Highlighting Weasel Sentences for Promoting Critical Information Seeking on the Web

Fumiaki Saito[1], Yoshiyuki Shoji[2], and Yusuke Yamamoto[1]

[1] Shizuoka University, Johoku 3-5-1, Naka-ku, Hamamatsu, Shizuoka, Japan
{saito,yamamoto}@design.inf.shizuoka.ac.jp
[2] Aoyama Gakuin University, Fuchinobe 5-10-1, Chuo-ku, Sagamihara-shi, Kanagawa, Japan
shoji@it.aoyama.ac.jp

Abstract. This paper proposes a system that highlights weasel sentences while browsing webpages. The term weasel sentence is defined in the context of this paper as a quotation with an unknown or unidentifiable source. Following this definition, the system automatically detects weasel sentences in browsed webpages. Then, we investigate how highlighting weasel sentences affects the search behaviors and decision making of the users searching for information on the web. An online user study yielded the following results: (1) Highlighting the weasel sentences encouraged participants to invest more time in web browsing and to view a larger number of webpages. (2) The effect of (1) was more significant when participants were familiar with the search topics. (3) Web browsing elicited less change in the confidence of the search answers when participants were familiar with the given topics. The findings provide insights into how users can avoid gathering misleading on the web.

Keywords: Web browsing · Information credibility · Critical information seeking · Human factor · User interface.

1 Introduction

The credibility of online digital content is becoming a social problem. With the evolution of digital library technologies, people can create digital contents and casually search for them on the web. Currently, many people often rely on digital contents such as webpages for decision making on their future actions. Therefore, as stated by the ACRL Information Literacy Competency Standards [1], the credibility of online digital contents must be carefully checked. However, people frequently accept online digital contents as credible and also trust the technologies that search for them. For example, Nakamura and their colleagues found that people often believe that web search engines rank webpages by their credibility scores [18]. Pan et al. conducted a similar survey and reported that a lot of people choose a higher ranked webpage based on a greater trust in web search engine algorithms [19].

In general, people often think of the credibility of information as an objective quality like authenticity or accuracy. However, as researchers in social psychology indicate, information credibility is a subjective quality, and its interpretation depends on information receivers [6]. To prevent gathering of inaccurate information, users must be

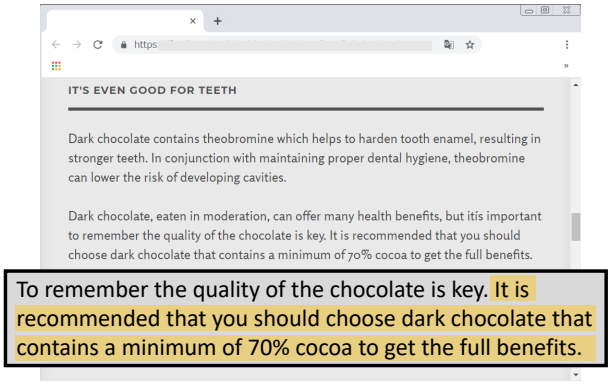

Fig. 1. Overview of our prototype system.

made aware of the existence of misinformation or misleading information among web contents and appreciate the value of critical information seeking.

In this paper, we propose a system that highlights weasel sentences in browsed webpages as a low-credibility signal that prompts critical information seeking (see Fig.1). Information credibility can be assessed by various measures such as authority, date of publication, coverage, and documentation[3]. In this paper, We focus on *the anonymous authority of claims*, highlighting the sentences which seem specific or meaningful but have unknown or unidentifiable sources. For example, in Wikipedia, weasel phrases such as "some people say" and "researchers believe" should be rewritten because they appear to be authorized without substantial evidence, and may therefore bias the readers[4]. In the Japanese Wikipedia, 6430 articles are manually tagged with the `{{Weasel}} and {{By whom}}` warnings and with rewrite recommendations[5]. In reality, a larger number of documents with weasel phrases reside on the web. Furthermore, web information is often viewed without sufficiently considering the information source [17], although researchers in information literacy call attention to information source for obtaining credible information [1]. This situation carries a severe risk of wrong decision making based on low-quality information in trusted web information. To tackle this problem, we build a detector of sentences with **weasel sentences** based on Wikipedia articles tagged with `{{Weasel}} and {{By whom}}` warnings and implement it on a web browser extension designed for promoting critical information seeking.

We also study whether the proposed system guides users' browsing/search behaviors toward critical information seeking on the web. In the fields of information retrieval and human-computer interaction, various search support systems have been proposed in several studies, such as evidence search systems [14], dispute suggestion systems [5], and systems to visualize credibility-related scores [20]. In studies on such systems, researchers often have evaluated their proposed systems from the aspects of user

[3] Berkeley Library's Evaluating resources: http://guides.lib.berkeley.edu/evaluating-resources
[4] Unsupported attributions: https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Words_to_watch
[5] The number is at the time of April 19th, 2018.

satisfaction or subjective usefulness. However, how the systems promote critical web information seeking at the behavioral level and the attitude level has been rarely explored. We discuss how the proposed system can affect the dwell time, the number of viewed pages, and confidence in decision making through an online user study.

The main contributions of this paper are as follows:

- We propose a method that detects weasel sentences in browsed webpages based on weasel annotation data in Wikipedia.
- Through an online user study, we show that highlighting weasel sentences increases both the time expended in reading webpages, and the number of webpages viewed.
- The highlighting effect depends on the search-topic familiarity of the users.

2 Related Works

2.1 Evaluation of quality and credibility of web information

Various methods that evaluate the quality and credibility of web information have been proposed. Dong et al. developed a method to evaluate web page information credibility, assuming that credible webpages have few false facts or claims [4]. Hasan et al. proposed a machine learning approach that evaluates the quality of Wikipedia articles from their readabilities and editing histories [7]. The model of Wang et al. classifies web information as true or false, using fact-checking websites on political topics [21]. Differently from these related works, we target the sentences in webpages, not the webpages themselves. Moreover, we adopt the anonymous authority of claims as a measure of sentence credibility.

2.2 Attitude for critical information seeking on the web

Even if people are aware of suspicious information, they often fail to judge its credibility due to the use of incorrect heuristics, i.e., cognitive bias [10]. Ieong et al. revealed the existence of domain bias whereby search users believe that relevant webpages are authorized by specific domains [9]. White et al. studied the relationship between beliefs about search topics and search behaviors [22]. 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 Promoting critical information seeking

Several methods that promote critical information seeking have also been proposed in the fields of human-computer interaction and information retrieval. Search Dashboard, proposed by Bateman et al., provides a UI that reflects search behaviors and summarizes search histories [2]. Liao et al. revealed that suggesting the opinion stance and expertise of the information sender can mitigate the so-called echo chamber effect, in which recommender systems present users only with contents that reflect their own views[15]. Yamamoto et al. proposed a "query priming" system that implicitly activates critical information seeking in web searches [24]. By highlighting the insufficiently accurate sentences, we aim to encourage users to consider the credibility of the information found in their document browsing.

3 Highlighting weasel sentences

This section describes our approach for automatically judging the anonymous authority of claims in documents. It then proposes a prototype system that highlights the weasel sentences during web browsing.

3.1 Classifier

Weasel sentence detection is treated as a binary classification problem on textual documents. Our proposed system vectorizes a given sentence and detects its anonymous authority status using a trained model.

Below we describe the (1) training data, (2) classification features, and (3) classification performance.

Training data Our weasel sentence detection focuses on the `{By whom}` tag in Wikipedia. Wikipedia volunteers tag the sentences in Wikipedia articles with various warnings that encourage other volunteers to improve the article quality. The `{By whom}` tag is one of such warning tags. The training dataset was derived from Japanese Wikipedia articles. After dividing the entire set of Wikipedia articles into sentences, we extracted the sentences labeled with the `{By whom}` tag as *weasel sentences (positive examples)*. Sentences not annotated with the `{By whom}` tag were then randomly extracted from the articles containing sentences with *weasel* tags. These sentences were accumulated as *non-weasel sentences (negative examples)*. We finally collected 2,236 positive examples and 2,236 negative examples for building an weasel sentence classifier.

Features For weasel sentence classification, our proposed system vectorizes the Wikipedia sentences as two feature types:

Bag-of-Words (BOW): Binary vectors representing the presence (1) or absence (0) of specific nouns, adjectives, and verbs in the sentences. In this study, the BOW vectors were created from terms appearing in more than two sentences. The target terms were extracted from the sentences by a Japanese morphological analyzer, `MeCab`[6].

Existence of weasel expressions: Binary vectors representing the presence (1) or absence (0) of specific weasel expressions in the sentences. Examples of weasel expressions are “some people think" and “researchers believe." To extract this feature, we prepared 27 weasel expressions listed in a Japanese Wikipedia article on weasel expressions[7] . Some of these expressions are listed in Table 1 .

Performance evaluation The usefulness of the above features in weasel sentence detection was experimentally evaluated on the dataset described in subsection 3.1. For this evaluation, we trained a support vector machine (SVM) classifier with a radial basis function kernel.

[6] MeCab: [http://taku910.github.io/mecab/]

[7] Weasel expression on Wikipedia (Japanese version): https://bit.ly/33lJ1hC

Table 1. Examples of weasel expressions used for weasel sentence detection.

Weasel expression
People are saying..., It has been claimed that..., Some people believeâĂę,
It is known that..., It has been mentioned that..., Researchers claimâĂę,
Critics claim..., I heard that..., There is criticism toâĂę

The best parameters C and γ of the SVM classifier, and the classification performance were estimated by 5-fold cross-validation. The precision, recall, F1-value, and accuracy were evaluated as 0.772, 0.748, 0.760, and 0.764, respectively.

3.2 Prototype system

Our prototype system was developed as a web browser extension for the user study. Once the extension is installed, any weasel sentences in the browsed webpages are highlighted by the system. Figure 1 is a screenshot of the system overview. The system flow is as follows:

1. When a user visits a webpage, the client system (browser) sends the webpage URL to our application programming interface (API) server.
2. The API server extracts the texts from the webpage and divides them into sentences. The end of each sentence is detected by the ".", "?", and "!"symbols.
3. The server classifies the sentences into weasel and non-weasel sentences using the trained model.
4. The server sends the weasel sentences to the client.
5. The client system highlights the weasel sentences in the browsed webpage. The highlights are presented as bold font against a yellow background.

3.3 Hypotheses

If the proposed system promotes critical information seeking during web browsing, then users should change their search/browsing behaviors to improve the integrity of their final decisions. This study poses the following hypotheses:

H1 The proposed system extends the time expended in web information seeking.
H2 The proposed system increases the number of visited webpages.
H3 The proposed system improves the users' confidence in their decisions made through web information seeking.
H4 The above-mentioned effects vary with users' familiarity with the search topics.

4 User study

This section describes the experimental design and the evaluation procedure of the proposed system.

Table 2. Search-task questions and topic familiarity. Numbers in parentheses represent the standard deviations among 188 participants.

Question	Mean of Familiarity (SD)
Does cinnamon help improve diabetes?	-2.53 (0.93)
Does vitamin C help prevent pneumonia?	-2.28 (1.00)
Can cocoa decrease blood pressure?	-2.05 (1.24)
Does garlic help improve and prevent a common cold?	-1.51 (1.33)

4.1 Participants

We recruited 250 participants through a Japanese crowdsourcing service, CrowdWorks[8]. After excluding the data of 67 participants who did not complete all tasks or spent an exceptionally long time on the tasks, the data from 188 participants were eligible for the analysis. Each participant received approximately $2 for their time.

4.2 Tasks

Each participant performed search tasks on four contentious medical questions (see Table 2). Medical issues were chosen because they are crucial to our life and require careful decision making. We re-used the four questions used in the user study of [24]. The participants were asked to search for the answers to each question using our experimental system, and provide a *yes* or *no* response to each question.

Before starting each task, we asked the participants how familiar they were with the task question. The answers were provided on a six-point Likert scale (from -3: completely unfamiliar to +3: completely familiar). Table 2 lists the mean topic familiarities of the participants with the four tasks. On average, the mean familiarity was under -1.50. This result confirms that most participants lacked sufficient knowledge to answer the task questions before participating in the user study.

4.3 Design and procedure

The effects of two factors (*UI condition* and *topic familiarity*) were examined by a between-subjects design. The UI condition (see subsection 4.4 for details) consisted of a *proposed* level, which highlighted the weasel sentences while viewing the web-pages, and a *controlled* level, which did not highlight the weasel sentences. The topic familiarity factor consisted of six levels as described in subsection 4.2. The participants were randomly allocated to one of the UI conditions. Consequently, 105 participants were allocated into the *proposed* group and 83 participants were categorized into the *controlled* group.

After signing a consent form on the crowdsourcing website, the participants progressed to our experimental website for the user study. As discussed above, the user study comprised four search tasks (see Table 2). The search task order was randomized

[8] CrowdWorks: https://crowdworks.jp/

for each participant to control the task ordering effect. Each search task consisted of three phases.

In the first phase, we asked the participants to indicate their familiarity with the target topics. The participants provided their responses on a six-point Likert scale as described in subsection 4.2. The participants then guessed their answers to the task questions and assessed their confidence in their answers (again on a six-point Likert scale from -3: completely unconfident to +3: completely confident). We defined this confidence assessment as the *pre-condidence* level.

The second phase of the user study was the *search phase*. In this phase, the participants researched their answers to each task using the experimental search system. Each search task was introduced by a brief instruction; for example,

> *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.*

After reading the description, the participants began searching by clicking the "Start search" button. Upon clicking the button, the participants were directed to our pre-prepared search engine result pages (SERPs) (see next subsection for details). The participants were asked to visit the webpages on each SERP, and to report their answers when satisfied with the result. The search process was not time-limited.

After the search phase, the participants reported their confidence levels in their answers (on a six-point Likert scale from -3: completely unconfident to +3: completely confident). We defined this confidence assessment as the *post-condidence* level.

4.4 Experimental system

When preparing the SERPs for the task topics, we imitated commonly used search engines such as Google and Bing. Each SERP displayed 30 fixed search results on the target topic, each comprising a title, a snippet, and a URL. Note that our SERPs did not provide a search box for modifying the initial query. The search results of the SERPs were displayed through a Bing web search API[9]. The search results on the tasks were pre-selected by inserting queries of the form "`{medical symptom name}` AND `{possible treatment}`" into the Bing API. Finally, we inserted the top-30 results for each task (e.g. "`blood pressure` AND `cocoa`") into the associated SERP.

When a participant clicked on a SERP title, the system opened a webpage in a different browser tab. All webpages listed on the SERPs were cached before disseminating the user study. The participant behaviors in the user study were monitored by a Javascript code embedded in each webpage. Furthermore, under the *proposed* UI condition, the weasel sentences in the webpages were highlighted in yellow.

The weasel sentences to be highlighted by the system were manually selected in advance. The manual selection was necessary because our classifier for weasel sentence detection was imperfect. By avoiding the incorrect classifications, we could purely examine the effects of highlighting the weasel sentences on the participants' behavior and decision making.

[9] https://azure.microsoft.com/services/cognitive-services/bing-web-search-api/

5 Analysis

The user study provided the behavioral data and pre/post-confidence levels in 752 sessions returned by 188 participants. This section describes the statistical analysis of the data.

5.1 Statistical Approach

The collected data were analyzed by the Bayesian approach. Although the Bayesian approach is less familiar than the frequentist approach, it was selected because it better handles the data uncertainty, yielding the probability distributions of the target parameters [11]. The Bayesian models were constructed using generalized linear mixed models (GLMMs) [13], extensions of linear models that model the target responses even when they follow a nonlinear distribution. GLMMs also distinguish the *fixed effects* caused by the experimental conditions from the *random effects* caused by variations among the random samples, such as participants and tasks. The Bayesian GLMMs have become popular tools for modeling various user behaviors in information-retrieval research and human-computer interactions, where they are replacing traditional ANOVA analysis [8, 11, 13].

In this study, we conducted the Bayesian GLMMs to study the effects of weasel sentence highlight, tasks, and participants, with probability interval. We developed Bayesian GLMMs of five response variables: *session time*, *dwell time in the SERPs*, *average dwell time per webpage*, *number of viewed pages*, and *confidence change* (described in Section 5.2) [10]. The fixed effects were *UI condition* and *topic familiarity*. The interactions between *UI condition* and *topic familiarity* also considered. The random effects were introduced by the subjects and tasks.

The *session time*, *dwell time in SERPs* and *average dwell time per webpage*, were assumed to follow Weibull distributions. Previously, Liu et al. reported that the dwell time in webpages follows this distribution [16]. Meanwhile, the *number of viewed pages* was assumed as Poisson-distributed, as appropriate for count data. As the link function for modeling the above-fixed variables on the GLMMs, we selected the log function. After observing the histogram, the *confidence change* was assumed to follow a normal distribution. As the link function for modeling the *confidence change*, we selected the identity function. The fixed and random effects were described by non-informative prior distributions.

We validated our hypotheses through two approaches. The first approach adopted the *high density interval (HDI)*, which summarizes the posterior distribution of a parameter such that every point inside the interval has higher credibility than any point outside the interval [12]. When zero lies outside the 95% (sometimes 90%) HDI of the coefficients of the target variable, Bayesian statistics interprets that the variable exerts a significant effect on the outcome. Meanwhile, frequentists conduct a significance test of the null hypothesis at the specified significance level (typically $\alpha = 0.05$, sometimes $\alpha = 0.1$) [12]. The other approach directly examines the posterior probabilities that our hypotheses are correct, given the data and their non-informative prior distributions.

[10] For Bayesian GLMMs, we used the `R` package `brms`[3].

Table 3. Coefficients for the generalized linear mixed model for session time, with mean, standard error, 90% HDI, and 95% HDI.

Variable	Mean	SE	90% HDI	95% HDI
Intercept	4.38	0.19	[4.02, 4.75]	[3.87, 4.87]
UI	0.29	0.16	[0.02, 0.56]	[-0.04, 0.61]
Familiarity	-0.14	0.04	[-0.19, -0.08]	[-0.20, -0.06]
UI * Familiarity	0.08	0.05	[0.003, 0.16]	[-0.01, 0.17]

5.2 Response Variables

To investigate hypotheses **H1**, **H2**, **H3**, and **H4** in subsection 3.3, we constructed GLMMs of the following five response variables.

Session time This variable measures the time expended by a participant during searching and browsing the webpages assigned to the task. The session time may lengthen when the searching and browsing tasks are more carefully performed. The session time of a participant completing a task was obtained by summing his/her dwell time in the associated SERP and his/her total dwell time in the webpages during the task. This measure was investigated in hypothesis **H1** and **H4**.

Dwell time in SERPs This variable measures the time expended by a participant in the SERP of a given task. This time may lengthen if the useful webpages are more carefully selected from the list of search results. This measure was evaluated in hypothesis **H1** and **H4**.

Average dwell time per webpage This variable measures the time expended (on average) by a participant on each webpage of the assigned task. This time may lengthen when any webpage is carefully browsed for quality-judgment information or evidence collection. This measure was investigated on hypothesis **H1** and **H4**.

Number of viewed webpages This variable measures the number of webpages viewed by a participant during a task. The webpage count will increase when the participant visits more webpages to collect and compare multiple pieces of evidence. This measure was evaluated in hypothesis **H2** and **H4**.

Confidence change This variable measures the extent to which the participants' confidence in their task answers (beliefs) changes after the web search. The post-confidence will increase if participants consider that the evidence strengthened or weakened their prior belief. This measure was evaluated in hypothesis **H3** and **H4**.

6 Results

6.1 Session Time

Table 3 shows the results of the Bayesian GLMM analysis on session time. The 90% HDI of the *UI condition*variable excluded zero (equivalent to $p < 0.10$ in the frequentist

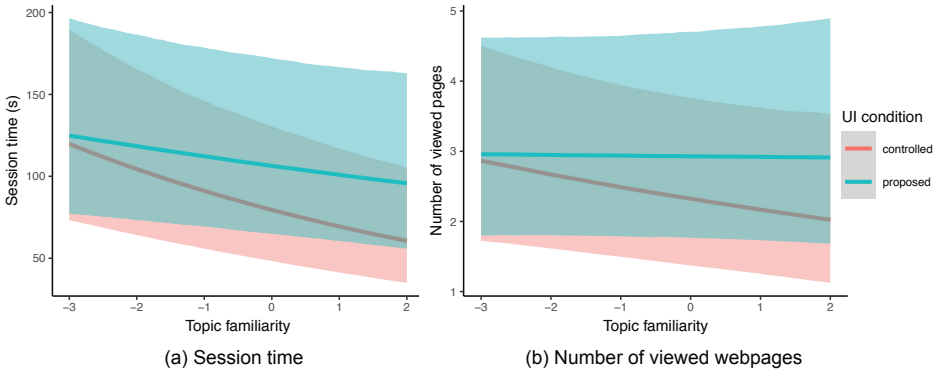

Fig. 2. Marginal effects of UI conditions at specific levels of topic familiarity for (a) session time and (b) number of viewed webpages. Green and orange lines indicate the predicted response values under *proposed* and *controlled* conditions, respectively. Colored backgrounds delineate the 95% credible intervals (CIs) of the predicted values.

Table 4. Coefficients of the generalized linear mixed model for dwell time on the SERPs, with mean, standard error, 90% HDI, and 95% HDI.

Variable	**Mean**	**SE**	**90% HDI**	**95% HDI**
Intercept	3.10	0.17	[2.78, 3.42]	[2.67, 3.53]
UI	0.19	0.16	[-0.09, 0.45]	[-0.12, 0.51]
Familiarity	-0.11	0.04	[-0.17, -0.04]	[-0.18, -0.03]
UI*Familiarity	0.06	0.05	[-0.02, 0.15]	[-0.04, 0.17]

approach). Furthermore, under the *proposed* condition, the session time was extended with a very high posterior probability (a 96.0% chance that the coefficient exceeded zero over the posterior distribution).

The 90% HDI of the interaction between *UI condition* and *topic familiarity* was also greater than zero. We confirmed a 96.1% chance that the coefficient exceeded zero over the posterior distribution. Figure 2a illustrates the marginal effects of the *UI condition* and topic familiarity. This figure indicates that (1) the participant session times were longer in the *proposed* condition than in the *controlled* condition, and (2) when participants were more familiar with a search topic, the *UI condition* exerted a larger effect than when participants were unfamiliar with the topic.

6.2 Dwell time on SERP and webpage

As indicated in Table 4, the 90% HDIs of the *UI condition* and the interaction between *UI condition* and *topic familiarity* contained zero (equivalently, the null hypothesis cannot be rejected at the $\alpha = 0.1$ significance level). In addition, there was insufficient evidence (insufficiently many high chances) that the coefficients of *UI condition* and the

Table 5. Coefficients of the generalized linear mixed model for average dwell time per webpage, with mean, standard error, 90% HDI, and 95% HDI.

Variable	Mean	SE	90% HDI	95% HDI
Intercept	3.65	0.16	[3.36, 3.95]	[3.26, 4.03]
UI	-0.09	0.16	[-0.36, 0.18]	[-0.42, 0.23]
Familiarity	-0.04	0.04	[-0.11, 0.03]	[-0.12, 0.05]
UI*Familiarity	-0.02	0.05	[-0.11, 0.08]	[-0.13, 0.09]

Table 6. Coefficients of the generalized linear mixed model for the number of viewed webpages, with mean, standard error, 90% HDI, and 95% HDI.

Variable	Mean	SE	90% HDI	95% HDI
Intercept	0.84	0.18	[0.48, 1.19]	[0.36, 1.35]
UI	0.23	0.14	[-0.01, 0.46]	[-0.06, 0.51]
Familiarity	-0.07	0.04	[-0.14, -0.004]	[-0.15, 0.01]
UI * Familiarity	0.07	0.05	[-0.01, 0.15]	[-0.03, 0.17]

interaction between *UI condition* and *topic familiarity* were positive (87.3% for the *UI condition*; 86.9% for the interaction).

As indicated in Table 5, the 90% HDIs of the *UI condition* and the interaction between *UI condition* and *topic familiarity* contained zero (in the frequentist approach, the null hypothesis could not be rejected at the $\alpha = 0.1$ significance level). Besides, there was insufficient evidence that the coefficients of *UI condition* and the interaction were positive (28.7% for the *UI condition*; 38.7% for the interaction).

6.3 Number of viewed webpages

Table 6 shows the GLMM results on the number of visited webpages. The 90% HDIs of *UI condition* and the interaction between *UI condition* and *topic familiarity* contained zero (in the frequentist approach, the null hypothesis could not be rejected at the $\alpha = 0.1$ significance level). However, the *proposed* condition encouraged the participants to visit more webpages with high probability (a 94.5% chance that the coefficient exceeded zero over the posterior distribution). Furthermore, there was a marginally high chance that the coefficient of the interaction between *UI condition* and *topic familiarity* exceeded zero over the posterior distribution (90.4%).

Figure 2b illustrates the marginal effects of *UI condition* and *topic familiarity* for the number of viewed webpages per task. This figure indicates that (1) the participants under the *proposed* condition viewed more webpages than those under the *controlled* condition, and (2) for participants familiar with the search topic, the *UI condition* exerted a larger effect than for participants unfamiliar with the topic.

Table 7. Coefficients of the generalized linear mixed model for confidence change, with mean, standard error, 90% HDI, and 95% HDI.

Variable	Mean	SE	90% HDI	95% HDI
Intercept	1.27	0.23	[0.85, 1.66]	[0.78, 1.78]
UI	-0.12	0.27	[-0.56, 0.32]	[-0.64, 0.40]
Familiarity	-0.37	0.07	[-0.49, -0.25]	[-0.51, -0.23]
UI * Familiarity	-0.17	0.09	[-0.32, -0.02]	[-0.35, 0.01]

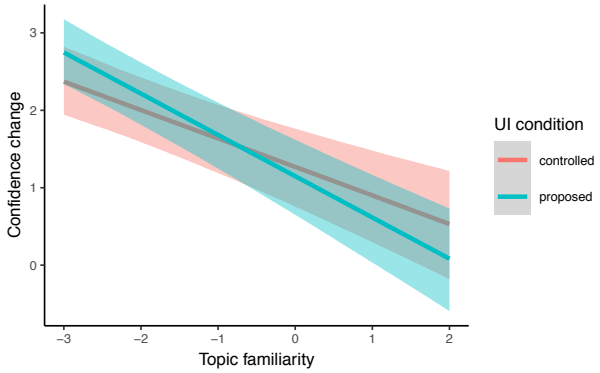

Fig. 3. Marginal effects of UI conditions at specific levels of topic familiarity for evaluating the confidence change. Green and orange lines indicate the predicted response values under*proposed* and *controlled* conditions, respectively. Colored backgrounds delineate the 95% CIs of the predicted values.

6.4 Confidence change

Table 7 shows the GLMM results of the confidence changes after the tasks. As indicated in the table, the 90% HDI of the *UI condition* variable included zero. Moreover, the probability of the coefficient exceeding zero was low over the posterior distribution (32.8%). However, as previously observed for the interaction between *UI condition* and *topic familiarity*, the 90% HDI of the interaction variable was less than zero (equivalent to $p < 0.1$ in the frequentist approach). Furthermore, we confirmed a 96.7% chance that the coefficient of the interaction was below zero over the posterior distribution.

Figure 3 illustrates the marginal effects of *UI condition* and *topic familiarity*. This figure indicates that when the participants were not familiar with the search topics, the confidence change was larger under the *proposed* condition than under the *controlled* condition. On the contrary, when the participants were familiar with the search topics, the confidence change was lower under the *proposed* condition than under the *controlled* condition.

7 Discussion

7.1 Detection of weasel sentences

As described in subsection 3.1, the precision, recall, F-value, and accuracy of our classifier all exceeded 0.7, indicating that our classifier outperformed a simple classifier that randomly judges whether or not sentences are weasel ones.

Rather than discussing the development of a high-performance classifier, we investigated the effect of the classifier on user behaviors during web browsing. Therefore, we simply employed the SVM classifier as a baseline method. However, the performance of weasel-sentence detection could be improved. For example, our classifier neglected the order and semantics of the terms in the sentences. In the future, we have a plan to employ a deep neural network-based method for higher performance. A further study of how the performance of weasel sentence detection affect user behavior should be conducted.

7.2 Effect of weasel sentences during web browsing

Here, we discuss hypotheses **H1**, **H2**, **H3**, and **H4** in terms of the study results.

For **H1**, we examined the session time, dwell time in SERPs, and average dwell time per webpage during each search task. According to the analytical results, the *UI condition* affected the session time (subsection 6.1). However, the behavioral analysis revealed no influence of *UI condition* on the dwell time in SERPs or on the average dwell time per webpage (subsection 6.2). From these results, we cannot conclude whether the participants spent longer in the SERPs or the webpages under the proposed condition. However, we can conclude that highlighting the weasel sentences promoted longer web browsing in a particular session level. Therefore, we consider that **H1** was supported.

For **H2**, we examined the number of viewed webpages during each search task. Highlighting the weasel sentences influenced the number of viewed webpages (subsection 6.3), meaning that participants using the proposed system visited more webpages than those using the controlled system. From this result, we consider that highlighting the weasel sentences promotes browsing for comparison with other webpages or additional verification, thus supporting **H2**.

For **H3**, we examined the confidence difference in the participants' answers before and after the search tasks. The effect of highlighting the weasel sentences depended on the participants' familiarity with the researched topic. According to Fig.3, if the participants were unfamiliar with the search topic, their confidence change was more influenced by web browsing under the *proposed* condition than under the *controlled* condition. On the other hand, if the participants were familiar with the search topic, their confidence change was lower under the *proposed* condition than under the *controlled* condition. From these results, we consider that **H3** was partially supported; specifically, it was supported when the participants were unfamiliar with the search topics.

In the **H4** evaluation, participants who were more familiar with the search topics were more influenced by weasel-sentence highlighting than those with less knowledge of the topic (Fig. 2a and 2b). Furthermore, as discussed for Fig.3, highlighting the weasel sentences exerted less effect on the confidence change when the participants were familiar with the topic. These results support hypothesis **H4**.

From the above discussion, we conclude that highlighting the weasel sentences encouraged the participants to view more webpages, and consequently extend their time in web browsing, when they were somewhat familiar with the search topics. Before the user study, we expected that when the participants saw the highlighted weasel sentences in a webpage, they would read the text more carefully over a longer time than when the weasel sentences was not highlighted. However, this expectation was not supported by our results: the average dwell time per webpage was no longer under the proposed condition than under the controlled condition. We surmised that the participants judged the highlighted text as poor-quality, and consequently abandoned it. The actual interpretations and usages of the highlighted sentences should be investigated through participant interview or eye-tracking analysis.

Interestingly, we found that when the participants reported more knowledge of the search topic, there was less difference between their pre- and post-confidence levels than those of participants reporting low knowledge of the topic, despite using our proposed system. One interpretation is that when the participants with high topic familiarity looked at highlighted weasel sentences, the sentences might strength their prior beliefs in their task answers. This interpretation indicates that if people use our system with incorrect prior beliefs, they may be prompted into wrong decision making. As discussed in [25, 23], several studies have claimed that user characteristics such as topic familiarity and critical thinking capability can affect critical information seeking on the web. Therefore, we should conduct a deeper analysis of the relationship between user characteristics and the effect of the proposed method. Moreover, we should study a better method to support unbiased and critical information seeking on the web.

8 Conclusion

We proposed a system that highlights the weasel sentences on webpages during web browsing. Furthermore, we studied how highlighting such sentences affected the search behaviors and decision making of people searching for web information.

The findings of the online user study are summarized as follows: (1) Weasel-sentence highlighting encouraged the participants to extend their web browsing time and view more documents. (2) The effect of finding (1) was more significant when the participants were familiar with the search topics. (3) When participants with more topic familiarity used our system, their prior belief in their search-task answers was less influenced by web browsing than that of users unfamiliar with the topic. The proposed search-interaction design can enhance user engagement in critical information seeking on the web.

Acknowledgments

This work was supported in part by Grants-in-Aid for Scientific Research (18H03243, 18H03244, 18H03494, 18KT0097, 18K18161, 16H02906) from MEXT of Japan.

References

1. American Library Association, Association for College and Research Libraries: Information Literacy Competency Standards for Higher Education. Tech. rep. (2000)

2. Bateman, S., Teevan, J., White, R.W.: The search dashboard: How reflection and comparison impact search behavior. In: Proc. of CHI 2012
3. Bürkner, P.C.: brms: An R package for Bayesian multilevel models using Stan. Journal of Statistical Software **80**(1), 1–8 (2017)
4. Dong, X.L., Gabrilovich, E., Murphy, K., Dang, V., Horn, W., Lugaresi, C., Sun, S., Zhang, W.: Knowledge-based trust: Estimating the trustworthiness of web sources. Proceedings of the VLDB Endowment **8**(9), 938–949 (2015)
5. Ennals, R., Trushkowsky, B., Agosta, J.M.: Highlighting Disputed Claims on the Web. In: Proc. of WWW 2010
6. Fogg, B.J., Tseng, H.: The Elements of Computer Credibility. In: Proceedings of CHI 1999
7. Hasan Dalip, D., André Gonçalves, M., Cristo, M., Calado, P.: Automatic quality assessment of content created collaboratively by web communities: A case study of wikipedia. In: Proc. of JCDL 2009
8. Hofmann, K., Mitra, B., Radlinski, F., Shokouhi, M.: An eye-tracking study of user interactions with query auto completion. In: Proc. of CIKM 2014
9. Ieong, S., Mishra, N., Sadikov, E., Zhang, L.: Domain bias in web search. In: Proc. of WSDM 2012
10. Kahneman, D.: Thinking, fast and slow. Macmillan (2011)
11. Kay, M., Nelson, G.L., Hekler, E.B.: Researcher-centered design of statistics: Why bayesian statistics better fit the culture and incentives of hci. In: Proc. of CHI 2016
12. Kruschke, J.: Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan. Academic Press (2014)
13. Lee, J., Walker, E., Burleson, W., Kay, M., Buman, M., Hekler, E.B.: Self-experimentation for behavior change: Design and formative evaluation of two approaches. In: Proc. of CHI 2017
14. Leong, C.W., Cucerzan, S.: Supporting Factual Statements with Evidence from the Web. In: Proc. of CIKM 2012
15. Liao, Q.V., Fu, W.T.: Expert voices in echo chambers: effects of source expertise indicators on exposure to diverse opinions. In: Proc. of CHI 2014
16. Liu, C., White, R.W., Dumais, S.: Understanding web browsing behaviors through weibull analysis of dwell time. In: Proc. of SIGIR 2010
17. Metzger, M.J.: Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. Journal of the American Society for Information Science and Technology **58**(13), 2078–2091 (2007)
18. Nakamura, S., Konishi, S., Jatowt, A., Ohshima, H., Kondo, H., Tezuka, T., Oyama, S., Tanaka, K.: Trustworthiness analysis of web search results. In: Proc. of ECDL 2007
19. Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., Granka, L.: In Google We Trust: Users' Decisions on Rank, Position, and Relevance. Journal of Computer-Mediated Communication **12**(3), 801–823 (2007)
20. Schwarz, J., Morris, M.: Augmenting Web Pages and Search Results to Support Credibility Assessment. In: Proc. of CHI 2011
21. Wang, W.Y.: Liar, liar pants on fire: A new benchmark dataset for fake news detection. In: Proc. of ACL 2017
22. White, R.: Beliefs and biases in web search. In: Proc. of SIGIR 2013
23. Yamamoto, T., Yamamoto, Y., Fujita, S.: Exploring People's Attitudes and Behaviors Toward Careful Information Seeking in Web Search. In: Proc. of CIKM 2018
24. Yamamoto, Y., Yamamoto, T.: Query Priming for Promoting Critical Thinking in Web Search. In: Proc. of CHIIR 2018
25. Yamamoto, Y., Yamamoto, T., Ohshima, H., Kawakami, H.: Web Access Literacy Scale to Evaluate How Critically Users Can Browse and Search for Web Information. In: Proc. of WebSci 2018