USING GAZE TRACKING TO TACKLE DUPLICATE QUESTIONS ON COMMUNITY BASED QUESTION ANSWERING WEBSITES: A CASE STUDY OF IFIXIT

A Thesis

presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Computer Science

by

Pankti Gandhi

June 2018
COMMITTEE MEMBERSHIP

TITLE: Using Gaze Tracking to Tackle Duplicate Questions on Community based Question Answering Websites: A Case Study of iFixit

AUTHOR: Pankti Gandhi

DATE SUBMITTED: June 2018

COMMITTEE CHAIR: Franz Kurfess, Ph.D.
Profession of Computer Science

COMMITTEE MEMBER: Sandrine Fischer, Ph.D.
Human-Computer Interaction Researcher

COMMITTEE MEMBER: Clark Turner, Ph.D.
Professor of Computer Science

COMMITTEE MEMBER: Sara Bahrami, Ph.D.
Assistant Professor of Computer Science
ABSTRACT
Using Gaze Tracking to Tackle Duplicate Questions on Community based Question Answering Websites: A Case Study of iFixit

Pankti Gandhi

The number of unanswered questions on Community based Question Answering (CQA) websites has increased significantly due to the rising number of duplicate questions. This is a serious problem, one that could lead to the decline of such beneficial websites.

This thesis presents novel avenues that use gaze tracking technology and behavioral testing to tackle this problem. Based on prior studies on web search behaviors, we assumed that adding contextual information (snippets) to proposed related questions displayed on the ‘Ask a Question’ page of the CQA website iFixit would improve the asker experience and reduce their tendency to post a new duplicate question. The first lab experiment where this web page was redesigned and compared to the original one was conducted on 8 participants. Results confirmed that participants were more likely to find an answer to their question on the redesigned page. A second experiment, conducted remotely and on a larger sample of 74 participants, aimed to discover strategic attributes that increase the perceived similarity of question pairs. These attributes were used in the third lab experiment (20 participants) to redesign and assess the snippets from Experiment 1. Results indicated that snippets containing ‘symptom(s)’ and ‘cause(s)’ attributes constitute an incremental improvement over basic snippets: they are perceived as slightly more relevant and require significantly less gaze fixations on the asker’s part.
ACKNOWLEDGMENTS

Thanks to my family and friends for all their encouragement and support.
TABLE OF CONTENTS

LIST OF FIGURES ......................................................... viii

CHAPTER

1 Introduction ......................................................... 1
  1.1 Motivation ......................................................... 2
  1.2 Research Questions ............................................... 4
  1.3 Contributions ..................................................... 4
  1.4 Thesis Structure .................................................. 5

2 Background and Related Work .................................... 6
  2.1 Community Based Question Answering Websites .................. 6
    2.1.1 The Decline of CQA Websites ............................... 7
    2.1.2 Measures to Reduce Duplicate Content .................... 9
  2.2 Context in Information Retrieval Systems ...................... 10
  2.3 Gaze Tracking ................................................... 11
    2.3.1 Prior Gaze Tracking Studies ................................. 12

3 Experiments ......................................................... 15
  3.1 Experiment 1 ..................................................... 15
    3.1.1 Motivation .................................................... 15
    3.1.2 Materials ...................................................... 17
    3.1.3 Procedure .................................................... 20
    3.1.4 Results ....................................................... 22
    3.1.5 Implications/Takeaways ..................................... 27
  3.2 Experiment 2 ..................................................... 27
    3.2.1 Motivation .................................................... 27
    3.2.2 Materials ...................................................... 28
    3.2.3 Procedure .................................................... 31
    3.2.4 Results ....................................................... 32
    3.2.5 Implications/Takeaways ..................................... 34
  3.3 Experiment 3 ..................................................... 35
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>The evolution of answering success for new questions posted each month (Generated from dataset provided by iFixit)</td>
</tr>
<tr>
<td>2.1</td>
<td>Overview of CQA websites</td>
</tr>
<tr>
<td>2.2</td>
<td>Title dissimilarity leads to duplication</td>
</tr>
<tr>
<td>2.3</td>
<td>Description snippet of a Google search result</td>
</tr>
<tr>
<td>2.4</td>
<td>Mean time to complete search task for each task type, broken down by snippet length</td>
</tr>
<tr>
<td>2.5</td>
<td>Mean number of search results fixated for each task type, broken down by snippet length</td>
</tr>
<tr>
<td>3.1</td>
<td>Search results on iFixit website</td>
</tr>
<tr>
<td>3.2</td>
<td>Answers forum on iFixit website</td>
</tr>
<tr>
<td>3.3</td>
<td>The original design of the proposed related questions (control_1)</td>
</tr>
<tr>
<td>3.4</td>
<td>The redesigned proposed related questions (top_text_1)</td>
</tr>
<tr>
<td>3.5</td>
<td>Experiment Design</td>
</tr>
<tr>
<td>3.6</td>
<td>Number of questions found relevant by users</td>
</tr>
<tr>
<td>3.7</td>
<td>Positive judgement for Satisfaction and Findability</td>
</tr>
<tr>
<td>3.8</td>
<td>Click likeliness per question position</td>
</tr>
<tr>
<td>3.9</td>
<td>Heat Map showing users’ gaze behavior for control_1</td>
</tr>
<tr>
<td>3.10</td>
<td>Heat Map showing users’ gaze behavior for top_text_1</td>
</tr>
<tr>
<td>3.11</td>
<td>Question pairs countered-balanced following a Latin Square design</td>
</tr>
<tr>
<td>3.12</td>
<td>A screenshot of the web application</td>
</tr>
<tr>
<td>3.13</td>
<td>Question Similarity</td>
</tr>
<tr>
<td>3.14</td>
<td>Reaction Times (RT) as a function of Judgment and Question Similarity</td>
</tr>
<tr>
<td>3.15</td>
<td>3DS top_text</td>
</tr>
<tr>
<td>3.16</td>
<td>3DS sympt_cause</td>
</tr>
<tr>
<td>3.17</td>
<td>Experiment 3 design</td>
</tr>
<tr>
<td>3.18</td>
<td>Number of questions found relevant by users</td>
</tr>
</tbody>
</table>
3.19 Fixation Count ......................................................... 42
3.20 Heatmap for sympt_cause ............................................ 43
3.21 Heatmap for top_text .................................................. 43
3.22 Click likeliness per question position .............................. 44

4.1 One of the few instructions on the iFixit website for asking better
    questions [2] ................................................................. 46
4.2 Tooltip analyzing the quality of question asked .................... 47
The web has changed the way people search and share information. Search engines have facilitated the process of obtaining information to the point where just typing a few descriptive words on the home page finds and displays a list of matching websites ranked by relevance. However, these search results may not provide an exact solution to the users problem and it may be time-consuming to review all the results, especially if there is no guarantee of finding a suitable answer. Community based Question Answering (CQA) websites aim to alleviate this issue by offering users an opportunity to rapidly and efficiently obtain the desired knowledge.

CQA is an Internet-based crowd-sourcing service that enables users to post their questions and obtain the answers from other users later [16]. A large number of these questions cannot be answered by standard information retrieval tools easily. The users in such a forum can be divided in three groups: 1) users who only ask questions, 2) users who only answer questions and 3) users who ask and answer questions [12]. The asker posting a question lacks knowledge of a specific topic and searches for an expert on the same topic to provide the desired knowledge. In this way, an asker is querying a topic and the experts providing the knowledge about this topic are the source of information, thus replacing other sources like documents or databases. There are broadly two types of CQA websites; general CQA systems like Quora [8] and Yahoo! Answers [11] as well as domain specific CQAs like StackOverflow [10] and iFixit [6]. iFixit is a wiki-based site that teaches people how to fix almost anything. It provides a space where users can ask any question as long as it is related to making devices last longer. Answers to these questions are collaboratively managed by users on the website.
1.1 Motivation

(a) The number of new questions.

(b) The relative proportion of questions according to how they were answered.

Figure 1.1: The evolution of answering success for new questions posted each month (Generated from dataset provided by iFixit)
According to the dataset provided by iFixit, the number of questions asked each year increased from 11228 in 2013 to 36341 in 2017 with 2016 being the year with the highest number of questions asked (36816) (see Figure 1.1a). In 2014, the number of unanswered questions surpassed the number of questions that have an accepted answer. The proportion of these unanswered questions among all new questions, rapidly increased (see Figure 1.1b). In 2014, it was only 7.34 percent, whereas in 2016, it was 31.56 percent.

Despite efforts to prevent asking questions that have already been answered, the site contains duplicate questions. We found that although exact duplicates are very rare, many duplicate questions exist that are asked in different ways. In order to combat duplicate questions, most CQA websites suggest proposed related questions when a user tries to post a question. This is to ensure that the asker has already checked preexisting questions before posting a new one. These suggestions typically contain the question title and number of answers. However, question titles may often not accurately describe what the problem is. For example, the title of a question could be ‘My phone does not turn on’ for which the description of the question states that the reason for this was because it was dropped in water. Possible answers for this question would revolve around drying the phone. For another question with the title ‘My iPhone screen shows blue lines’ the description could also be that the reason for this was because it was dropped in water. Again, possible solutions would involve best tactics to dry out the phone. Thus, more often than not, the title of a question does not effectively convey enough information about the question. In such cases, if only the title of the question is suggested, an asker may not click on it as they might feel that it is not relevant. Due to this, our thesis aims to explore if adding additional contextual information to the question title in proposed related questions could aid users in deciding whether a suggested question might contain the answer that they are looking for. This could likely reduce the number of duplicate questions asked on
the website.

1.2 Research Questions

In this research, we aim to answer three questions.

1. Does adding a snippet to a question title in proposed related questions enhance the asker’s experience in terms of perceived relevance and satisfaction of such questions when recommended?

2. Can certain strategic attributes increase the perceived similarity of a question pair?

3. Does adding such strategic attributes to snippets improve the perceived relevance of related questions and likeliness to click on them?

1.3 Contributions

The main contributions of this thesis are summarized below:

- Ascertained that adding a snippet to the question title has a positive impact on the number of questions found relevant by askers, the number of questions they would click on, their satisfaction with the proposed set of questions and their likeliness to find an answer.

- Discovered that a question pair containing both cause(s) and symptom(s) increases an asker’s propensity to identify them as duplicates and in most cases, it also takes them the least amount of time to make that decision.

- Established that askers find related questions slightly more relevant with lesser fixation in stimuli where the snippets contain symptom(s) and/or causes(s)
over stimuli whose snippets contain the first two to three lines of the question description.

1.4 Thesis Structure

The rest of the thesis is organized as follows:

In Chapter 2, prior related work and essential background knowledge is presented

Chapter 3 discusses the three experiments we conducted. This chapter is divided into 3 sections with one experiment being described in each section. All three sections comprise of 5 subsections: Motivation, Materials, Procedure, Results and Implications/ Takeaways

Chapter 4 is the conclusive chapter where we summarize the work done in this thesis and possibility of further extensions
Chapter 2

BACKGROUND AND RELATED WORK

In this chapter, prior related work done is reviewed in order to point out the many contributions of previous researchers and to place the contributions of this thesis in the proper context. The related work for this effort spans three topics - Community based Question Answering websites, the use of context in information retrieval systems and gaze tracking. In each case, we first provide some essential background information, followed by a discussion of previous research related to our own. We conclude by summarizing how our research builds on this previous work.

2.1 Community Based Question Answering Websites

Community based Question Answering (CQA) services can be described as dedicated platforms for users to answer other users questions, resulting in a community where users share and interactively give ratings to questions and answers [14]. These websites enable users to ask questions that are subjective, open-ended as well as those that could benefit from expert opinions.

Answerers are rewarded for the quality and quantity of their answers. These rewards are generally in the form of points and reputation on the website. A social network based on real identities motivates users to improve their reputation as it can be viewed by any other user visiting their profile. A key to the success of CQA services is the quality and timeliness of the answers that users get. From the perspective of askers and answerers, askers care about the quality and timeliness of the answers to their questions, and answerers care about the quality of the questions they view and the effort to find appropriate questions to answer. A majority of the questions posted
on such websites are answered fast enough to nearly obviate the need to perform a conventional web search, as evidenced by an ever-increasing quantity of duplicate questions. This is a serious problem, one that could lead to the decline of such beneficial websites.

Figure 2.1: Overview of CQA websites

2.1.1 The Decline of CQA Websites

Srba and Bielikova [30] conducted a case study of Stack Overflow which is considered as one of the most popular CQA sites. Their research confirmed the community perception that the quality of the website is deteriorating due to large amount of low quality content created by ‘undesired groups of users’. These users include the help vampires, the noobs and the reputation collectors. Help vampires are those who are only interested in getting answers to their questions. They don’t return the help they’ve received back to the community. Noobs are low expertise users who overload the system with a lot of low quality content thereby making it difficult for experts to
find unique and interesting questions. Reputation collectors try to answer as many questions as possible to increase their site reputations. On one hand, they help experts by answering the uninteresting questions. On the other hand, they mutually reinforce help vampires and noobs and motivate them to ask more low-quality questions. Users often don’t try to search for an answer before posting questions because they know that reputation collectors will provide them with a tailor-made answer in a very short amount of time. Their data indicates that the proportion of deleted or unanswered questions rose from 22 percent in 2011 to 40 percent in 2014 and the proportion of active users in the same period decreased from 15 percent to 5 percent.

Ahasanuzzaman et al. [13] aimed to determine why duplicate questions are asked by users. They wanted to know if this was an accidental mistake, if the existing answers are not helpful and whether users search for the answer before asking a question. They sampled 600 duplicate questions in total comprising of 100 questions from each year between 2009 and 2014 inclusive to avoid any bias. These questions were manually analyzed and compared with their master questions. They discovered that one of the reasons for duplicate questions is that the title of the duplicate question does not match the master question despite having similar content. Thus, many users asked the same question again, as the title of the previous question did not indicate a clear relationship between the two questions (see Figure 2.2). This raised the question of whether users who view the related questions feel the same way. As only the titles of questions are displayed, is it possible that they are ignored because the titles do not aptly convey the description? Would displaying some contextual information along with the related question make users want to find out more about them?
2.1.2 Measures to Reduce Duplicate Content

Quality is one of the most important aspects of most CQA sites. However, duplicate questions make it hard to keep quality high. A writer will write a great answer to a question once, but if there are multiple versions of that same question, writing the same answer multiple times becomes a tedious task. Besides this, duplicate questions make site maintenance harder, waste resources that could have been used to answer other questions, and cause users to unnecessarily wait for answers that are
already available [31]. To tackle this, CQA websites automatically suggest potential
duplicates before a question is added, and if the user goes ahead and posts a duplicate
question, other users can redirect duplicate questions to consolidate them [15]. On the
iFixit website, three moderator votes are required to merge a question as duplicate.

Tools exist for automating the process of detecting duplicate questions. For ex-
example, DupPredictor and Dupe [31] [13] use a discriminative model classifier together
with BM25 scoring function. This works by feeding different features for each pair
of preprocessed questions into the binary classification model in order to determine
if the pair of questions are duplicates or not. These models can also tell the level of
duplication, which is a probabilistic value that can be utilized to deduce how likely
the new question is a duplicate of another question. This value is subsequently used
to present a ranked list of results to a user.

Nevertheless, it would be preferable to prevent duplicates from occurring. This
can be achieved by properly identifying when a new question is similar to existing
ones and providing this information to the asker in a way that encourages them to
take the time to examine existing similar questions instead of adding a new question.

2.2 Context in Information Retrieval Systems

Most modern search engines such as Google provide a snippet for each search result in
the search engine results page (SERP). A snippet is a short summary of the content
of a website and is generated based on the search term (see Figure 2.3).

Prior work in question-answering suggests that additional textual context to an
answer provides a substantial boost to user preference and performance. Lin et al. [24]
conducted a user study to explore the question of how much text a question answering
system should return to the user. They tested four different interface conditions that
varied in length and discovered that overall, users preferred a paragraph-sized chunk of
Figure 2.3: Description snippet of a Google search result

text over just an exact phrase as the answer to their questions. They found that when users research a topic, increasing the amount of text returned to users significantly decreases the number of queries that they pose to the system, suggesting that users utilize supporting text to answer related questions.

The success of snippets in SERPs motivates us to investigate the possibility of adding snippets to recommended questions in CQA websites like iFixit. Our hypothesis is that providing relevant contextual information with the question title could decrease the tendency to post duplicate questions.

2.3 Gaze Tracking

Gaze tracking is a sensor technology that enables a device to know exactly where a user’s eyes are focused. Gaze-tracking methodologies have been extensively used in the domain of Web search because gaze can be used as a proxy for a user’s attention. While many techniques rely on the explicit actions of users (e.g., mouse clicks, query streams or diary reports), gaze tracking can yield much more detailed moment-by-
moment observations about how users interact with information [14]. Gaze-tracking and studying variations in gaze patterns emerged as a field of research in the late 19th century. Since then, gaze-tracking has evolved, and has now become a widely-used technique to analyze how humans interact with computers. Eye tracking data tracks the motion of the eyes and yields two types of events. Fixations occur when a person focuses their vision on a point, while changes in eye position are referred to as saccades [27]. Following standard practice, analysis is focused on where and when fixations occur. Fixations are extracted using velocity threshold identification (I-VT, [23]) as implemented in SMI BeGaze [9]. These fixations are used to generate heat maps. Heat Maps show fixation hits related to the color scale between blue (less hits) and red (most hits) (for an example of a heatmap see Figure 3.9). Heat maps are useful because they use color to communicate relationships between data values that would be much harder to understand if presented numerically in a spreadsheet.

2.3.1 Prior Gaze Tracking Studies

Goldberg et al. [17] were among the pioneers of investigating the gaze patterns of users when browsing different types of web-pages. Salojrvi et al. [29], Granka et al. [18] and Joachims et al. [21] were among the first to suggest that eye-tracking can be used to capture and understand user attention, and therefore infer relevance from behavioral observations. Guan and Cutrell [19] showed that people tend to look at the top-ranked results only, and when they do not find their target they either click on the first result or reformulate their query. Apart from search results, gaze tracking has also been applied to examine other components of result. For instance, Hofmann et al. [20] conduct an in-depth user study of user interactions with Query Auto Completion in web search.

Cutrell and Guan [14] presented a study using eye tracking techniques to inves-
tigate how varying the amount of information in web search results affected user performance on two task types: navigational and informational. Informational tasks required the user to find specific information that could be found in one or more places. As navigational tasks are not relevant to this thesis, we will limit our discussion of results to informational tasks. All web search queries were submitted to a special server for MSN Search. This enabled them to control the snippet length of each search result. They presented search results with three different snippet lengths: short, medium or long. In their manipulations, short snippets usually contained a single line of words, medium snippets about two to three lines, and long snippets typically six to seven lines of words. They conducted an experiment on 18 participants such that each of 12 search tasks (6 different tasks of each type) was counterbalanced across participants to ensure that every task was seen with every snippet length.

To investigate the effect of task type and snippet length on how people search, they performed a 2 (Task Type) x 3 (Snippet Length) x 2 (Repetition) repeated measures multivariate analysis of variance (RM MANOVA).

As illustrated in Figure 2.4, informational tasks showed an improvement of 24 seconds in the amount of time taken to complete a task with long snippets. This result supports the notion that more information in the snippet may help searchers determine whether a given site is likely to have the information they are interested in.

Gaze measures provided by eye tracking showed that searchers looked at one third fewer results with long snippets than when they were given short snippets (see Figure 2.5). Maximum number of search results were viewed when accompanied by a medium length snippet.
Figure 2.4: Mean time to complete search task for each task type, broken down by snippet length

Figure 2.5: Mean number of search results fixated for each task type, broken down by snippet length.

To sum up, prior work in adding additional contextual information to Question Answering systems suggests that when users research a topic, increasing the amount of text returned to users significantly decreases the number of queries that they pose to the system, suggesting that users utilize supporting text to answer related questions. Cutrell and Guan [14] used gaze tracking technology to verify that adding information to the contextual snippet in search results significantly improved performance for informational tasks.

This thesis builds upon the prior work discussed here by testing the effect of adding a two to three line snippet along with the question title in the proposed related questions. If we can ascertain that adding additional contextual information improves user satisfaction with proposed related questions, the next step will be to determine if snippets with different types of contextual information make a difference to the number of questions found relevant by the user as well as which questions they would click on.
Chapter 3

EXPERIMENTS

3.1 Experiment 1

3.1.1 Motivation

This experiment was conducted to answer the first research question, i.e., to examine if adding a snippet to a question title enhances the asker’s experience. Our motivation came from prior work in question-answering that suggests that adding textual context to an answer provides a substantial boost to user preference and performance [24]. While this was established for Search Engine Result Pages (SERPs) of websites like Google, where each search result contains a two to three line description, it has yet to be measured for CQA websites like iFixit.

iFixit currently has an inconsistent user interface across its website where adding a snippet to the question title is concerned. It shows the first line of the question description along with the question title on its search result page (see Figure 3.1). However, it does not display a snippet on the answers forum (see Figure 3.2) or when suggesting potential relevant questions on the page where users ask questions (see Figure 3.3). Till date, iFixit has not evaluated whether adding a snippet to the question title has any benefits. Thus, our contribution of experimenting and measuring this hypothesis in lab conditions could encourage them to keep their user interface uniform across the website by either adding or removing the snippets from all pages.
Figure 3.1: Search results on iFixit website

Figure 3.2: Answers forum on iFixit website
In sum, Experiment 1 attempts to transpose Cutrell and Guan’s [14] snippet manipulation from SERPs to CQAs. However, adding additional contextual information to each related question could pose a tradeoff for searchers. On one hand, adding more information enables users to decide if a link to another question is likely to contain an answer for the question they asked. On the other hand, for related questions that are irrelevant to the user’s question, adding additional information could result in cognitive noise thereby potentially harming performance. With this in mind, this experiment was designed with the following assumptions about adding a snippet to iFixit’s ‘Ask a Question’ page:

- It should increase the number of recommended questions perceived as relevant
- It should make the asker more inclined to click beyond the first and second recommended questions
- It should boost the asker’s satisfaction with the proposed set of questions
- It should improve the likeliness to find an answer among the proposed questions

3.1.2  Materials

The experiment took place in a standard office setting. The SensoMotoric Instruments (SMI) REDn Scientific eye tracker [3] was used with a 22 inch monitor to track the user’s gaze. This eye tracker is attached to the monitor and is therefore, non-invasive. It has a gaze sampling frequency of 60 Hz, with an accuracy of up to a 0.4 visual angle. All on-screen material was presented in full-screen mode, with a resolution of 1920 x 1080 px. The SMI Experiment Suite Scientific was used to design the experiment and analyze the data. It has the following two components:

1. **SMI Experiment Centre**: SMI Experiment Center allows designing complex experimental scenes involving text, questionnaires, images, videos, PDFs,
or dynamic websites and merge several elements with the Composite Editor. However, the Composite Editor cannot have an image and a question on the same screen. Thus, for each question of the stimuli, three screens were involved. The first screen asked the question, the second screen presented the stimulus and the third screen reiterated the question but this time with answer choices. These stimuli were presented to the user in a randomized order.

2. SMI BeGaze: SMI BeGaze is the analysis component of SMI Experiment Suite Scientific. It provides a comprehensive spectrum of smart functions for analysis and visualization of the data collected in SMI Experiment Center. For instance, results are visualized by Scan Paths or Heat Maps and can be exported as a video or an image for documentation and presentations.

All stimuli are of the ‘Ask a Question’ section on the iFixit website. There were ten stimuli in total for five randomly selected questions - five from the original website (control) and five redesigns (top_text). The control stimuli include the device selected by the user, a text box to enter the question’s title and a box with proposed related questions. There are usually two to five related questions displayed in the box. Each related question has a question title, along with information about when it was answered and how many people answered it. These five stimuli were named control_1 - control_5 (Figure 3.3 shows the control_1 stimuli). The redesigned stimuli were created in Adobe Photoshop. They were named top_text_1 - top_text_5 (Figure 3.4 shows the top_text_1 stimuli). These stimuli included all the information from the control stimuli along with top two to three lines of the recommended question. The recommended questions displayed in control stimuli and top_text stimuli for the same question title are different. This is due to two reasons. Many questions that are actually recommended on the page either do not have any description or their description is too short to display as a snippet. Second, as this experiment has a
within subject design, having different questions in the two versions of stimuli would prevent carryover effects. Carryover effects could occur if, for example, the user saw the control_1 stimuli first and remembered the answers to the four questions asked along with it. If they are subsequently presented with top_text stimuli with the exact same recommended questions, they may click the same answers without reading the snippet.

![Ask a Question](image)

Figure 3.3: The original design of the proposed related questions (control_1)
3.1.3 Procedure

Participants were selected using snowball sampling. Overall, 8 participants were recruited, all of whom were students at Cal Poly. As this experiment was conducted...
in the Computer Science building, there was a bias towards students in the Computer Science department. However, the fields of Business Administration, Printed Electronics and Electrical Engineering were also represented.

The experiment was conducted in the following steps. After being introduced to the study, the participants read and signed a consent form. The eye tracker was calibrated to the participant. They then had to complete the survey on the screen. This process took between 20 to 30 minutes per participant.

The survey was divided to a pre-experiment questionnaire, analysis of the ten stimuli and a post experiment questionnaire. In the pre-experiment questionnaire, participants were asked about their previous experience with CQA websites, their ability to read content on the screen as well as if they would be willing to repair their device in case it was broken. Next, the ten stimuli were presented in a randomized order.

For each stimulus, participants were asked the following four questions:

1. How many out of these 5 proposed questions do you think are relevant? (Response range from 0 to 5)
2. Which of these questions would you click on? (Response range from 0 to 5)
3. How satisfying is the set of proposed questions? (Likert scale from 1 to 7)
4. How likely are you to find your answer among the proposed questions? (Likert scale from 1 to 7)

After the experiment, participants answered a short questionnaire with questions about demographic information and their experience with this experiment. Of interest was the answer to the question ‘Which of the two presentations do you prefer’. The participant had to choose between ‘The one with the description’ and ‘The one
without the description’.

**Figure 3.5: Experiment Design**

### 3.1.4 Results

All participants were aware that their gaze as well as all interactions with the system were being recorded. They were 24 years old on average and 62.5% male. Based on the pre-experimental questionnaire, all participants could read the screen comfortably with or without their glasses/contacts. A majority of them reported familiarity with CQA websites such as StackOverflow, Yahoo! Answers and Quora. Very few participants reported using iFixit before.

The effect of adding a snippet (*Snippet_factor*) to the question title was measured using within subject ANOVAs on four dependent variables [1]. These variables were number of relevant questions (*relevance*), satisfaction with the SERP (*satisfaction*), likeliness to find answer (*findability*) and position of questions one would click on (*answer position*). *Snippet_factor* comprises of two conditions: the control where the SERPs only provided the titles for each recommended question and top_text where the first two to three lines of the question description were added to the title. All ANOVAs reached significance as we now detail.

As shown in Figure 3.6, adding a snippet was confirmed to improve SERP relevance, i.e., the number of recommended questions that participants perceived as relevant from 2.321 (SE = 0.293) to 3.150 (SE = 0.377). This increase was significant \( F(1, 7) = 19.567, p = 0.003 \) and the effect size was high \( \eta_p^2 = 0.737 \). This result strongly supports the hypothesis that adding top_text snippet to titles should suffice
for askers to better spot the results that are relevant to them.

The *satisfaction* and *findability* dependent variables were recoded to count any participant selection of the ‘Very satisfying’, ‘Satisfying’ or ‘Somewhat Satisfying’ on the one hand, and ‘Very Likely’, ‘Likely’ and ‘Somewhat Likely’ on the other hand. As illustrated in Figure 3.7, top_text was found to increase *satisfaction* from 3.250 (SE = 0.366) to 4.000 (SE = 0.267). This result was significant \( F(1, 7) = 9.000, p = 0.020 \) and the effect size was high \( \eta^2_p = 0.563 \). Top_text also improved the *findability* from 3.125 (SE = 0.398) to 4.375 (SE = 0.263). This result was also significant \( F(1, 7) = 11.667, p = 0.011 \) and the effect size was high \( \eta^2_p = 0.625 \). This analysis supported the assumption that askers are more satisfied with the redesigned stimuli and also feel more likely to find an answer to their question when contextual information is presented.

Finally, position of the question one would click on was analyzed using a repeated measure (Position 1 through 5 and none) ANOVA (see Figure 3.8). Again, this analysis revealed a main effect of Snippet_factor \( F(1, 7) = 6.906, p = 0.34, \eta^2_p = 0.494 \) as well as of Position \( F(5, 35) = 6.482, p = 0.000, \eta^2_p = 0.481 \). Interestingly,
Figure 3.7: Positive judgement for *Satisfaction* and *Findability*

the two variables interacted significantly ($F(5.35) = 3.294, p = 0.000, \eta^2_p = 0.544$).

Top_text snippet selectively lifted participants’ consideration for Positions 3 and 4. This result presents two interesting points:

1. More users selected ‘I would not click on any of these’ for the control stimuli as opposed to the redesigned stimuli. It is now easy to understand why CQA websites have so many duplicate questions. A lot of proposed questions that are actually relevant have titles that the users feel would not answer their question.

2. Users found more questions relevant to their query in the top_text stimuli vs the control stimuli. For the top_text stimuli, Question 4 was selected by users as the question which is most likely to contain their answer. This means that they viewed Questions 1-5 and made a decision based on the data contained in the description. For the control stimuli, most users selected questions 1-3.
Figure 3.8: Click likeliness per question position

The position results can also be illustrated with the heat maps generated for control 1 (Figure 3.9) and top_text 1 (Figure 3.10). In those figures, the presence of green for all 5 proposed related questions indicates that users view the title and related information of all questions. Yellow indicates an increased amount of fixation and red indicates the areas where users fixated the most. This result coincides with the findings in [22] which states that most people employ a linear strategy in which each result is evaluated in turn. Most gaze activity was directed at the first few items. Items towards the end of the list got the least amount of user attention.

In Figure 3.10, the presence of the colors green, yellow and red for the first three results demonstrates that users fixate the most on these results. However, all 5 results were viewed by users. This is a positive result because if users view the information along with the title, they can better determine if a suggestion is likely to contain the information they are interested in.
Figure 3.9: Heat Map showing users’ gaze behavior for control_1

Figure 3.10: Heat Map showing users’ gaze behavior for top_text_1
Regarding the post test survey, all participants reported that they would prefer having the description under the question title to provide some context about the question. The next question asked participants to give their suggestions on how to improve the related questions section. The most popular suggestion was to use keywords related to the description or to highlight the keywords in the description as they often tended to gloss over the text.

3.1.5 Implications/Takeaways

This experiment showed that while users spent more time analyzing the increased information on the redesigned webpages, their satisfaction was higher. For all stimuli, participants reported that they were more likely to find an answer to their question when contextual information is displayed along with the question title of proposed related questions.

3.2 Experiment 2

3.2.1 Motivation

Upon investigating the questions posted on iFixit, it was found that majority of them typically contain at least one, if not both of the below attributes:

1. Symptom(s)
   
   Symptoms are signs which indicate that the device is malfunctioning. Possible examples of symptoms include ‘reboots on its own’, ‘does not display the current date’ etc.

2. Cause(s)
Cause(s) describe the reason for device malfunction. Possible examples of causes include ‘left in a heated car for two days’, ‘dropped it in water’ etc.

This experiment aims to answer research question two by exploring if either of these attributes alone, or a combination of both increases how similar two posts look to a user. The stimuli for testing this issue were pairs of device malfunction posts, referred here as ‘questions’, taken from iFixit, that could pass as duplicates as they share the same cause(s), symptom(s) or both.

3.2.2 Materials

The three conditions for the question similarity factor were symptom_only, cause_only and symptom+cause. To find pairs that satisfy all the conditions, we picked six types of devices (e.g., iPad, PS4, Macbook, etc.) and first searched iFixit for pairs of questions that had identical cause(s) and symptom(s). Subsequently, we took one of those two questions, and found two other questions: one with the same cause(s) but different symptom(s) and the other with the same symptom(s) but different cause(s). Thus, for each device, there were three pairs of questions satisfying one similarity condition each. Symptom(s) and cause(s) in the question posts were manually identified and their correctness was justified by the author. Presenting a participant with the three pairs of a 3DS malfunction for example, would cause carry over effects. This can be avoided by presenting any given participant with one condition per device, and counter-balancing the remaining conditions for this device from that participant to the next two participants (see Figure 3.11).

In addition to being presented with question pairs of varied similarity, participants were instructed how they should assess these pairs. This second factor, called judgment, was administered between subject meaning that each participant was assigned to one of three groups. One group was asked to rate the overall similarity between
two questions, another group was asked to compare how similar the symptom(s) look in two questions and the last group was asked to compare how similar the cause(s) look in two questions. These experiment stimuli (i.e. the question pairs) and judgment instructions (i.e. how similar are these two malfunctions?; see Figure 3.12 for an instance of the cause condition) were administered by means of a web application created using the Flask framework [28].

For the front end, we coded a template using the Bootstrap framework. Bootstrap is the most popular HTML, CSS, and JS framework for developing responsive, mobile first projects on the web [26]. Every component in Bootstrap consists of an HTML structure, CSS declarations and if needed, some JavaScript code. The template for this experiment was stored in a templates folder in Flask’s application package in order to keep the logic of our application separate from the layout of the web pages. The stimuli as well as the answer forms were dynamically loaded into the template using a basic controller that fetched predefined data from a .csv file. Each participant’s answers were saved in a separate log file along with timestamps associated with each answer. This experiment was deployed on Amazon Web Services and the link to the experiment was distributed to participants using Amazon Mechanical Turk.

![Figure 3.11: Question pairs countered-balanced following a Latin Square design](image-url)
To prepare the data for further statistical analysis, some python code was written that took a folder of folders containing log files as input and loaded the data into a Pandas DataFrame [7]. From each log file, the stimuli, answers and the timestamps associated with each answer were imported. Some additional variables that were calculated from this data and added to the dataframe were:

- **Reaction Time**
  
  This variable is calculated for each answer by subtracting the timestamp of that answer with the preceding one.

- **Likert Scale**
  
  A numerical value is allotted to each answer and stored in this variable as per the table below. If the answer is none of the values in the table then the variable remains empty.
• HIT Exclusion

This variable is assigned the value ‘Y’ if the reaction time for an answer is too short or too long. This is to ensure that the participant was attentive while answering and did not randomly answer the question without reading. If the HIT is valid, it is not assigned any value.

• Participant Exclusion

This variable contains the value ‘Y’ if the number of unique answers for a participant is less than 4 or if they have more than 3 HIT Exclusions. Else it is not assigned any value.

Once the dataframe was prepared, it was exported to a .csv file. Any participant that had the value ‘Y’ for Participant Exclusion was removed from the file.

### 3.2.3 Procedure

In order to obtain a large and diverse set of participants, this experiment was conducted on Amazon Mechanical Turk. Each user’s participation and the data obtained was processed anonymously. Overall, 74 valid participants were obtained. Each participant was assigned to one of three between subjects group (judgment). The first group containing 27 participants analyzed overall post similarity i.e. posts that contain both symptom(s) and cause(s). The second group containing 23 participants...
analyzed similarity of posts based on symptom(s) alone and the third group containing 24 participants analyzed similarity of posts based on cause(s) alone.

The experiment began with written information as to what device malfunctions are and how they often involve some descriptions of symptoms, causes and actions taken by the asker to fix the issue. Then, instructions were given to judge pairs either in terms of overall, symptoms or causes similarity (judgment factor). Subsequently, participants were given two training trials to familiarize them with the procedure. Following this, they were presented with the six question pairs previously described and asked to rate how similar they seemed to them, on the Likert scale depicted in Figure 3.12. This entire process took at most four minutes per participant.

3.2.4 Results

The effect of question similarity (symptom+cause, symptom_only, cause_only) and judgment (overall, symptoms, causes) on Reaction Time and Likert Rating were analyzed with 2-way ANOVAs.

**Likert Rating**

A significant main effect of question similarity on Likert Rating was found ($F(2, 142) = 12.377, p = 0.00, \eta^2_p = 0.148$). As illustrated in Figure 3.13, Likert Rating was higher for [question similarity, symptom+cause] ($M = 3.656, SE = 0.93$) than for [question similarity, cause_only] ($M = 3.092, SE = 0.127$) and [question similarity, symptom_only] ($M = 2.907, SE = 0.120$). This result supports the notion that a combination of symptom(s) and cause(s) increases perceived similarity of question pairs.
The main effect of judgment on Likert Rating was not significant \( (F(2, 71) = 0.472, p = .626, \eta_p^2 = 0.013, power = 0.124) \); neither was the interaction between judgment and question similarity \( (F(4, 142) = 0.644, p = 0.632, \eta_p^2 = 0.018, power = 0.207) \).

**Reaction Time**

When it came to the time it takes to judge the question pairs, neither judgment \( (F(2, 71) = 1.861, p = .163, \eta_p^2 = 0.050, power = 0.375) \) nor question similarity \( (F(2, 142) = 1.510, p = .224, \eta_p^2 = 0.021, power = 0.317) \) were significant.

The two factors did interact significantly \( (F(4, 142) = 2.766, p = 0.030, \eta_p^2 = 0.072) \). However, the effect size for this result was small, thereby suggesting that this interaction is not practically meaningful (depicted in Figure 3.14).

It seemed like participants asked to judge the pairs on an overall basis took a little longer and possibly found it a little difficult to deal with question pairs that shared similar causes but discrepant symptoms. It also appeared as though participants instructed to judge pairs either in terms of their cause or of their symptom similarity,
took longer to make their decision when both similarities were fulfilled. However, to reiterate, this result is too weak to consider that judgment and/or question similarity really affected participants’ reaction times.

### 3.2.5 Implications/Takeaways

This experiment suggested that, regardless of what the asker has in mind (see the judgment factor), they will find a question pair containing the same cause(s) and symptom(s) to be more similar than posts that match only one of these attributes. A question or post recommendation strategy that surfaces any available symptom(s) AND cause(s) attributes should thus improve the asker experience, by increasing the perceived similarity or relevance without impacting the time it takes to make such a judgment.
3.3 Experiment 3

3.3.1 Motivation

This experiment was conducted to answer research question three. Once it was ascertained that a question pair containing both symptom(s) and cause(s) increases their perceived similarity, we wanted to determine if these two attributes are worth testing on the ‘Ask a Question’ page and whether they would improve the asker’s experience. The snippets found in Experiment 1 to improve relevance, number of questions clicked, satisfaction and likeliness to find an answer were contrasted with snippets that include symptom(s) and/or cause attributes. Experiment 3 was very analog to Experiment 1 as participants assessed redesigned ‘Ask a Question’ pages for iFixit under gaze tracking conditions. Experiment 3 narrowed down the 4 assessment questions from Experiment 1 (How many out of these 5 proposed questions do you think are relevant?, Which of these questions would you click on?, How satisfying is the set of proposed questions?, How likely are you to find your answer among the proposed questions?) to 2: ‘How many out of these 5 proposed questions do you think are relevant?’ was asked because its effect size was the highest ($\eta_p^2 = 0.737$) and ‘Which of these questions would you click on?’ was asked for its valuable insight into the number and positions of questions clicked.

3.3.2 Materials

This experiment used the same eye tracking apparatus and method of creating stimuli that was used in Experiment 1. This time, two types of snippets were created in reference to the results of the previous two experiments.

1. Snippets with top_text (top_text, Figure 3.15) were very similar to those from Experiment 1: they contained the first two to three lines from the descrip-
tion of each recommended question. We manually stripped any cause(s) and symptom(s) statements from these snippets.

2. **Snippets with Cause(s) and/or Symptom(s)** (sympt,cause. Figure 3.16) were made of two to three lines from the description of the recommended question that contained any cause(s) and/or symptom(s). As described earlier, symptoms are signs that indicate that the device is malfunctioning. Possible examples of symptoms include ‘reboots on its own’, ‘does not display the current date’ etc. Causes describe what caused or causes the device to malfunction. Possible examples of causes include ‘left in a heated car for two days’, ‘dropped it in water’ etc.

Snippets with Cause(s) and/or Symptom(s) reflect the outcome from Experiment 2 that cause(s) and symptom(s) interplay in the perceived similarity of questions. They were designed around the following rules:

- If the question title does not contain either the symptom(s) or cause(s) then the snippet must contain the cause(s) and/or symptom(s)

- If the question title contains the symptom(s) then the snippet must contain the cause(s)

- If the question title contains the cause(s) then the snippet must contain the symptom(s)

However, there are two stimuli that do not follow these rules. These stimuli only have symptoms in the question title as well as the snippet. The absence of cause(s) is a reality in a large number of questions posted on CQA websites as people do not always know what is causing a device to malfunction.
Figure 3.15: 3DS top-text
Figure 3.16: 3DS sympt._cause
3.3.3 Procedure

Overall, 23 participants were recruited. As this experiment was conducted in the Computer Science building at Cal Poly, there was a bias towards university students.

All materials involved in the experiment were reviewed and approved by the Cal Poly Institutional Review Board (IRB). These materials include the stimuli used in this experiment along with the accompanying questions and an informed consent form to receive agreement to participate from the participants.

The experiment was conducted in the following steps. After being introduced to the study, the participants read and signed an informed consent form. The eye tracker was calibrated to the participant. Participants were then given a training trial to familiarize them with the procedure. They were subsequently presented with a series of stimuli and corresponding questions. The experiment took 15 minutes per participant on average.

All stimuli were redesigned versions of the ‘Ask a Question’ section on the iFixit website. This section includes the device selected by the user, a text box to enter the question’s title and a box with proposed related questions. There are usually two to five related questions displayed in the box. Each related question has a question title, along with information about when it was answered and how many people answered it. The redesigned stimuli included all the information from the original page and two to three lines of contextual information (snippet) per suggestion. There were 12 stimuli in total for six questions - six stimuli containing snippets with top_text and six stimuli containing snippets with sympt_cause. These stimuli were equally divided among two groups of participants as per Figure 3.17 so that each participant would only see one version of the page, either top_text or sympt_cause. The order of stimuli within each group was randomized.
For each stimulus, participants were asked the following 2 questions:

1. How many out of these 5 proposed questions do you think are relevant? (Response range from 0 to 5)

2. Which of these questions would you click on? (Response range from 0 to 5)

A limitation of using the SMI Experiment Center, is that the current version does not allow an image and a question to be placed side by side. However, it does allow an image and text to be placed on the same page. Therefore we had to put a stimuli with the question in text next to it and ask participants to answer the question on the next page.

### 3.3.4 Results

The effect of adding different types of snippets (*Snippet_factor*) to the question title was measured using within subject ANOVAs on four dependent variables. These variables were number of relevant questions (*relevance*), trial duration (*duration*), fixation count (*fixation*) and position of questions one would click on (*answer position*). *Snippet_factor* comprises of two conditions: sympt_cause where the snippet contains two to three lines from the description of the question that contain the cause(s) and/or the problem(s) that led to the question.
symptom(s) versus top_text where the snippet comprises of the first two to three lines of the question description.

Three participants were removed from the analysis: Two because their relevance answers lacked variability (they repeatedly picked two out of the six options), and one because they showed some fast but random answering for one stimulus. This leaves us with 20 valid results: 11 participants in Group 1 and 9 participants in Group 2.

Figure 3.18 presents the results for number of relevant questions. It was found that while sympt_cause slightly increased the number of recommended questions that participants perceived as relevant from 2.682 (SE = 0.177) to 3.062 (SE = 0.156), there was no significant difference between the two conditions ($F(1,18) = 4.206, p = 0.055, \eta^2_p = .189, power = 0.492$). Similarly, trial duration ($F(1,19) = 1.448, p = 0.244, (\eta^2_p = .071, power = .208)$) did not reach significance.

![Figure 3.18: Number of questions found relevant by users](image)

As illustrated in Figure 3.19, sympt_cause decreased the amount of time users fixate on recommended related questions from 4.544 (SE = 0.646) to 2.846 (SE =
0.278). This result was significant \((F(1, 19) = 12.774, p = 0.002, \eta^2_p = .402)\). This result supports the hypothesis that sympt\_cause snippets enable askers to spot the results that are relevant to them faster than top\_text snippets.

![Fixation Count](image)

**Figure 3.19: Fixation Count**

The fixation results can also be illustrated with the heat maps generated for sympt\_cause (Figure 3.20) and top\_text (Figure 3.21) of the iPod stimuli. The presence of higher amount of green, yellow and red colors in the top\_text stimuli indicates that users fixate more on these results as opposed to sympt\_cause. However, higher fixations do not indicate that askers would click on more recommended questions. Upon reviewing the questionnaire results, it was found that on average, the number of questions that askers would click on was slightly higher for sympt\_cause (2.56 for sympt\_cause vs 2.37 for top\_text). Thus, askers found the recommended questions slightly more relevant in sympt\_cause stimuli and did not have to fixate on them as much in order to make that decision.
Finally, position of the question one would click on was analyzed using a repeated measure (Position 1 through 5 and none) ANOVA (see Figure 3.22). This analysis revealed that the interaction between snippet factor \( (F(1,19) = 2.786, p = 0.111, \eta^2_p = 0.128, power = 0.354) \) and answer position \( (F(5,95) = 37.450, p = 0.00, \eta^2_p = 0.663, power = 1.000) \) did not yield a statistically significant result \( (F(5.95) = 0.168, p = 0.974, \eta^2_p = 0.009, power = 0.088) \).

### 3.3.5 Implications/Takeaways

This experiment indicates that askers found the recommended questions slightly more relevant with lesser fixation in stimuli where the snippets contain symptom(s) and/or...
Figure 3.22: Click likeliness per question position

causes(s) over stimuli whose snippets contain the first two to three lines of the question description. Thus, sympt\_cause constitutes an incremental improvement over top\_text.
4.1 Summary of Results

This thesis presents the first application of gaze tracking methodologies to tackle duplicate questions on Community based Question Answering (CQA) websites.

Based on prior studies on web search behaviors, we assumed that adding contextual information (snippets) to proposed related questions displayed on the ‘Ask a Question’ page of the CQA website iFixit would improve the asker experience and reduce their tendency to post a new duplicate question. The first lab experiment where this web page was redesigned and compared to the original one was conducted on 8 participants. Results confirmed that participants were more likely to find an answer to their question on the redesigned page.

A second experiment was conducted remotely via Amazon Mechanical Turk on a larger sample of 74 participants which aimed to discover attributes that increased the perceived similarity of question pairs. It was discovered that askers find a question pair containing the same ‘cause(s)’ and ‘symptom(s)’ to be more similar than posts that match only one of these attributes.

The third lab experiment, conducted on 20 participants, used these attributes to redesign and assess the snippets from Experiment 1. Results indicated that snippets containing ‘symptom(s)’ and ‘cause(s)’ attributes constitute an incremental improvement over basic snippets i.e they are perceived as slightly more relevant and require significantly less gaze fixations on the asker’s part.
4.2 Future Work

The work done in this thesis was exploratory and there are several directions in which the results identified could be applied to the iFixit case, more specifically with revisions to the ‘Ask a Question’ page. iFixit already provides instructions that state how the ‘symptoms’, ‘causes’ and ‘action taken’ attributes examined in this thesis help formulate a ‘good question’.

Experiments 2 and 3 of this thesis formally back up such common-sense guidelines with behavioral insights. They showed higher perceived similarity between questions and question relevance when both symptoms and causes are present, suggesting that guidelines revolving around such symptoms cause attributes are worth enforcing. Unfortunately, these guidelines are only accessible through a link that is not prominent on the webpage. Thus, a majority of askers may not even view the instructions before posting a question.

One way to remedy this would be to adopt the approach used by websites like StackOverflow where askers are allowed to ask a question only after they have read the instructions and clicked on a check box to accept that they have read them. Currently, iFixit incites users to formulate ‘strong’ question titles. However, as Figure 4.1 shows, the rules behind this mechanism do not reflect the aforementioned guidelines.

Be specific.

If you ask a vague question, you’ll get a vague answer. If you provide details and symptoms, we can write you a relevant, personalized answer. Questions like ‘My iPod is broken’ aren’t likely to get you a response. If you tell us what is broken, how it happened, and what remedies you’ve already tried, we’re happy to help. Photos help, too!

Figure 4.1: One of the few instructions on the iFixit website for asking better questions [2]

One quick fix here could be to add to the tooltip a mention to add symptom(s)
and cause(s) to the title. It would be interesting to A/B test such a UI redesign in terms of the clicks on recommended questions a given popular question topic gets.

A more engineering-intensive fix would be to tie the tooltip to the presence/absence of symptom(s) or cause(s) in the question title. This direction entails an automatic extraction of attributes like causes and symptoms, for example by having a few people manually label them in a large set of iFixit posts and then training a supervised classifier. A range of causes and symptoms could hence be extracted as a basis to enforcing the tooltip about question strength showed in Figure 4.2, or even color-code these attributes of interest not only on iFixit’s ‘Ask and question’ page but also in the websites’ actual question/answer pages.

In Google SERPs, if a snippet contains the search terms that were typed in, these are made bold to enhance the user’s experience. It is likely that snippets with highlighted terms are easier to check and are clicked more often. This idea can be adopted to snippets in CQA websites like iFixit and tested to determine if doing so results in askers clicking on more recommended questions.

Figure 4.2: Tooltip analyzing the quality of question asked

Finally, it should be noted that snippet heatmaps from SERPs studies [19] [14]
as well as experiments 1 and 3 of this thesis show that askers tend to fixate on top results longer and more often while the last result is barely looked at. A quick remote study like the one conducted in this thesis could examine the optimal number of recommended questions that should be displayed as well as how many lines of snippet are needed to maximize askers decision to click on them.

The ‘Ask a question’ webpage on various CQA websites could possibly look very different in the future. Experiments like the ones conducted in this paper could be instrumental in guiding web developers on the best layout to tackle duplicate questions.
BIBLIOGRAPHY


[4] iFixit Question - is my iPhone ruined.

[5] iFixit Question - Phone wont turn on.


[8] Quora - A place to share knowledge and better understand the world.


APPENDICES

Appendix A

TECHNOLOGIES USED IN EXPERIMENT 2

• Bootstrap

Bootstrap is the most popular HTML, CSS, and JS framework for developing responsive, mobile first projects on the web [26]. Every component in Bootstrap consists of an HTML structure, CSS declarations and if needed, some JavaScript code. Bootstrap was used in this experiment to design the template for the web application.

• Flask

Flask is a micro web framework written in Python and based on the Werkzeug toolkit and Jinja2 template engine [28]. It is classified as a microframework because it does not require particular tools or libraries. Flask was used in this experiment to populate the bootstrap template with stimuli and data from a .csv file as well as to convert these populated templates into a web application.

• Amazon Web Services

Amazon Web Services (AWS) offers cloud web hosting solutions that enable individuals and organizations to post a website or web application onto the Internet. Our web application was hosted using AWS.
• Pandas

Pandas is an open source python package providing fast, flexible, and expressive data structures that facilitate data manipulation and analysis [7]. This library is built upon the functionality provided by NumPy. Pandas dataframes were used to combine the data from multiple log files in experiment 2 in order to perform statistical analysis.