Asking Clarifying Questions: To benefit or to disturb users in Web search?

Jie Zou\textsuperscript{a,b}, Aixin Sun\textsuperscript{a,b,*}, Cheng Long\textsuperscript{a,b,*}, Mohammad Aliannejadi\textsuperscript{c}, Evangelos Kanoulas\textsuperscript{c}

\textsuperscript{a} Singtel Cognitive and Artificial Intelligence Lab for Enterprises@NTU, 50 Nanyang Ave, 639798, Singapore
\textsuperscript{b} School of Computer Science and Engineering, Nanyang Technological University, 50 Nanyang Ave, 639798, Singapore
\textsuperscript{c} Informatics Institute, University of Amsterdam, Science Park 904, Amsterdam 1098 XH, The Netherlands

A R T I C L E I N F O

Keywords:
User study
Clarifying Questions
Information-seeking systems

A B S T R A C T

Modern information-seeking systems are becoming more interactive, mainly through asking Clarifying Questions (CQs) to refine users’ information needs. System-generated CQs may be of different qualities. However, the impact of asking multiple CQs of different qualities in a search session remains underexplored. Given the multi-turn nature of conversational information-seeking sessions, it is critical to understand and measure the impact of CQs of different qualities, when they are posed in various orders. In this paper, we conduct a user study on CQ quality trajectories, i.e., asking CQs of different qualities in chronological order. We aim to investigate to what extent the trajectory of CQs of different qualities affects user search behavior and satisfaction, on both query-level and session-level. Our user study is conducted with 89 participants as search engine users. Participants are asked to complete a set of Web search tasks. We find that the trajectory of CQs does affect the way users interact with Search Engine Result Pages (SERPs), e.g., a preceding high-quality CQ prompts the depth users to interact with SERPs, while a preceding low-quality CQ prevents such interaction. Our study also demonstrates that asking follow-up high-quality CQs improves the low search performance and user satisfaction caused by earlier low-quality CQs. In addition, only showing high-quality CQs while hiding other CQs receives better gains with less effort. That is, always showing all CQs may be risky and low-quality CQs do disturb users. Based on observations from our user study, we further propose a transformer-based model to predict which CQs to ask, to avoid disturbing users. In short, our study provides insights into the effects of trajectory of asking CQs, and our results will be helpful in designing more effective and enjoyable search clarification systems.

1. Introduction

Information-seeking systems are expected to deliver information to users efficiently and effectively in order to satisfy users’ information needs. Benefiting from the recent developments in dialogue systems, the conversational information-seeking system is increasingly becoming a vital factor in information retrieval (IR). Such systems are now enabled to ask their users Clarifying Questions (CQs), in order to enhance their ability to understand the users’ underlying information needs and retrieve the right information for them.
Asking CQs in information-seeking systems has been formalized and studied in different ways in the recent IR literature, including CQ generation (Zamani, Dumais et al., 2020; Zamani, Lueck et al., 2020), next CQ selection (Aliannejadi et al., 2019; Hashemi et al., 2020), and CQ engagement level prediction (Sekulic et al., 2021). Although these studies have demonstrated the significance and applicability of CQs, their main focus was on the development of effective algorithms. That is, studying the extent to which CQs affect users' performance and satisfaction in a search session is still underexplored, while equally important. Along this line, Zamani, Mitra et al. (2020) present a user study, demonstrating that CQs have both functional and emotional benefits. However, when generating CQs by a machine learning model in an information-seeking system, the model inevitably generates CQs of different qualities (Zamani, Dumais et al., 2020). For instance, for a user who issues a query “Welcome to the Jungle” (which can be a film name or a song name) looking for the corresponding film, the system may generate a high-quality CQ like “Do you refer to the movie or the song?” to the user. The system may also ask a low-quality CQ such as “What song information are you looking for?”. Accordingly, Wang and Ai (2021) point out that posing low-quality CQs, instead of directly showing search results, brings in the risk of user dissatisfaction. They model the risk of CQs using reinforcement learning. However, the authors do not provide a systematic study on how CQs of different qualities could impact user behavior and satisfaction. Zou et al. (2022a), on the other hand, conduct a large-scale user study, demonstrating the effect of asking off-topic CQs on the users’ interactions with the search engine and compare that with asking good or at least relevant CQs. Although they show that asking a good CQ leads to improved user performance and satisfaction, they study CQs in isolation of each other, and therefore it is unclear how a series of CQs affects each other and the user in a search session. In a search session, CQs are not independent, each of which contribute to the overall structure of the conversation. Therefore, CQs may affect each other, while the order of CQs and their quality may affect the user behavior and satisfaction in different ways.

For example, what is the impact of displaying a low-quality CQ after a high-quality CQ and vice versa, on user satisfaction and behavior interacting with the Search Engine Result Pages (SERPs)? Because of the interactive nature of conversations, the system can infer the relevance of a CQ by the user’s response (Zou et al., 2022a). Therefore, knowing the relative impact of high-quality CQs after low-quality CQs (or vice versa) can lead to significantly improved choices in conversational information-seeking systems. In other words, while Zou et al. (2022a) demonstrate the effectiveness of asking CQs in Web search, they do not provide insights into how system designers can utilize user feedback for a better search experience. For instance, what if the system asks a CQ and receives negative feedback from the user? Should the system avoid asking, or take the risk of asking another CQ? In a different scenario, if the system asks a CQ and receives positive feedback, should that affect the system's decision on asking another CQ or not?

To fill this research gap, in this work we explore the effect of various CQ trajectories in terms of CQ quality in information-seeking systems. Although the effect of question order has been studied in some domains such as social sciences (DeMoranville & Bienstock, 2003; Schuman & Presser, 1996), the effect of CQ trajectories of CQ quality has not been studied for information-seeking systems to the best of our knowledge. In more details, we aim to explore to what extent the trajectory of CQs of different qualities affects user search behavior and satisfaction. We first investigate the query-level impact of showing different-quality CQs on user behavior to understand the immediate impact caused by the adjacent query. In addition to the query-level impact, we also explore the session-level impact of different quality trajectories of CQs on user behavior and satisfaction, to understand the effect of showing different-quality CQs across the entire session. The quality of conversational information-seeking systems asking CQs is usually evaluated based on the entire session, and thus trajectories are important at the session level. This way we are able to capture the impact at the session level and compare it with query-level impact. To this end, we conduct a user study involving 89 participants, asking them to complete a set of Web search tasks, following a standard laboratory-based user study setup (Edwards & Kelly, 2017; Harvey & Pointon, 2017). In particular, we simulate various quality trajectory patterns that a user and a system would encounter where we study the effect of these trajectory patterns on user behavior and satisfaction.

Our user study leads to the following findings:

- Users interact with SERPs less extensively when they have seen preceding low-quality CQs before high-quality CQs. Similarly, users interact with SERPs more extensively if they have seen preceding high-quality CQs before low-quality CQs.
- A low-quality CQ leads to low search performance and user satisfaction. Asking a follow-up high-quality CQ improves search performance and user satisfaction, alleviating the earlier negative impact caused by asking a low-quality CQ.
- Compared with always showing CQs, only showing high-quality CQs while hiding other CQs leads to a better search experience: better search performance and better user satisfaction with shorter sessions and lower time efforts. That is, always showing CQs may be risky and low-quality CQs do disturb users.

Based on the above findings, we further develop a transformer-based model on top of user behavior data, named TranShow. TranShow predicts which CQs we should show to users, in order not to disturb users. To the best of our knowledge, TranShow is the first effort to utilize user behavior data for predicting whether a system should ask a CQ or not.

The rest of this paper is organized as follows. In Section 2, we introduce the related work. In Section 3, we explain the details of our method and study design. In Section 4, we describe the research questions, and provide a detailed analysis as well as a discussion of the results. Section 5 expresses our CQ showing prediction model and performance analysis, while Section 6 concludes the paper.

---

1 The quality of CQs measures the degree of clarifying user intents and usefulness for completing the search task.

2 A search session is an entire session for a user completing a search task, which may involve multiple actions from the user.
3. Method

In this section, we present our method, including the study protocol, study system, CQs and trajectory patterns, behavior and satisfaction measures, and participants.

3.1. Study protocol

We have developed an online Web system to conduct the user study. The study protocol is shown in Fig. 1 as follows:

(1) Participants are presented with detailed instructions regarding this study. They are also trained to be familiar with the system through a detailed video.

(2) We ask the participants to complete a demographic survey on their gender, age, career field, English language proficiency, and highest education level completed.
J. Zou et al.

(1). Detailed Instruction

(2). Demographic survey

(3). Task selection

Suppose you'd like to take vacation, and you finally decide to go to the city of Bellevue, Washington. Since you have never been there before, you would like to search online to gather some information about Bellevue.

(4). Interacting with search system

(5). Post-task Questionnaire

End

Fig. 1. Study protocol overview. An example Web search task is shown in Step (3), and the search interface is shown in Step (4). After completing Step (5), users may choose to complete another task from Step (3).

(3) Participants are presented with a list of Web search tasks. They are guided to read through all the task descriptions. Then they select a Web search task with which they feel most comfortable, to relieve task assignment bias (Ho & Vaughan, 2012).

(4) Participants start to search for relevant information for the selected task, either by submitting queries themselves or by interacting with presented CQs. Once they find some relevant results, they can bookmark the results as relevant.

(5) After the participant completes a Web search task, he/she is asked to complete a post-task questionnaire, on his/her experience and satisfaction, including the overall satisfaction rating, and perceived helpfulness.

(6) Participants are asked to consider selecting another Web search task to complete, by returning to step (3), or they can finish the study participation.

The flow of study processing for asking CQs in Web search is also provided in Fig. 2. All participants are informed that their data is securely encrypted and is only used in a collective manner. We did not collect any data to breach their privacy. Moreover, this study is approved by the ethics committee of the institute.

3.2. Study system

We follow Zou et al. (2022a) to mimic the search interface of a commercial search system (Bing.com) (Zamani, Dumais et al., 2020) and use ChatNoir3 Web search engine (Bevendorff et al., 2018; Potthast et al., 2012) to produce SERPs. ChatNoir is a widely used Web search engine indexing the entire ClueWeb09/12 corpus (Hagen et al., 2017; Vakkari et al., 2019). As shown in step (4) in Fig. 1, the system includes a typical search interface and embeds a CQ pane at the top of the search page. Users are allowed to reformulate their queries by themselves or interact with the CQ pane (if presented).

We design four search tasks derived from the Text Retrieval Conference (TREC) Web Track 2009–2012,4 with an expanded description to offer participants a search context as done in previous work (Borlund, 2003; White et al., 2007). These search tasks are about collecting information regarding certain topics (e.g., a tourist city, an airport, a movie, a famous quote, or a fun fact) from various Web sources. An task example is shown in step (3) in Fig. 1.

3.3. CQs and trajectory patterns

In the absence of a reasonable model to automatically generate CQs and candidate answers of different qualities, we manually construct a pool of CQs along with their candidate answers beforehand, same as other studies in the community (Aliannejadi et al., 2020, 2019; Ren et al., 2021; Zou et al., 2022a).

3 https://www.chatnoir.eu.
5 From our post-hoc analysis, distribution of the selected search tasks is balanced. There is no obvious preference for any particular task. Furthermore, 93% of participants indicate that the task description is very clear.
Fig. 2. The flow of study processing for asking CQs in Web search. We aim to explore the effect of CQ trajectories (query-level and session-level) in terms of CQ quality on user behavior and satisfaction. CQ quality, query-level and session-level trajectory patterns of CQs are described in Section 3.3. User behavior and satisfaction measures are introduced in Section 3.4.

Table 1
Example CQs of different qualities. The generated CQs are from the search task “Bellevue Washington”.

<table>
<thead>
<tr>
<th>Quality</th>
<th>Examples of CQs &amp; answers</th>
</tr>
</thead>
</table>

In this work, we use three-level CQ quality categories similar to Zamani, Lueck et al. (2020) and Zou et al. (2022a): (L) low-quality, (M) mid-quality, and (H) high-quality. The CQ generation pipeline is as follows. First, two expert annotators generate CQs in three quality categories as well as candidate answers, for each search task. In case of disagreement, they would discuss and agree on a better formulation. Second, the expert-generated CQs are deployed in a pilot study, and all CQs are labeled by the actual users independently, again in the three CQ quality categories. Third, we remove those CQs with conflict category labels between expert annotators and pilot study users. Hence, only the CQs with consistent quality labels are used in the actual study, to ensure the label quality. In the end, we have six CQs with their respective candidate answers for each search task. Each CQ has at most five answers, with each answer corresponding to a reformulated query for the next turn, following Bing’s setting (Zamani, Mitra et al., 2020; Zou et al., 2022a). When generating and labeling the three-level CQ quality categories, we provide detailed guidelines and examples for each of the low-, mid-, and high-quality CQs to expert annotators and pilot study users. Specifically, the guideline indicates that high-quality CQs should meet the following criteria (Zamani, Dumais et al., 2020): (1) highly helpful for completing the search task, (2) correctly clarifying user intents, and (3) being fluent and grammatically correct. If a CQ fails to meet some of these criteria but is still an acceptable CQ to be asked, it should be labeled as mid-quality. Otherwise, low-quality labels should be provided by expert annotators and pilot study users. The general principle in constructing such CQ quality categories is to cover a variety of CQ quality categories and explore the potential effects of CQ trajectories under these quality categories. High-quality CQ is designed as the highest CQ quality which is highly useful and can aid users, whereas low-quality CQ is the lowest CQ quality which is usually useless or off-topic so that may elicit user dissatisfaction. CQ examples under each CQ quality category are shown in Table 1.

To study trajectory effects in information-seeking systems, we also define a series of query-level and session-level trajectory patterns, as shown in Table 2. Specifically, we define the following nine trajectory patterns to study the query-level impact of
to ensure they are not filtered out wrongly. (a minimal expected threshold for a trustworthy worker Han et al., 2020). The removed participants are further checked manually reading the textual descriptions of Web search tasks. We then filter out participants who read the task descriptions in less than 10 s descriptions to make sure that they have read and understood the instructions, and (ii) we evaluate the time participants spent

89 participants are recruited after quality control filtering.

Before the actual study, we ran a pilot study with 23 participants, to iterate over the experimental design and confirm the quality labels of CQs. For the actual user study, participants were split into two groups. One group of participants completes the search task with a plain interface, i.e., without CQs shown in the search session. Another group of participants completes the search task with a CQ pane, similar to the one in step (4) in Fig. 1. If a participant was assigned to the group with a CQ pane, CQs would be shown by following a randomly assigned CQ trajectory pattern, following the assignment setup of related literature (Collins-Thompson et al., 2016; Harvey & Pointon, 2017; Kelly, 2009; Kelly & Azzopardi, 2015). For each trajectory pattern, the CQ to be shown to the user was randomly selected from CQs under the corresponding CQ quality category for a certain search task (Collins-Thompson et al., 2016; Harvey & Pointon, 2017; Kelly, 2009; Kelly & Azzopardi, 2015). Users can answer multiple CQs in one search session. When a CQ is shown to the user, the user can choose to reformulate their query again or click on a CQ answer. If the user clicks on a CQ answer, the answer is concatenated to the user’s query and resubmitted to the search engine, following Bing’s setting (Zamani, Mitra et al., 2020). For participants in the group with a CQ pane, once they issue a new query, or answer a CQ, a new CQ is selected to be displayed to them.

3.4. Behavior and satisfaction measures

In our user study, we use the user behavior and satisfaction measures, following Kelly and Azzopardi (2015) and Zou et al. (2022a). Specifically, regarding behavior measures, we include the following metrics: the CQ engagement rate (click-through rate), number of SERP scrolls, number of SERP hovers, number of SERP clicks, number of queries issued, number of query terms, number of results marked relevant (# bookmarks), number of correct bookmarks (# hit), and SERP quality measured by nDCG@10 (normalized discounted cumulative gain from rank 1 to 10), dwell time on SERPs per query, and the overall task time for a Web search task. Regarding satisfaction, we use explicit feedback collected through the post-task questionnaires, including the overall satisfaction rating, and user-perceived helpfulness. Specifically, in the post-task questionnaires, the options for perceived helpfulness include “positive”, “negative” and “neutral”. The overall satisfaction rating is on a scale from 1 to 5. The query-level measures are calculated based on the current search page while the session-level measures are calculated accumulatively throughout the full search session.

3.5. Participants

We recruit participants through an academic crowdsourcing platform, called Prolific. Before the actual study, we ran a pilot study with 23 participants, to iterate over the experimental design and confirm the quality labels of CQs. For the actual user study, 89 participants are recruited after quality control filtering.

To ensure data quality from crowdsourcing participants, we perform quality control filtering to avoid including results from some less trustworthy participants. Specifically, we deploy two quality checks: (i) we ask participants questions regarding the study descriptions to make sure that they have read and understood the instructions, and (ii) we evaluate the time participants spent reading the textual descriptions of Web search tasks. We then filter out participants who read the task descriptions in less than 10 s (a minimal expected threshold for a trustworthy worker Han et al., 2020). The removed participants are further checked manually to ensure they are not filtered out wrongly.

Participants were paid around 3.33 pounds each to complete the study. Their demographic data is briefed as follows:

- Gender: 55 females, 30 males, 4 non-binary.
- Age: 12 participants are in the age group 18–24, 41 are in 25–34, 21 are 35–44, and 15 are older than 44 years old.
- Career field: 21 in science, computers and technology, 11 in management, business and finance, 10 in education and social services, 13 in healthcare, 9 in arts and communications, 4 in law and law enforcement, and 1 in trades and transportation; 20 do not specify.
- English language proficiency: 82 native, 5 proficient, and 2 beginners.
- Highest education level completed: 11 high school, 23 college, 33 bachelor’s, 10 master’s, and 2 doctorate; 10 do not specify.

From our post-hoc analysis, the distribution of each trajectory pattern is balanced.

https://www.prolific.co/.
4. Results

The analysis is based on 897 search requests made between our system and 89 crowd workers on 241 search sessions. The statistics of the collected data are detailed in Table 3. Unless otherwise reported, we perform t-tests (Kim, 2015) and one-way analysis of variance (ANOVA) for statistical analysis in this study, assuming the independence of different groups (Xie et al., 2018). Specifically, we use t-tests for comparisons between two groups only, and use one-way ANOVA and Tukey’s HSD tests for comparisons with more than two groups, following past user studies (Turpin & Scholer, 2006). The presence or absence of CQs is a between-subjects variable, whereas the CQ quality category is a within-subjects variable.

Through analyzing the collected data, we answer the following research questions:

• **RQ1:** To what extent do different trajectory patterns affect user behavior at the query level?
• **RQ2:** To what extent do different trajectory patterns affect user behavior and satisfaction at the session level?
• **RQ3:** Does always showing CQs benefit or disturb users?

4.1. Query-level impact of trajectory (RQ1)

We first investigate the query-level impact of CQ trajectory patterns on user search behaviors. The results of CQ engagement are plotted in Fig. 3(a), and results of other user behavior measures categorized by different query-level trajectory patterns are reported in Table 4.

**CQ engagement.** Observe from Fig. 3(a), the engagement of first showing CQs is higher than others. A possible reason is that users focus more on their first seen CQs. Showing better preceding CQs (mid- or high-quality CQs) before low-quality CQs increases the CQ engagement of low-quality CQs, indicating CQs might affect user interactions of the following CQs, and low-quality CQs are beneficial from the last seen mid- or high-quality CQs.

**SERP scrolls and hovers.** SERP scrolls and hovers are valuable signals of user behavior in Web search (Huang et al., 2012). Table 4 shows that the number of scrolls and hovers increase from L2L to H2L, from L2M to H2M, and from L2H to H2H. This observation for scrolls and hovers indicates that users scan a SERP more extensively when they have seen a preceding high-quality CQ and users scan a SERP less extensively when they are shown a preceding low-quality CQ. That is, users are more confident about SERP when they have seen a preceding high-quality CQ.

**SERP clicks.** We observe that the number of SERP clicks on the H2L condition is more than that of the L2L condition. Also, we see that the number of SERP clicks on H2H is higher than that on L2H, suggesting that the quality of a preceding CQ plays a role in the number of SERP clicks on the current CQ. That is, asking a preceding high-quality CQ leads to higher SERP clicks for the current CQ, no matter what the quality of the current CQ is.

**Dwell time.** For SERP dwell time, we see L2L < M2L < H2L, L2M < M2M < H2M, and L2H < M2H < H2H. This result again confirms that users spend more time and scan a SERP more extensively when they have seen a preceding high-quality CQ, while...
users spend less time and scan a SERP less extensively when they are shown a preceding low-quality CQ. A reason is that users tend to interact with the high-quality CQ when they see a preceding high-quality CQ, leading to an improved SERP quality. The improved SERP quality then makes users locate more relevant SERP results, leading to an increase in dwell time.

**SERP quality.** For SERP quality, measured by nDCG@10, we observe L2L < H2L, L2M < H2M, and L2H < H2H. This indicates that showing a preceding low-quality CQ before a high-quality CQ decreases SERP quality while showing a preceding high-quality CQ before a low-quality CQ increases SERP quality. A possible reason is query reformulation, i.e., users reformulate their queries toward better queries when they have seen a preceding high-quality CQ.

**Bookmark quality.** Bookmark quality, in terms of # bookmarks and # hits, reflects user search performance for finding relevant information. From Table 4, we observe that both the number of bookmarks and the number of hits decrease from H2H to M2H to L2H. This demonstrates that showing a preceding low-quality CQ before a high/mid-quality CQ lowers user search performance for finding relevant information, as a preceding low-quality CQ leads to low SERP quality. In contrast, showing preceding high/mid-quality CQs before low-quality CQs increases the number of bookmarks.

**Summary.** Overall, as for the query-level impact of different trajectories of asking CQs on user behavior, we observe the trend that adding preceding low-quality CQs before high-quality CQs lowers user behavior measures. Adding preceding high-quality CQs before low-quality CQs leads to increases in user behavior measures. That is, when users have seen preceding low-quality CQs before high-quality CQs, they interact with SERPs less extensively, including less CQ engagement, mouse hovering behavior, mouse scrolling behavior, SERP clicking behavior, bookmark behaviors (number of bookmarks and number of hits), and spend less dwell time. In contrast, users interact with SERPs more extensively when they have seen preceding high-quality CQs before low-quality CQs.

### 4.2. Session-level impact of trajectory (RQ2)

We now explore the session-level impact of CQ trajectory patterns on user search behaviors. In addition, we investigate how user satisfaction for the whole session is affected by CQ trajectory patterns.9

---

9 Users perceive satisfaction from the whole search session and they only rate satisfaction measures after they complete a search session. Therefore, we explore the effects of CQ trajectory on user satisfaction from the session level only.
4.2.1. User behavior

We analyze user behavior in a session where we display a high-quality CQ after a low-quality CQ (or vice versa) and see how it affects various user behavior measures on a session-level scale. The results of CQ engagement are shown in Fig. 3(b) and the results of other user behavior measures are listed in Table 5. To make the comparison easier, we report the results of four more descriptive trajectory patterns, i.e., Lb4L, Lb4H, Hb4L, and Hb4H, in Table 5.

CQ Engagement. From Fig. 3(b), we observe that a low-quality CQ leads to a low CQ engagement, and showing a high-quality CQ after a low-quality CQ in a session increases CQ engagement. This observation that users prefer to engage with high-quality CQs, is in line with previous work (Zou et al., 2022a; Zou, Kanoulas et al., 2020). We see a considerable difference in CQ engagement when comparing sessions of Lb4L and Lb4H. Starting with a low-quality CQ and followed by higher-quality CQs lead to better engagement. This suggests that there is a good chance for a system to recover from an initial low-quality CQ by asking CQs of higher quality. Therefore, systems can take into account user feedback on a low-quality CQ, and make a more informed decision on showing the next CQs.

SERP scrolls and hovers. From Table 5, we observe that numbers of SERP scrolls and hovers increase from Lb4L to Lb4H, and from Hb4L to Hb4H. This indicates that users are more engaged with SERP when CQs of higher quality are posed to them after a preceding CQ. We observe a somewhat symmetric difference in terms of hover and scroll when comparing the four conditions and keep the initial CQ fixed ([Lb4L - Lb4H] vs. [Hb4L - Hb4H]). The effect of showing a high-quality CQ after a low-quality CQ is an inverse of showing a low-quality CQ after a high-quality CQ. Therefore, as much as the system can recover from a low-quality CQ (by asking a high-quality CQ), it faces the risk of losing user interest by asking low-quality CQs (after a high-quality CQ).

SERP clicks. The number of SERP clicks increases from Lb4L to Lb4H, and from Hb4L to Hb4H. The trend is similar to scrolls and hovers. Therefore, the same symmetric effect exists in this case too.

User queries. Also, numbers of user queries and query terms increase from Lb4L to Lb4H, and from Hb4L to Hb4H.

Session time. The session time increases from Hb4L to Hb4H while remaining a similar number between Lb4L and Lb4H. Here, we observe an interesting phenomenon: a considerable drop in session time when a low-quality CQ is shown after a high-quality CQ (Hb4L). This can be due to the fact that users find most of the relevant results when they interact with high-quality CQs; therefore, when they see a low-quality CQ, there is less motivation to interact with the SERP, leading to a shorter session time.

SERP quality. The SERP quality increases from Lb4L to Lb4H while remaining a similar number between Hb4L and Hb4H.

Bookmark quality. Numbers of marked results and hits increase from Lb4L to Lb4H, and from Hb4L to Hb4H. We see a higher impact when comparing the difference between Hb4L and Hb4H, as opposed to the difference between Lb4L and Lb4H. There is more risk in showing a low-quality CQ after a high-quality CQ, compared with the gain one receives when showing a high-quality CQ after a low-quality CQ. This can be due to the first impression of users when they interact with a low-quality CQ, and that impacts their behavior and interest all over the session.

Summary. As for session impact, compared with showing low-quality CQs, showing follow-up high-quality CQs leads to better user behavior; users spend similar time to find more relevant information (# bookmarks and # hits). This suggests that while a low-quality CQ leads to low user search performance and user satisfaction, asking a follow-up high-quality CQ improves search performance and user satisfaction, alleviating the earlier negative impact caused by the low-quality CQ. After showing high-quality CQs, showing follow-up high-quality CQs leads to finding more relevant information, but also more time, compared with showing follow-up low-quality CQs. However, we observe that even though the user effort may not change much in these two cases, their efficiency is different. There are various possible reasons, e.g., a bad first impression, or misleading CQs that affect the user’s next query in a session. Therefore, even though the system can recover from a low-quality CQ, it still cannot fully recover because of the impact of the low-quality CQ has on users’ behavior and interest.

4.2.2. User satisfaction

Observe from Table 6 that, when there are low-quality CQs in a session (Lb4L, Lb4H, and Hb4L), user satisfaction ratings are lower than no CQs showing. This confirms that low-quality CQs lead to low user satisfaction. Showing high-quality CQs after low-quality CQs in a session improves the user satisfaction ratings, and user-perceived helpfulness (i.e., higher positive percentage and lower negative percentage) (Lb4L vs. Lb4H). Showing low-quality CQs after high-quality CQs in a session lowers the user satisfaction ratings, and user-perceived helpfulness (Hb4L vs. Hb4H). In short, when a low-quality CQ leads to low user satisfaction, a follow-up high-quality CQ improves user satisfaction and alleviates the earlier negative impact of the low-quality CQ.

Table 6
User satisfaction measures by different session trajectory patterns. Satisfaction are represented by its means followed by standard deviations in parenthesis. User-perceived helpfulness are represented by ratio of positive/negative ratings.

<table>
<thead>
<tr>
<th></th>
<th>No CQs</th>
<th>Lb4L</th>
<th>Lb4H</th>
<th>Hb4L</th>
<th>Hb4H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Satisfaction</td>
<td>3.33(1.20)</td>
<td>2.97(1.16)</td>
<td>3.04(1.32)</td>
<td>3.28(1.20)</td>
<td>3.33(1.14)</td>
</tr>
<tr>
<td>Helpfulness (%)</td>
<td>25.00/50.00</td>
<td>47.95/46.58</td>
<td>56.36/40.00</td>
<td>64.00/26.67</td>
<td>65.38/26.92</td>
</tr>
</tbody>
</table>
4.3. Session-level impact of always showing CQs (RQ3)

Next, we explore the session-level impact on user behavior in cases where a system would always show CQs to users. Our goal is to demonstrate the importance of effective prediction of clarification needs, in a search session, and the negative effect it may bring in on the user’s experience.

4.3.1. User behavior

Table 7 reports session-level behavior measures for always vs. not-always showing CQs. For the condition of always showing CQs, we constantly show a CQ (low- or mid- or high-quality CQ) for each search request in the search session. For the condition of not-always showing CQs, we show only one category of CQs (low- or mid- or high-quality CQ) and hide CQs for search requests if they are not low/mid/high-quality CQs in the search session, respectively. For instance, for search sessions showing high-quality CQs only (‘H-only’), we show the CQ if it is a high-quality CQ and hide the CQ if it is not a high-quality CQ.

CQ engagement. We observe that CQ engagement decreases for low-quality CQ from always showing CQs (48.05%) to not-always showing CQs (28%), while increases for high-quality CQ from always showing CQs (50.41%) to not-always showing CQs (54.10%). Showing various low-quality CQs leads to more engagement from users, which can also increase the risk of dissatisfaction because these CQs are more likely to be useless.

SERP scrolls and hovers. Observe from Table 7, numbers of SERP scrolls and hovers increase from low-quality CQs to mid-quality CQs to high-quality CQs. Compared with L and M (always showing), L-only and M-only (not-always showing) achieve a lower number of SERP scrolls and hovers, respectively. H and H-only contain similar numbers of SERP scrolls and hovers. This again confirms that showing more CQs lead to more user engagement and bears more cost to users, leading to the risk of user dissatisfaction.

SERP clicks. Number of SERP clicks changes slightly from low-quality CQs to mid-quality CQs to high-quality CQs. M-only and H-only (not-always showing) achieve a higher number of SERP clicks than M and H (always showing), respectively. Despite a higher CQ engagement as well as hovering and scrolling, users click on fewer results. This suggests that even though users are engaged more with SERP and spend more time, their performance to click the interesting results is lower when CQs are always shown to them.

User queries. Numbers of user queries and query terms increase from low-quality CQs to mid-quality CQs to high-quality CQs. L-only, M-only, and H-only (not-always showing) achieve a lower number of user queries and query terms than L, M, and H (always showing), respectively.

Session time. Session time increases from low- to high-quality CQs. L-only, M-only, and H-only (not-always showing) achieve lower session time than L, M, and H (always showing), respectively. As mentioned earlier, more engagement with CQs as well as SERP leads to more time spent on SERP.

SERP quality. SERP quality increases from low-quality CQs to mid-quality CQs to high-quality CQs. L-only, M-only, and H-only (not-always showing) achieve lower SERP quality than L, M, and H (always showing), respectively. This is perhaps because engaging with more CQs leads to longer sessions and more SERPs, hence having higher nDCG.

Bookmark quality. Number of marked results and hits increase from low-quality CQs to high-quality CQs. L-only and M-only (not-always showing) achieve a lower number of marked results and hits than L and M (always showing), respectively, while H-only achieves a higher number of marked results and hits than H. Looking at session time, we notice a big difference in the time spent on pages for the two conditions (always showing vs. not-always showing). Therefore, we compute the rate of gain for a fair comparison of user performance. We follow Aliannejadi, Azzopardi et al. (2021) and define rate of gain as the number relevant documents per second, formally, rate of gain = #hits/session time. As reported in the last row of Table 7, we see that all of the values are higher for not-always showing condition, indicating that users are more efficient when fewer CQs were shown to them.

Summary. Mostly user behavior metrics increase from low-quality CQs to mid-quality CQs then to high-quality CQs (L → M → H). This indicates that users from sessions involving high-quality CQs interact with SERPs more deeply and spend more time, and find more relevant information. Compared with low-quality CQs (L) or mid-quality CQs (M), showing low-quality CQs only or showing mid-quality CQs only (M-only) lowers user behavior measures (L vs. L-only & M vs. M-only). Compared with sessions containing low-quality CQs or mid-quality CQs, users from sessions showing low-quality CQs only or showing mid-quality CQs only interact with SERPs less deeply and spend less time, finding less relevant information. Higher efficiency is observed when CQs are not always shown to users. Moreover, the highest improvement in the rate of gain is observed for high-quality CQs (H vs. H-only), suggesting that a selective posing of high-quality CQs would help users achieve a much higher rate of gain. This is in line with Aliannejadi, Kiseleva et al. (2021) where the necessity of predicting clarification needs in a conversational session is highlighted. This result also suggests that hiding low-quality and mid-quality CQs for a session is beneficial. Our finding is in agreement with Wang and Ai (2021), that low-quality CQs could bring risk to users, suggesting that we should model the risk of low-quality CQs and predict the necessity of showing CQs.
This is in line with the findings of Aliannejadi, Azzopardi et al. (2021) some cases always showing low-quality CQs lead to a higher number of hits in a session, it comes with the cost of nearly twice as engagement with the CQs (see Fig. 3); therefore, there is a higher risk when users engage with lower-quality CQs. Even though in session by taking into account various factors, e.g., query ambiguity (Guo et al., 2021). Showing more CQs clearly leads to higher to model the risk of asking a CQ (Wang & Ai, 2021) and predict the clarification need (Aliannejadi, Kiseleva et al., 2021) in a RQ3: Does always showing CQs benefit or disturb users?

Can lead to a higher rate of gain, which reiterates the importance of predicting the risk (Wang & Ai, 2021) of asking a low-quality failure. Moreover, as observed from Table 5, for all CQ qualities, an informed decision on the trajectory of showing CQs to the users if a higher-quality CQ is shown after a lower-quality CQ, giving the opportunity for improved user experience in case of an initial = CQ engagement for Lb4H

A high-quality CQ after a low-quality CQ would affect the user's satisfaction of the session as a whole? From Fig. 3(b), we see the This is crucial to understand as it sheds light on various aspects of user behavior and its relation to the RQs, e.g., how much asking a high-quality CQ after a low-quality CQ would affect the user's satisfaction of the session as a whole? From Fig. 3(b), we see the immediate effect of asking multiple CQs in the RQ1, in this RQ we focus on the effect of CQ trajectories at a session-wide level. This is crucial to understand as it sheds light on various aspects of user behavior and its relation to the RQs, e.g., how much asking a high-quality CQ after a low-quality CQ would affect the user's satisfaction of the session as a whole? From Fig. 3(b), we see the same pattern during a session, i.e., asking a high-quality CQ after a low-quality CQ leads to more engagement. As we compare the CQ engagement for Lb4H = 61% and LbL = 12%, we see that the system can gain up to 5 times higher user engagement in a session if a higher-quality CQ is shown after a lower-quality CQ, giving the opportunity for improved user experience in case of an initial failure. Moreover, as observed from Table 5, for all CQ qualities, an informed decision on the trajectory of showing CQs to the users can lead to a higher rate of gain, which reiterates the importance of predicting the risk (Wang & Ai, 2021) of asking a low-quality CQ.

RQ2: To what extent do different trajectory patterns affect user behavior and satisfaction at the session level? After studying the immediate effect of asking multiple CQs in the RQ1, in this RQ we focus on the effect of CQ trajectories at a session-wide level. This is crucial to understand as it sheds light on various aspects of user behavior and its relation to the RQs, e.g., how much asking a high-quality CQ after a low-quality CQ would affect the user's satisfaction of the session as a whole? From Fig. 3(b), we see the same pattern during a session, i.e., asking a high-quality CQ after a low-quality CQ leads to more engagement. As we compare the CQ engagement for Lb4H = 61% and LbL = 12%, we see that the system can gain up to 5 times higher user engagement in a session if a higher-quality CQ is shown after a lower-quality CQ, giving the opportunity for improved user experience in case of an initial failure. Moreover, as observed from Table 5, for all CQ qualities, an informed decision on the trajectory of showing CQs to the users can lead to a higher rate of gain, which reiterates the importance of predicting the risk (Wang & Ai, 2021) of asking a low-quality CQ.

RQ3: Does always showing CQs benefit or disturb users? From Table 7, we clearly see that systems should employ a method to model the risk of asking a CQ (Wang & Ai, 2021) and predict the clarification need (Aliannejadi, Kiseleva et al., 2021) in a session by taking into account various factors, e.g., query ambiguity (Guo et al., 2021). Showing more CQs clearly leads to higher engagement with the CQs (see Fig. 3); therefore, there is a higher risk when users engage with lower-quality CQs. Even though in some cases always showing low-quality CQs lead to a higher number of hits in a session, it comes with the cost of nearly twice as much time as they spend in a session with fewer shown CQs. This is in line with the findings of Aliannejadi, Azzopardi et al. (2021)
where they discuss the trade-off between showing more results vs. asking CQs and find that asking more CQs does not always return the same amount of gain (i.e., lower rate of gain).

4.5. Implications

In this section, we also highlight the implications arising from this study, including both theoretical and practical implications.

**Theoretical implications** Our study advances the understanding of user interactions with CQs in information-seeking systems. Specifically, our results contribute to the understanding of how and why different trajectories of CQs with different qualities affect user behavior and satisfaction at the query level and session level. To the best of our knowledge, our work is the first effort to discuss the effects of trajectories of asking CQs, and thus enriches the literature on asking CQs in information-seeking systems. Unlike ad-hoc retrieval where a single query would be enough, conversational information-seeking systems by asking CQs are usually evaluated based on a whole session, and thus trajectories are crucial in this setting. By quantifying user behavior and satisfaction caused by asking CQs of different quality, we further disentangle the relationships between the quality of CQs and various aspects of user behavior and satisfaction. Also, our results reinforce the importance of asking CQs in information-seeking systems. We confirm that asking high-quality CQs is beneficial. However, we also provide insights into the extent to which asking low-quality CQs disturbs users and what negative impacts low-quality CQs bring to users. Moreover, our study provides viewpoints on how multiple CQs in a search session affect each other, and how a trajectory of CQs can be used for a better user experience. For example, we confirm that when the system receives negative feedback for the first low-quality CQ, the system can take follow-up actions such as asking CQs of higher quality to regain user satisfaction.

**Practical implications** Practitioners and researchers can gain practical lessons from these findings in terms of the effect of different trajectories to ask CQs at the query and session levels. Our results have design implications that will allow practitioners to support CQs, and help them to decide in which situations to show CQs so that users can benefit from asking for CQs. As asking CQs in information-seeking systems is still far from widespread adoption, the system providers need a better understanding of user interaction with CQs of different quality. Our study provides a foundation for the industry to develop a normative framework to evaluate the adoption potential of new services regarding asking CQs. Also, our results are able to support the development of CQ-based algorithms. Our study sheds light on asking a CQ with uncertain quality as it may bring risks. Although this is inevitable as the system may provide risky CQs unintentionally, our study provides motivations for modeling risk and urge the researchers to design suitable risk-aware algorithms based on asking CQs (Wang & Ai, 2021). Similarly, users may choose noisy answers when interacting with CQs, our results prompt the researchers to design noise-tolerant algorithms on the basis of asking CQs. Moreover, we investigate the effects of always showing CQs, indicating that always showing CQs leads to a higher user dissatisfaction risk. This motivates the problem of CQ showing prediction. Future research for CQ-based systems would benefit from incorporating CQ showing prediction into their models. Last, our study discovers the relationships between user behavior measures and CQ quality, which can be used for assessing CQ quality. We demonstrate that user behavior and satisfaction are affected by the quality of CQs. This would advance the research on algorithms for generating CQs and selecting CQs in information-seeking systems; the algorithm should incorporate a quality-aware module, either from CQ quality assessment models or from real CQ quality labels (e.g., three-level CQ quality labels in the real dataset for asking CQs in information-seeking systems, MIMICS Zamani, Lueck et al., 2020), when generating and selecting CQs.

5. CQ showing prediction

As we demonstrate that always showing CQs is not optimal, it is necessary to predict whether the system should or not display a CQ to users, to avoid disturbing users with low-quality CQs. To this end, we develop a transformer-based algorithm on top of the user study data to predict the necessity of showing CQs.

5.1. Model

We adopt the Transformer framework to predict CQ Showing. We suppose there is a set of CQs, denoted by \( C = \{c_1, c_2, \ldots, c_n \} \). Each \( C \), \( c_i \in C \) has a set of corresponding candidate answers \( A_i \). Given an initial query \( q \), a selected CQ \( c \) along with its candidate answers \( A \), a list of search results \( S \), and user behavior data \( Y \), we aim to predict the CQ showing signal of the current selected CQ \( c \) and its candidate answers \( A \).

\[
P(q, c, A, S, Y) = \text{Sigmoid} \left( \text{Dropout} \left( W \times h_{\text{CLS}} + b \right) \right) \in [0, 1].
\]  
\[
h_{\text{CLS}} = \text{ALBERT}(q, c, A, S, Y),
\]

where \( q, c, A, S, \) and \( Y \) are embeddings of \( q, c, A, S, \) and \( Y \), respectively. The selected CQ \( c \) is a CQ selected by the system intended to show to a user before the CQ showing prediction. For search results \( S \), we use SERP titles in this paper. For user behavior data \( Y \), we use the query position (i.e., \( i \)th query) and mouse hovering time on the previous showing CQs before the current search query in the search session. If there are no previous showing CQs or no mouse hovering behavior, it is set to zero. We use \( q, c, A, S, \) and \( Y \) as input factors, as the quality of \( q, c, A, \) and \( S \) contain cues on how necessary to show the CQ pane, while \( Y \) demonstrates the user preference to view the CQ pane. Our ablation experiments in Section 5.2 also demonstrate the importance and contribution of these factors. The model outputs the probability of showing the selected CQs based on the input representations. In our implementation,
we use ALBERT (Lan et al., 2019), a lite BERT, as our pre-trained language model for generating embeddings of $q$, $c$, $A$, $S$, and $Y$. ALBERT takes the tokenized query $q$, CQs $c$, candidate answers $A$, SERP titles $S$, behavior information like the query position and mouse hovering time on the previous showing CQs $Y$ as input, separated by the Separation $SEP$ token. A classification token $CLS$ is also inserted at the beginning of the input sequence. In this way, the hidden-state representation of the classification token $CLS$ in the last layer $h_{CLS}$ is further passed to produce the prediction. The prediction is performed by a Sigmoid function with a linear layer and a dropout layer in between to produce an output probability over CQ showing. During training, we minimize Binary Cross Entropy loss as the nature of the task is a binary classification problem.

We train our model using Adam optimizer (Kingma & Ba, 2014) and Pytorch with a learning rate of $10^{-4}$. The hidden dimension is 768, as in the pre-trained language model ALBERT. The maximum sequence length is set to 512 tokens. We use 5-fold cross-validation to evaluate our model. In each fold, the dataset is split into training and testing set by the ratio of 8:2. For ground truth generation, we use the SERPs in which the showed CQ receives a click-through as the positive CQ showing class while others as the negative CQ showing class.

5.2. Performance

From the results reported in Table 9, we observe that the performance is much higher than the straightforward baselines: always showing CQs and always hiding CQs. Hence, CQ showing is predictable and our model to predict CQ showing is effective.

To evaluate the importance of different factors, we further conduct an ablation study comparing our TranShow model with its ablation variants. We refer the TranShow by removing the information of query, CQ pane (i.e., CQs and candidate answers), SERP titles, query position, and CQ hovering time as “-w/o query”, “-w/o CQ pane”, “-w/o SERP titles”, “-w/o query position”, and “-w/o hovering”, respectively. Based on the results in Table 9, all factors contribute to the final performance. Among all factors, the CQ pane and user behavior data of mouse hovering time on the previous showing CQs are the most important factors as they contribute most to the prediction performance.

6. Conclusion and limitations

In this paper, we conduct a user study to investigate the effects of trajectory of showing CQs in information-seeking systems. We explore the effect of asking different trajectory patterns showing different quality categories of CQs on user behavior and satisfaction at the query- and session-level. Moreover, we propose a transformer-based model, named TranShow, to predict CQ showing to alleviate the disturbance that low-quality CQs might bring to users.

From our analysis, we learn that low-quality CQs disturb users and systems should ask CQs only when necessary. Always showing CQs may be risky and systems should hide CQs with uncertain quality. When a low-quality CQ is shown to users, the system still has the chance to redeem user interest by asking follow-up high-quality CQs. Furthermore, future research for CQ-based systems should incorporate CQ showing prediction or risk modeling of CQs into their models, instead of simply asking many CQs.

One limitation of this study is the small number (i.e., four) of Web search tasks used. Using a limited number of experimental tasks to control factors while maintaining the cost at a reasonable level is a widely used setting in many studies in the community (e.g., Cole et al., 2015; Collins-Thompson et al., 2016; Edwards & Kelly, 2017; Han et al., 2015; Harvey & Pointon, 2017; Kelly et al., 2015; Kelly & Azzopardi, 2015; O’Brien et al., 2020; Zou et al., 2022a). Nevertheless, we leave the analysis for more Web search tasks as future work. Also, the number of participants is limited, while a similar setting has been adopted in some previous laboratory user study setups (Cole et al., 2015; Collins-Thompson et al., 2016; Edwards & Kelly, 2017; Han et al., 2015; Harvey & Pointon, 2017; Kelly & Azzopardi, 2015; Kiesel et al., 2018; Kim, 2008; Liu et al., 2019; Vtyurina et al., 2017; Xie et al., 2017; Yuan et al., 2014; Zou et al., 2022a). It would be beneficial to extend the user study with more participants from different user groups in the future.

We propose a transformer-based model, TranShow, to predict whether or not to show CQs, based on the collected user study data. We understand that, as a controlled laboratory user study, the data is limited to train a large model. To this end, we utilize the pre-trained language model to alleviate this limitation and the experimental results demonstrate that our proposed model is effective. Nevertheless, it is worth collecting more training data from a large number of users to train a prediction model in the future.

Last, in this paper, we mainly focus on exploring the effects of trajectory patterns with the length of two (i.e., each trajectory pattern involves two CQs). We leave the trajectory patterns with the length of more than two for future work.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Always showing</td>
<td>0.494</td>
<td>0.661</td>
</tr>
<tr>
<td>Always hiding</td>
<td>0.506</td>
<td>0.000</td>
</tr>
<tr>
<td>TranShow</td>
<td>0.755</td>
<td>0.796</td>
</tr>
<tr>
<td>-w/o query</td>
<td>0.722</td>
<td>0.767</td>
</tr>
<tr>
<td>-w/o CQ pane</td>
<td>0.497</td>
<td>0.500</td>
</tr>
<tr>
<td>-w/o SERP titles</td>
<td>0.728</td>
<td>0.788</td>
</tr>
<tr>
<td>-w/o query position</td>
<td>0.742</td>
<td>0.794</td>
</tr>
<tr>
<td>-w/o hovering</td>
<td>0.517</td>
<td>0.681</td>
</tr>
</tbody>
</table>
CRediT authorship contribution statement

Jie Zou: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. Aixin Sun: Conceptualization, Formal analysis, Funding acquisition, Supervision, Writing – review & editing. Cheng Long: Conceptualization, Formal analysis, Funding acquisition, Supervision, Writing – review & editing. Mohammad Aliannejadi: Conceptualization, Formal analysis, Software, Writing – review & editing. Evangelos Kanoulas: Conceptualization, Formal analysis, Writing – review & editing.

Data availability

Data will be made available on request.

Acknowledgments

This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from Singapore Telecommunications Limited (Singtel), through Singtel Cognitive and Artificial Intelligence Lab for Enterprises (SCALE@NTU). This study is also supported by the NWO Smart Culture - Big Data/Digital Humanities (314-99-301), the NWO Innovative Research Incentives Scheme Vidi (016.Vidi.189.039), and the H2020-EU.3.4. - SOCIETAL CHALLENGES - Smart, Green, And Integrated Transport (814961). All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References


