell time per disputed web page, the analytical results show a significant difference between the three UIs ($F(2, 260)$ = 4.13, p^* < .05, η^2 = 3.08e-2). Moreover, as shown in Figure 7, the post-hoc pairwise tests indicate that SERP-d made participants view each disputed web page longer than Control and Page-d at a significance level of 0.05 (41.4 s > 25.7 s, p^* = .022; 41.4 s > 27.0 s, p^* = .038). Figure 6 shows that total dwell time for the disputed web pages were shorter than that of non-disputed web pages. However, the results on Figure 7 indicate that once participants using SERP-d visited disputed web pages, they viewed the pages more carefully than participants using Control and Page-d. Regarding non-disputed web pages, we found no significant difference between the three UIs ($F(2, 170)$ = 1.83, p = .164, η^2 = 2.10e-2), although the mean average dwell time for SERP-d appeared greater than that of Control and Page-d (SERP-d = 35.1s, Control = 20.8s, Page-d = 25.8s).

4.3 Search task answer

To analyze how SERP-d and Page-d affected the participants' final decisions, we evaluated the 310 answers reported as effective health methods. We compared the reported answers with the disputed topics the system highlighted during the tasks by manually classifying them into two groups: *disputed answers* and *non-disputed answers*. During the classification process, we removed five answers that were too ambiguous to judge.

Figure 8 shows the ratio of participant answers classified as *disputed answer* and *non-disputed answer*. According to Figure 8, more participants using Page-d and SERP-d reported non-disputed answers than those using Control (69.5%, 69.4% > 56.9%). A Pearson's χ^2 test revealed marginal trends toward significance between Page-d and Control (χ^2 = 3.57, p = 5.88e-2 < 0.1, Cramer's V = 1.31e-1) and between SERP-d and Control (χ^2 = 3.36, p = 6.67e-2 < 0.1, Cramer's V = 1.30e-1). These results indicate that SERP-d and Page-d were likely to influence the participants' final decisions in such a way to avoid suspicious topics.

4.4 Usability of disputed topic suggestion

To investigate the participants' subjective evaluations and use of

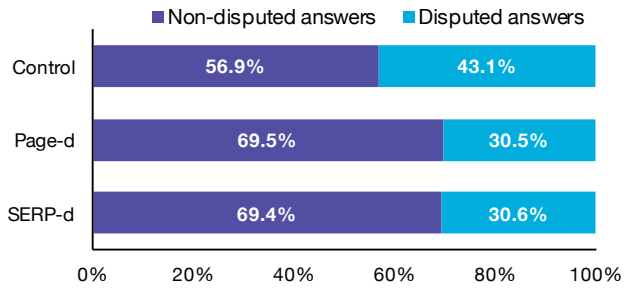


Figure 8: Ratio of participant answers contained/not contained in disputed topic list

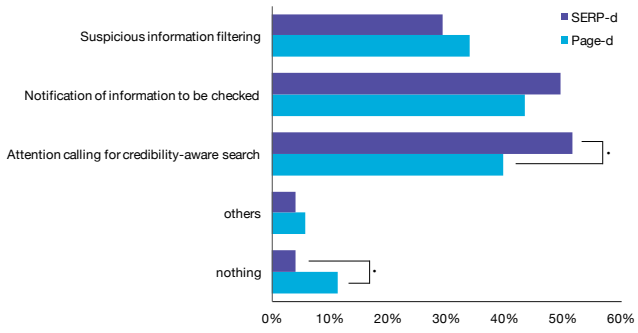


Figure 9: Ratio of participants who thought each factor was useful (marginal trends toward significance are illustrated by lines)

Table 3: Mean and SEM of usefulness rating score for SERP-d and Page-d

UI condition	SERP-d	Page-d
Mean usefulness (SEM)	4.02 (9.30e-2)	3.74 (1.01e-1)

Table 4: Features of SERP-d UI and Page-d UI

Effect	SERP-d	Page-d
Prevention of suspicious topic	x	
Reinforcing critical info. seeking	x	
Seeking complementary information	x	x

Page-d and SERP-d, we analyzed the answers to the post questionnaire. Table 3 shows how useful SERP-d and Page-d were to search for effective health methods. The usefulness of SERP-d and Page-d were 4.02 and 3.74, respectively, and there was statistical significance between the two UIs ($F(1, 203) = 4.23$, $p^* < .05$, $\eta^2 = 2.04\text{e-}2$). This indicates that both UIs were considered useful on average. Another finding is that the participants thought that SERP-d was more useful for searching for effective health methods than Page-d.

Figure 9 shows which factors of SERP-d and Page-d participants considered useful for search tasks. The two most useful factors of SERP-d and Page-d were attention calling for credibility-aware search (SERP-d = 51.5%; Page-d = 39.6%) and notification of information to be checked (SERP-d = 49.5%; Page-d = 43.4%). With respect to the attention calling aspect, there was a marginal trend toward significance between participants using the two UIs ($\chi^2 = 2.92$, $p < 8.74\text{e-}2 < 0.1$, Cramer's $V = 1.19\text{e-}1$). Fewer participants indicated that suspicious information filtering was a useful factor than the previous two aspects for both of the UIs (SERP-d = 29.3%; Page-d = 34.0%). 11.3% of participants using Page-d and 4.04% of participants using SERP-d indicated that the UI was not useful. Regarding this "nothing useful" factor, there was a marginal trend toward significance between participants using the two UIs ($\chi^2 = 3.77$, $p < 5.22\text{e-}2 < 0.1$, Cramer's $V = 1.36\text{e-}1$). Note that Page-d was considered less useful than SERP-d.

5. DISCUSSION

Through our study of 92 participants' behaviors and subjective evaluations, we obtained the following findings for disputed topic suggestion in web search.

1. Participants using SERP-d often spent more time performing tasks than when using Control and Page-d, especially when viewing SERPs and non-disputed web pages.

2. When scanning a search results list with SERP-d, participants clicked more search results without disputed topics than those with disputed topics.

3. Once participants visited web pages whose disputed topics were highlighted by SERP-d, they often spent more time viewing them than with the other UIs.

4. With respect to search behaviors (i.e., dwell time and click-through ratio), there was no difference between Page-d and Control.

5. SERP-d and Page-d made fewer participants select disputed topics as their task answers than Control.

6. Page-d and SERP-d obtained good evaluations from participants on average.

From these findings, we examine the potential of the two dispute suggestion styles from three aspects: (1) preventing users from browsing suspicious topics, (2) reinforcing careful information seeking, and (3) seeking complementary information for decision making. These are summarized in Table 4.

Disputed topic suggestion in SERPs notifies searchers which search results contain suspicious information before they click the results. It can be presumed from some of our findings that, rather than simply filtering out suspicious information, the SERP-d suggestion style has a significant impact on searcher consideration of information credibility on the web. As shown in Figure 4, SERP-d prevented participants from clicking disputed web search results. In addition, SERP-d appeared to make the participants more careful in web search. This can be interpreted from three findings: (a) participants using SERP-d spent more time performing tasks, (b) they selected web search results more carefully for a longer time (Figure 6), and (c) SERP-d encouraged them to read disputed web pages more carefully once visiting them as well as non-disputed pages (Figure 7). These results indicate that SERP-d influenced the entire search process so that the participants examined the credibility of web pages, regardless of whether the pages contained disputed topics on their SERPs. Therefore, we conclude that disputed topic suggestion in SERPs promotes careful information seeking.

On the other hand, it can be presumed that disputed topic suggestion on visited web pages could be used as complementary information for decision making rather than strongly restraining checking suspicious information and promoting careful information seeking. If the Page-d suggestion style prevented checking suspicious information, participants using Page-d would have spent less time dwelling on disputed web pages than with Control because the participants would have immediately left the pages after seeing disputed topic suggestions. However, the average dwell time on disputed web pages for Page-d was as long as that for Control (Figure 7). Furthermore, if Page-d promoted careful information seeking as SERP-d does, the participants with Page-d would have spent more time performing tasks and browsing disputed Web pages than with Control (Figures 6 and 7). On the other hand, Figure 8 shows that more participants using Page-d reported non-disputed answers than those using Control. Furthermore, the post questionnaire revealed that some participants said Page-d was useful for filtering out suspicious information (Figure 9). These results indicate that the participants could use suggested disputed topics on Page-d to reduce uncertainty on their final answers for the tasks. Consequently, we consider that the Page-d suggestion works as complementary information to support a judgment on whether browsed information is credible.

Our study has revealed that, even when suggesting the same disputed topics, the timing of suggestion might influence participant search behaviors. Both SERP-d and Page-d can prevent searchers from making the decision which might contain suspicious information. However, SERP-d can motivate searchers to carefully and slowly scrutinize web pages using threat appeal. Page-d is not powerful enough to provoke mental movement towards more careful information seeking. In summary, the results of our study indicate that disputed topic suggestion before visiting web pages can encourage searchers toward careful information seeking more easily than after.

Our study provides insights on search interaction design to enhance user consideration of information credibility in web search. However, there are issues to address in future. The first is the effect of individual variability on disputed topic suggestion. Our study indicates that SERP-d made participants more careful in web search and browsing on average than Page-d and Control. However, we found that SERP-d did not influence some participants' behaviors. One possible reason is related to the participants' prior beliefs or preferences about search topics. For example, White et al. reported that, if searchers have confident belief in topics initially, they are more likely to maintain their belief after a search [25]. Moreover, according to Liao et al., even if people confirm that opposing opinions exist, they will often preferentially select information that supports their prior beliefs [12]. Considering these studies, it is possible that disputed topic suggestion in SERPs does not always influence searchers' search attitude and behavior because they might have strong prior belief or preference. To develop methods that promote careful information seeking, we must study the relationship between prior searcher belief and the effects of disputed topic suggestion.

The second issue is related to the visibility of disputed topics suggestion in SERPs. In our study, we used artificial SERP settings, where web search results with disputed topics appeared more frequently in higher ranked positions; thus, disputed topics attracted more participant attention. However, in reality, commercial search engines order search results using proprietary algorithms. Therefore, even if disputed topics exist in a list of search results, searchers do not always see them. To enhance searcher engagement in careful search, we must consider other fore-alarming methods easy to implement on search engines like [26], in addition to our disputed topic highlighting method.

6. CONCLUSION AND FUTURE WORK

In this paper, we have studied how disputed topic suggestion influences search behavior and decision making to obtain credible information from the web when *scanning a search results list* and *browsing web pages*. We conducted an online study with 92 participants to evaluate the two styles of disputed topic suggestion. The results show that, even when suggesting the same disputed topic, the timing of suggestion influences participant behaviors. We found that disputed topic suggestion in SERPs encourages people to consider information credibility, prevent them from consuming suspicious information, and make them spend more time searching the web. On the other hand, the results indicate that disputed topic suggestion in web pages does not change people's search behaviors; however, it can help find complementary information to support decision making. Thus, we suggest that focusing on fore-alarming before accessing information is important in the design of support systems to enhance searcher willingness to engage in careful information seeking.

However, several issues require future consideration. The first is the effect of individual variability on disputed topic suggestion. Second, we must examine methods to gather disputed topics from the web and examine methods to implement disputed topic suggestion as fore-alarming into search engines under practical constraints. Furthermore, providing complementary information to resolve disputes is another important issue.

To obtain correct web information and support effective decision making, it is important to strengthen people's credibility judgment process and the development of machine-based credibility judgment systems. We believe that this work will contribute to the promotion of careful information seeking on the web.

7. ACKNOWLEDGMENTS

This work was supported in part by Grants-in-Aid for Scientific Research (#15H01718) from MEXT of Japan.

8. REFERENCES

[1] B. T. Adler and L. de Alfaro. A Content-Driven Reputation System for the Wikipedia. In *Proceedings of the 16th International Conference on World Wide Web (WWW 2007)*, pages 261–270, 2007.

[2] S. Akamine, D. Kawahara, Y. Kato, T. Nakagawa, K. Inui, S. Kurohashi, and Y. Kidawara. Wisdom: A web information credibility analysis system. In *Proceedings of the ACL-IJCNLP Software Demonstrations (ACLDemos 2009)*, pages 1–4, 2009.

[3] C. Castillo, M. Mendoza, and B. Poblete. Information Credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web (WWW 2011)*, pages 675–684, 2011.

[4] C. L. A. Clarke, E. Agichtein, S. Dumais, and R. W. White. The Influence of Caption Features on Clickthrough Patterns in Web Search. In *Proceedings of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2007, pages 135–142. ACM, 2007.

[5] N. Diakopoulos and I. Essa. Modulating Video Credibility via Visualization of Quality Evaluations. In *Proceedings of the 4th Workshop on Information Credibility (WICOW 2010)*, pages 75–82. ACM, 2010.

[6] X. L. Dong, E. Gabrilovich, K. Murphy, V. Dang, W. Horn, C. Lugaresi, S. Sun, and W. Zhang. Knowledge-based Trust: Estimating the Trustworthiness of Web Sources. *Proceedings of the VLDB Endowment*, 8(9):938–949, 2015.

[7] R. Ennals, B. Trushkowsky, and J. M. Agosta. Highlighting Disputed Claims on the Web. In *Proceedings of the 19th International Conference on World Wide Web (WWW 2010)*, pages 341–350, 2010.

[8] A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating Information from Disagreeing Views. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM 2010)*, pages 131–140. ACM, 2010.

[9] S. Ieong, N. Mishra, E. Sadikov, and L. Zhang. Domain Bias in Web Search. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM 2012)*, pages 413–422. ACM, 2012.

[10] D. Kahneman. A perspective on judgment and choice: mapping bounded rationality. *American psychologist*, 58(9):697, 2003.

[11] C. W. Leong and S. Cucerzan. Supporting Factual Statements with Evidence from the Web. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM 2012)*, pages 1153–1162. ACM, 2012.

[12] Q. V. Liao and W.-T. Fu. Beyond the Filter Bubble: Interactive Effects of Perceived Threat and Topic Involvement on Selective Exposure to Information. In *Proceedings of the 31th ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2013)*, pages 2359–2368. ACM, 2013.

[13] G. Lindgaard, C. Dudek, D. Sen, L. Sumegi, and P. Noonan. An Exploration of Relations Between Visual Appeal, Trustworthiness and Perceived Usability of Homepages. *ACM Transactions on Computer-Human Interaction*, 18(1):1–30, 2011.

[14] J. Maddock, K. Starbird, H. J. Al-Hassani, D. E. Sandoval, M. Orand, and R. M. Mason. Characterizing Online Rumoring Behavior Using Multi-Dimensional Signatures. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW 2015)*, pages 228–241. ACM, 2015.

[15] M. Metzger, A. Flanagin, K. Eyal, D. Lemus, and R. McCann. Credibility for the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment. *Communication yearbook*, 27:293–336, 2003.

[16] M. R. Morris, J. Teevan, and K. Panovich. 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)*, pages 1739–1748. ACM, 2010.

[17] S. Nakamura, S. Konishi, A. Jatowt, H. Ohshima, H. Kondo, T. Tezuka, S. Oyama, and K. Tanaka. Trustworthiness Analysis of Web Search Results. In *Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2007)*, pages 38–49. Springer, 2007.

[18] J. Pasternack and D. Roth. Latent Credibility Analysis. In *Proceedings of the 22nd International Conference on World Wide Web (WWW 2013)*, pages 1009–1020, 2013.

[19] P. Pirolli, E. Wollny, and B. Suh. So You Know You're Getting the Best Possible Information: A Tool that Increases Wikipedia Credibility. In *Proceedings of the 27th ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2009)*, pages 1505–1508. ACM, 2009.

[20] R. W. Rogers. *Social psyhophysiology*, chapter Cognitive and physiological processes in fear appeals and attitude change: A revised theory of protection motivation, pages 153–176. Guilford Press, 1983.

[21] J. Schwarz and M. Morris. Augmenting Web Pages and Search Results to Support Credibility Assessment. In *Proceedings of the 29th ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2011)*, pages 1245–1254. ACM, 2011.

[22] E. Sillence, P. Briggs, L. Fishwick, and P. Harris. Trust and Mistrust of Online Health Sites. In *Proceedings of the 22th ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2004)*, pages 663–670. ACM, 2004.

[23] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum*, 33(1):6–12, 1999.

[24] B. Suh, E. H. Chi, A. Kittur, and B. A. Pendleton. Lifting the Veil: Improving Accountability and Social Transparency in Wikipedia with WikiDashBoard. In *Proceeding of the 26th ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2008)*, pages 1037–1040. ACM, 2008.

[25] R. W. White. Beliefs and biases in web search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2013)*, pages 3–12. ACM, 2013.

[26] Y. Yamamoto. Disputed Sentence Suggestion Towards Credibility-Oriented Web Search. In *Proceedings of the 14th Asia-Pacific international conference on Web Technologies and Applications (APWeb 2012)*, pages 34–45. Springer, 2012.

[27] Y. Yamamoto and K. Tanaka. Enhancing Credibility Judgment of Web Search Results. In *Proceedings of the 29th ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2011)*, pages 1235–1244. ACM, 2011.