Mitigating Position Bias in Review Search Results with Aspect Indicator for Loss Aversion

Mitigating Position Bias in Review Search Results with Aspect Indicator for Loss Aversion

Hiroki Ihoriya, Masaki Suzuki, and Yusuke Yamamoto

Shizuoka University, Hamamatsu, Japan
{ihoriya,suzuki,yamamoto}@design.inf.shizuoka.ac.jp

Abstract. Conventional review websites display a list of item search results with average rating scores (i.e., star ratings). We propose a method of designing snippets that encourage users to search items on review websites more carefully. The proposed snippets include aspect indicators that identify negative aspects if the item has a good star rating and vice versa. We expect the aspect indicators will help mitigate biases due to ranking position and star ratings by making users feel a "loss" if they do not carefully examine items. Our user study showed that the proposed method of including aspect indicators for loss aversion made participants spend more time searching a list of search results and checking items with worse star ratings, especially when searching hospitals. In contrast, showing aspect indicators that conformed to star ratings caused shortsighted review searches.

Keywords: information retrieval, cognitive bias, human factor, search user interaction

1 Introduction

Many people use review websites to purchase products and services, which we refer to as *items* in this paper. Similar to conventional web search engines such as Google, review websites provide users with a function to rank and list items. A typical search engine results page (SERP) lists items with their average customer ratings (i.e., *star ratings*) as well as a *snippet* including their name and description. Although star ratings are helpful for quickly understanding other customers' satisfaction with items, the ratings themselves and ranking lists based on them often cause a cognitive bias in users. This leads to shortsighted decision-making on review websites; users often choose items only because they have good star ratings or high-ranking positions in review searches.

Researchers have confirmed various cognitive biases that occur during web searches and have reported that these biases often negatively influence information seeking and outcomes [3]. One of the most famous cognitive biases is the *position bias*, in which people often preferentially click higher-ranked items on a list of search results [4]. As another example of bias, White [18] reported that users prefer positive information to negative information. As noted above, most review websites rank and list items along with star ratings on SERPs. Therefore, good

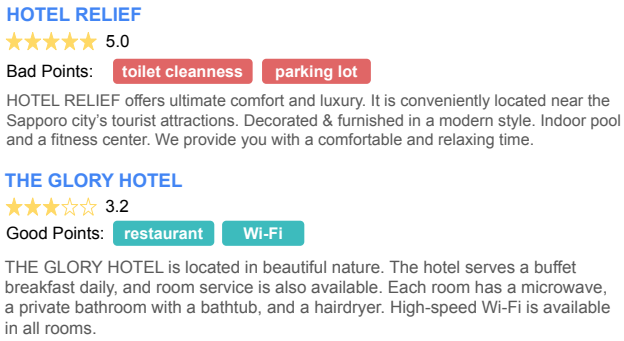

Fig. 1. Overview of snippets with aspect indicators for loss aversion in hotel review search.

star ratings can promote the effect of position bias, which can cause users to prioritize higher-ranked items and make shortsighted decisions without detailed exploration, even though such items may have flaws that are unacceptable to the user. For example, a user may be searching a hotel review website, and toilet cleanliness is a critical consideration for them. Some hotels may gain good star ratings from other customers and be ranked high, even though their toilets are not clean. However, conventional SERPs do not display such negative information. Therefore, the user may be biased by the ranking position and star rating and select a high-ranked hotel with unclean toilets.

In this paper, we propose a method of designing snippets that mitigate the biases of the ranking position and star rating to encourage users to search items on review websites more carefully. Our method is inspired by the *loss aversion rule*, which explains the tendency of people to prefer avoiding the pain of losing more to acquiring an equivalent gain. In our proposed method, the snippets display *aspect indicators* that identify negative aspects of items with good star ratings and vice versa as Figure 1 shows. Such snippets should help make users feel a "loss" if they do not carefully examine items and mitigate the position bias when searching the review website.

Our contributions can be summarized as follows:

- We designed snippets with aspect indicators for loss aversion to mitigate the position and star rating biases of review searches.
- We performed a user study to confirm the effectiveness of presenting snippets with aspect indicators relative to snippets without aspect indicators.
- We demonstrated the negative effects of aspect indicators that conform to star ratings rather than display contrasting information.

2 Related work

2.1 Cognitive bias in information retrieval

Many studies have focused on cognitive biases in information retrieval. White [18] investigated the influence of users' prior beliefs on their web search behavior and showed that users prefer to view information that supports their prior beliefs and are reluctant to view information that supports the opposite belief. Baeza-Yates [4] demonstrated a position bias in user searches where higher rankings on SERPs tend to result in higher click rates. Craswell et al.[7] developed a probabilistic model of how position bias occurs in the real world. Yue et al.[21] reported on the influences of the attractiveness of the title, URL, summary, and other aspects of items on SERPs on users' clicking behavior. Their results revealed that items with attractive titles can bias users' clicking behavior. Ieong et al.[10] demonstrated the existence of a domain bias, where users tended to trust webpages published in a specific domain. Lindgaard et al.[12] found that people are more likely to trust a webpage that looks attractive.

2.2 Changing behaviors and attitudes during web searches

Several studies have investigated interactions that change the behaviors and attitudes of web search users. Agapie et al.[1] proposed a search user interface (UI) that places a halo around a query box if a user inputs long queries. They reported that their proposed interface encourages users to input longer queries for better search results. Harvey et al.[9] demonstrated that providing examples of high-quality queries can help users formulate queries more effectively. Scott et al.[5] proposed the Search Dashboard system to visualize the search and browsing behaviors of web search users. Their experimental results showed that their system improves the search behavior of users.

Some researchers have studied methods of promoting careful information seeking. Munson et al.[15] proposed a web browser extension that indicates whether the user's browsing history is politically balanced. Hamborg et al.[8] developed NewsBird, which aggregates international news from various perspectives. Liao et al.[11] showed that indicating the opinion stance and expertise of the information sender could mitigate the echo chamber effect. Yamamoto et al.[19] proposed the `Query Priming` system, which inserts queries that evoke critical thinking during query completion/recommendation in a search system. They showed that this system helps make web search users more likely to correct their queries and visit websites that provide relevant data and evidence-based information. Yamamoto et al.[20] also proposed the `Personalization Finder` web browser extension, which reveals the effects of web search personalization and promotes careful web searching.

2.3 Product search

With the proliferation of e-commerce and review sites, product search is a branch of information retrieval that has gained increasing attention. Ai et al.[2] proposed

an explanatory search model that focuses on the gap between the search system and the customer's perception of product relevance. Empirical experiments showed that their proposed model could produce reasonable explanations for search results. To predict the overall rating on review websites more accurately, Cheng et al.[6] presented a novel aspect-aware latent factor model that estimates the importance of an aspect of an item for a user. Qiu et al.[16] used online experimental data from Airbnb users to investigate whether the average star rating or number of reviews was more important for indicating users' trustworthiness. Their results showed that the relative effectiveness of ratings and reviews varies greatly depending on the strength of the differentiating power of reputation. To measure the actual quality of a product, McGlohon et al.[14] analyzed 8 million product reviews collected from multiple review sites.

The approaches presented in these studies are helpful for users who proactively search for products and services on review websites. However, if users have low motivation to scrutinize products, they tend to make shortsighted decisions based on superficial information such as ranking positions and star ratings. Therefore, promoting a more careful approach to information seeking on review websites is necessary.

3 Proposed method: Aspect indicators for loss aversion

3.1 Overview

Our proposed method mitigates the bias of ranking position and star ratings on SERPs in review websites by extending snippets to display aspect indicators (i.e., positive/negative aspects of items). The intent is to make users feel a "loss" if they do not scrutinize items more carefully.

Figure 1 illustrates an example where the proposed method is applied to search results in a hotel review website. Aspect indicators are presented along with the snippet of each item based on an analysis of review comments about each item. For example, Hotel Relief has an excellent star rating, but the aspect indicator suggests that some users have complained about its toilet cleanliness. The proposed method determines the polarity of aspect indicators according to the degree of the star rating for an item. If an item has a better star rating than the average rating of items in the search result list, negative aspect indicators are displayed with its snippet to make the user aware of disadvantages. In contrast, when an item has a worse star rating than average, positive aspect indicators are displayed to make the user aware of advantages. For example, the Glory Hotel has a star rating of 3.2, which is lower than average. Thus, the proposed method displays a positive aspect indicator suggesting that the hotel serves an excellent dinner. In this case, we expect that users looking for gourmet food will check the hotel's information in more detail because they do not want to miss the possibility of a good dinner.

3.2 Algorithm to find aspect indicators

Once a review website returns a list of items to a query, the following procedure is adopted to display aspect indicators along with snippets of the items:

1. Analyze the sentiment expressed about entities within review comments about items matching a given query.
2. Determine characteristic entities for each item as aspect indicator candidates considering their frequency in review comments.
3. Calculate the sentiment score of the characteristic entities obtained in Step 2, and classify the entities as positive or negative aspect indicators.
4. Display some of the selected aspect indicators along with each snippet considering the star rating of the item.

In Step 1, the system extracts entities from review comments about searched items and analyzes the sentiment score for each entity. Although various methods have been proposed for entity sentiment analysis [13], we adopted Google Natural Language API[1] to review comments about items listed for a given query. The API then returns a list of entities with sentiment scores ranging from -1.0 to 1.0. For example, an entity sentiment analysis on the comment "This hotel's breakfast was good" would extract "breakfast" as an entity with a sentiment value of 0.75. In Step 2, the system determines characteristic entities for each item as aspect indicator candidates. We used term frequency–inverse document frequency (TF-IDF) weighting to calculate the feature scores of entities extracted in Step 1. Let N be the number of items in a search result list for a query, i_k be the k-th item in the list, and c_{i_k} be a set of review comments about item i_k. Then, the system calculates the TF-IDF score $tfidf(e, i_k)$ of the entity e for item i_k as follows:

$$tfidf(e, i_k) = tf(e, i_k) \times idf(e) \tag{1}$$

$$idf(e) = log\frac{N}{df(e)} \tag{2}$$

where $tf(e, i_k)$ is the frequency of e in review comments about i_k and $df(e)$ is the number of items for which e appears in review comments. For each item, we calculate the TF-IDF values of all entities appearing in its review comments, and we use the top 10 entities as aspect indicator candidates.

In Step 3, we further refine the indicator candidates. Our aim is to present negative or positive aspects of items so that users will feel loss if they do not check them. Therefore, the indicator candidates are classified based on the sentiment intensity in the review comments. The system calculates the average sentiment scores of each indicator candidate appearing in the review comments for each item. When the average sentiment score of an aspect indicator is greater or less than 0.0, the indicator is considered as positive and negative, respectively.

[1] Google Cloud Natural Language API: https://cloud.google.com/natural-language/docs/analyzing-entity-sentiment

In Step 4, the system determines whether to display positive or negative aspect indicators for each item by considering the average rating score (i.e., star rating). If an item has a better star rating than the average star rating for all items returned to a query, then negative aspect indicators are presented along with the snippet. Similarly, positive aspect indicators are presented if the star rating is worse than average.

3.3 Hypothesis

We expect our proposed method mitigates the bias induced by the ranking position and star rating by making users feel a "loss" if they do not scrutinize items more carefully on review websites. We built the following hypotheses:

H1 Aspect indicators for loss aversion encourage users to spend more time reviewing searches more carefully and to compare more items.
H2 Aspect indicators for loss aversion encourage users to check items with low rankings or bad star ratings.

We speculated that the effect of aspect indicators for loss aversion would vary depending on search topics. We expected the aspect indicators to be more effective for critical topics such as health than casual topics such as travel. Therefore, we built an additional hypothesis:

H3 The presentation of aspect indicators for loss aversion is more effective for searches of hospitals than searches of hotels.

4 User study

We performed an online user study to evaluate the effectiveness of the proposed method of aspect indicators for loss aversion in review searches. The user study was conducted in Japanese on June 22 and 23, 2021.

4.1 Participants

We recruited 560 participants via Lancers.jp, which is a Japanese crowdsourcing service.[2] Before the user study, we also explained our data collection policy, and participants proceeded only if they agreed that we could use the data collected during the search tasks. We excluded 95 participants from the analysis as outliers because they did not complete the tasks or spent an unusually long time performing the tasks.[3] Finally, we analyzed 465 participant responses. All participants who completed the tasks received 150 JPY (approximately 1.50 USD). On average, the participants finished all tasks within 9.18 min.

[2] https://lancers.jp/
[3] We used the 1.5x interquartile range rule to identify outlier participants.

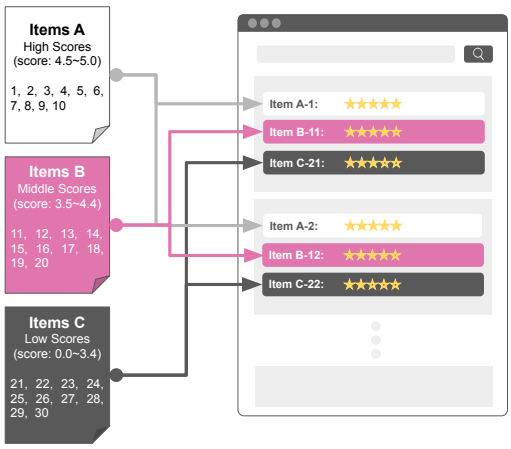

Fig. 2. Manipulation of the search result list.

4.2 Search tasks

We prepared two topics for the search tasks: hospitals and hotels. According to Hypothesis **H3**, the effectiveness of the aspect indicators varies depending on the importance of the topic. Therefore, we set hospitals as a critical topic and hotels as a non-critical topic.

For each topic, we prepared three search tasks, including a practice task. For the hospital topic, we asked the participants to search for the best hospital where they would want to go for respiratory medicine and cardiology. For the hotel topic, we asked the participants to search for the best hotel to stay in Hokkaido and Tokyo.[4] Each participant performed either the hospital or hotel search task in the user study.

4.3 Search system

We developed a search system to monitor participant behavior during the user study. The system displayed two types of pages: SERPs and detailed pages each linked to an item on the SERP. The SERPs presented a list of 30 items matching a query. Each search result comprised three components common to review search results: the title (item name), star rating (average score of customer ratings), and item description (i.e., snippet). To collect the information displayed in the search results, we crawled two popular review websites[5] before the user study. Then, we collected information on hotels and hospitals relevant to search tasks.

For the user study, we disabled the search system to prevent participants from changing the query for each search task so that a fixed list of 30 items would be

[4] Hokkaido and Tokyo are popular travel destinations in Japan.
[5] caloo.com for hospitals and jalan.net for hotels (both websites are in Japanese).

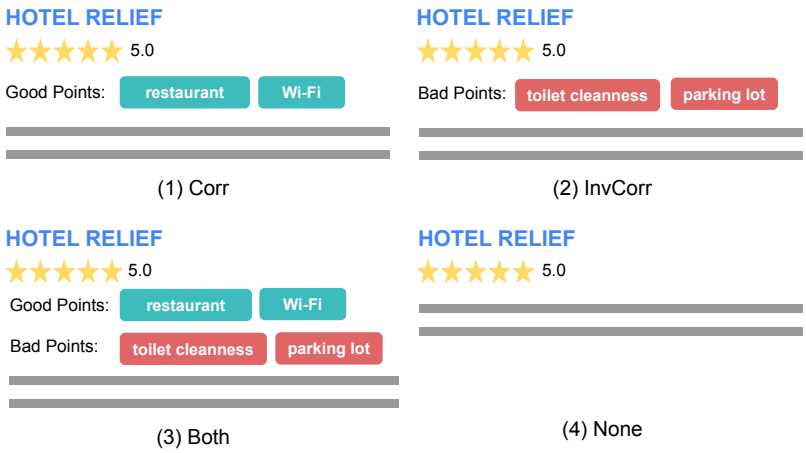

Fig. 3. Four SERP UIs for user study.

displayed. We manipulated the ranking order of the list in the following steps. First, we ordered the collected items by star rating and divided the items into three groups: top 10, middle 10, and bottom 10. Then, we created a list of the items by alternately allocating items from the three groups as described in Figure 2. One purpose of the user study was to examine whether aspect indicators can mitigate the tendency of users to ignore search results with bad star ratings and prefer high-ranked items with good star ratings. We expected that this ranking manipulation would give participants more opportunities to see items with average or bad star ratings on the list.

The participants could view detailed information on items by clicking each search result in the SERPs. The system displayed a copy of the webpages that we crawled before the study as detailed information on items. We disabled hyperlinks in the detailed pages to prevent the participants from leaving the website of the user study.

Figure 3 illustrates the four SERP UIs that we developed for the user study: `Both`, `Corr`, `InvCorr`, and `None`. The `Both`, `Corr`, and `InvCorr` UIs are extensions of the `None` UI. The `None` UI was the baseline and displayed only titles, star ratings, and item descriptions. The `Corr` UI displayed positive aspect indicators along with item descriptions for items with good star ratings (better than average) and negative aspect indicators for items with bad star ratings (worse than average). In contrast, the `InvCorr` UI displayed positive aspect indicators along with item descriptions for items with bad star ratings and negative aspect indicators for items with good star ratings. Finally, the `Both` UI displayed both positive and negative aspect indicators for each item regardless of the star rating.

Table 1. Participant allocation.

	Search UI			
Search topic	`Corr`	`InvCorr`	`Both`	`None`
Hotel	61	59	58	60
Hospital	56	61	53	57

4.4 Design and procedure

We adopted a 2 × 4 factorial design to examine the effects of two factors: the search topic and search UI condition. The search topic factor had two levels: hospital and hotel. The search UI condition had four levels: `Both`, `Corr`, `InvCorr`, and `None`.

After the participants agreed to a consent form on Lancers.jp, they moved to our website for the user study. Then, we randomly allocated a search UI condition and search topic to each participant. Table 1 presents the allocation.

First, the participants read a description of the task flow and search system. We also explained our data collection policy. The participants proceeded with the user study only if they agreed that we could use the data collected during the search task. Next, we asked the participants to perform a practice task to familiarize them with the search system. Then, each participant performed two main search tasks with either the hospital or hotel topic. The task order was randomized for each participant. Before each main search task, each participant was presented with the task scenario. An example task scenario for the hospital topic is presented below:

> *Please assume the following case. One day, you are told that your mother has lung cancer. To help her, you visit a review website to find a hospital offering good treatment regarding respiratory medicine. Check the displayed hospital search results and check several reviews about hospitals. When you come to a satisfactory decision, stop the search, and report your decision.*

Then, each participant clicked a "start search" button to display the list of search results. The participants browsed the list to formulate their answers. Once the participants reached their decision, they reported their best choice on our study website. Then, we asked the participants to explain the reason for their decision.

5 Results

We analyzed 465 participant responses to examine the effect of aspect indicators for loss aversion. We employed a non-parametric one-way analysis of variance (ANOVA) (Kruskal-Wallis test) for the UI factor separate from the search topics

Table 2. Statistics on the search behavior for the hotel search task. The mean and standard deviation (SD) of each response are shown (significance level at ***: 0.001, **: 0.01, and *: 0.05).

Metric	UI condition				p-value
	`Corr`	`InvCorr`	`Both`	`None`	
Session time (s)	252.9 (156.0)	264.1 (150.0)	289.7 (195.9)	251.2 (163.4)	0.71
SERP dwell Time (s)	55.8 (56.6)	68.8 (59.1)	80.8 (74.1)	63.3 (94.2)	*
Average page dwell time (s)	83.6 (66.1)	94.5 (66.6)	104.8 (67.0)	83.6 (59.0)	0.37
Page view count	4.03 (3.13)	3.91 (4.04)	3.52(2.75)	3.75(3.15)	0.80
Minimum click depth	2.85 (3.34)	5.91 (5.89)	5.41 (5.91)	4.66 (4.32)	***
Average score of clicked items	4.31 (0.18)	4.06 (0.33)	4.16 (0.21)	4.20 (0.21)	***
Score of decided item	4.35 (0.17)	4.11 (0.35)	4.17 (0.24)	4.23 (0.25)	***

Table 3. Statistics on search behavior for the hospital search task. The mean and standard deviation (SD) of each response are shown (significance level at ***: 0.001, **: 0.01, and *: 0.05).

Metric	UI condition				p-value
	`Corr`	`InvCorr`	`Both`	`None`	
Session time (s)	292.6 (159.2)	276.8 (159.2)	344.6 (196.3)	238.9 (133.3)	*
SERP dwell time (s)	67.5 (52.8)	74.5 (83.5)	91.5 (80.6)	46.4 (44.8)	***
Average page dwell time (s)	82.2 (77.2)	86.4 (59.7)	93.7 (65.2)	77.9 (50.5)	0.37
Page view count	4.56 (3.74)	4.15 (3.65)	4.73 (3.47)	4.48 (3.56)	0.53
Minimum click depth	2.79 (3.67)	3.67 (4.31)	3.93 (4.57)	2.34 (2.84)	0.11
Average score of clicked items	4.06 (0.15)	3.92 (0.32)	3.95 (0.23)	4.08 (0.17)	***
Score of decided item	4.04 (0.23)	3.93 (0.32)	3.9 (0.35)	4.04 (0.26)	*

because the collected data did not follow a normal distribution. We used the Benjamini-Hochaberg FDR test [17] for multiple comparison tests in the post hoc analysis. Effects were considered significant at the significance level $\alpha = 0.05$.

5.1 Session time

To investigate the effort that participants put into the tasks, we first analyzed how long participants spent on search tasks (i.e., session time). As indicated in Table 2, we did not observe significant differences between the four UIs for the hotel search tasks ($p = 0.71$). On the other hand, Table 3 indicates statistically significant differences between the four UIs for the hospital search tasks ($p < 0.05$). The post hoc analysis indicated that the mean session times of participants using the `Both` UI was statistically longer than the times of participants using the `None` UI for the hospital search tasks (mean: 344.6 vs 238.9; $p < 0.05$).

In summary, these results demonstrate that showing both positive and negative aspects along with each search result encouraged participants to spend more time on the hospital search tasks than the plain SERPs.

5.2 Dwell times on SERPs and detailed pages

Next, to investigate how much time participants spent examining a list of item search results, we analyzed the dwell times on SERPs. As indicated in Tables

2 and 3, we observed significant differences between the four UIs for both the hotel ($p < 0.05$) and hospital ($p < 0.001$) search tasks. The post hoc analysis showed that the UIs did not have a significant effect on the hotel search tasks. In contrast, the mean SERP dwell times of the `Corr` UI (67.5 s), `InvCorr` UI (74.5 s), and `Both` UI (91.5 s) were statistically longer than that of the `None` UI (46.4 s) for the hospital search tasks (`Corr-None`: $p < 0.05$; `InvCorr-None`: $p < 0.05$; `Both-None`: $p < 0.001$).

We also analyzed the average dwell times on detail pages to investigate how carefully participants read them. We did not observe significant differences between the four UIs for both the hotel ($p = 0.37$) and hospital ($p = 0.37$) search tasks.

In summary, these results indicate that SERPs with any aspect indicators made participants spend more time viewing the search results than SERPs without aspect indicators for hospital search tasks. However, the indicators did not affect the time spent viewing detailed pages.

5.3 Page views

To investigate whether participants checked multiple information sources, we analyzed how many detailed pages participants viewed (i.e., clicked) during the search tasks (i.e., page views). As indicated in Tables 2 and 3, we observed no significant differences between the four UIs in the hotel ($p = 0.80$) and hospital ($p = 0.53$) search tasks.

5.4 Minimum click depth

To investigate whether the participants had a position bias, we examined the ranks of the search results that the participants clicked on to analyze the minimum search result rank (i.e., minimum click depth). We interpreted a greater minimum click depth to mean that the participant checked search results lower on the list.

As presented in Tables 2 and 3, the ANOVA results indicated significant differences between the four UIs in the hotel search tasks ($p < 0.001$). The post hoc analysis revealed that the `Corr` UI had a smaller mean minimum click depth (2.85) than the `InvCorr` UI (5.91), `Both` UI (5.41), and `None` UI (4.66) (`Corr-InvCorr`: $p < 0.001$; `Corr-Both`: $p < 0.01$; `Corr-None`: $p < 0.01$). However, we observed no significant differences between the four UIs in hospital search tasks ($p = 0.11$).

These results indicate that, when the displayed aspect indicators conformed to the star ratings of the items (i.e., the `Corr` UI), the participants tended to click shallower search results than for the other UIs in hotel search tasks.

5.5 Rating scores of clicked items

We investigated how aspect indicators affected the star ratings of clicked items (i.e., average score of clicked items). Tables 2 and 3 indicate significant differences

between the four UIs in the hotel ($p < 0.001$) and hospital ($p < 0.01$) search tasks.

The post hoc analysis of the hotel search tasks revealed that the `Corr` UI had a statistically greater average score of clicked items (4.31) than the `InvCorr` UI (4.06), `Both` UI (4.16), and `None` UI (4.20) (`Corr-InvCorr`: $p < 0.001$; `Corr-Both`: $p < 0.001$; `Corr-None`: $p < 0.01$). Meanwhile, the post hoc analysis of the hospital search tasks revealed that the `InvCorr` UI and `Both` UI had statistically smaller average scores of clicked items (3.92 and 3.95, respectively) than the `None` UI (4.08) (`InvCorr-None`: $p < 0.05$; `Both-None`: $p < 0.01$). Furthermore, the `Corr` UI had a statistically larger average score of clicked items (4.06) than the `Both` UI (3.95) ($p < 0.05$).

These results suggest that, when the participants performed hotel search tasks by using the `Corr` UI, they tended to click items with better star ratings than with other UIs. In contrast, when the aspect indicators contrasted with the star ratings of items for hospital search tasks, the participants tended to view items with worse star ratings than when no aspects were displayed on the SERPs.

We also examined the star ratings of items that the participants selected for each task (i.e., decided item score). Tables 2 and 3 indicate significant differences between the four UIs for the hotel ($p < 0.001$) and hospital ($p < 0.05$) search tasks. The post hoc analysis of the hotel search tasks revealed that the `Corr` UI had a statistically larger decided item score (4.35) than the `InvCorr` UI (4.11), `Both` UI (4.17), and `None` UI (4.23) (`Corr-InvCorr`: $p < 0.001$; `Corr-Both`: $p < 0.001$; `Corr-None`: $p < 0.01$). Meanwhile, the post hoc analysis of the hospital search tasks revealed that the `Both` UI had a statistically smaller decided item score (3.90) than the `None` UI (4.04) ($p < 0.05$).

These results show that the `Corr` UI made participants select items with better star ratings as their final answer than other UIs for the hotel search tasks. Furthermore, displaying aspect indicators that contrasted with the star ratings (i.e., the `InvCorr` and `Both` UIs) encouraged participants to select items with worse star ratings for the hospital search tasks.

6 Discussion

The results of the user study revealed that the participants tended to spend more time examining the list of item search results with the `InvCorr` UI and `Both` UI than with the `None` UI for hospital search tasks (Section 5.2). In addition, the `InvCorr` UI and `Both` UI did not encourage participants to check more items than the `None` UI for both the hotel and hospital search tasks (in Section 5.3). The click-through analysis revealed that the `InvCorr` UI and `Both` UI promoted viewing detailed pages about items with worse star ratings than `None` UI for hospital search tasks (in Section 5.5). In particular, the participants using the `Both` UI in hospital search tasks tended to select items with worse star ratings as their final decision than those using the `None` UI.

The `InvCorr` UI and `Both` UI displayed aspect indicators that contrasted with the star ratings of items to make users feel loss aversion. The results of the user study led to the following interpretations of their effectiveness:

1. For hospital search tasks, participants using the `InvCorr` UI and `Both` UI felt "the loss of not viewing hospital pages with low scores and positive aspects" and "the loss of viewing hospital pages with high scores and negative aspects."
2. Participants spent a longer time viewing the hospital list to think carefully about which hospital to check.
3. They considered hospitals with bad star ratings as candidates if presented with positive aspects.

On the other hand, although participants tended to spend more time on SERPs in the hospital search tasks when using the `Corr` UI rather than the `None` UI, the participants also clicked items with better star ratings when using the `Corr` UI rather than the `Both` UI.

The `Corr` UI displayed aspect indicators conforming to the star ratings of items (i.e., positive aspects for items with above-average star ratings and negative aspects for items with below-average star ratings). The results of the user study led to the following interpretations:

1. The `Corr` UI provided information that reinforced the star ratings of items.
2. Consequently, the participants using the `Corr` UI spent less time viewing the item lists and jumped to items with higher-ranking positions.

Unlike the `InvCorr` UI and `Corr` UI, the `Both` UI displayed both positive and negative aspects for items regardless of the star rating. We think that the `Both` UI provided the participants with useful information for decision-making while mitigating the biases caused by the ranking position and star rating. Thus, we concluded that the `Both` UI could promote more careful information seeking than the `InvCorr` UI.

The analytical results suggest that the `InvCorr` UI and `Both` UI were more effective for the hospital search tasks than for the hotel search tasks. One possible reason for this result may be the degree of risk of the decision. Obviously, hospital selection would have a more significant impact than hotel selection. Thus, the `InvCorr` UI and `Both` UI would make the participants feel a greater loss from shortsighted information seeking for hospitals than for hotels.

In summary, displaying aspect indicators that contrast with star ratings could mitigate biases caused by position ranking and star ratings. Consequently, users would spend more time viewing a list of hospital search results and checking hospitals with lower ranking or worse star ratings, although the aspect indicators did not seem to encourage them to compare detailed pages more often. Thus, we concluded that **H1**, **H2**, and **H3** were partially supported.

6.1 Limitations

One limitation of our study concerns the statistical analysis. We adopted a 2×4 factorial design to examine two factors: search topics and search UIs.

We employed a non-parametric one-way ANOVA for search topics because we assumed that hospital search tasks would be considered more crucial than hotel search tasks. However, we needed to statistically analyze the difference between the search topics for a more rigorous discussion. In future studies, we plan to apply statistical models such as generalized linear mixture modeling to analyze the effect of search topics.

Another limitation was our approach to finding aspect indicators for loss aversion. We used TF-IDF weighting to determine aspect indicator candidates for items (i.e., hotels and hospitals). In addition, we used the Google Natural Language API to determine whether the extracted aspect candidates are positive or negative. We think that this approach can be improved to find better aspect indicators.

7 Conclusion

We proposed a method of mitigating the biases of the ranking position and star rating to encourage users to search items on review websites more carefully. Our proposed method extends snippets to display negative aspects of items with good star ratings and vice versa. Thus, the snippets provide information that users could miss if they only looked at star ratings.

The results of an online user study indicated that aspect indicators contrasting with star ratings could encourage users to spend more time viewing a list of hospital search results and to check hospitals with lower rankings or worse star ratings. On the other hand, we found that aspect indicators conforming to star ratings could cause shortsighted information seeking, especially for hotel search tasks. These results imply that snippets should display both positive and negative aspects of items to mitigate biases and let users know additional features of items.

This study had several issues that should be improved upon. We need to examine for which search topics the proposed method would be effective more rigorously. Furthermore, we should improve the method of extracting aspect indicators for loss aversion.

Acknowledgements This work was supported in part by Grants-in-Aid for Scientific Research (18H03244, 21H03554, 21H03775) from MEXT of Japan.

References

1. Agapie, E., Golovchinsky, G., Qvarfordt, P.: Leading people to longer queries. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 3019–3022. CHI 2013, ACM, New York, NY, USA (2013)
2. Ai, Q., Zhang, Y., Bi, K., Croft, W.B.: Explainable product search with a dynamic relation embedding model. ACM Transactions on Information Systems (TOIS) **38**(1), 1–29 (2019)

3. Azzopardi, L.: Cognitive biases in search: A review and reflection of cognitive biases in information retrieval. In: Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (CHIIR 2021). p. 27–37 (2021)
4. Baeza-Yates, R.: Bias on the web. Communications of the ACM **61**(6), 54–61 (2018)
5. Bateman, S., Teevan, J., White, R.W.: The Search Dashboard: How Reflection and Comparison Impact Search Behavior, p. 1785–1794. Association for Computing Machinery, New York, NY, USA (2012), https://doi.org/10.1145/2207676.2208311
6. Cheng, Z., Chang, X., Zhu, L., Kanjirathinkal, R.C., Kankanhalli, M.: Mmalfm: Explainable recommendation by leveraging reviews and images. ACM Transactions on Information Systems (TOIS) **37**(2), 1–28 (2019)
7. Craswell, N., Zoeter, O., Taylor, M., Ramsey, B.: An experimental comparison of click position-bias models. In: Proceedings of the 2008 international conference on web search and data mining. pp. 87–94 (2008)
8. Hamborg, F., Meuschke, N., Gipp, B.: Matrix-based news aggregation: Exploring different news perspectives. In: Proc. of the 17th ACM/IEEE Joint Conference on Digital Libraries. pp. 69–78. JCDL '17 (2017)
9. Harvey, M., Hauff, C., Elsweiler, D.: Learning by example: Training users with high-quality query suggestions. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 133–142. SIGIR 2015, ACM (2015)
10. Ieong, S., Mishra, N., Sadikov, E., Zhang, L.: Domain bias in web search. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining. p. 413–422. WSDM '12, Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2124295.2124345, https://doi.org/10.1145/2124295.2124345
11. Liao, Q., Fu, W.: Expert voices in echo chambers: Effects of source expertise indicators on exposure to diverse opinions. In: Proc. of the 32nd SIGCHI Conference on Human Factors in Computing Systems. pp. 2745–2754. CHI '14, ACM (2014)
12. Lindgaard, G., Dudek, C., Sen, D., Sumegi, L., Noonan, P.: An exploration of relations between visual appeal, trustworthiness and perceived usability of homepages. ACM Trans. Comput.-Hum. Interact. **18**(1) (May 2011). https://doi.org/10.1145/1959022.1959023, https://doi.org/10.1145/1959022.1959023
13. Liu, Q., Zhang, H., Zeng, Y., Huang, Z., Wu, Z.: Content attention model for aspect based sentiment analysis. In: Proceedings of the 2018 World Wide Web Conference (WWW 2018). p. 1023–1032 (2018)
14. McGlohon, M., Glance, N., Reiter, Z.: Star quality: Aggregating reviews to rank products and merchants. In: Fourth international AAAI conference on weblogs and social media (2010)
15. Munson, S., Lee, S., Resnick, P.: Encouraging reading of diverse political viewpoints with a browser widget. In: Proc. of the Seventh International AAAI Conference on Weblogs and Social Media. pp. 419–428. ICWSM '13 (2013)
16. Qiu, W., Parigi, P., Abrahao, B.: More stars or more reviews? In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 1–11 (2018)
17. Thissen, D., Steinberg, L., Kuang, D.: Quick and easy implementation of the benjamini-hochberg procedure for controlling the false positive rate in multiple comparisons. Journal of educational and behavioral statistics **27**(1), 77–83 (2002)
18. White, R.: Beliefs and biases in web search. In: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. pp. 3–12 (2013)

19. Yamamoto, Y., Yamamoto, T.: Query priming for promoting critical thinking in web search. In: Proceedings of the 2018 Conference on Human Information Interaction & Retrieval. p. 12–21. CHIIR '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3176349.3176377, https://doi.org/10.1145/3176349.3176377
20. Yamamoto, Y., Yamamoto, T.: Personalization finder: A search interface for identifying and self-controlling web search personalization. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020. p. 37–46. JCDL '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3383583.3398519, https://doi.org/10.1145/3383583.3398519
21. Yue, Y., Patel, R., Roehrig, H.: Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. In: Proceedings of the 19th international conference on World wide web. pp. 1011–1018 (2010)