PlanHelper: Supporting Activity Plan Construction with Answer Posts in Community-based QA Platforms

CHENGZHONG LIU, The Hong Kong University of Science and Technology, China
ZEYU HUANG, The Hong Kong University of Science and Technology, China
DINGDONG LIU, The Hong Kong University of Science and Technology, China
SHIXU ZHOU, The Hong Kong University of Science and Technology, China
ZHENHUI PENG, Sun Yat-sen University, China
XIAOJUAN MA*, The Hong Kong University of Science and Technology, China

Community-based Question Answering (CQA) platforms can provide rich experience and suggestions for people who seek to construct Activity Plans (AP), such as bodybuilding or sightseeing. However, answer posts in CQA platforms could be too unstructured and overwhelming to be easily applied to AP construction, as validated by our formative study for understanding relevant user challenges. We therefore proposed an answer-post processing pipeline, based on which we built PlanHelper, a tool assisting users in processing the CQA information and constructing AP interactively. We conducted a within-subject study (N=24) with a Quora-like interface as the baseline. Results suggested that when creating AP with PlanHelper, users were significantly more satisfied with the informational support and more engaged during the interaction. Moreover, we performed an in-depth analysis on the user behaviors with PlanHelper and summarized the design considerations for such supporting tools.

CCS Concepts: • **Human-centered computing** → **Interactive systems and tools**; **Empirical studies in HCI**.

Additional Key Words and Phrases: Activity Plan Construction, CQA Platforms, Information Digest Support

ACM Reference Format:
Chengzhong Liu, Zeyu Huang, Dingdong Liu, Shixu Zhou, Zhenhui Peng, and Xiaojuan Ma. 2022. PlanHelper: Supporting Activity Plan Construction with Answer Posts in Community-based QA Platforms. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 454 (November 2022), 26 pages. https://doi.org/10.1145/3555555

1 INTRODUCTION

Constructing activity plans is a regular yet non-trivial routine in people's daily lives. People generally refer to rich online User-Generated Content (UGC) such as websites or blogs to learn experiences from others [1, 2, 32, 56]. Among vast online resources, Community-Based Question

*Corresponding author

Authors' addresses: Chengzhong Liu, chengzhong.liu@connect.ust.hk, The Hong Kong University of Science and Technology, Hong Kong, China; Zeyu Huang, zhuangbi@connect.ust.hk, The Hong Kong University of Science and Technology, Hong Kong, China; Dingdong Liu, dliuak@connect.ust.hk, The Hong Kong University of Science and Technology, Hong Kong, China; Shixu Zhou, szhouav@connect.ust.hk, The Hong Kong University of Science and Technology, Hong Kong, China; Zhenhui Peng, pengzhh29@mail.sysu.edu.cn, Sun Yat-sen University, Zhuhai, China; Xiaojuan Ma, mxj@cse.ust.hk, The Hong Kong University of Science and Technology, Hong Kong, China.

2573-0142/2022/11-ART454 $15.00
https://doi.org/10.1145/3555555

Answering (CQA) platforms such as *Quora*[1] and *Yahoo! Answers*[2] have become a major channel for people to collect helpful suggestions from others to address their personal needs [19, 48]. These platforms have accumulated millions of posts that feature personal experience sharing, such as making a sightseeing itinerary, bodybuilding, self-learning a musical instrument, etc. [27, 40]. In this paper, we focused on the information-seeking type of questions, e.g., *"What are some tips for starting bodybuilding?"*[3], under which the answer posts are useful for building the activity plans.

Despite the rich answer posts that can help with activity planning, it is still challenging for users to distill and aggregate the relevant information from the posts composed by people with diverse experience and expertise [35]. Those posts are generally unstructured [82] and carry no guarantee to fit the individual preferences of the users [62]. To ease the information extraction challenge, common mainstream CQA platforms, e.g., *Quora*, employ methods including predicting the best answer [16, 71] and ranking the entire list of possible answers based on the contextual relevance to the question [34, 82]. These mechanisms focus primarily on selecting a subset of responses to alleviate the reading burden, but their results may contain redundant information while missing some other key points [34, 82]. Some Natural Language Processing (NLP) researchers showed that document summarization techniques could alleviate this issue. They have postulated that condensing answers to an information-seeking question into a comprehensive summary is a practical approach for users to identify the information necessary for constructing their activity plans [33, 72]. Nevertheless, these works generally aim to maximize informational coverage [70], in the process of which users' personal interests are largely overlooked [53].

Previous Human-Computer Interaction (HCI) studies have also proposed many tools to support plan building with task-related information. For example, Allen et al. [3] proposed strategies based on paper and pens to manage and organize multiple personal activities. William et al. [29] facilitated the creation of a document-like, rich-text project plan by outlining task components with references to other associated information resources. Christian et al. [62] developed an automated tool to mine solutions to questions of high specificity from contextual knowledge. However, these studies relied on either highly relevant information (e.g., daily schedule or email communication about the intended task) [20, 29] or well-structured documents (e.g., API documentation) [62] to produce a relatively fixed outcome. Therefore, such methods are not directly applicable to the activity plan construction in the CQA context. Furthermore, little is known about what practical challenges users would encounter during this process and how to design a tool to mitigate them.

To fill this gap, we conducted a Formative Study with 13 users who had prior experiences of developing activity plans with CQA platforms to understand the challenges during such a process. The findings revealed users' demand to organize the information structurally, accommodate personal needs, and maintain engagement during plan development. Based on the derived Design Requirements, we designed PlanHelper[4], an interactive system empowered by a computational pipeline of NLP techniques [19, 23, 28, 48]. Following previous works [19, 48, 66], we denoted the following taxonomy: an *aspect* is an abstract concept user may consider for their plans; a *sub-aspect* is a more concrete topic keyword that belongs to an aspect; a *proposition* is a sentence in an answer post containing key opinions or summarizing a portion of adjacent text, which shall belong to a sub-aspect. PlanHelper's pipeline then **1)** decomposes the posts and identify the propositions, **2)** clusters them and extract sub-aspect keywords, and finally **3)** groups those sub-aspects into higher-level aspects. The pipeline supports PlanHelper's functionalities to allow users to flexibly

[1]https://www.quora.com
[2]https://answers.yahoo.com/
[3]https://www.quora.com/What-are-some-tips-for-starting-bodybuilding
[4]Open-sourced at https://github.com/fhfuih/PlanHelper-CSCW2022

browse the processed answer posts, exploring (sub-)aspects of interests and recording propositions accordingly to draft activity plans.

We evaluated PlanHelper regarding users' perception of the tool, behaviors during activity plan construction, the usability and usefulness of the tool through a within-subject experiment, with a Quora-like interface as the baseline. We recruited 24 frequent CQA users for the study, each joining two separate and counterbalanced plan construction sessions of different activities where they had little prior experience. The results suggested that the participants were significantly more satisfied and more engaged when constructing activity plans with PlanHelper than with the Quora-like baseline. In addition, they also perceived PlanHelper to be significantly more useful without reduced usability compared to the conventional system. We further identified the behavior patterns that emerged in the plan construction process with PlanHelper and obtained corresponding explanations to derive critical design implications for future improvement.

The key contributions of this work are threefold:

- PlanHelper, a proof-of-concept interactive system supporting users to build activity plans;
- An NLP pipeline to process CQA answer posts and an associated hierarchical structure to organize the information digest;
- In-depth evaluation of PlanHelper and derived design considerations.

2 RELATED WORK

2.1 Activity Plans and CQA Platforms

Planning is a broad topic receiving substantial attention in the HCI community. Still, often it is grouped with another term to describe a specific type of planning behavior under a given circumstance [42], e.g., *actionable plans* [2], *project plans* [29]. In this work, we used the term *activity plan* to refer to a detailed proposal for achieving the goal of an activity that usually spanned multiple aspects, came with different alternatives, and had no fixed guidelines to follow. Thus, deciding a plausible course of actions (or even non-actions) for the intended activity (i.e., activity plan construction in the scope of this paper) generally required the collection of sufficient information concerning the activity, generation of alternatives and assessment criteria, and evaluation of choices. HCI scholars also proposed other definitions of planning. Agapie et al. defined *actionable plans* for conceptually less complex routines with clear schedules to follow [2], while our *activity plans* did not emphasize time scheduling and a fine-grained step-by-step decomposition of the actions. Some work concerned scenarios other than daily activities, Kaur et al. [32] and Rahman et al. [56] regarded an *action plan* as a step-by-step listing of specific work, e.g., writing articles. Jones et al. defined a *project plan* as a representation of decomposed information related to a project [29].

CQA platforms are an increasingly popular category of QA-oriented forums [41]. Empirical studies have suggested the richness of helpful information and professional knowledge embedded in the contents on CQA platforms [5, 40, 50, 70, 76] and the diversity of perspectives and opinions [19, 41, 48, 66]. These traits make CQA platforms an excellent source for users to gather enough information to plan their activities. Nevertheless, users still perceive various challenges reading CQA posts; for example, user-generated posts are mostly unstructured [82], the authors' expertise varies [35], and some information may not fit the particular need [62]. Hence, additional supporting tools are needed to help mitigate these challenges and improve users' efficiency condensing the information on CQA platforms to their activity plans.

2.2 Activity Planning and Supporting Systems

Existing HCI works have tried to tackle customized activity planning from various approaches. Agapie et al. avoided the high cost of inviting domain experts to tailor plans using crowdsourcing instead [1, 2]. Lund et al. proposed a less structured and more flexible system to help users plan and manage their time [42]. MixTAPE was another system that directly generates plans from the relevant information collected, which could be later adjusted manually [56]. However, we argue that it is a different process to draft activity plans from CQA contents, as users are attaining domain knowledge from others' experiences and stories and drawing their plans simultaneously. Compared with directly tailoring plans from expertise or a knowledge base, we have shifted the design focus to assisting users in digesting and filtering the valuable information from those unstructured sources generated by other common people.

2.3 Augmentation Techniques for CQA Platforms

Analyzing and extracting the information is a non-trivial process for CQA users, as the UGC in the CQA platforms varies largely in quality [4, 5, 21]. To address the hardship of locating high-quality information from CQA posts, many text-analysis works have been applied, ranging from predicting the best answer [16, 71], ranking the answers by relevance [34, 82], to summarizing threads into concise takeaways [33, 66, 72]. However, considering the uniqueness of each user's situation and needs, there can hardly be a one-fits-all "best" answer when users compile their activity plans [53]. The summarization process can also miss some points that may be valuable for some users [34, 82]. Thus, we argue that a more plausible approach should maximize the information coverage and hand over the information selection process to users themselves.

Apart from algorithmic text analysis works, researchers have explored various visual interfaces to analyze a thread from multiple angles, including the topic and central idea summarization and visualization of a single post [15], the entire thread [78], and each author [25, 51]. However, mere summarization is not enough for users to plan their activities from CQA contents due to the different individual interests and contexts. From another perspective, Hoque et al. provided visual hints of each post's relevance and usefulness and the richness of information in an entire thread [26]. But like many CQA text analytic works, it only helped filter answer posts, not supporting users' digestion of each post's content. To assist users in digesting the CQA posts and converting them to their activity plans, we referred to the works in both directions and enabled users to understand each post effectively and select the most relevant information from the thread.

2.4 UGC Information Management

Information management tools are useful for individuals to harness their knowledge by improving summarization efficiency [7, 8]. Due to the unstructured nature of UGC and the difference among individuals, users often need careful management and assembling of valuable information pieces before constructing their customized activity plans [20]. Note-taking is a popular and practical choice among the commonly mentioned means to manage information [6]. Jones et al. proposed a system supporting information management, especially for personal planning [29], proving the effectiveness of hierarchical structures in a personal planning note. Rahman et al.'s MixTAPE [56] also offered an intelligent note-taking system to cater to various stakeholders in team project planning. Some designs have gone beyond linear notes and helped users note down concept maps within a specific context [44, 69, 75]. But since our UGC source (i.e., CQA posts) differed from theirs in domain knowledge and structures, and users expected different goals of managing information (i.e., making their activity plans instead of team project plans), the information management tool should be also carefully redesigned to fit in the specific contexts [53].

Table 1. Identified challenges in using CQA answer posts as informational support

Categories	Challenges	# mentions
Extract and Organize Answer Posts	**C1**. Maintain a structured and clear mindset	13/13
	C2. Digest disorganized and diverse user posts	10/13
	C3. Keep a clear and systematic note	8/13
Accommodate Personal Need	**C4**. Select suitable candidates for plan details	10/13
	C5. Evaluate alternatives of plan details efficiently	9/13
	C6. Establish the most suitable activity plan	12/13
Maintain Engagement Level	**C7**. Ease the cog. load of digesting scattered info.	10/13
	C8. Resolve information redundancy w/o losing focus	7/13

3 FORMATIVE STUDY

3.1 Participants and Procedure

With the approval of our institution's IRB, we recruited 13 participants (7 male, 5 female, 1 prefer not to say; age range 19-30). All are frequent users of CQA platforms (5 using every day, 5 using 4-6 days a week, 3 using at least once a week) and have been using them to make customized activities plans. We conducted a semi-structured interview with each participant to learn about their experience of digesting posts on CQA platforms to assist activity planning. More specifically, we asked how frequently they referred to CQA platforms to aid the construction of their activities plans, what challenges they faced when processing relevant information from user posts, what were their current practices to compile the information to fit their individual interests, and what kind of technical support would be ideal for facilitating such practices. We also invited the participants to share anecdotal examples to help articulate their thoughts in contexts. Each interviewee spent around 30 minutes in our study and received $6 as compensation.

Parallel to the interview, two authors used the thematic analysis [10] to inductively code and theme critical challenges from the interview materials. Following the rule of saturation [63], we monitored the codebook during the process and identified data saturation after the first ten interviews. We continued the last three interviews to ensure that no new themes emerged and the formative study was comprehensive.

3.2 Findings

Most participants agreed that CQA platforms support them to make plans based on experiences and/or suggestions from other people; such information is otherwise scattered in results directly returned by the searching engines and is thus hard to distill (10/13). The participants have used CQA platforms to make arrangements for a wide range of activities, including but not limited to traveling to a foreign country, skincare, fitness training, college application, and exam preparation. For example, one participant using CQA platforms to make a skincare plan said that she needed first to analyze what factor the author of the post focused on, i.e., moisturizing or eye treatment, to see if it addressed her need. Then after this filtering, she needed to compare different suggestions from the answer posts to compile a suitable skincare plan. Based on the interview results, the inductive thematic analysis [10] concluded eight critical challenges in three categories during the activity plan construction process: organizing relevant information extracted from answers structurally, comparing possible options to accommodate personal needs, and maintaining the engagement level. Table 1 gives a holistic view of the challenges.

3.2.1 Extract and Organize Answer Posts. According to all participants, effectively digesting the user posts to extract the most relevant content is the first step of leveraging CQA platforms for compiling an activity proposal. However, it is often challenging to keep the minds clear due to an overwhelming number of aspects and alternative suggestions to consider (**C1**). For example, in a thread asking for bodybuilding tips, some answers mention specific workout methodologies while others talk about diets and other lifestyle recommendations. Among those discussing diet, some posts suggest taking supplements while others are against it. Also, most content on CQA platforms is user-generated and not well-organized [5, 72]. Thus, it is also hard to extract key information from a single post (**C2**). Many participants (8/13) take notes to organize the relevant information identified and keep a clear mindset in monitoring the plan development. Still, they found it challenging to structure their notes systematically at the beginning, before they read enough information and developed a systematic view of the topic themselves. Eventually, the notes would become messy and unclear when more contents are added (**C3**). Especially, two participants, anticipating the notes to become disorganized, were reluctant to start taking notes. The participants with note-taking behaviors also concerned about the extra workload and the context-switching issue of note-taking. Therefore, they desired a tool that could assist in organizing the notes and optimizing the note-taking workflow. Regarding how to make a clear note, another two participants explicitly mentioned that drawing mind-maps in the note is very helpful. One of them added that *"taking notes in different parts of the answers and then compiling them to mind-maps is a very efficient way to digest information."*

3.2.2 Accommodate Personal Need. Rather than abstract ideas and concepts, a detailed plan usually consists of concrete and executable instructions customized to individual contexts, interests, and needs. As noted by many participants, pointers to such instructions are often scattered across several posts on CQA platforms (10/13). They find it common that their personal interests could not be fully covered by one post, and thus they need to combine information from multiple answers to cover the aspects they care. The participants would constantly context-switch, jumping back and forth among posts with different writing styles or structures looking for all necessary information addressing their personal needs. Therefore, significant difficulty is introduced when summarizing and selecting suitable candidates of their plan details (**C4**). After gathering as many details as possible, evaluating and comparing their feasibility is another non-trivial yet frustrating process for the majority of the participants (**C5**). As every individual has particular needs and contexts, some details may not suit them well. And nearly all of them expressed the desire to have the activity plan suit themselves as much as possible (e.g., expected expense, individual capability, etc.) (**C6**).

3.2.3 Maintain Engagement Level. Most participants identified the loss of focus during the plan construction as a significant issue. Most of them (10/13) had experienced getting disoriented and mentally exhausted after trying to read and sort through numerous CQA posts. One participant explicitly mentioned that *"designing the activity plan is already a time-consuming process, and the distraction caused by irrelevant information in the posts would make it worse."* As most participants pointed out, the cognitive load was generally introduced by trying to digest the scattered information (**C7**), which is the primary threat to their engagement. In addition, some participants explicitly expressed that redundant information in the user posts was also a significant distraction during the plan construction as reading similar content was boring (**C8**).

4 DESIGN AND IMPLEMENTATION OF PLANHELPER

4.1 Design Requirements

To support the activity plan construction with the user posts on the CQA platforms, we presented our system: PlanHelper. Based on the surveyed Related Work and our Formative Study, the current setup of the mainstream CQA systems does not provide enough informational support for users to construct their activity plans [19, 48]. Therefore, to provide such support, we derived the following design requirements to facilitate the plan construction process:

DR1 **Present answer post in an organized way.** Mining valuable information from UGC, e.g., information on CQA platforms is usually very challenging due to the unstructured and unorganized nature of the data source [49]. Yet, presenting the data source in a more well-organized way can mitigate the challenge of digesting such information [62]. In Formative Study, many users called for more structured representation to deal with the unstructured posts and diverse suggestions from different users (**C2**) to maintain a clear mindset regarding the domain knowledge (**C1**) and offload the cognitive burden of reading them (**C7**). In addition, the organization of answers should also help resolve redundant information, another irritating challenge users face (**C8**).

DR2 **Record useful information structurally.** Previous work suggests that people would be unfocused when exposed to the knowledge they are unfamiliar with [44]. Taking high-quality notes could be a way to address this, as the notes, particularly those with a clear logic flow, can help people remember and digest newly learned information better [9, 30]. In the Formative Study, participants also pointed out their preference of note-taking to manage the knowledge acquired to improve the plan quality (**C3**, **C4**). Keeping track of the information furthermore helps them maintain a clear mindset and digest subsequent posts with ease (**C1**, **C7**).

DR3 **Support customized editing of notes.** People's behaviors in note-taking are very diverse based on their needs and habit [9]. The notes to oneself serve as the crucial support in making activity plans addressing personal needs [74]. Thus, enough freedom is necessary in note-editing to support users in organizing and comparing details in their preferred manners (**C4**, **C5**). Ultimately, users can customize their activity plan with ease by cherry-picking the information that suits them the best (**C6**).

DR4 **Provide holistic and summative visualization of the content.** In the Formative Study, participants have complained about the complexity of a CQA thread as an aggregation of people's different views on a topic. Such complexity leaves participants unconfident whether they have fully understood the information in the threads. Since an efficient representation of complex subjects can facilitate people's understanding process [17], a summative representation of the thread content should ease the cognitive load processing the scattered information in CQA posts (**C6**).

4.2 User Interface

Following the Design Requirements, we designed an interactive user interface (Figure 1 and 2) to augment the current CQA content reading and activity planning experience. An algorithm pipeline (Figure 3) is also developed to support the UI functionalities, which are elaborated in section 4.3. The user interface consists of three parts as described below, completing an integrated workflow from information digest to organization and then customizing activity plans.

4.2.1 The Answer Pane. The answer pane is the main area to show the question and answers. It was built on top of Quora's basic features and interface (e.g., users can expand and view each answer and the metadata) and enhanced with highlighted propositions, the aspect each post covers,

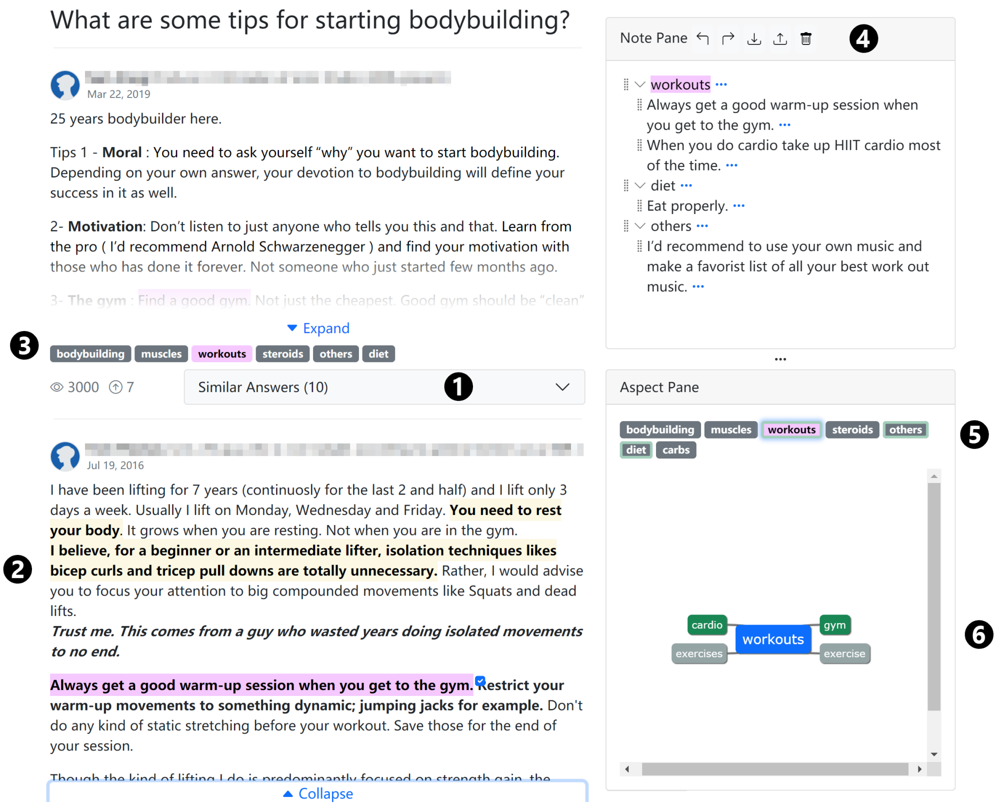
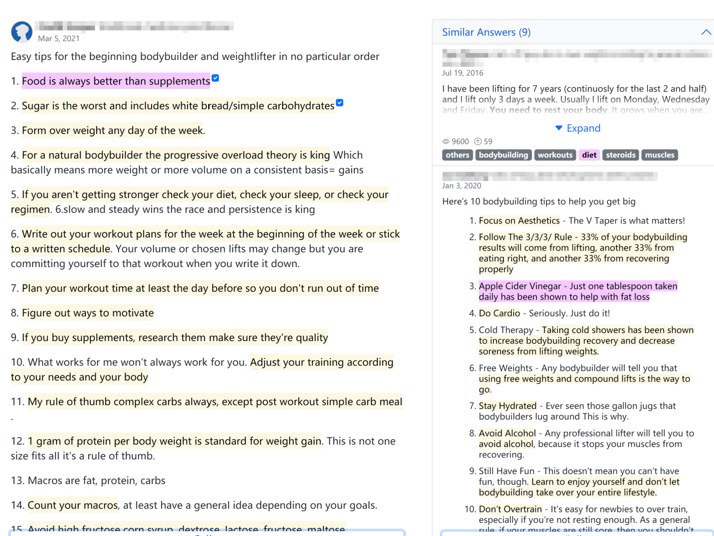


Fig. 1. The interface of PlanHelper. The explanation of each component (①-⑥) is in section 4.2.



(a) An expanded main answer (b) An expanded similar answer

Fig. 2. More interface details of the answer pane of PlanHelper

and the short answer posts similar to each better-developed post (i.e., *main answers*, see section 4.3.3 for detailed definition of *similar answers* as well) (**DR1**). Only the main answers are displayed directly in the answer pane and ranked by the new aspect each answer introduces. Each main answer cluster of similar answers is displayed in a collapsed list below and ranked by similarity, which can be expanded and viewed manually (Figure 1, ①). All other features mentioned below apply to both main answers and similar answers.

After expanding an answer, all its propositions extracted by the pipeline are highlighted (Figure 1, ②). The set of aspects covered by all the propositions in an answer is also displayed below the answer's body (Figure 1, ③). By clicking on an aspect, users can assign a customized color to this aspect, used as the background color of this aspect and the highlight color of all its propositions in all answers (**DR1**).

4.2.2 The Note Pane. This pane is where users collect and organize propositions from the answers according to their own needs (**DR2, DR3**). After expanding an answer to see the proposition in highlight, users can click on a proposition to add it to the note pane (Figure 1, ④). The note pane adopts a two-level list structure, automatically grouping propositions by their pre-computed aspects (**DR2**). Users can drag and drop to reorder the aspects as a whole, reorder propositions within an aspect, or move propositions to another aspect. Users can also collapse an aspect's proposition list, edit an aspect word or a proposition, color an aspect in the note pane, and undo/redo any previous actions (**DR3**). To quickly refer to the original text, users can hold the **ctrl** key (or the **command** key on *macOS*) and click on a proposition in the note pane to jump to the original text in the answer pane or vice versa. Finally, users can download the note pane data as a textual file for further reference and upload the data to restore the previous work.

4.2.3 The Aspect Pane. It contains an overview of all aspects covered in this thread (i.e., by all posts) and the sub-aspects associated with each aspect. It also indicates to users whether an aspect or a sub-aspect is covered in their notes to hint at their reading progress (**DR4**). On the top, it lists all aspects covered by the entire answer thread (Figure 1, ⑤). If a proposition is present in the note pane, its corresponding aspect in the aspect pane will have a green indicator. Clicking on an aspect would display a concept overview of all sub-aspects clustered to this aspect in a mind-map (Figure 1, ⑥). Similarly, a sub-aspect node is in green if at least one proposition belonging to it is currently present in the note pane. A modal window would pop up listing all propositions associated with this sub-aspect in the thread to give a clear overview (**DR4**) by further clicking on a sub-aspect node. Users can click to add them to the note pane quickly.

4.3 Pipeline Process CQA Answer Posts

4.3.1 Thread Selection and Proposition Labeling. To support PlanHelper design, we selected two Question-Answers (QA) threads from Quora to construct our database. The preparation of the database can be split into two steps:

Thread Selection. Our Formative Study discovered that traveling and bodybuilding are two everyday activities that users usually consult CQA platforms for guides and plans. Moreover, the content in these two topics involves a large amount of personal experiences and opinions. We picked two popular threads under these two topics, which contain more than 10k views, 60 followers, and 30 answers, as our contexts for Evaluation of PlanHelper. The selected QA treads are *"What do I need to know for my first stay in Paris? . . . "* [5] and *"What are some tips for starting bodybuilding?"*[6].

[5] https://www.quora.com/What-do-I-need-to-know-for-my-first-stay-in-Paris-I-will-be-there-from-June-to-August-For-example-laws-language-tips-places-customs-travel-hacks-food-night-life-dress-people-etc
[6] https://www.quora.com/What-are-some-tips-for-starting-bodybuilding

Table 2. The average ROUGE scores (in percentage) of each thread.

QA Thread	ROUGE-1	ROUGE-2	ROUGE-L
Bodybuilding[7]	73.26%	68.52%	71.40%
Sightseeing[8]	75.05%	70.92%	72.86%

The contents we crawled from these two QA threads include the questions, the answers, and the titles of "related questions" recommended by Quora. The metadata attached to each answer of the threads was also crawled, including the author's name and description, the posting date, and the number of views and upvotes. In this process, a total of 93 pieces of answers were crawled. To make the information load balanced such that the comparison was fair, we kept only 30 top-voted answers for each question. With the selected 30 answers for sightseeing and bodybuilding, the proposition extraction algorithm (see section 4.3.2 below) identified 150 and 141 propositions (word count: 6825 and 7239) respectively, indicating a roughly balanced information load of each thread.

Proposition Labeling. Two researchers first independently labeled the propositions in the selected QA threads to provide the ground truth to train the corresponding NLP model of the information processing pipeline. To evaluate the agreement level between two researchers, we adopted the ROUGE criteria [38], a common NLP practice to measure the text difference. Specifically, we concatenated the labeled propositions separately for each answer to form a paragraph and calculated the ROUGE scores between the paragraphs from two researchers. We summarized the average ROUGE of the answers in two threads in Table 2. The results showed that in two threads, the ROUGE scores reached around .7 in all common metrics and all > .7 in ROUGE-L, indicating a good level of agreement. Finally, a third researcher discussed with the two researchers who labeled the propositions to resolve the conflict, consequently finalizing the ground truth propositions. In the end, 86 pieces of answers were labeled and discussed, and 469 propositions were identified.

4.3.2 Proposition Extraction & Aspect Generation. To ensure the generalizability of the pipeline, we developed an NLP model based on ground truth propositions to automatically extract propositions from CQA answers. The propositions would be subsequently clustered into various sub-aspects and aspects. The whole NLP pipeline is illustrated in Figure 3 and elaborated below.

Preprocessing. To preprocess the answer posts for model training, we firstly adopted SymSpell[9], a state-of-the-art typo correction package to fix typos, e.g., mistyping "L" for "I". We then split long paragraphs (more than 300 words) into smaller ones to standardize the text input, reducing the computational cost and improving the performance.

Proposition Extraction. We approximated proposition extraction as a language generation task and adopted the pre-trained "sshleifer/distilbart-cnn-12-6" model [79], which is based on the state-of-art model BART [37]. It took a preprocessed paragraph as input, and it output a set of proposition sentences concatenated by a special separator token. We fine-tuned the model on our dataset with high regularization and length penalty to ensure that the output sentences used a similar vocabulary to the original text. As shown in Table 3, the model output achieved a high ROUGE-L [38] similarity score of 83.51% on the test set compared to the ground truth. Specifically, since the output was not identical to the original text's sentences, we fuzzy-matched each output sentence to the best correspondent sentence in the paragraph as the predicted proposition.

[9]https://github.com/wolfgarbe/SymSpell

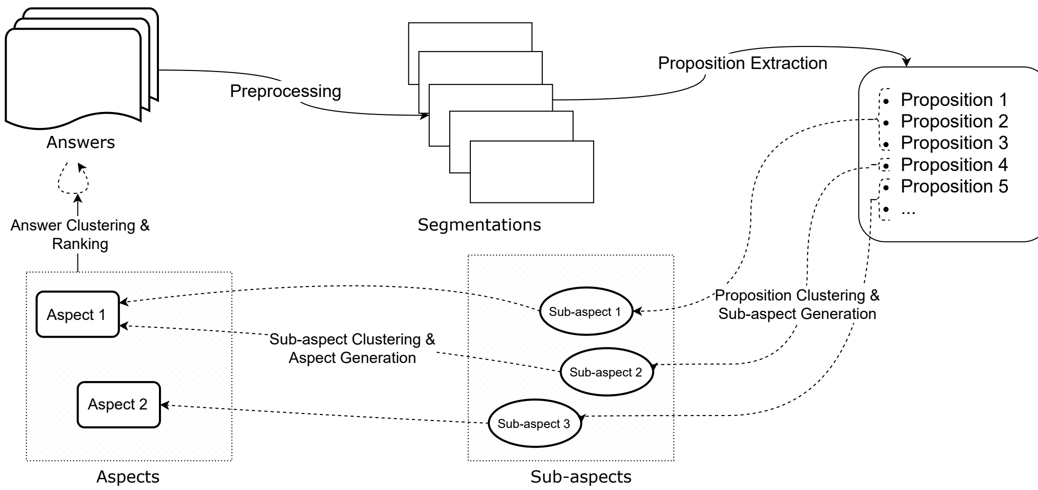


Fig. 3. Proposition extraction and re-structure pipeline of PlanHelper.

Table 3. Performance of Extraction Algorithm (ROUGE scores in percentage).

ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-L-sum
84.17%	82.31%	83.51%	83.83%

Proposition Clustering and Sub-Aspect Generation. After extracting propositions from answers, the propositions were further clustered by the topics they covered. We firstly used a pre-trained model of sentence BERT, namely "paraphrase-mpnet-base-v2" [58] to encode propositions to vectors in high dimensions. Then we used the Affinity Propagation algorithm [18] to cluster the semantic embeddings in the vector space and output around 30 clusters. Finally, we concatenated all propositions into a paragraph for each cluster and fed it into the pre-trained keyBERT model [23] to generate the keywords, which is used as the sub-aspect for this proposition cluster.

Sub-Aspect Clustering and Aspect Generation. Although sub-aspects served as concrete topic keywords to represent the propositions, users would be overwhelmed to digest and conceptualize all of them simultaneously. We therefore clustered them into more generalized conceptual keywords, a.k.a. aspects. Recent studies on knowledge base discovery, especially ConceptNet [39], provide external knowledge support to further group topics. We used ConceptNet-numberbatch[10] [67], to embed the sub-aspect words and cluster them. In each aspect cluster, all the propositions were concatenated into a paragraph and then fed into the pre-trained keyBERT model [23] to generate the most suitable aspect keyword. Since ConceptNet includes commonsense knowledge, the clusters in the ConceptNet space were larger than the sentence BERT space during the sub-aspect generation. Consequently, the aspect keywords in this step were also more abstract. However, if the distance between a sub-aspect word and its aspect word is too long in the ConceptNet embedding space to be considered as one cluster, this sub-aspect word is therefore excluded from its aspect and categorized into a special *"others"* aspect to collect the outliers.

4.3.3 Answers Clustering and Ranking. Based on their aspect and sub-aspect coverage, we clustered answers in the thread to reduce the cognitive load in reading redundant information, which is

[10] https://github.com/commonsense/conceptnet-numberbatch

common in CQA platforms [34, 82]. We further defined *main answer* and *similar answer*. The *main answers* were more comprehensive, covering more aspects, and displayed directly in the answer pane. Each main answer had a set of similar answers covering fewer and mostly overlapping aspects, which was visually collapsed in the answer pane.

Since digesting a question thread on CQA platforms is cognitively exhausting (see Formative Study), the main answers were selected to ensure that users can effectively acquire information in most aspects by reading a minimal number of answers. We used the greedy approach to determine the least number of answers covering all aspects in the given thread and denoted them as the main answers. In the two threads of our study, *bodybuilding* and *sightseeing*, each thread has six main answers. The main answers were ranked by the number of aspects they contained in descending order; if this number were identical, the answer with a higher word count would rank higher.

We considered the remaining answers as *similar answers*, and we calculated their similarity to each main answer based on the number of overlapping aspects. Each similar answer would be assigned to the most similar main answer. After the assignment, each main answer would only keep the top 10 similar answers, since *Google*, the dominant Internet search engine [73], also recommended 10 results displayed per page[11] to accommodate user reading habits. The similar answers that were not in the top 10 of the corresponding main answer were reassigned to the most similar main answers that did not reach this limit. Under each main answer, its similar answers were ranked firstly based on the similarity and then the word count.

5 EVALUATION

5.1 Hypotheses and Measurements

Previous research suggests that sufficient information support is crucial when people elaborate on the details to decide their activity plans [81]. As found in our Formative Study and other literature, enough informational support during the activity plan building process could make people more confident and thus perform better in various tasks [36, 59]. Moreover, especially in the CQA systems, users desire to get valuable content to support their construction of activity plans with less cognitive load [52]. Other studies indicate that an engaging experience would boost people's performance [57, 83]. Furthermore, considering that a competitive tool should be both functional and usable [24, 31], yet there exists a trade-off between functionality and usability [22, 31], we made the following hypotheses:

- **H1** Compared to the baseline, PlanHelper significantly improved users' satisfaction with the activity plan construction process (**H1**). This included the user satisfaction level on the informational support (*H1a*), the confidence of the coverage of the information (*H1b*), the confidence in concluding the best activity plan with the provided information (*H1c*), and the cognitive cost in constructing the activity plan (*H1d*).
- **H2** Compared to the baseline, users of PlanHelper were more engaged in designing their activity plans (**H2**). To be specific, PlanHelper improved users' concentration (*H2a*), sense of ecstasy (*H2b*), clarity (*H2c*), doability (*H2d*), sense of serenity (*H2e*), timelessness feeling (*H2f*), and intrinsic motivation (*H2g*) during the activity plan construction process.
- **H3** Compared to the baseline, PlanHelper's additional features reduced the usability (**H3**).
- **H4** Compared to the baseline, PlanHelper was significantly more useful (**H4**).

We surveyed related literature to design the questionnaire to test the aforementioned hypotheses. To test **H1**, we derived questions from various previous studies: general satisfaction level during the task (*H1a*) [31], self-reported confidence to conduct a task (*H1b & H1c*) [13], and subjective

[11] https://www.google.com/preferences

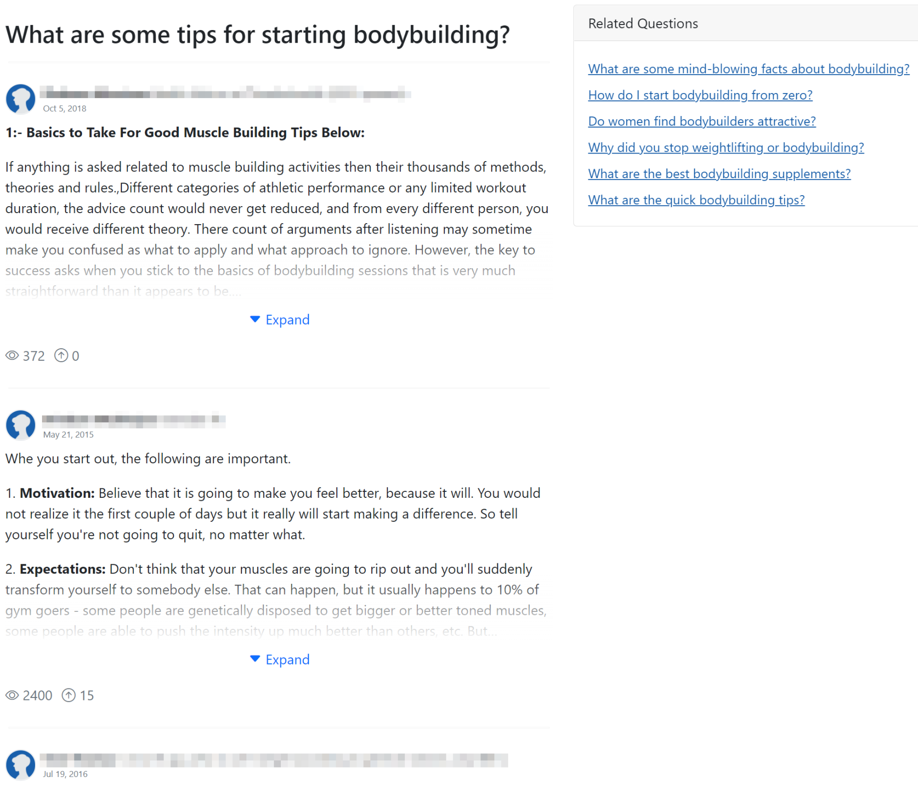


Fig. 4. The Quora-like baseline system

cognitive load of the task (*H1d*) [12]. The dimensions of **H2** are defined based on Brien's theoretical model [46] regarding the flow theory for a positive experience [14]. **H3** is tested with the standard System Usability Scale (SUS) questionnaire [11], and **H4** is tested with reference to [31].

5.2 The Quora-like CQA Baseline System

To test the hypotheses above, we conducted a within-subject study to compare PlanHelper to a baseline system with 24 users frequently using CQA platforms to build customized activity plans. Moreover, to minimize the potential influences of other factors, such as UI design style [31] and extra information in the comment section to the answers, the baseline system is implemented to emulate Quora in features but resemble PlanHelper in UI components and styles.

The Quora-like baseline system (shown in Figure 4) excluded the "similar answer" clustering and proposition highlights in the answer pane, listing all answers in Quora's original order (i.e., ranked by upvotes[12]). The note pane, the aspect pane, and all related functionalities of PlanHelper were removed. The original positions of the panes were replaced by Quora's own "related questions" list. As a result, the baseline system had similar functionality to Quora. Users could browse answers and the associated metadata (e.g., dates, upvotes, author profiles) and see the titles of related questions on the right. To control the amount of information attained, users cannot click and jump to the

[12]https://qr.ae/pGbtvE

Table 4. Demographics of all the participants. This table includes participants' ID, gender, age, their Use Patterns, and time they spent on the activity design tasks (*Similar*: Difference < 25%, see section 6.3.2).

ID	Gender	Age	Use Pattern	Time (in minutes)		
				Quora	PlanHelper	Difference
P1	Male	19	Balanced	36	39	Similar
P2	Female	21	Aspect	26	45	Increase
P3	Female	20	Note	28	60	Increase
P4	Female	21	Note	23	34	Increase
P5	Male	26	Balanced	14	10	Decrease
P6	Female	26	Balanced	14	10	Decrease
P7	Male	24	Aspect	13	21	Increase
P8	Female	23	Note	15	16	Similar
P9	Male	26	Aspect	8	14	Increase
P10	Female	21	Aspect	8	20	Increase
P11	Male	20	Aspect	19	22	Similar
P12	Female	21	Balanced	7	12	Increase
P13	Male	22	Note	22	13	Decrease
P14	Male	22	Aspect	22	24	Similar
P15	Male	20	Note	24	53	Increase
P16	Male	23	Balanced	9	34	Increase
P17	Male	22	Balanced	16	8	Decrease
P18	Female	22	Balanced	15	20	Increase
P19	Female	21	Note	21	18	Similar
P20	Female	21	Note	15	10	Decrease
P21	Female	21	Balanced	5	9	Increase
P22	Male	20	Balanced	12	11	Similar
P23	Female	21	Balanced	13	14	Similar
P24	Female	21	Balanced	12	10	Similar

related questions in the baseline system. Also, comments to the answers were excluded to ensure consistency with PlanHelper.

5.3 Participants and Procedure

With the approval of our institution's IRB, we recruited 24 frequent CQA users (13 female, 11 male; age range 19-26, M = 21.8, SD = 1.9; CQA usage frequency: 9 daily, 11 4-6 days a week, 4 at least once a week; details summarized in Table 4) through online advertising, social media, and word-of-mouth at a local university. The inclusion criteria were that participants are not experienced in the given context but interested in exploring it and making an activity plan about it. The user study lasted 60-90 minutes, and each participant received $15 for completing the tasks and interviews.

During the experiment, each participant was asked to complete two activity plan design tasks with a roughly balanced informational load, as mentioned in section 4.3.1. In each task, they were prompted with the context information that they answered in the screening questionnaire. To be specific, the prompts are:

- You are going to go sightseeing in Paris. How would you plan with the provided posts?
- You are going to make a bodybuilding plan. How would you plan with the provided posts?

After briefing the context, the task, and PlanHelper's features (if using PlanHelper), participants were asked to start the interaction with the system and present an activity plan whenever they felt ready. They were not allowed to use any information outside the system, except searching for the meaning of a specific term. However, they can use any word processor to take notes to support the activity plan construction process. Once they finished the plans, they were asked to present the design to the interviewer. Participants would consequently fill a feedback questionnaire after presenting the activity plan coherently for each task. The two tasks were completed using PlanHelper and the baseline system separately. We formed four combinations to counterbalance the experiment using Latin Square: **1)** Sightseeing (baseline) - Bodybuilding (PlanHelper), **2)** Bodybuilding (baseline) - Sightseeing (PlanHelper), **3)** Bodybuilding (PlanHelper) - Sightseeing (baseline), and **4)** Sightseeing (PlanHelper) - Bodybuilding (baseline).

At the end of the study, an interview was conducted to collect their comments to complete the quantitative results. Specifically, we asked about their overall impression of PlanHelper, comments on the specific features and the explanations of their behaviors when interacting with the systems.

6 RESULTS

After collecting the participants' ratings on the satisfaction level, engagement level, usability, and usefulness of the interaction with PlanHelper and the baseline system respectively during the activity plan construction process, we interviewed the participants about how their plans were constructed with the given tools. All tests were measured in a 7-point Likert scale, with 1 for the most negative impression (e.g., not useful at all) and 7 for the most positive (e.g., very useful). We performed Wilcoxon signed-rank test [80] to assess the difference in the participants' ratings regarding various factors of the two systems. The test affirmed that the quantitative results did not suffer from the context difference and the ordering. Table 5 summarizes the statistical results of the hypotheses proposed in section 5.1. We also conducted a post-hoc power analysis on the experiment results. Assuming normal distributions, $\alpha = 0.05$, and one-tail tests, all hypotheses with significant results achieved a power level of > 0.9. Hence, the sample size was adequate.

6.1 Perception of the PlanHelper System

6.1.1 User Satisfaction. Compared to the baseline, participants were significantly more satisfied with the informational support from PlanHelper ($W = 13.00, p = 8.40 \times 10^{-5}$); *H1a* is accepted. Participants were also significantly more confident that they had covered as much useful information as possible in designing their activity plans ($W = 13.50, p = 2.70 \times 10^{-5}$); *H1b* is accepted. Moreover, they were also significantly more confident that they had constructed better activity plans with PlanHelper than the Quora-like baseline system ($W = 9.00, p = 3.71 \times 10^{-4}$); *H1c* is accepted. Finally, regarding the cognitive cost, participants reported they were significantly less tired when using PlanHelper ($W = 13.50, p = 5.25 \times 10^{-6}$) and perceived significantly lower cognitive load during the design of activity plans ($W = 4.00, p = 3.07 \times 10^{-5}$); *H1d* is accepted. So far, **H1** is fully accepted.

Overall, nearly all participants (19/24) explicitly expressed that with PlanHelper, the activity plan construction process was more efficient, and they were able to produce more detailed and concrete activity plans. Three participants said that PlanHelper gave them a starting point and the big picture of the activity plan to be constructed (P2, P4, P12). Another two participants noted that PlanHelper helped filter the useless and redundant information (P1, P2). Further analyses of individual features of PlanHelper are addressed in section 6.2.3.

6.1.2 User Engagement. As shown in Table 5, participants' engagement level was significantly improved in all dimensions; thus, **H2** is fully accepted. Five participants (P4, P5, P7, P11, P24) highlighted that the PlanHelper helped them concentrate of the activity plan construction task

Table 5. The statistical user feedback with Baseline and PlanHelper, where the p-values (-: p > .100, +: .050 < p < .100, *:p < .050, **:p < .010, ***:p < .001) is reported.

Category	Factor	Baseline Mean/S.D.	PlanHelper Mean/S.D.	Statistics W	Statistics p	Statistics Sig.	Hypotheses
Satisfaction	Information Support	4.42/1.19	5.88/0.73	13.00	8.40E-05	***	H1a acc.
	Information Coverage	3.79/1.41	5.17/1.03	13.50	2.70E-04	***	H1b acc.
	Confidence	3.58/1.32	4.92/1.04	9.00	3.71E-04	***	H1c acc.
	Tiredness	3.04/1.27	5.38/1.11	13.50	5.25E-06	***	H1d-1 acc.
	Cognitive Load	3.33/0.99	5.46/1.00	4.00	3.07E-05	***	H1d-2 acc.
Engagement	Concentration	4.29/1.17	6.08/0.49	0.00	3.75E-05	***	H2a acc.
	Sense of Ecstasy	2.71/1.17	5.92/0.86	0.00	1.10E-05	***	H2b acc.
	Clarity	4.17/1.31	5.71/0.68	0.00	1.72E-04	***	H2c acc.
	Doability	4.13/1.39	5.63/1.11	14.00	8.97E-05	***	H2d acc.
	Sense of Serenity	3.21/1.22	5.33/1.21	9.50	1.01E-04	***	H2e acc.
	Timelessness Feeling	3.92/1.35	5.50/1.00	4.00	1.70E-04	***	H2f acc.
	Intrinsic Motivation	4.04/1.40	5.54/0.82	15.00	3.34E-04	***	H2g acc.
Usability		63.40/15.73	69.86/10.02	96.50	2.07E-01	-	H3 rej.
Usefulness		3.29/1.43	5.83/0.69	4.50	2.10E-05	***	H4 acc.

by presenting information in a well-structured way. P7 especially gave credit to the aspect pane, which *"helped me to easily find information that addressed my personal interest, thus making me concentrated and engaged."* P5 even reported that *"I felt completely immersed in the process, and I did not even notice that I had recorded so much useful information related to my activity plan."*

6.2 System Usability and Features of PlanHelper

6.2.1 System Usability. PlanHelper achieved an SUS rating of $Mean = 69.86$ ($SD = 10.02$), compared to the baseline system's rating of $Mean = 63.40$ ($SD = 15.73$). In addition, the Wilcoxon signed-rank test indicated no significant difference between the two systems ($W = 96.50, p = 2.07 \times 10^{-1}$), which means the additional features of PlanHelper did not reduce the usability compared to the baseline system. As such, **H3** is rejected.

6.2.2 Usefulness of the Features. Finally, ratings on the overall usefulness demonstrated that PlanHelper was significantly more useful than the baseline system ($W = 4.50, p = 2.10 \times 10^{-5}$), which means **H4** is accepted. During the interview, users reported that reading the information with the baseline system was *"boring and tiring"* (P4, P12-14, P16, P19, P20, P24), although the UI of the baseline system was *"neat, clean and easy to use"* (P11, P13, P18, P24).

We further asked participants to evaluate the usefulness of the features of PlanHelper, which is summarized in Table 6. To grow a deeper understanding of how participants used the features to construct their activity plans, we further analyzed the use pattern of PlanHelper in section 6.3.1 from our observations and participants' interviews.

6.2.3 Additional Comments on the Features. This section summarized how PlanHelper features helped participants construct their activity plans, according to the relevant user comments.

Aspect Driven Categorization. Overall, participants acknowledged that the aspect-driven approach assisted them a lot in digesting CQA answers. Three participants reported that the provided aspects were helpful in *"suggesting what they should consider when making activity plans"* (P4, P6, P10). Moreover, many participants regarded it as an efficient and novel way to digest information (P2, P4-6, P9-11, P14-16, P20, P21, P24); some participants even abandoned the way

Table 6. User ratings on PlanHelper features (*Algorithm* stands for features developed based on our proposed *Pipeline*, while *UI* and *UX* stand for *User Interface* and *User Experience* based features).

Type	Feature	Mean	S.D.
Algorithm	Answer ranking	5.13	1.45
	Similar answers clustering	5.17	1.40
	Proposition clustering with aspects	5.88	0.80
UI	Proposition highlighting	5.71	1.23
	Aspect coloring	5.71	1.27
	Concept overview in aspect pane	5.54	1.10
UX	Editing note pane	5.50	1.29
	Show original	5.75	1.15
	Sub-aspect listing	5.29	1.40
	Undo and redo	4.92	1.44
	Download and upload	5.25	1.36

that they did in the baseline system, e.g., using mainly the aspect pane without reading original answers carefully in PlanHelper (P2, P14). In addition, P20 adopted a unique reading strategy: she read the propositions in answers aspect-by-aspect with the help of the coloring feature.

List of Propositions of Sub-aspects. Some participants reported that viewing propositions by sub-aspects had helped them find missed-out information. Based on participants' reading habits, such a feature either enabled them to discover information in uncovered aspects/sub-aspects (P2, P6) or complete relevant information of already-noted aspects/sub-aspects (P12, P16).

Color Propositions by Aspects. Many participants found that coloring propositions and corresponding aspects effectively facilitated their reading (P2, P5, P11, P12, P16, P19, P20, P22, P23). Among them, three participants (P19, P20, P23) reported that colored propositions helped them find relevant information faster because they read the colored texts instead of the complete answers, and *"the reading burden is reduced"* (P19). Furthermore, P19 reported that various colors in the answer posts had made the system *"vivid"* and *"fun to use"*, making her less bored and more engaged using PlanHelper than the baseline system.

6.3 Behavior During Activity Plan Construction

To facilitate the derivation of the Design Considerations, we observed participants' behaviors without interruption during their interaction. We recorded participants' use patterns of PlanHelper, their preferences of the system's features, the time users spent constructing their activity plans, and any other particular habits worth noting. After the participants finished constructing the plans, we also asked them to review their use patterns and why they behaved so. We discovered three different use patterns of the participants' interaction with PlanHelper (see section 6.3.1) and concluded three factors that affected the construction time of the activity plans (see section 6.3.2). Nevertheless, we did not observe any correlation between activity plan construction time and other recorded factors such as use patterns during the process.

6.3.1 Use Patterns.

Balanced Attention. Around a half of the participants (11/24) put nearly equal attention to the note pane and the aspect pane of PlanHelper during the activity plan construction. Typical user behavior was to **1)** read the first several answers ranked on the top, then **2)** add the propositions they were interested in, and finally **3)** go over the aspect pane to add more propositions to their personal need (P6, P12, P16, P17, P22). During the process, some used aspects coloring to help them quickly go over the answers that ranked low.

Besides this general behavior, three participants (P1, P5, P21) balanced their attention because they were initially highly interested in the aspect pane but subsequently realized that they were unlikely to construct a complete activity plan primarily by it. Hence, they later shifted attention to the note pane and tried to add some propositions from the answer posts. One of them reported that *"I'm quite disappointed by the accuracy of the relationships among aspects, sub-aspects, and corresponding propositions. Therefore, I had to give up my initial intention of heavily using aspect pane."* (P1). Another two participants (P18, P23) used the aspect pane as a mere checking tool to ensure all useful information had been recorded in the note pane. Although they put similar attentions on both panes, they only added a few propositions with the aspect pane. *"I have gone over nearly all the answers, but I still wanted to make sure I covered all the information I needed."* (P18).

Preferring Note Pane. About one-fourth of the participants (7/24) preferred to use the note pane rather than the aspect pane, and they spent nearly all their time interacting with the note pane. Nevertheless, we did not observe a general pattern for this behavior. Some reported that this was purely a user habit issue: e.g., *"It was simply my habit that I did not want to use the aspect pane."* (P8). P19's habit included manually adjusting each proposition's aspect with her own words right upon adding it to the note pane. She did so even before she viewed the aspect pane or the aspect overview below each answer, making the provided aspects useless.

Similar to participants with **Balanced Attention.**, those who mainly used the note pane might have a high expectation towards the accuracy of the aspect pane: *"The aspect pane is not accurate and thus untrustworthy"* (P13); *"algorithm behind this pane really confused me"* (P15).

Another reason could be that the participant had a different set of *"aspects"* in their mind. P15 also pointed out that he was very demanding of the accuracy of the aspects because otherwise, it would be pretty hard to convince himself to give up the *"aspects in my own mind"*.

Preferring Aspect Pane. Another one-fourth of the participants (6/24) preferred to use the aspect pane over the note pane of PlanHelper. The majority of the participants (P2, P10, P11, P14) in this style would firstly **1)** read one or two answers ranking at the top to get familiar with the context and then **2)** use the aspect pane to find the propositions they wanted. Contrary to some participants who doubted the accuracy of the aspect pane, participants in this style generally trusted the algorithm's accuracy. Therefore, they added most of the propositions from the aspect pane and used the note pane only as an index to help them find the original location of the propositions in the answers. *"I believed that in the aspect pane, all key information had been extracted"*, said P2. Another supporting reason behind this pattern was to save some efforts extracting information. P14 even said, *"I was just too lazy . . . I would indeed suffer from confusion caused by . . . (imperfect accuracy and loss of the context) . . . , but this can be mitigated by first adding them to the note pane and then jumping to their original location (to double-check)."*

6.3.2 Time of Activity Plan Construction. We measured participants' time to construct activity plans using each system without interrupting or urging them. About half of the participants (11/24) spent more than 25% of their time constructing activity plans with PlanHelper than the Quora-like baseline system; another one-fourth (5/24) of the participants spent less than 25% time with PlanHelper than the baseline system; the last one-third (8/24) spent similar time on both systems. We interviewed the users at the end of the study and concluded three subjective factors that affected the time spent in activity plan construction with the two systems: **1)** The organized information on PlanHelper helped some participants digest information faster, thus *reducing* the time in plan construction with PlanHelper. Some participants referred to the structure and categorization of the aspect pane to organize their notes (P6, P17). **2)** The support provided by PlanHelper encouraged some participants to digest more, thus *increasing* the time in plan construction with PlanHelper. With the help of aspects and highlights, some participants spent extra time to ensure they did not miss anything (P3, P4, P16, P18). *"I felt like people have a tendency to process high-quality information that is well-structured. PlanHelper provided such an excellent pipeline in organizing the information"*, reported P4. **3)** In the *baseline* system, less organized information made people less interested or patient to read, thus *reducing* the time in plan construction. Two participants (P4, P16) felt less motivated to read carefully, missing much detailed information in their notes. Another two (P12, P18) felt less attentive or even reluctant to read the less-voted posts.

7 DISCUSSION

7.1 Design Considerations

In section 6.3, we conducted an in-depth analysis of notable user behavior related to the current design of PlanHelper. In this subsection, we derived several design considerations (**DC**'s) from the analysis results for future works on supporting activity planning with answer posts in CQA platforms, or more broadly, with online UGC.

*7.1.1 Ease the Reading Burdens (**DC1**).* According to our Formative Study, users need to consume sufficient answer posts to develop a satisfactory plan with high confidence. In the baseline condition, three participants reported that they gave up reading many less-voted posts because they felt overwhelmed by unrelated or redundant information. In contrast, the PlanHelper system successfully mitigated this problem by extracting and highlighting the core contents, a method proven to facilitate material processing and understanding [47, 61]. To make information distillation from an extensive collection of documents more efficient and less tiresome, PlanHelper allows participants to locate and understand key points of each answer with the highlighted propositions, which more than half of the participants (13/24) found useful. PlanHelper also provides multiple ways to assure users what they have covered (e.g., by marking propositions or checking the aspect pane), making them more motivated to read [60]. In brief, to support activity plan construction, the system should process large-volume UGC well to ease the reading burdens of the users.

*7.1.2 Support Information Harvesting with Note-Taking (**DC2**).* Harvesting useful information is not a trivial task for people when they need to construct activity plans [19, 29, 52, 66, 82]. Moreover, users' preferences vary; e.g., in our study, some participants preferred practical information (P10, P14), while others preferred personal experience (P17, P20). Such findings impose higher demands on supporting users' note-taking, which is a practical way for people to offload the cognitive pressure in harvesting information [43, 61]. Following the design suggestions in [29], PlanHelper supported effective information harvesting via note-taking by allowing users to add notes quickly and reorganize them via drag-and-drop. With these features, our participants generally acknowledged that PlanHelper had helped them harvest information more effectively. Especially, P7 & P20 seldom took notes during the activity plan construction before but were educated to do so during the study. They later found it particularly useful.

*7.1.3 Avoid "Over-Guidance" by the System (**DC3**).* It is worth noting that two participants thought that PlanHelper tends to make them ignore the original answers and focus on the highlighted propositions only. Another participant also expressed the concern that during the activity plan construction with PlanHelper, users' mindsets could be constrained by the given aspects. Such feedback demonstrated the possibility that users may be *"over-guided"* by PlanHelper to some extent. Therefore, for future designs of the relevant tools, one should consider what kind of AI assistance is appropriate to support information digest. Specifically, as suggested by P24, one possible solution is to modularize different forms of information summarization support, e.g., to enable proposition highlights but disable aspect extraction.

In summary, **DC1 & DC2** reflect the core expected functionalities desired by users, and **DC3** implies the necessity of carefully calibrating issues in the design of Human-AI collaboration.

7.2 Generalizability of PlanHelper

7.2.1 Accommodation of More Diverse UGC Posts. As CQA platforms often contain redundant or similar question threads, PlanHelper can extend its answer posts input to the "related questions" threads recommended by the CQA systems themselves [77]. Moreover, in the future, we would experiment PlanHelper's generalizability on UGC outside CQA platforms such as forum posts or blogs. Some participants (P13, P17, P18, P19) mentioned their interests in these platforms as a complement to CQA content when they plan their activities. Yet we acknowledged the potential challenges in processing them with our proposed pipeline, as UGC may vary drastically in lengths and language styles compared to those in CQA platforms.

7.2.2 Compatibility of Other Contexts to Use PlanHelper. The two representative activities we evaluated in the user study shall demonstrate the effectiveness of PlanHelper for activity plans, as long as relevant posts contain enough executable suggestions and clear aspects that the pipeline can summarize. The plan construction process can be conceptualized as a non-programmed decision-making [65], and PlanHelper aims to assist users in the intelligence (i.e., information collection and synthesis), design (i.e., generation of alternatives), and choice (i.e., evaluation and selection of an alternative) stages. Therefore, PlanHelper can be generalized to support other types of decision-making tasks based on the UGC. Taking *TOEFL*[13] test preparation mentioned by P6: it could involve aspects such as anxiety, booklet, practice, vocabulary, etc. [54]. The potential sub-aspects of an aspect, i.e., anxiety, might include the doubt of continuation, time management, anxiety during tests, etc., which are common for test preparation [45].

In addition to activity plans, PlanHelper might also be helpful to organize posts related to questions that asked for opinions. One participant (P21) specifically mentioned that PlanHelper might help her synthesize opinions on specific social issues, e.g., *whether mandatory vaccination is morally justified.* She suggested that the relevant aspects could be public health, personal freedom, and economic impact. Questions in these domains generally attract answers with arguments from various viewpoints, which many CQA readers may find inspiring. In this case, users can pin the insightful arguments among the answer posts listed in the "answer pane" to the "note pane" while enjoying an overview of the viewpoints in the "aspect pane". Although the inherent difference in language choices and features may require additional tuning on the model, it also suggests the potential compatibility of the text processing pipeline in other contexts from a conceptual view.

7.3 Contributions to CQA

The thriving of the mainstream CQA platforms is based on the crowdsourced intelligence of the community members. With our proposed text processing pipeline and the associated "proposition — sub-aspect — aspect" data structure, PlanHelper has demonstrated its capabilities in significantly improving readers' experience in constructing activity plans. Apart from the readers, the platform stakeholders and the writers may also benefit from this data structure. For CQA platforms, the answer post ranking mechanisms could utilize our proposed structure to support more effective information digestion. With the extracted propositions and summarized sub-aspects & aspects, the platforms could recommend less redundant information to prevent users from getting bored.

[13] https://www.ets.org/toefl/

For the writers on CQA platforms, such a structure is also useful for providing a comprehensive overview of the question thread before they write. In this way, they can focus more on the rarely addressed aspects and refrain from writing repetitive content.

8 LIMITATIONS AND FUTURE WORK

In this section, we examined the limitations of our study and summarized the future work based on the study results and the following discussion.

8.1 Study Settings

In our Formative Study and User Study, the participants are mainly young adults (age range 19-30) and such demographic distributions may introduce some bias. More studies on diverse user groups could be applied to improve the accessibility and inclusiveness of the PlanHelper and even the common settings of mainstream CQA platforms. Furthermore, the selection of activities (and subsequently question threads) may be another source of bias. Although the selection was based on Formative Study results and considered representative, we would like to also experiment with our system on other activities and CQA questions for future work. It may further evaluate the generalizability of our CQA answer post processing pipeline and the system.

8.2 Human Effort Engaged in Threads Coding

Despite the high automation in the proposed pipeline, human effort was still involved in the proposition labeling process [37, 79]. Such human efforts were inevitable as to our best knowledge, no existing CQA corpora could fulfill the requirements. While we acknowledged the cold-start problem of PlanHelper, we addressed the scalability issue in the following two directions.

For the activity with pre-labeled propositions, our pipeline is capable to be scaled to a large number of the relevant CQA posts with no further proposition labeling, as the model (section 4.3.2) is already trained to identify the proposition structures. In case of encountering outlier propositions, human-in-the-loop tuning can be adopted to refine the model to enhance the performance.

We also foresaw PlanHelper's potential scalability to a larger range of activities. The training dataset can be derived from existing corpora or crowdsourcing methods. Transfer learning [55] may be also experimented to adapt the pre-trained model of one activity to another.

8.3 Future Work

Apart from extending the user study to broader coverage of question threads & user groups (section 8.1) and minimizing the reliance on manual coding (section 8.2), we would like to present more insightful directions for future work.

8.3.1 Generalizing the Pipeline. One main contribution of our work is the "proposition — sub-aspect — aspect" structure from the characteristics of non-factoid CQA [66] threads and the pipeline to mine such structure. Thus, it would be interesting to see how it can be applied in other logically interconnected but loosely structured UGC beyond CQA posts. Apart from the suggestions for planning daily activities, future work can also look into how the structure can be applied to other contexts such as opinion sharing mentioned in section 7.2.2.

8.3.2 Mining Multimodal Content. Since UGC allows a large degree of freedom in modality [27, 66], future work can investigate the incorporation of multimodal analysis in supporting the information extraction and organization on CQA platforms. Specifically, in the future, PlanHelper could be designed to cover the comments sections, multimedia (e.g., images), and different formats set by the author (e.g., **bold** & *italic*) on CQA platforms. Such information often tends to complete the context or help the author convey and emphasize ideas [82]. We have identified several algorithmic attempts [27, 64, 68], but fewer on related HCI venues.

9 CONCLUSION

In this paper, we presented PlanHelper, a proof-of-concept supporting tool to help users of CQA platforms to construct their activity plans. Based on the Formative Study results, we proposed an NLP pipeline to process the unstructured answer posts on the CQA platforms, and an interactive interface on top. A within-subject user study showed that PlanHelper significantly improved users' satisfaction level and engagement during the activity plan construction process compared to the Quora-like interface. We further concluded use patterns of PlanHelper during the activity plan construction process to grow a deeper understanding of how participants interacted with PlanHelper. We summarized design considerations based on our user study, providing insights for the future work to design and build activity plan supporting tools like PlanHelper.

ACKNOWLEDGMENTS

Many thanks to the anonymous reviewers for their insightful suggestions. This work is supported by the Hong Kong General Research Fund (GRF) with grant No. 16204420.

REFERENCES

[1] Elena Agapie, Bonnie Chinh, Laura R. Pina, Diana Oviedo, Molly C. Welsh, Gary Hsieh, and Sean Munson. 2018. *Crowdsourcing Exercise Plans Aligned with Expert Guidelines and Everyday Constraints*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173898

[2] Elena Agapie, Lucas Colusso, Sean A. Munson, and Gary Hsieh. 2016. PlanSourcing: Generating Behavior Change Plans with Friends and Crowds. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 119–133. https://doi.org/10.1145/2818048.2819943

[3] David Allen. 2015. *Getting things done : the art of stress-free productivity*. Penguin Books, New York, New York.

[4] Ricardo Baeza-Yates. 2009. User Generated Content: How Good is It?. In *Proceedings of the 3rd Workshop on Information Credibility on the Web* (Madrid, Spain) *(WICOW '09)*. Association for Computing Machinery, New York, NY, USA, 1–2. https://doi.org/10.1145/1526993.1526995

[5] Antoaneta Baltadzhieva and Grzegorz Chrupała. 2015. Question Quality in Community Question Answering Forums: A Survey. *SIGKDD Explor. Newsl.* 17, 1 (Sept. 2015), 8–13. https://doi.org/10.1145/2830544.2830547

[6] Aaron Bauer and Kenneth Koedinger. 2006. *Evaluating the Effect of Technology on Note-Taking and Learning*. Association for Computing Machinery, New York, NY, USA, 520–525. https://doi.org/10.1145/1125451.1125563

[7] Ofer Bergman, Richard Boardman, Jacek Gwizdka, and William Jones. 2004. Personal Information Management. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems* (Vienna, Austria) *(CHI EA '04)*. Association for Computing Machinery, New York, NY, USA, 1598–1599. https://doi.org/10.1145/985921.986164

[8] Michael S. Bernstein, Max Van Kleek, m. c. schraefel, and David R. Karger. 2007. *Management of Personal Information Scraps*. Association for Computing Machinery, New York, NY, USA, 2285–2290. https://doi.org/10.1145/1240866.1240995

[9] Peter Brandl, Christoph Richter, and Michael Haller. 2010. *NiCEBook: Supporting Natural Note Taking*. Association for Computing Machinery, New York, NY, USA, 599–608. https://doi.org/10.1145/1753326.1753417

[10] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa arXiv:https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa

[11] John Brooke. 1996. SUS: A quick and dirty usability scale. *Usability evaluation in industry* 189 (1996), 4–7.

[12] Yujia Cao, Mariët Theune, and Anton Nijholt. 2009. Modality Effects on Cognitive Load and Performance in High-Load Information Presentation. In *Proceedings of the 14th International Conference on Intelligent User Interfaces* (Sanibel Island, Florida, USA) *(IUI '09)*. Association for Computing Machinery, New York, NY, USA, 335–344. https://doi.org/10.1145/1502650.1502697

[13] Benjamin Collier and Julia Bear. 2012. Conflict, Criticism, or Confidence: An Empirical Examination of the Gender Gap in Wikipedia Contributions. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (Seattle, Washington, USA) *(CSCW '12)*. Association for Computing Machinery, New York, NY, USA, 383–392. https://doi.org/10.1145/2145204.2145265

[14] Mihaly Csikzentmihaly. 1990. *Flow: The psychology of optimal experience*. Vol. 1990. Harper & Row, New York, USA.

[15] Kushal Dave, Martin Wattenberg, and Michael Muller. 2004. Flash Forums and ForumReader: Navigating a New Kind of Large-Scale Online Discussion. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work* (Chicago, Illinois, USA) *(CSCW '04)*. Association for Computing Machinery, New York, NY, USA, 232–241. https://doi.org/10.1145/1031607.1031644

[16] Hualei Dong, Jian Wang, Hongfei Lin, Bo Xu, and Zhihao Yang. 2015. Predicting Best Answerers for New Questions: An Approach Leveraging Distributed Representations of Words in Community Question Answering. In *2015 Ninth International Conference on Frontier of Computer Science and Technology*. 13–18. https://doi.org/10.1109/FCST.2015.56

[17] Haakon Faste and Honray Lin. 2012. *The Untapped Promise of Digital Mind Maps*. Association for Computing Machinery, New York, NY, USA, 1017–1026. https://doi.org/10.1145/2207676.2208548

[18] Brendan J Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *science* 315, 5814 (2007), 972–976.

[19] Shogo Fujita, Tomohide Shibata, and Manabu Okumura. 2020. Diverse and Non-redundant Answer Set Extraction on Community QA based on DPPs. In *Proceedings of the 28th International Conference on Computational Linguistics*. 5309–5320.

[20] Leonardo Galicia. 2007. *Supporting Proactive Planning of Multiple Activities*. Association for Computing Machinery, New York, NY, USA, 1649–1652. https://doi.org/10.1145/1240866.1240874

[21] Arpita Ghosh and Preston McAfee. 2011. Incentivizing High-Quality User-Generated Content. In *Proceedings of the 20th International Conference on World Wide Web* (Hyderabad, India) *(WWW '11)*. Association for Computing Machinery, New York, NY, USA, 137–146. https://doi.org/10.1145/1963405.1963428

[22] Nancy C. Goodwin. 1987. Functionality and Usability. *Commun. ACM* 30, 3 (March 1987), 229–233. https://doi.org/10.1145/214748.214758

[23] Maarten Grootendorst. 2020. KeyBERT: Minimal keyword extraction with BERT. https://doi.org/10.5281/zenodo.4461265

[24] Tom Hewett, Mary Czerwinski, Michael Terry, Jay Nunamaker, Linda Candy, Bill Kules, and Elisabeth Sylvan. 2005. Creativity support tool evaluation methods and metrics. *Creativity Support Tools* (2005), 10–24.

[25] Enamul Hoque and Giuseppe Carenini. 2015. ConVisIT: Interactive Topic Modeling for Exploring Asynchronous Online Conversations. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (Atlanta, Georgia, USA) *(IUI '15)*. Association for Computing Machinery, New York, NY, USA, 169–180. https://doi.org/10.1145/2678025.2701370

[26] Enamul Hoque, Shafiq Joty, Luis Marquez, and Giuseppe Carenini. 2017. CQAVis: Visual Text Analytics for Community Question Answering. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces* (Limassol, Cyprus) *(IUI '17)*. Association for Computing Machinery, New York, NY, USA, 161–172. https://doi.org/10.1145/3025171.3025210

[27] Jun Hu, Shengsheng Qian, Quan Fang, and Changsheng Xu. 2018. Attentive interactive convolutional matching for community question answering in social multimedia. In *Proceedings of the 26th ACM international conference on Multimedia*. 456–464.

[28] Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A Challenge Dataset and Effective Models for Aspect-Based Sentiment Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 6280–6285. https://doi.org/10.18653/v1/D19-1654

[29] William Jones, Predrag Klasnja, Andrea Civan, and Michael L. Adcock. 2008. The Personal Project Planner: Planning to Organize Personal Information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 681–684. https://doi.org/10.1145/1357054.1357162

[30] Vaiva Kalnikaitundefined and Steve Whittaker. 2008. Cueing Digital Memory: How and Why Do Digital Notes Help Us Remember?. In *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction - Volume 1* (Liverpool, United Kingdom) *(BCS-HCI '08)*. BCS Learning & Development Ltd., Swindon, GBR, 153–161.

[31] Youwen Kang, Zhida Sun, Sitong Wang, Zeyu Huang, Ziming Wu, and Xiaojuan Ma. 2021. MetaMap: Supporting Visual Metaphor Ideation through Multi-Dimensional Example-Based Exploration. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 427, 15 pages. https://doi.org/10.1145/3411764.3445325

[32] Harmanpreet Kaur, Alex C. Williams, Anne Loomis Thompson, Walter S. Lasecki, Shamsi T. Iqbal, and Jaime Teevan. 2018. Creating Better Action Plans for Writing Tasks via Vocabulary-Based Planning. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 86 (nov 2018), 22 pages. https://doi.org/10.1145/3274355

[33] Mostafa Keikha, Jae Hyun Park, and W Bruce Croft. 2014. Evaluating answer passages using summarization measures. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 963–966.

[34] Md Tahmid Rahman Laskar, Jimmy Xiangji Huang, and Enamul Hoque. 2020. Contextualized Embeddings based Transformer Encoder for Sentence Similarity Modeling in Answer Selection Task. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 5505–5514. https://www.aclweb.org/anthology/2020.lrec-1.676

[35] Long T. Le, Chirag Shah, and Erik Choi. 2016. Evaluating the Quality of Educational Answers in Community Question-Answering. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries* (Newark, New Jersey, USA) *(JCDL '16)*. Association for Computing Machinery, New York, NY, USA, 129–138. https://doi.org/10.1145/2910896.2910900

[36] Pierre Le Bras, David A. Robb, Thomas S. Methven, Stefano Padilla, and Mike J. Chantler. 2018. *Improving User Confidence in Concept Maps: Exploring Data Driven Explanations*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173978

[37] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

[38] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[39] Hugo Liu and Push Singh. 2004. ConceptNet—a practical commonsense reasoning tool-kit. *BT technology journal* 22, 4 (2004), 211–226.

[40] Qiaoling Liu, Tomasz Jurczyk, Jinho Choi, and Eugene Agichtein. 2015. Real-Time Community Question Answering: Exploring Content Recommendation and User Notification Strategies. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (Atlanta, Georgia, USA) *(IUI '15)*. Association for Computing Machinery, New York, NY, USA, 50–61. https://doi.org/10.1145/2678025.2701392

[41] Yuanjie Liu, Shasha Li, Yunbo Cao, Chin-Yew Lin, Dingyi Han, and Yong Yu. 2008. Understanding and Summarizing Answers in Community-Based Question Answering Services. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1* (Manchester, United Kingdom) *(COLING '08)*. Association for Computational Linguistics, USA, 497–504.

[42] John R. Lund and Jason Wiese. 2021. *Less is More: Exploring Support for Time Management Planning*. Association for Computing Machinery, New York, NY, USA, 392–405. https://doi.org/10.1145/3461778.3462133

[43] Tamas Makany, Jonathan Kemp, and Itiel E Dror. 2009. Optimising the use of note-taking as an external cognitive aid for increasing learning. *British Journal of Educational Technology* 40, 4 (2009), 619–635.

[44] Xiaojun Meng, Shengdong Zhao, and Darren Edge. 2016. HyNote: Integrated Concept Mapping and Notetaking. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (Bari, Italy) *(AVI '16)*. Association for Computing Machinery, New York, NY, USA, 236–239. https://doi.org/10.1145/2909132.2909277

[45] Silke Neuderth, Burkhard Jabs, and Armin Schmidtke. 2009. Strategies for reducing test anxiety and optimizing exam preparation in German university students: a prevention-oriented pilot project of the University of Würzburg. *Journal of Neural Transmission* 116, 6 (2009), 785–790.

[46] Heather O'Brien. 2016. *Theoretical Perspectives on User Engagement*. Springer International Publishing, Cham, 1–26. https://doi.org/10.1007/978-3-319-27446-1_1

[47] JongHwan Oh, SungJin Nam, and Joonhwan Lee. 2014. Generating Highlights Automatically from Text-Reading Behaviors on Mobile Devices. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI EA '14)*. Association for Computing Machinery, New York, NY, USA, 2317–2322. https://doi.org/10.1145/2559206.2581176

[48] Adi Omari, David Carmel, Oleg Rokhlenko, and Idan Szpektor. 2016. Novelty based ranking of human answers for community questions. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 215–224.

[49] Ina O'Murchu, John G Breslin, and Stefan Decker. 2004. Online Social and Business Networking Communities.. In *ECAI Workshop on Application of Semantic Web Technologies to Web Communities*, Vol. 107.

[50] Aditya Pal, F. Maxwell Harper, and Joseph A. Konstan. 2012. Exploring Question Selection Bias to Identify Experts and Potential Experts in Community Question Answering. *ACM Trans. Inf. Syst.* 30, 2, Article 10 (May 2012), 28 pages. https://doi.org/10.1145/2180868.2180872

[51] Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. *Supporting Comment Moderators in Identifying High Quality Online News Comments*. Association for Computing Machinery, New York, NY, USA, 1114–1125. https://doi.org/10.1145/2858036.2858389

[52] Maria Soledad Pera and Yiu-Kai Ng. 2011. A Community Question-Answering Refinement System. In *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia* (Eindhoven, The Netherlands) *(HT '11)*. Association for

Computing Machinery, New York, NY, USA, 251–260. https://doi.org/10.1145/1995966.1995999

[53] Rodney Peters. 2001. Exploring the Design Space for Personal Information Management Tools. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems* (Seattle, Washington) *(CHI EA '01)*. Association for Computing Machinery, New York, NY, USA, 413–414. https://doi.org/10.1145/634067.634309

[54] Donald E Powers. 1988. PREPARING FOR THE SAT®: A Survey of Programs and Resources. *ETS Research Report Series* 1988, 2 (1988), i–31.

[55] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019).

[56] Sajjadur Rahman, Pao Siangliulue, and Adam Marcus. 2020. MixTAPE: Mixed-Initiative Team Action Plan Creation Through Semi-Structured Notes, Automatic Task Generation, and Task Classification. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 169 (oct 2020), 26 pages. https://doi.org/10.1145/3415240

[57] Anna Reid and Ian Solomonides. 2007. Design students' experience of engagement and creativity. *Art, Design & Communication in Higher Education* 6, 1 (2007), 27–39.

[58] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[59] Soo Young Rieh, Erik Choi, Katie Sagan, and Jennifer Colby. 2017. Beyond Questioning and Answering: Teens' Learning Experiences and Benefits of Social Q&A Services. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17 Companion)*. Association for Computing Machinery, New York, NY, USA, 295–298. https://doi.org/10.1145/3022198.3026350

[60] Hitinui Robert, Koichi Kise, and Olivier Augereau. 2017. Real-Time Wordometer Demonstration Using Commercial EoG Glasses. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers* (Maui, Hawaii) *(UbiComp '17)*. Association for Computing Machinery, New York, NY, USA, 277–280. https://doi.org/10.1145/3123024.3123183

[61] Nirmal Roy, Manuel Valle Torre, Ujwal Gadiraju, David Maxwell, and Claudia Hauff. 2021. Note the Highlight: Incorporating Active Reading Tools in a Search as Learning Environment. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (Canberra ACT, Australia) *(CHIIR '21)*. Association for Computing Machinery, New York, NY, USA, 229–238. https://doi.org/10.1145/3406522.3446025

[62] Christian Severin Sauer and Thomas Roth-Berghofer. 2012. Solution Mining for Specific Contextualised Problems: Towards an Approach for Experience Mining. In *Proceedings of the 21st International Conference on World Wide Web* (Lyon, France) *(WWW '12 Companion)*. Association for Computing Machinery, New York, NY, USA, 729–738. https://doi.org/10.1145/2187980.2188193

[63] Benjamin Saunders, Julius Sim, Tom Kingstone, Shula Baker, Jackie Waterfield, Bernadette Bartlam, Heather Burroughs, and Clare Jinks. [n.d.]. Saturation in Qualitative Research: Exploring Its Conceptualization and Operationalization. 52, 4 ([n. d.]), 1893–1907. https://doi.org/10.1007/s11135-017-0574-8 arXiv:29937585

[64] Rajiv Ratn Shah. 2016. Multimodal Analysis of User-Generated Content in Support of Social Media Applications. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval* (New York, New York, USA) *(ICMR '16)*. Association for Computing Machinery, New York, NY, USA, 423–426. https://doi.org/10.1145/2911996.2912032

[65] Herbert A Simon. 1966. Theories of decision-making in economics and behavioural science. In *Surveys of economic theory*. Springer, 1–28.

[66] Hongya Song, Zhaochun Ren, Shangsong Liang, Piji Li, Jun Ma, and Maarten de Rijke. 2017. Summarizing Answers in Non-Factoid Community Question-Answering. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (Cambridge, United Kingdom) *(WSDM '17)*. Association for Computing Machinery, New York, NY, USA, 405–414. https://doi.org/10.1145/3018661.3018704

[67] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. 4444–4451. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972

[68] Avikalp Srivastava, Hsin-Wen Liu, and Sumio Fujita. 2019. Adapting Visual Question Answering Models for Enhancing Multimodal Community Q&A Platforms. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (Beijing, China) *(CIKM '19)*. Association for Computing Machinery, New York, NY, USA, 1421–1430. https://doi.org/10.1145/3357384.3358000

[69] Chien-Lin Tang, Jingxian Liao, Hao-Chuan Wang, Ching-Ying Sung, and Wen-Chieh Lin. 2021. ConceptGuide: Supporting Online Video Learning with Concept Map-Based Recommendation of Learning Path. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 2757–2768. https://doi.org/10.1145/3442381.3449808

[70] Yla R. Tausczik, Aniket Kittur, and Robert E. Kraut. 2014. Collaborative Problem Solving: A Study of MathOverflow. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Baltimore,

Maryland, USA) *(CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 355–367. https://doi.org/10.1145/2531602.2531690

[71] Qiongjie Tian, Peng Zhang, and Baoxin Li. 2013. Towards predicting the best answers in community-based question-answering services. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 7.

[72] Mattia Tomasoni and Minlie Huang. 2010. Metadata-Aware Measures for Answer Summarization in Community Question Answering. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Uppsala, Sweden, 760–769. https://www.aclweb.org/anthology/P10-1078

[73] Daniel Trielli and Nicholas Diakopoulos. 2019. *Search as News Curator: The Role of Google in Shaping Attention to News Information.* Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300683

[74] Max G. Van Kleek, Michael Bernstein, Katrina Panovich, Gregory G. Vargas, David R. Karger, and MC Schraefel. 2009. *Note to Self: Examining Personal Information Keeping in a Lightweight Note-Taking Tool.* Association for Computing Machinery, New York, NY, USA, 1477–1480. https://doi.org/10.1145/1518701.1518924

[75] Jorge J. Villalon and Rafael A. Calvo. 2008. Concept Map Mining: A Definition and a Framework for Its Evaluation. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 03 (WI-IAT '08)*. IEEE Computer Society, USA, 357–360. https://doi.org/10.1109/WIIAT.2008.387

[76] Gang Wang, Konark Gill, Manish Mohanlal, Haitao Zheng, and Ben Y. Zhao. 2013. Wisdom in the Social Crowd: An Analysis of Quora. In *Proceedings of the 22nd International Conference on World Wide Web* (Rio de Janeiro, Brazil) *(WWW '13)*. Association for Computing Machinery, New York, NY, USA, 1341–1352. https://doi.org/10.1145/2488388.2488506

[77] Kai Wang, Zhao-Yan Ming, Xia Hu, and Tat-Seng Chua. 2010. Segmentation of Multi-Sentence Questions: Towards Effective Question Retrieval in CQA Services. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Geneva, Switzerland) *(SIGIR '10)*. Association for Computing Machinery, New York, NY, USA, 387–394. https://doi.org/10.1145/1835449.1835515

[78] Furu Wei, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle X. Zhou, Weihong Qian, Lei Shi, Li Tan, and Qiang Zhang. 2010. TIARA: A Visual Exploratory Text Analytic System. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Washington, DC, USA) *(KDD '10)*. Association for Computing Machinery, New York, NY, USA, 153–162. https://doi.org/10.1145/1835804.1835827

[79] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[80] Robert F Woolson. 2007. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials* (2007), 1–3.

[81] Zhiwei Zeng, Xiuyi Fan, Chunyan Miao, Cyril Leung, Chin Jing Jih, and Ong Yew Soon. 2018. Context-Based and Explainable Decision Making with Argumentation. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (Stockholm, Sweden) *(AAMAS '18)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1114–1122.

[82] Xiaodong Zhang, Sujian Li, Lei Sha, and Houfeng Wang. 2017. Attentive Interactive Neural Networks for Answer Selection in Community Question Answering. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (San Francisco, California, USA) *(AAAI'17)*. AAAI Press, 3525–3531.

[83] Mengxiao Zhu, Yoav Bergner, Yan Zhang, Ryan Baker, Yuan Wang, and Luc Paquette. 2016. Longitudinal Engagement, Performance, and Social Connectivity: A MOOC Case Study Using Exponential Random Graph Models. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (Edinburgh, United Kingdom) *(LAK '16)*. Association for Computing Machinery, New York, NY, USA, 223–230. https://doi.org/10.1145/2883851.2883934

Received January 2022; revised April 2022; accepted May 2022