A Study on Understanding and Encouraging Alternative Information Search

Suppanut Pothirattanachaikul
Graduate School of Informatics
Kyoto University

Abstract

Search engines become a main information resource where searchers promptly obtain the information they need in a matter of seconds. According to the survey, approximately 57.% of the participants answered that one motivation for using Web search engines is to find a way or means to solve their goal. Consistent with this survey, search engines tend to develop relevance-oriented functionalities to make the searchers satisfy their needs with less effort. Although search engines enable searchers to promptly satisfy their information needs, they might not aware of the accuracy of the obtained information. For example, a searcher who wants to lose weight may issue "green tea extract" as a query to a search engine and mostly receive positive information regarding the green tea extract even though the benefits and effectiveness of the green tea extract to weight loss is remained questioning by some researchers. Based on this situation, search engines may lead their users to inaccurate information, and as a result, causing them with fatal consequences.

To accurately obtain information, previous studies suggested that searchers should expose themselves to the information that are inconsistent with their beliefs and search carefully (spending more time on a search task, checking the author's information in a document during a search, etc.). Unfortunately, searchers are likely to exhibit confirmation bias; where they seek information to confirms their beliefs via Web search engines. For example, the searchers who believe in the effectiveness of green tea extract might select only positive information about green tea extract. Consequently, they refrain from diverse and careful searches for information and do not consider the information that are not consistent with their beliefs even such information could also satisfy their information need. This research aims to mitigate the confirmation bias in Web searches. To mitigate confirmation bias in Web searches, previous studies suggested that people should be exposed to information that are not consistent with their beliefs. In this study, we regarded such information as alternative information (i.e. the information that help achieving the same goal as the input query and are not consistent with the searchers' beliefs). We tackled the alternative information search problem to understand the effect of alternative information to the search behaviors and encourage them to interact with the alternative information. This problem can be separated into three parts. First, we **mined the alternative information** (e.g., the information that achieves the same goal as the given query). Second, we **analyzed the effects of alternative information on search behaviors and beliefs**. We were curious whether searchers prefer to interact with alternative information, which are different from the input query. Understanding the effect of alternative information on search behaviors can help us design the search system that prevents the

searchers from confirmation bias and maintains their satisfaction with the system. Finally, we **designed an interface of the question-answering system that encourages the interaction between searchers and alternative information** as suggested by previous studies that people are likely to consider questions related to the alternative hypothesis when such questions are provided.

Acknowledgement

I would like to express my sincere appreciation to Professor Masatoshi Yoshikawa for his valuable comments and suggestions along with this research. He always supported my research through his great research vision and foresight.

I am grateful to my thesis committee members Professor Keishi Tajima and Professor Shinsuke Mori for their comments and suggestions on this thesis.

I would also like to thank Dr. Yusuke Yamamoto of Shizuoka University for valuable discussions and comments.

I would like to thank the secretaries of Yoshikawa lab, Ms. Rika Ikebe, and Mrs. Nakahara Yoko for their kindly help.

Special thanks should be given to Associate Professor Makoto P. Kato of Tsukuba University who gave me numerous encouragement and suggestions. I will not forget the opportunity he gave me to make a presentation at Tsinghua University.

I would like to express my sincere gratitude to Emeritus Professor Katsumi Tanaka for every opportunity he provided. I will always remember his lessons and cheerful encouragement.

I would like to express special thanks to Associate Professor Takehiro Yamamoto of the University of Hyogo, who I would like to regard as my other older brother. I will always remember every lesson you taught especially the value of being a researcher. I hope that I would be able to work with him in the future.

I also would like to express my additional thanks to my close friend Hideaki Anazawa for his tremendous help during my stay in Japan.

I would like to thank my brother Pornchai Pothirattanachaikul, who introduced me to the computer and game console (we played Sonic the Hedgehog 3 on Sega Mega Drive together in 1995). I also would like to thank my sister Pimpisa Pothirattanachaikul for her support.

Last but not least, I would like to express a very special thanks to my partner, Wiradee Imrattanatrai for providing cheerful encouragement during my Ph.D. study and showing the value of hard work. She always helps me to choose the right path and I am not able to achieve anything without her encouragement.

Finally, I am deeply grateful to my parents Chukasaem Pothirattanachaikul and Runglawan Pothirattanachaikul for their love and support through the long journey of Ph.D. study.

Our beliefs constitute a large part of our knowledge of the world - Nils J. Nilsson

Contents

Preface v

1 Introduction **1**
1.1 Background 1
1.2 Approach 2
1.2.1 Mining Alternative Information 3
1.2.2 Understanding How and When People Search for the Alternative Information 3
1.2.3 Encouraging the Searchers to Interact with Alternative Information . 3
1.3 Difference between existing literature and ours 4
1.4 Thesis Structure 4

2 Related Work **7**
2.1 Task-Oriented Web Search 7
2.2 Connecting cQA with Web Search 8
2.3 Comparative Entity Mining 8
2.4 Confirmation Bias 8
2.5 Credibility 9
2.6 Question Answering System 9
2.7 Bias, Credibility, and Search Behaviors 9

3 Mining alternative information **11**
3.1 Introduction 11
3.2 Problem Definition 13
3.2.1 Alternative Action 13
3.2.2 Alternative Action Mining Problem 14
3.2.3 Relation to Existing Work 14
3.3 Approach 15
3.3.1 Technical Challenge 15
3.3.2 Method Overview 16
3.3.3 Retrieve Question-Answer Pairs 17
3.3.4 Extract Candidate Actions from Answers 17
3.3.5 Measure Alternativeness between Actions 18
3.3.6 Measure Effectiveness of Action by Community Evaluation 19

3.3.7 Generate Diversified Ranked List . . . 19
3.4 Experimental Setup . . . 20
3.4.1 Dataset . . . 20
3.4.2 Proposed and Baseline Methods . . . 20
3.4.3 Test Collection Construction . . . 22
3.4.4 Evaluation Metric . . . 23
3.5 Experimental Results . . . 24
3.5.1 Comparison with Baselines . . . 24
3.5.2 Effect of Parameters . . . 27
3.5.3 Effect of Domain . . . 27
3.5.4 Examples . . . 28
3.6 Limitations . . . 28
3.7 Summary . . . 29

4 Understanding how and when people search for the alternative information 31
4.1 Introduction . . . 31
4.2 Concept . . . 32
4.2.1 Critical Thinker in Web Search . . . 32
4.2.2 Factors Investigated in This Study . . . 33
4.2.3 Controlled Documents Preparation . . . 33
4.3 Method . . . 35
4.3.1 Experimental Design . . . 35
4.3.2 Procedure . . . 35
4.3.3 Search System . . . 38
4.3.4 Participants . . . 40
4.4 Results . . . 40
4.4.1 Perception of Credibility . . . 41
4.4.2 Data Annotation . . . 41
4.4.3 Search Behavior . . . 42
4.4.4 Document's Opinion and Credibility on Search Behavior (RQ1) . . . 44
4.4.5 Belief Dynamics (RQ2) . . . 45
4.4.6 Search Behaviors and Belief Dynamics (RQ3) . . . 47
4.5 Discussion . . . 49
4.5.1 Implications . . . 49
4.5.2 Limitations . . . 49
4.6 Summary . . . 50

5 Encourage the searchers to interact with the alternative information 53
5.1 Introduction . . . 53

5.2 Material . . . 55
5.2.1 Tasks . . . 55
5.2.2 Questions and Answers . . . 57
5.3 Method . . . 57
5.3.1 Overview . . . 57
5.3.2 Dependent Variables . . . 59
5.3.3 Hypothesis . . . 60
5.3.4 Experiment 1 . . . 61
5.3.5 Experiment 2 . . . 62
5.3.6 Search System . . . 63
5.3.7 Participants . . . 64
5.4 Results . . . 64
5.4.1 Search Behavior . . . 64
5.4.2 Effect of Question and Answer on Search Behavior and Beliefs (RQ1) 65
5.4.3 Effect of Alternative Question on Search Behaviors and Beliefs (RQ2) 69
5.5 Discussion . . . 72
5.5.1 Implications . . . 72
5.5.2 Limitations . . . 73
5.6 Summary . . . 74

6 Conclusions **75**
6.1 Summary . . . 75
6.2 Future Directions . . . 77

List of Figures

1.1 "Exercise" and "eat healthy" can be regarded as the alternative information to the input query "green tea pills" since they can achieve the same search task *find the information related to weight loss* 2
1.2 Overview of our approaches 5

3.1 Example question-answer pairs in cQA corpus. Our method extracts alternative actions by using question-answer structure. 12
3.2 Example structure among actions. In the figure, actions a_i (*take a sleeping pill*) and a_j (*stroll before bedtime*) are called *alternative actions* when they share the same action a_g (*solve one's sleeping problem*) as their goal. Note that the actions "take a sleeping pill" and "drink chamomile tea" are also alternative actions, since they share the other action "cure one's anxiety." . . . 13
3.3 Comparison between existing work [Liu et al. 2014, Yamamoto et al. 2012, Yang and Nyberg 2015, Yilmaz et al.] and our study. Most existing studies addressed the problem of finding *sub* actions, whereas our study finds *alternative* ones. 15
3.4 Method overview. 16

4.1 Screenshots of our controlled documents with (a) low credibility and (b) high credibility (translated from Japanese). With the exception of visual styles, both controlled documents show the same content in their articles. 34
4.2 Screenshot of our search system. The controlled document was inserted at the second rank of search results. 37
4.3 Search Result Manipulation Overview. 39

5.1 Search engine result page (SERP) where *People also ask* is displayed above organic search results. 54
5.2 Screenshot of our search system (translated from Japanese). *People also ask* was presented above organic search results. The answer was revealed when the question was clicked on. 63

List of Tables

3.1 Manually predefined terms for retrieving question-answer pairs. 17

3.2 Example queries used in experiment (EnCollection). 22

3.3 D#-nDCG@k for each test collection (highest values among methods are in bold). 25

3.4 D#-nDCG@8 for different λ and α for both JACollection and EnCollection (highest value in bold). 26

3.5 D#-nDCG@8 of RelDivQA ($\lambda = 0.4$, $\alpha = 0.5$ for JaCollection, $\lambda = 0.7$, $\alpha = 0.6$ for EnCollection) for different domains. 27

3.6 Example top 2 alternative actions retrieved by proposed and baseline methods for queries "chamomile tea" and "youth hostel". Relevant actions are in bold. 28

4.1 Search tasks used in this study. Sources of controlled documents are also shown. 36

4.2 Demographics of participants. 40

4.3 Mean (M) and standard deviation (SD) for each controlled document condition (C_1: inconsistent and high credibility, C_2: inconsistent and low credibility, C_3: consistent and high credibility, and C_4: consistent and low credibility) and statistical testing results of each search behavior of the model with different fixed effects (***: significance level at 0.001, **: 0.01, and *: 0.05). Coefficient (β), standard error (SE), t-statistic (t), and p-value (p) of each fixed effect are reported. The χ^2 and p-values of the model which is statistically significant compare to the null model are also reported. 43

4.4 Participants' behavior according to opinion types. 44

4.5 Participants' behavior according to credibility level. 46

4.6 Percentage of belief dynamics for each controlled document condition. (For each prior belief, the polar that the majority of participants leaned towards in posterior belief is highlighted in gray). The presentation of this table is inspired by the paper [White 2013]. 46

4.7 Distribution of the clicked documents' opinion for each task (excluding controlled documents) where TaskID refers to task questions in Table 4.1. 47

4.8 Mean (M) and standard deviation (SD) for each type of belief dynamics and statistical testing result of each search behavior with the belief dynamic as dependent variables (*: significance level at 0.05 ,***: significance level at 0.001). Coefficient (β), standard error (SE), and z-statistic (z) are also reported. 48
4.9 Distribution of prior knowledge for each task where TaskID refers to task questions in Table 4.1. 50

5.1 Search tasks translated from Japanese. Candidate remedies and their Cochrane review URLs are shown. 56
5.2 Questions and example answers that would be contained in People also ask (translated from Japanese). As the opinion was not manipulated in Experiment 2, the opposing answers were not required for the alternative question. 58
5.3 Demographic data of participants. 65
5.4 Mean (M) and standard deviation (SD) for each experimental condition in Experiment 1 (C_1: control, C_2: belief-consistent first, and C_3: belief-inconsistent first) and statistical testing results of each search behavior of the model with different fixed effects (***: significance level at 0.001, **: 0.01, and *: 0.05). Coefficient (β), standard error (SE), t-statistic (t), and p-value (p) of each fixed effect are reported. The χ^2 and p-values of the model which is statistically significant compared to the null model are also reported. 66
5.5 Participants' behavior according to whether questions were presented. (Experiment 1) 67
5.6 Participants' behaviors according to whether they encountered answers. (Experiment 1) 67
5.7 Participants' behaviors according to the answer's opinion they first encountered. (Experiment 1) 68
5.8 Percentage of beliefs for Experiment 1 grouped by whether participants first encountered with belief-consistent answer. (for each prior belief, the polar that most participants leaned toward in posterior belief is highlighted in gray). . . 69
5.9 Mean (M) and standard deviation (SD) for each experimental condition in Experiment 2 (C_1: control, C_2: rank 1, and C_3: rank 4) and statistical testing results of each search behavior of the model with different fixed effects (***: significance level at 0.001, **: 0.01, and *: 0.05). Coefficient (β), standard error (SE), t-statistic (t), and p-value (p) of each fixed effect are reported. For the variable alternative question, β_a represents the effect of the alternative question while β_r represents the effect of its rank. The χ^2 and p-values of the model which is statistically significant compared to the null model are also reported. 70

5.10 Participants' behavior according to whether an alternative question was presented. (Experiment 2) 71
5.11 Participants' behaviors according to whether they encountered answers. (Experiment 2) 72
5.12 Participants' behaviors according to whether they encountered an alternative answer. (Experiment 2) 73
5.13 Percentage of beliefs for Experiment 2 grouped by whether the alternative question was placed at rank 1 or rank 4. 74

1 Introduction

1.1 Background

Web searchers often use a Web search engine to find a way or means to achieve his/her real-world goal. For example, a searcher who wants to lose weight may issue the query "green tea extract," intending to find a good pills to *solve his/her obesity problem*. According to a survey on 1,000 Web searchers, reported by Nakamura *et al.* [Nakamura et al. 2007], approximately 57.5% of the searchers answered that one motivation for using Web search engines is to find a way or means to solve their goal. Nowadays, search engines provide several useful functions such as personalized search, entity card, question-answering system, etc. Such functionalities make searchers easily satisfy their need and finish the task early.

Even recent advancements in search engines enable searchers to promptly obtain information that satisfies their information need, they might not aware of the accuracy of their query. For example, searchers who want to lose weight might not aware that green tea extract is ineffective [Chalasani et al. 2014, Jurgens et al. 2012], and issue *green tea extract* as a query to find an online marketplace that sells the green tea extract. Moreover, search engines are designed to be a relevance-oriented system rather than encouraging searchers to find accurate information making the majority of the search results likely to be biased towards the query. This causes the problem called *filter bubble*, where searchers are isolated by an algorithm, which unintentionally presents the information based on user preferences. Based on a *green tea extract* situation, the filter bubble might lead users to information that is not accurate; the consequence thereof may result in disastrous decisions.

To obtain the information accurately, previous studies suggested that searchers should expose themselves to diverse information [Garrett and Resnick 2011, Liao and Fu 2014, Liao et al. 2015, Resnick et al. 2013] and carefully verify the information by spending more time on a search task, checking the author information in a document during a search, *etc* [Yamamoto et al. 2018a,b]. Unfortunately, searchers are likely to exhibit confirmation bias; where they seek information to confirms their beliefs via Web search engines [White 2013, 2014, White and Hassan 2014, White and Horvitz 2015]. For example, the searchers who believe that green tea extract helps lose weight might select only positive information about the green tea extract, even though researchers argue that it is ineffective and harmful [Chalasani et al. 2014, Jurgens et al. 2012]. Consequently, they are likely to refrain from diverse and careful searches for information and not likely to consider the information that is not consistent with their beliefs even such information also satisfies their information need.

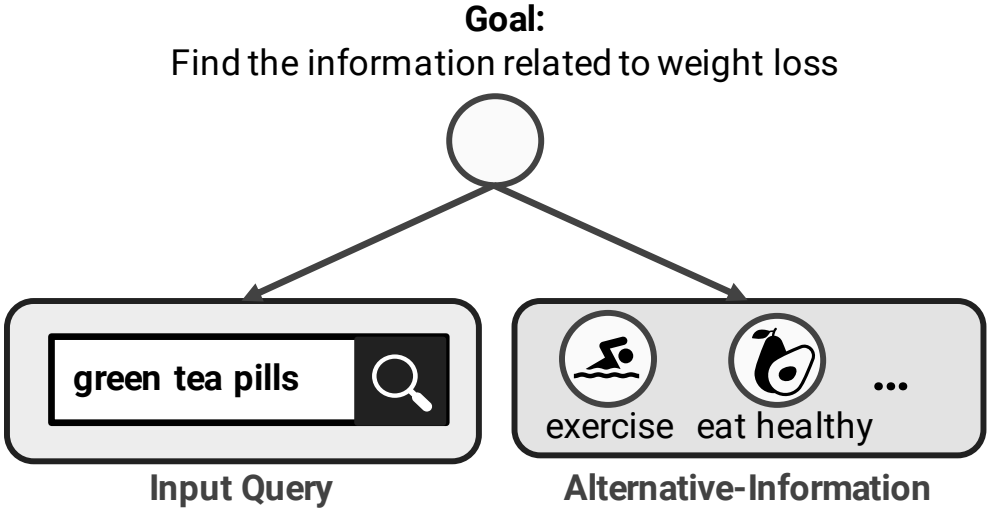

Figure 1.1: "Exercise" and "eat healthy" can be regarded as the alternative information to the input query "green tea pills" since they can achieve the same search task *find the information related to weight loss*

In order to mitigate confirmation bias, previous studies suggested that people should be exposed to alternative or opposing information that they might not have chosen previously [Buchs et al. 2004, Hernandez and Preston 2012, Nemeth et al. 1992, Sunstein 2017]. In line with these studies, we proposed the *alternative information search* problem. An alternative information search is defined as the circumstance where the searchers are encouraged to expose themselves to the *alternative information*: the information that achieves the same goal of a search task but searchers are not expected to encounter. For example, "exercise" and "eat healthy" can be regarded as the alternative information to the query "green tea pills' as they achieve the search task "find the information related to weight loss" as the input query but searchers might not initially come up with due to the lack of their knowledge (Figure 1.1). By suggesting alternative-information, searchers might be able to notice different solutions and make an improved decision to solve their obesity problem.

1.2 Approach

Figure 1.2 shows the overview of our approaches to solving the *alternative information search* problem. To encourage searchers to expose themselves to *alternative information*, we address the approach of *mining alternative information*, *understanding the effect of alternative*

information to search behaviors and beliefs, and *encouraging searchers to interact with alternative information*. The overview of each approach is described below.

1.2.1 Mining Alternative Information

In task-oriented search, searchers might not aware of alternative information that could help achieve the same goal behind their query as they often invest their beliefs in the initial query. Thus, they may not consider the alternative information although it can solve the same problems as the query. Although the current search engines provide query suggestions for supporting searchers to reformulate their query, it is difficult for the searcher to find alternative information. In this study, we regarded the *alternative information* as the *action* that achieves the same goal as a query (alternative-action hereinafter). For example, an action *"drink a cup of hot milk"* can be regarded as the alternative action to the query *"take sleeping pills"* as it achieves the same goal *"sleep well"* as the query. We developed the system to extracted the alternative-action from Q&A corpus and ranked such actions based on how well they can satisfy the same goal as the input query.

1.2.2 Understanding How and When People Search for the Alternative Information

Although we successfully extract good alternative actions regarding to the input queries, we are curious whether searchers prefer to interact with alternative actions. Therefore, it is necessary to understand how and when searcher search for the alternative information. We analyzed the effects of alternative information to the search behaviors and beliefs. In this study, the alternative information was defined as information that is inconsistent with the searchers' beliefs prior to the search but useful to achieve the search task. For example, searchers might believe that *"green tea extract"* is helpful for the solving weight loss problem. However, *"green tea extract"* is harmful for ones health. Therefore, suggesting the information that inconsistent with searchers beliefs in this case might help them avoid the an inaccurate information and achieve the goal effectively by looking for another solutions. We focused on two *factors* of the alternative information searchers encountered during their searches: (1) *opinion* that is either consistent or inconsistent with the searchers' beliefs prior to the search task, and (2) *credibility*, which represents the believability of the information. We examined the extent to which these factors encouraged or discouraged searchers from searching carefully and altering their beliefs after searches.

1.2.3 Encouraging the Searchers to Interact with Alternative Information

To support the interaction between searchers and alternative information, the interface that encourages the searchers to interact with alternative information is necessary. An efficient practice to encourage people to consider an alternative is to present alternatives as questions as suggested by Trope and Bassok that people are likely to consider questions related to the alternative hypothesis when such questions are provided [Trope and Bassok 1982, 1983].

Recently, major search engines provided the functionality known as *People also ask*, which presents a set of questions and answers that are relevant to specific queries above the search results. However, the effect of *People also ask* on the confirmation bias still be questioned. To solve this problem, two experiments were conducted to investigate the effect of *People also ask* on search behaviors and *beliefs* (*i.e.,* how people's beliefs change after the search process). For the first experiment, alternative information was regarded as the information that is inconsistent with the searchers' beliefs. For the second experiment, the alternative information was regarded as the *alternative remedy*; remedy that solves the same problem as the remedy mentioned in the search task *(i.e., target remedy)*. By doing so, we are able to learn and design question answering system that encourage the interaction between searchers and alternative information.

1.3 Difference between existing literature and ours

Existing studies [Liu et al. 2014, Yamamoto et al. 2012, Yang and Nyberg 2015, Yilmaz et al.] addressed the problem of finding *sub*-tasks (or *sub*-goals) of a given query. For example, given the query "sleeping problem", many of the existing studies were focused on finding "take a sleeping pill" or "stroll before bedtime", both of which can achieve the goal behind the query "solve one's sleeping problem." Such work can be seen as a method for supporting searchers to find more concrete solutions for achieving the action represented by the query.

In contrast, our study is different from these previous studies in a way that ours is to help searchers to find alternative solutions for the goals behind the query. We believe that our study provides another type of supports for a searcher in a task-oriented Web search and is as important as the existing studies.

Our work can also be viewed as the extension to the previous confirmation bias studies in the following ways. First, while the previous studies [Liao and Fu 2014, Liao et al. 2015, Schwind and Buder 2012, Schwind et al. 2012] revealed that people's confirmation bias can be mitigated by presenting information inconsistent with their preference, the detailed relationship between their search behaviors and decisions when encountered with inconsistent information during a search is still unclear. Second, we consider both the opinion of a search result and its credibility, which enables us to better understand their relationship. Third, we also consider the relationship between people's belief dynamics and their detailed search behaviors to better understand how their search behaviors relate to their decisions.

1.4 Thesis Structure

The remainder of this thesis is structured as follows:

- Chapter 2
  We overview existing studies related to our research presented in this thesis.

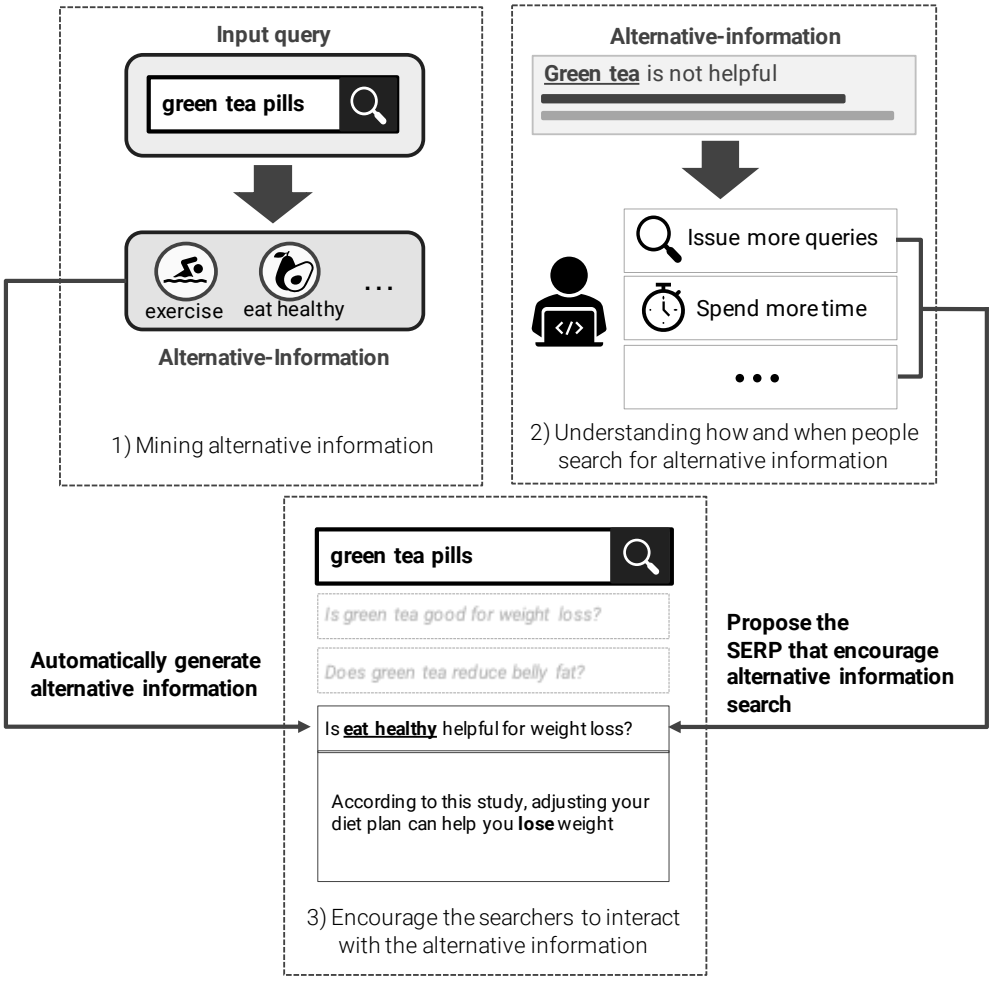

Figure 1.2: Overview of our approaches

- Chapter 3

  We proposed the *alternative information mining* problem, where a system is required to find alternative information for a given query. An alternative information for a query is defined as an action that can achieve the same goal as the query. For example, given the query "sleeping pills," our objective is to find alternative actions such as "have a cup of hot milk" or "stroll before bedtime," both these alternative information can achieve the same goal behind the query, i.e., "solve the sleeping problem." To tackle the problem, we propose leveraging a community Q&A (cQA) corpus for mining alternative actions. We propose a method to compute how well two actions can be alternative actions by using

a question-answer structure in a cQA corpus. We conducted experiments to investigate the effectiveness of our method using two newly built test collections, each containing 50 queries. The experimental results indicated that our proposed method significantly outperformed two types of baselines, one used the conventional query suggestions and the other extracted alternative-actions from the Web documents, in terms of D#-nDCG@8.

- Chapter 4
  This chapter examined the extent to which *opinion* and *credibility* encouraged or discouraged searchers from searching carefully. By understanding the effect of such factors to the search behaviors, we would be able to design a search system that encourages the people to consider alternative information and search carefully. We conducted a user study in which 260 participants were asked to perform health-related search tasks while controlling a search result with different opinions and credibility levels. The results revealed that search engines could prevent users from polarization and thus, help them to obtain accurate information, by presenting documents that are inconsistent with users' beliefs on the higher-rank of the results.
- Chapter 5
  This chapter proposed the investigation on the impact of collective questions and answers displayed on Search engine result pages (SERP), known as *People also ask* on searchers' behaviors and beliefs. Two experiments were conducted in which participants were asked to perform health-related search tasks. In both experiments, items in *People also ask* were manipulated. Experiment 1 focused on the effect of *question*, *answer*, and answer's *opinion*. Experiment 2 focused on the effect of the *alternative question*, i.e., a question related to the solution that can achieve the same goal as a query. The findings suggest that *People also ask* might not help mitigate confirmation bias as participants are likely to spend less effort on the search process (i.e., issue fewer queries) when they first encounter a belief-inconsistent answer unlike when they encounter a belief-inconsistent document within the search results. An additional experiment is required to validate that participants who first encounter a belief-inconsistent answer are more likely to alter their beliefs as the number of such participants was inadequate.
- Chapter 6
  We summarize the research addressed in this thesis and offer several directions to be explored in future work.

2 Related Work

2.1 Task-Oriented Web Search

Hassan *et al.* studied on supporting the *complex search task*, in which a searcher has to accomplish several subtasks to satisfy his/her information need [Hassan and White 2012, Hassan Awadallah et al. 2014]. They proposed a method that includes grouping queries into the same task through the query log mining and query syntactic analysis. Wang *et al.* proposed a method for extracting task names from the microblog corpus for supporting the complex search task [Wang et al. 2014]. Jones and Klinkner proposed the *mission-goal* hierarchical relationship [Jones and Klinkner 2008] between information needs, and proposed a method for classifying a pair of queries into the same mission/goal (referred to as *task* in their study) or not. Aiello *et al.* also proposed a clustering algorithm that clusters missions into underlying topics [Aiello et al. 2011]. Although, in the present study, a hierarchical relation is assumed between actions as in these previous studies; our study focuses on the users' real-world behavior rather than on other types of searches such as covering many aspects of a topic.

The studies that are most relevant to the present study are those by Yamamoto *et al.* [Yamamoto et al. 2012] and Yang *et al.* [Yang and Nyberg 2015]. Yamamoto *et al.* defined the *goal-subgoal* relationship and proposed a method for clustering queries into subgoals by leveraging the sponsored search data. Yang *et al.* defined the *task-subtask* relationship and proposed a method for connecting search queries with task descriptions written in wikiHow[1]. Although they used the different terminologies for defining the hierarchical relationship, both definitions were based on the *is-achieved-by* relationship. In this study, we also use the is-achieved-by relation to define alternative actions. Recently, TREC and NTCIR attempted to tackle the task-oriented Web search [Liu et al. 2014, Yilmaz et al.]. For example, in the TaskMine subtask of the NTCIR-11 IMine task, a system was asked to retrieve a set of subtasks for a given task.

The key difference between the above studies and ours is that most of the existing studies focused on finding *sub* tasks for a given query. For example, given the query "lose weight," the desired outputs are "do physical exercise" and "control calorie intake," each of which can achieve the query [Liu et al. 2014]. On the other hand, we attempt to find alternatives to a given query (See Section 3.2.3). This enables us to suggest a searcher with other solutions for solving his goal, which has not been addressed by the existing work.

[1] http://wikihow.com

2.2 Connecting cQA with Web Search

Recent studies indicated that a cQA corpus can be used to improve the performance of a Web search. Adi *et al.* [Omari et al. 2016] proposed a method for ranking answers of a cQA corpus based on their novelty, in order to improve Web search by displaying the answer in SERP. Yamamoto *et al.* [Yamamoto et al. 2011] used a cQA corpus as a resource for extracting adjective facets and use them as query suggestions to support Web searchers.

Liu *et al.* extensively analyzed the behavior logs obtained from a Web search engine and a cQA service, and revealed the typical patterns when a Web searcher gives up his/her search and asks a question in the cQA service. According to their study, a searcher who issues a query containing terms such as "how," "can," or "do," *etc.*, tended to ask a question. Another study showed that one popular type of questions on a cQA service is a how-to question [Harper et al. 2009]. With regard to these studies, Webner *et al.* focused on a Web search query related to how-to information, and proposed a method for extracting its answer from the cQA corpus [Weber et al. 2012]. These results suggest that a cQA corpus contains much how-to information, which can be effectively used for mining alternative actions.

2.3 Comparative Entity Mining

Suggesting alternative actions to a searcher may allow his/her to *compare* several solutions for making an appropriate decision. Some researchers tackled the problem of finding comparative entities from the data. Jindal and Liu studied on identifying comparative sentences in a text corpus [Jindal and Liu 2006]. Shasha *et al.* [Li et al. 2010] extended their work and proposed a method to extract comparative entities from the questions in Yahoo! Answers. Tsukuda *et al.* [Tsukuda et al. 2014] proposed a method to extract co-ordinate entities, which share one or more common hypernyms with a given query by using the hypernym-hyponym dictionary. The focuses of the above work were on finding comparative *entities* (e.g., "iPod" and "PSP" [Li et al. 2010] or "Cristiano Ronaldo" and "Lionel Messi" [Tsukuda et al. 2014]), whereas our work deals with *actions* (e.g., "take a sleeping pill" and "stroll before bedtime"). For some task-oriented searches, suggesting comparative entities are firmly beneficial to a searcher since he/she can find another solution to achieve his/her goal. However, as in the example of "sleeping pills" mentioned in Section 3.1, there are task-oriented searches that cannot be wholly solved by suggesting the comparative entities.

2.4 Confirmation Bias

Confirmation bias is the phenomenon where people seek information that confirms their predispositions without considering contradictory opinions [Jonas et al. 2001, Nickerson 1998]. The effect of confirmation bias can be explained by the defense mechanism, cognitive dissonance, in which people prefer supportive information over opposing information to avoid mental discomfort [Festinger 1957]. This kind of behavior may lead people to make

disastrous decisions because they may seek the information that deviates from the truth, or they may overlook the risks of confirmatory information [Janis 1982, Nemeth and Rogers 1996, White 2013].

In terms of social interactions, previous studies have suggested that people tend to ask questions to which a positive answer is considered as strongly confirmatory and a negative answer is considered as weakly disconfirmatory [Dardenne and Leyens 1995, Skov and Sherman 1986, Snyder and Swann 1978, Swann et al. 1982]. In addition, typically, people prefer confirmatory answers over disconfirmatory answers [Heritage 2013].

2.5 Credibility

In this study, we follow the definition of credibility by Fogg, who simply defined it as the *believability* of information and its sources [Fogg 2003]. Fogg and Tseng proposed a taxonomy on the credibility of web information. It is comprised of four categories based on items used to judge the credibility of information: presumed credibility, earned credibility, surface credibility, and reputed credibility [Fogg and Tseng 1999]. We employed surface credibility to prepare documents with different credibility levels (See Section 4.2.3). To evaluate information carefully, credibility awareness is essential. Metzger et al. found that credibility awareness impacts information seeking behaviors in which people are likely to be a critical information consumer [Metzger et al. 2015]. Therefore, we considered credibility to be a major factor that could encourage people to search for information more carefully.

2.6 Question Answering System

Question answering systems have received significant attention in the information retrieval community. With a question answering system, searchers promptly obtain relevant information rather than going through a large number of web documents [Lin et al. 2003, Voorhees and Tice 1999]. Recently, major search engines have started to present a question answering system called *People also ask* in response to some queries where questions are selected by the search engine as a function of frequency the questions have been asked [Matias et al. 2017]. Although previous studies have explored the effect of question answering systems on search performance (i.e., decreased number of queries) [Belkin et al. 2000, Bernstein et al. 2012, Chilton and Teevan 2011, Lin et al. 2003], such studies did not consider the effect of question answering systems to confirmation bias in a search.

2.7 Bias, Credibility, and Search Behaviors

Previous studies found that people's search behaviors and decisions can be affected either by the bias in the search results or by the credibility of the sources. As for the bias in the search results, White and his colleagues [White 2013, 2014, White and Hassan 2014, White and Horvitz 2015] found that people's confirmation bias can be exacerbated during searches; they found that people's decisions after searching were more biased towards a medical remedy was

helpful even though there was evidence that it was not helpful. Moreover, they found that it was difficult to change people's beliefs after their decisions has been made. In addition to White's findings, Pogacar et al. [Pogacar et al. 2017] recently found that people were likely to make incorrect decisions when the search results were biased towards incorrect information. As for credibility, Kammerer et al. [Kammerer et al. 2013] found that people were more likely to spend less effort evaluating search results when the information sources seemed credible. Metzger [Metzger 2007] found that people were likely to believe visually aesthetic web information due to their limited time and cognitive capacity.

In order to mitigate confirmation bias and make more effective decisions, people should be exposed to alternative or opposing information that they might not have chosen previously [Buchs et al. 2004, Hernandez and Preston 2012, Nemeth et al. 1992, Sunstein 2017]. In line with this ideology, previous studies have found that recommending information inconsistent with the user's preference can stimulate divergent thinking and consideration of diverse information [Liao and Fu 2014, Liao et al. 2015, Schwind and Buder 2012, Schwind et al. 2012]. On the other hand, studies in pragmatics found that people tend to select questions that deviate from their hypothesis if such questions are provided [Trope and Bassok 1982, 1983]. Our study considered and extended these studies in the following ways. First, while the previous studies [Liao and Fu 2014, Liao et al. 2015, Schwind and Buder 2012, Schwind et al. 2012] revealed that the people's confirmation bias can be mitigated by presenting information inconsistent with their preference, the detailed relationship between their search behaviors and decisions when encountered with inconsistent information during a search is still unclear. Second, we consider both the opinion of a search result and its credibility, which enables us to better understand their relationship. Third, we also consider the relationship between people's belief dynamics and their detailed search behaviors to better understand how their search behaviors relate to their decisions. Forth, we investigate whether questions and answers can encourage careful information behaviors in the search context. Finally, while previous studies [Liao and Fu 2014, Liao et al. 2015, Schwind and Buder 2012, Schwind et al. 2012] revealed that presenting information that is inconsistent with people's preference can help mitigate the confirmation bias, this study wants to determine if such an effect persisted when contradictory information is presented as answers.

3 Mining alternative information

3.1 Introduction

In task-oriented search, the searcher faces the problem that he/she may not be aware of another existing solution that could help achieve the same goal behind the query. For example, for the searcher issuing the query "sleeping pills," other solutions such as "have a cup of hot milk" or "stroll before bedtime" exist as well, which can also help resolve the "solve his/her sleeping problem." Since the searcher often believes in the mean he/she initially comes up with, he/she may decide to take a sleeping pill without considering the other solutions that can solve the same problem. Although the current search engines provide query suggestions for supporting a searcher to reformulate his/her query, it is difficult for the searcher to find alternative solutions.

In this chapter, the *alternative action mining* problem is discussed, where a system is required to find alternative actions for a given query. An alternative action for a query is defined as an action that can solve the same problem (See Section 3.2). For example, given the query "sleeping pills," the objective is to find alternative actions such as "have a cup of hot milk" or "stroll before bedtime," both these alternative actions can achieve the same goal behind the query, i.e., "solve the sleeping problem." Mined alternative actions can be utilized for supporting a searcher in a task-oriented Web search. For example, by suggesting the alternative actions to the searcher issuing the query "sleeping pills," he/she is able to notice different solutions and make an improved decision on how to solve his/her sleeping problem.

We think that suggesting alternative-actions to a searcher is important in two situations. First is a situation in which a searcher lacks of the enough knowledge about the problem and only knows a few specific solutions. For such a searcher, providing the alternative-solutions may provide her new knowledge about the problem and an opportunity to explore diverse solutions before making a decision on choosing an actual solution to solve her problem. Second is a situation where a searcher already has the strong belief about the solution and does not consider the possibility of the other solutions even though she knows the existence of the other solutions. For example, a user may issue the query "sleeping pills" because she strongly believes that using sleeping pills is effective for solving the problem even though she knows that other solutions such as "stroll before bedtime" or "have a cup of hot milk" might solve the problem. We think that, by suggesting alternative-actions to her, we may raise

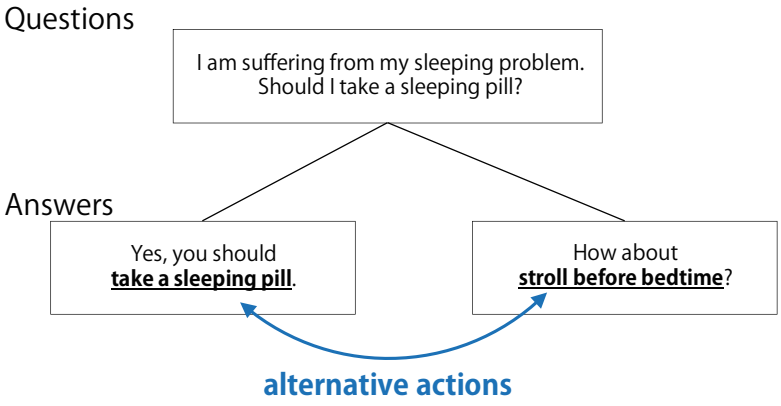

Figure 3.1: Example question-answer pairs in cQA corpus. Our method extracts alternative actions by using question-answer structure.

her awareness of the existence of the other solutions and encourage her to explore the other solutions instead of the solution she believed.

To tackle the alternative action mining problem, we propose leveraging a community Q&A (cQA) corpus. CQA services like Yahoo! Answers[1] or Baidu Zhidao[2] are widely used by people for solving their problems by communicating with other users. We hypothesize that the cQA corpus can be seen as an archival dataset comprising dialogues between questioners, who want to know the solutions to their problem, and respondents, who suggest good solutions for it. Figure 3.1 shows an example of a question-answer pair in a cQA corpus. The fundamental idea of using a cQA corpus is that, as can be seen in the figure, the two actions "take a sleeping pills" and "stroll before bedtime" are proposed by the respondents to satisfy the same goal of a questioner, which means "take a sleeping pills" and "stroll before bedtime" can be alternative actions. We also propose a method for computing how well two actions can be alternative actions using the question-answer structure of a cQA corpus. Our method constructs a question-action bipartite graph from a set of question-answer pairs and recursively computes how well two actions can be alternative actions (See Section 3.3).

We prepared two test collections, each containing 50 queries, for our evaluation. The experimental results using the test collections showed that our method outperformed the conventional query suggestions provided by the commercial search engines in terms of D#-nDCG.

[1] https://answers.yahoo.com/

[2] https://zhidao.baidu.com/

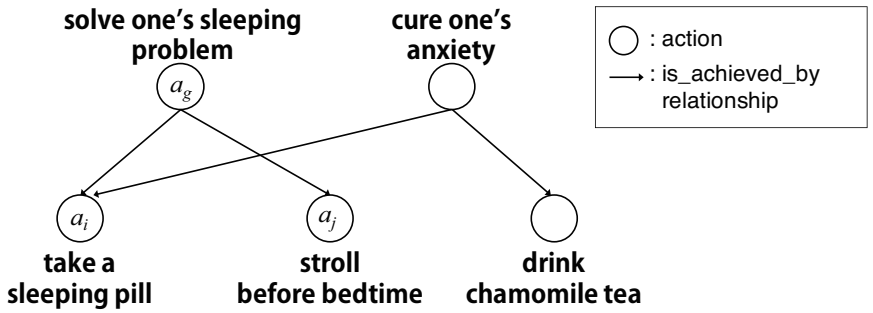

Figure 3.2: Example structure among actions. In the figure, actions a_i (*take a sleeping pill*) and a_j (*stroll before bedtime*) are called *alternative actions* when they share the same action a_g (*solve one's sleeping problem*) as their goal. Note that the actions "take a sleeping pill" and "drink chamomile tea" are also alternative actions, since they share the other action "cure one's anxiety."

3.2 Problem Definition

In this section, we define the *alternative action mining* problem addressed by the present study.

3.2.1 Alternative Action

Definition 1 (action): An *action* is an activity that a user wants to achieve. In our study, we represent an action as a verbal phrase, as in the work [Yamamoto et al. 2012]. For example, "take a sleeping pill," "have a cup of hot milk," and "solve one's sleeping problem" are actions.

Definition 2 (is-achieved-by relationship): For two actions a_i and a_g, we call a_g *is-achieved-by* a_i when achieving a_i helps to achieve a_g. Figure 3.2 illustrates the example actions for the is-achieved-by relationship. In the figure, action a_g is-achieved-by a_i since achieving "take a sleeping pill" helps to achieve "solve one's sleeping problem." For the convenience, in this study, we also call "a_g is a *goal* of a_i", whose meaning is "a_g is-achieved-by a_i."

Definition 3 (alternative actions relationship): For two actions a_i and a_j, we call a_i and a_j are *alternative actions* when they are different actions and there exists at least one other action a_g which is their common goal. As shown in Figure 3.2, the actions "take a sleeping pill" and "stroll before bedtime" are alternative actions since they share the same goal "solve one's sleeping problem." Also, actions "take a sleeping pill" and "drink chamomile tea" are alternative actions since they share the other goal "cure one's anxiety."

3.2.2 Alternative Action Mining Problem

As shown in Figure 3.2, the action "take a sleeping pill" can be used to achieve two different goals. Thus, issuing the query "sleeping pills" is hard to predict which goal the searcher wants to achieve, i.e., "solve his/her sleeping problem" or "cure his/her anxiety," and the desired alternative actions depend on it. To solve this ambiguity, we follow an approach similar to the one used in the search result diversification [Liu et al. 2014, Yamamoto et al. 2016], where, for a given query, the system is required to generate a *diversified* ranked list of documents satisfying as many different *search intents* behind the query as possible. The alternative action mining problem is defined as follows:

Alternative action mining problem: Given a query q, the alternative action mining problem refers to returning a diversified ranked list of k alternative actions $a_1, a_2, \ldots, a_k$ for that query, which can satisfy as many different goals of searchers issuing q.

Note that in this study we do not limit our query as a verbal phrase, and accept any form of a query as an input of our method. In this work, we assume that any query issued by a searcher implies its corresponding action. For example, for the query "sleeping pills," we assume that the implied action for the query is "take sleeping pills." So, for the query "sleeping pills", our objective is to automatically generate a ranked list of actions (e.g., "stroll before bedtime", "drink chamomile tea"), which can satisfy two different goals "solve one's sleeping problem" and "cure one's anxiety," of the action implied by the query "sleeping pills." The reason why we accept all forms of queries is that we think people may not use a verbal phrase such as "take sleeping pills" as a query when using a Web search engine and noun queries such as "sleeping pills" is much popular. Note that while our method accepts any forms of queries, its outputs (i.e., alternative-actions) are the form of the verbal phrase.

3.2.3 Relation to Existing Work

To clarify our problem, Figure 3.3 illustrates the relation of our problem to the existing work [Liu et al. 2014, Yamamoto et al. 2012, Yang and Nyberg 2015, Yilmaz et al.] in task-oriented Web search. As can be seen in the figure, most existing studies addressed the problem of finding *sub*-tasks (or *sub*-goals) of a given query. For example, given the query "sleeping problem", many of the existing studies were focused on finding "take a sleeping pill" or "stroll before bedtime", both of which can achieve the goal behind the query "solve one's sleeping problem." Such work can be seen as a method for supporting searchers to find more concrete solutions for achieving the action represented by the query.

In contrast, our study is different from these previous studies in a way that ours is to help searchers to find alternative solutions for the goals behind the query. We believe that our study provides another type of supports for a searcher in a task-oriented Web search and is as important as the existing studies.

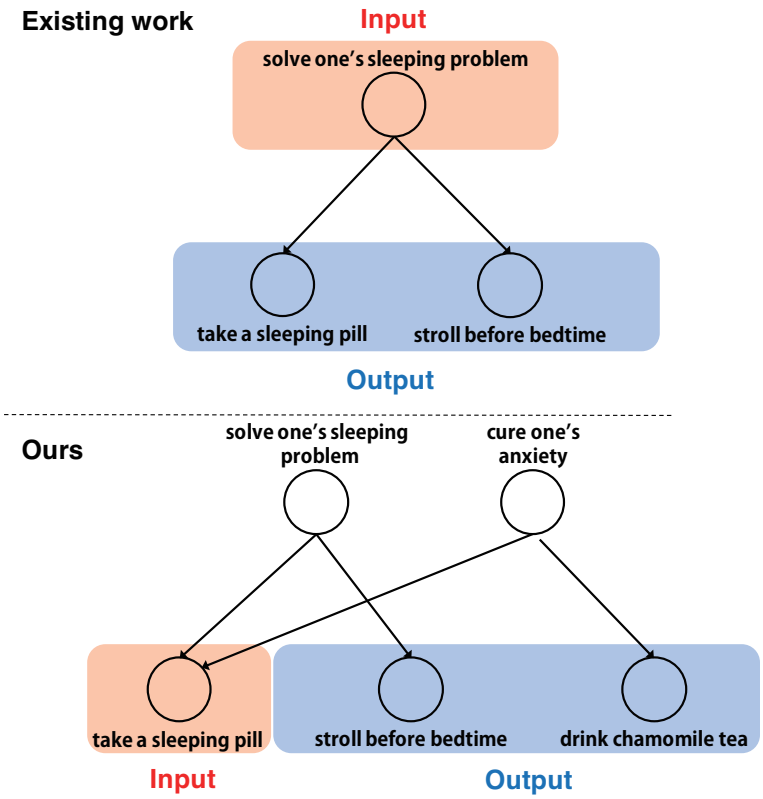

Figure 3.3: Comparison between existing work [Liu et al. 2014, Yamamoto et al. 2012, Yang and Nyberg 2015, Yilmaz et al.] and our study. Most existing studies addressed the problem of finding *sub* actions, whereas our study finds *alternative* ones.

3.3 Approach

In this section we explain our proposed method, which utilizes a cQA corpus to automatically find alternative actions to a given query.

3.3.1 Technical Challenge

One key challenge of the alternative action mining problem is measuring how well two actions a_i, and a_j are alternative actions. In other words, we have to compute how well the two actions can achieve the same goal. In this study, we refer to the strength of this relationship as *alternativeness* between a_i and a_j, denoted as $\mathrm{alt}(a_i, a_j)$. If we could measure the alternativeness between query q and action a_i, $\mathrm{alt}(q, a_i)$, the rank of the action for the query should be basically determined by its value. The difficulty in measuring the alternativeness

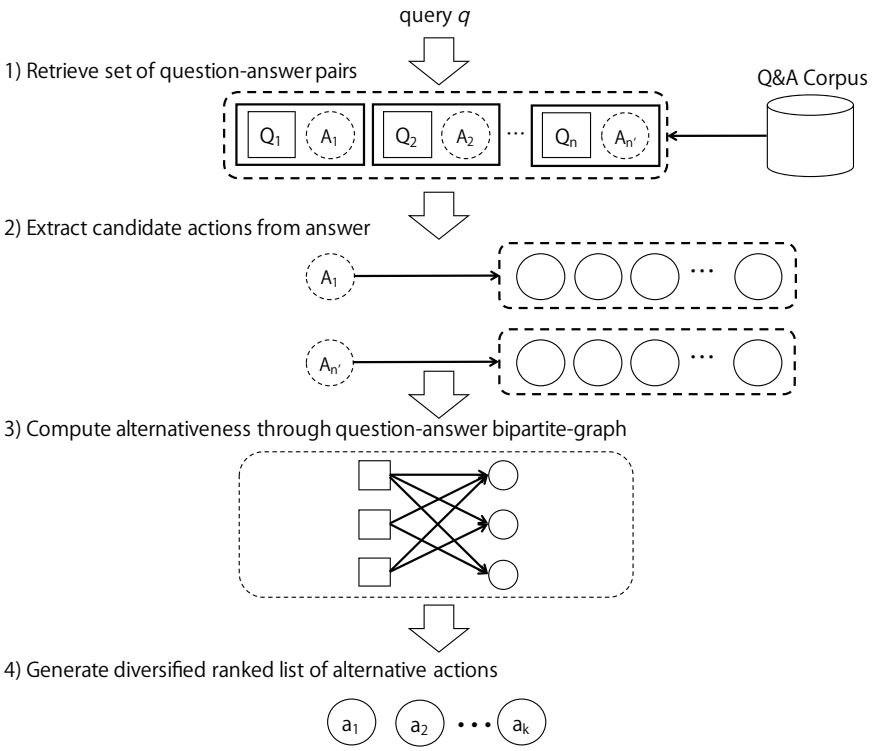

Figure 3.4: Method overview.

is that using textual similarity (e.g., the Levenshtein distance) or semantic similarity (e.g., word embedding [Mikolov et al. 2013]) is not effective. For example, the two actions "take a sleeping pill" and "stroll before bedtime" should have high alternativeness even if they are neither textually nor semantically similar.

3.3.2 Method Overview

Given a query q, our objective is to retrieve a ranked list of k alternative actions $a_1, a_2, \ldots a_k$ that satisfy as many goals behind the query as possible. Figure 3.4 shows the overview of our method. Given a query q, our first step is to retrieve a set of question-answer pairs related to the query from the community Q&A corpus. Then, we extract the candidate actions from the retrieved answers. Our third step, which is the core of our method, is to compute the alternativeness between actions through the question-action bipartite graph. Finally, we apply the search result diversification algorithm to the actions and generate a diversified ranked list of alternative actions for the query.

Table 3.1: Manually predefined terms for retrieving question-answer pairs.

Terms
effect, should I, disadvantage, try, want to,
worst, need to, begin, beginner, why, risk of, prefer, alternative

3.3.3 Retrieve Question-Answer Pairs

Given a query q, we first retrieve a set of question-answer pairs from a cQA corpus. We hypothesize that some questions are likely to receive many suggestions by the respondents because of their content. For example, when a question contains "*Should I* use sleeping pills?" in its text, its answers will likely to contain many actions other than "take a sleeping pill" because the questioner is unsure about his/her idea. Thus, retrieving such questions may help us find alternative actions from their answers.

To this end, we manually prepare some terms that are likely to indicate that a questioner is unsure about his/her idea. Table 3.1 lists up the terms we prepared. By using these terms we retrieve questions and answers related to the query. More specifically, we first retrieve answers containing q. We then obtain the questions from these answers. Next, we rank the questions by using the terms show in Table 3.1 by the Okapi BM25 algorithm and obtain the top n questions ($n = 10,000$ in our experiments). Finally, we retrieve all the answers of these questions, and obtain a set of question-answer pairs for the n questions.

3.3.4 Extract Candidate Actions from Answers

After we obtain the set of question-answer pairs, we extract candidate actions from their answers by applying the standard text-chunking approach using Conditional Random Field (CRF) to extract verbal phrases from the answers. We prepare 500 sentences by sampling answers in the cQA corpus, and annotate the verbal phrases in the sentences. We then learned the classifier which classifies terms in a sentence into a verbal phrase or not. We use the standard 19 features for text chunking [Okazaki 2007] including bag-of-words and parts-of-speech of the target word and its surroundings. We apply the learnt classifier to the texts of the answers and extract a set of actions.

In this step, different verbal phrases are treated as different actions even though they are semantically similar. For example, when CRF outputs "do exercise" and "do exercise everyday" from the answers, we regard that we obtain two different actions "do exercise" and "'do exercise everyday." Also, when CRF outputs two identical verbal phrases "do exercise" and "do exercise" from the answers, we regard that we obtain one action "do exercise." We expect that semantically similar actions such as "do exercise" and "do exercise everyday" can be removed by applying the MMR algorithm described in Section 3.3.4.

3.3.5 Measure Alternativeness between Actions

Obtaining the set of actions from the previous step, we compute the alternativeness between actions. As we mentioned in Section 3.3.1, it is hard to compute the alternativeness between two actions simply using the usual textual or semantic similarity. To compute the alternativeness between two actions we make the following two hypotheses:

- **H1 (question → action):** If two questions are likely to *represent the same goal*, actions in their answers are likely to be *alternative actions*.
- **H2 (action → question):** If two actions are likely to be *alternative actions*, questions of the answers containing these actions are likely to *represent the same goal*.

Take, as an example, two different questions "*I am suffering from my sleeping problem. What should I do*" and "*How can I sleep well?*" We can expect that answers to these different questions are intended to satisfy the same goal, which we can obtain **H1**. Also, for two answers containing the different actions "take a sleeping pill" and "have a cup of hot milk", we can expect that their questions are likely to address the same problem, which we can obtain **H2**.

Since **H1** and **H2** are recursive – the alternativeness between two actions depends on how likely two questions represent the same goal, and this depends on the alternativeness between actions in their answers – we apply the SimRank algorithm [Jeh and Widom 2002], which is designed to compute the similarity between nodes on a graph, on the question-action bipartite graph. We first prepare the question-action bipartite graph from the questions and actions extracted in the previous steps (shown in Figure 3.4, step 3). Let $Q = \{Q_i\}_{i=1}^{n}$ be the set of questions retrieved by the step described in Section 3.3.3, $A = \{a_j\}_{j=1}^{m}$ be the set of actions extracted in the step described in Section 3.3.4, and $\mathcal{A} = A \cup \{q\}$ be the union of these actions and query, we construct a bipartite graph $G = (Q \cup \mathcal{A}, E)$, where $E \subseteq Q \times \mathcal{A}$ and edge $e_{ij} = (Q_i, a_j) \in E$ in G represents that action a_j appears in at least one answer of question Q_i.

Let $\mathrm{sim}_{\mathrm{goal}}(Q_i, Q_j)(Q_i, Q_j \in Q)$ represent how well two questions represent the same goal, and $\mathrm{alt}(a_i, a_j)(a_i, a_j \in \mathcal{A})$ represent how well two actions are alternative actions. If $Q_i = Q_j$, the initial value for $\mathrm{sim}_{\mathrm{goal}}(Q_i, Q_j)$ is set to 1, otherwise the initial value for $\mathrm{sim}_{\mathrm{goal}}(Q_i, Q_j)$ is 0. We use the same condition to initialize the value for $\mathrm{alt}(a_i, a_j)$. After we assign the initial values to $\mathrm{sim}_{\mathrm{goal}}(\cdot,\cdot)$ and $\mathrm{alt}(\cdot,\cdot)$, we update these two measures by iteratively computing the following two formulae:

$$\mathrm{sim}_{\mathrm{goal}}(Q_i, Q_j) = \frac{C}{|O(Q_i)||O(Q_j)|} \sum_{k=1}^{|O(Q_i)|} \sum_{l=1}^{|O(Q_j)|} \mathrm{alt}(O_k(Q_i), O_l(Q_j)) \,, \tag{3.1}$$

$$\mathrm{alt}(a_i, a_j) = \frac{C}{|I(a_i)||I(a_j)|} \sum_{k=1}^{|I(a_i)|} \sum_{l=1}^{|I(a_j)|} \mathrm{sim}_{\mathrm{goal}}(I_k(a_i), I_l(a_j)) \tag{3.2}$$

where C is a constant value and $O(Q_i)(\subseteq \mathcal{A})$ is a set of out-neighbors of Q_i and $I(a_i)(\subseteq Q)$ is a set of in-neighbors of a_i. We use $C = 0.8$ and the values of $\mathrm{alt}(\cdot,\cdot)$ obtained after the five iterations, as suggested in [Jeh and Widom 2002].

From the alternativeness between query q and action a_i $\mathrm{alt}(q,a_i)$, we can know that how well an action a_i can be the alternative actions for a query q, which is used to determine the relevance of the action to the query. Moreover, the alternativeness between two actions $\mathrm{alt}(a_i,a_j)$ indicates how well two actions share the same goal; high $\mathrm{alt}(a_i,a_j)$ means they are similar in terms of their goals and low $\mathrm{alt}(a_i,a_j)$ means they are dissimilar. This information can be used for the diversification of the ranked list.

3.3.6 Measure Effectiveness of Action by Community Evaluation

To improve the performance of measuring the relevance between query and action, we also propose utilizing the quality of answers evaluated by the community in the service. Most cQA services enable their users to evaluate the quality of answers by, for e.g., selecting best answers or up-voting good answers. Our idea is that such evaluations by the community can help find actions that many people believe in their effects.

We compute the effectiveness of action a_i as the probability that an answer containing a_i be selected as a best answer:

$$\mathrm{effect}(a_i) = \frac{|\mathrm{BestAnswers}(a_i)| + \theta}{|\mathrm{Answers}(a_i)| + 2\theta}\ , \tag{3.3}$$

where $\mathrm{BestAnswers}(a_i)$ and $\mathrm{Answers}(a_i)$ represent the set of best answers and answers in the question-answer pairs retrieved by the step in Section 3.3.3, respectively, and θ is the Laplace smoothing parameter ($\theta = 8$ in our experiments).

3.3.7 Generate Diversified Ranked List

Once we compute the alternativeness between actions and their effectiveness, we generate a ranked list of actions. As described in Section 3.2, the purpose of the ranked list is to achieve as many different goals behind the query as possible. To this end, we apply the result diversification technique to diversify the ranked list. We apply the Maximal Marginal Relevance (MMR) algorithm [Carbonell and Goldstein 1998] to generate the diversified ranked list of actions. MMR iteratively chooses the relevant items considering both the relevance and diversity. Letting $A = \{a_j\}_{j=1}^{m}$ be the set of candidate actions to be ranked, MMR selects a^r, an action ranked at the r-th position using:

$$a^r = \underset{a \in A_q \setminus S^{r-1}}{\arg\max} \left[\lambda \cdot \mathrm{rel}(q,a) - (1-\lambda) \max_{a' \in S^{r-1}} \mathrm{alt}(a,a') \right]\ , \tag{3.4}$$

where

$$\mathrm{rel}(q,a) = \alpha \cdot \mathrm{alt}(q,a) + (1-\alpha) \cdot \mathrm{effect}(a)\ . \tag{3.5}$$

S^{r-1} denotes a set of $r-1$ actions that MMR has already selected, λ is a parameter balancing the relevance and the diversity, and α is a parameter balancing the alternativeness and the effectiveness. By applying the MMR algorithm, we obtain the diversified ranked list of k alternative actions for the query.

3.4 Experimental Setup

We address the following research questions by conducting the experiments: (1) Does our proposed method outperform the query suggestions provided by the commercial search engines in terms of the providing alternative actions for a query? (2) Can our method work for the cQA corpora on different services? (3) How do the parameters λ, which balance the relevancy and diversity, and α, which combine the alternativeness and effectiveness, affect the performance? (4) For what kinds of queries does our method work effectively? We prepared two test collections for Japanese and English, which we refer to **JaCollection** and **EnCollection**.

3.4.1 Dataset

We use the corpus archived in Yahoo! Chiebukuro[3], which is the most popular community Q&A service in Japan, for JaCollection. We build the search system on Elasticsearch to retrieve questions and answers from the corpus. We also use the data archived in Reddit[4] as another cQA corpus for EnCollection to investigate whether our method works on different data. Reddit is one of the most popular online communities, where users communicate by making posts and giving comments to them. We use the APIs[5] provided by Reddit to retrieve posts and their comments.

One difference between Yahoo! Chiebukuro and Reddit, which affects our method, is that Yahoo! Chiebukuro allows users to vote for the best answer whereas Reddit does not have option. Instead, Reddit allows users to provide a positive or negative voting to a comment. Thus, when computing Equation (3.3) for the Reddit data, instead of computing the best answer probability we compute the probability that a comment receives positive votes.

3.4.2 Proposed and Baseline Methods

To measure the effectiveness of our method, we prepare the following methods:

- **Query Suggestion (QS)**: We extract the query suggestions from a Web search engine to investigate whether the current query suggestions provide alternative actions. We use the query suggestions provided by the two commercial search engines, which we refer to as **QS1** and **QS2**, respectively. The reason why we use the query suggestions as our

[3] http://chiebukuro.yahoo.co.jp/

[4] https://www.reddit.com/

[5] https://www.reddit.com/dev/api/

baselines is we expect that some of the query suggestions are "parallel move" of the query. According to the study by Boldi *et al.* [Boldi et al. 2009], parallel move is one type of query reformulation in which a searcher reformulates her query from one aspect of a topic to something related but not equivalent (e.g., from "kyoto travel" to "osaka travel"). As query suggestions in part relying on the query log of searchers, we expected that "parallel move" query suggestions can be alternative-actions of a query.

- **RelDivWeb**: This is another baseline method, which extracts alternative-actions from the Web documents. First, it extracts the top 500 Web search results from the Bing Web search API[6] and obtained the contents of the 500 Web documents. We used the existing Python libraries[7] to extract the main contents of the Web documents. We then applied the same CRF classifier trained in Section 4.4 to extract the set of candidate actions $A = \{a_1, \ldots, a_m\}$ from the main contents of the k Web documents. Finally, we ranked the actions by applying the MMR algorithm. MMR selects a^r, an action ranked at the r-th position using:

$$a^r = \arg\max_{a \in A \setminus S^{r-1}} \left[\lambda \cdot \mathrm{rel}(q,a) - (1-\lambda) \max_{a' \in S^{r-1}} \mathrm{sim}(a,a') \right] , \tag{3.6}$$

where

$$\mathrm{rel}(q,a) = \frac{\mathrm{freq}(a)}{\max_{a_i \in A} \mathrm{freq}(a_i)} , \tag{3.7}$$

$$\mathrm{sim}(a,a') = \frac{|D(a) \cap D(a')|}{|D(a) \cup D(a')|} . \tag{3.8}$$

In the above equations, $\mathrm{freq}(a)$ represents the frequency of an action a in the top 500 Web documents, and $D(a)$ represents the set of Web documents containing action a. Since a Web document does not have the hierarchical structure like a question-answer pair in cQA, we decided to use the frequency of an action for measuring relevance and co-occurrence for measuring similarity between actions. For parameter λ, we used the optimum value for EnCollection when evaluating JaCollection, and used the one for JaCollection when evaluating EnCollection.

- **RelOnlyWeb**: This is the same as RelDivWeb, except that we set $\lambda = 1.0$. Thus, RelOnlyWeb only considers the relevance but not the diversity of the ranked list.
- **RelDivQA**: This is our proposed method which generates a ranked list of actions based on Equations (3.4) and (3.8). Equations (3.4) and (3.8) contain two parameters λ and α. To fairly compare with the baselines and our method, we use the optimum

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

[7] https://github.com/yono/python-extractcontent (for Japanese documents), https://github.com/buriy/python-readability (for English documents)

Table 3.2: Example queries used in experiment (EnCollection).

Domain	**Queries**	
Health	kettlebell workout	acupuncture
	chamomile tea	pilates
Recreation	airbnb	uber taxi
	cheap flight	youth hostel
Education	coursera	vocational school
	public university	free certification

λ and α for EnCollection when evaluating JaCollection, In addition, we use the ones for JaCollection when evaluating EnCollection. The effects of these parameters are investigated in Section 3.5.2.

- **RelOnlyQA**: This is the same as RelDivQA, except that we set $\lambda = 1.0$. Thus, RelOnlyQA only considers the relevance but not the diversity of the ranked list.

3.4.3 Test Collection Construction

JaCollection used Yahoo! Chiebukuro and EnCollection used Reddit as the cQA corpus. We first prepare 50 queries to be used as the input to alternative action mining. We chose three domains (Health, Recreation and Education) to select queries. The reason why we chose these domains is, according to Donato *et al*., information needs in domains such as travel, health and education tend to be complex [Donato et al. 2010]. Hence, these are typical domains where it is important to show alternatives to a searcher. Note that these domains were also used in previous studies [Liu et al. 2014, Yamamoto et al. 2012] Table 3.2 shows example queries we use for EnCollection. Both test collections contain 23 queries from health, 14 from recreation and 13 from education.

We view our alternative mining problem as similar to the search result diversification. To evaluate our method, we need the following ground truth:

- **A set of goals $G^q = \{g_1^q, \ldots g_n^q\}$ for query q**, where n is the number of goals for q. E.g., for query "sleeping pills," $G = \{$"solve one's sleeping problem", "cure one's anxiety"$\}$.
- **Goal-level action relevance** $\mathrm{rel}^q(a, g)$, which represents how well an action a is an alternative action to the query q in terms of achieving its goal g.

For preparing the set of goals for each query, an assessor is asked to search the Web to familiarize with the query, and then to write down up to three goals in a verbal phrase representation. When writing down the goals for a query, the assessor was required to prepare

the goals for the action implied by the query. For example, for the query "sleeping pills", if the assessor guessed its implied action as "take sleeping pills", the assessor prepared the goals for the action "take sleeping pills."

In order to prepare goal-level action relevance, three assessors for each language are used in this experiment. We first pool the results of both baseline and proposed methods at the pool depth size at 10. For the proposed method, we generate the ranked result for each combination of the two parameters λ and α, by changing their parameters from 0.1 0.2, ..., to 1.0. Then, for each query, an assessor was asked to annotate goal-level action relevance for the pooled actions. The annotation is conducted in the following step. For each (query, goal, action), we ask the assessors to annotate its relevance with three-graded scores according to the following criteria:

- **highly relevant (2)**: action *a* strongly helps to achieve goal *g*, and also *a* is another solution which differs from the action represented by the query itself.
- **relevant (1)**: action *a* may help to achieve goal *g* and also *a* is another solution which differs from the action represented by the query itself.
- **irrelevant (0)**: otherwise. The cases when an action receives a relevance value of 0 are (a) the action does not help to achieve the goal, or (b) achieving the action solves the action indicated by the query, which means the action does not provide any alternative. (e.g., for the query "sleeping pill", "take Xanax" or "Xanax" (Xanax is a popular sleeping pill) was assigned the relevance of 0 since taking a Xanax achieves taking a sleeping pill).

Note that QS1 and QS2 often return a noun as their output. For such an output, the assessors were required to make a relevance judgement based on the action implied by the output, similar to when preparing the goals for noun queries.

The Fleiss' kappa coefficients [Fleiss and Cohen 1973] for these relevance judgments among three assessors were 0.290 (JaCollection) and 0.565 (EnCollection). We found that the assessors' judgments varied between relevances of 1 and 2, thus we decided to use the binary relevance rather than the graded relevance in the evaluation. The Fleiss' kappa coefficients for the binary relevance judgments were 0.368 (JaCollection), which is a fair agreement, and 0.626 (EnCollection), which is a substantial agreement. Finally, for each goal-level action relevance, we merge the results of three assessors as follows: (1) if two or more assessors give a relevance of 0, then we regard it as a relevance of 0, (2) otherwise, we regard it as a relevance of 1. As the results of this annotation step, we obtained 3,899 goal-level action relevance for JaCollection and 8,125 for EnCollection.

3.4.4 Evaluation Metric

We use D#-nDCG [Sakai and Song 2011], which was proposed by Sakai *et al.* as it intuitively evaluates a ranked-list in terms of both its diversity and relevance. Given a set of goals

$G^q = \{g_1^q, \ldots g_n^q\}$, let $p(g_i^q)$ represent the probability that a searcher issuing q has the goal g_i^q, for which we assumed the uniform probability $p(g_i^q) = \frac{1}{|G^q|}$ in our evaluation, and let $\text{gain}_i(a^r)$ be the goal-level gain value of the a^r, action at rank r returned by a method, which is defined as $\text{gain}_i(a^r) = 2^{\text{rel}^q(a^r, g_i)} - 1$. The global gain for this r-th ranked action is defined as:

$$\text{gg}(a^r) = \sum_{g_i \in G^q} p(g_i)\text{gain}_i(a^r) \ . \tag{3.9}$$

The ideal ranked list of actions is obtained by sorting all the pooled actions by the global gain. Let $\text{gg}^*(a^r)$ denote the global gain in this ideal list. D-nDCG at cutoff k is defined as:

$$\text{DnDCG@}k = \frac{\sum_{r=1}^{k} \text{gg}(a^r)/\log(r+1)}{\sum_{r=1}^{k} \text{gg}^*(a^r)/\log(r+1)} \ . \tag{3.10}$$

D#-nDCG is defined as a linear combination of D-nDCG and I-rec, which measures the recall of goals covered by the ranked list. Let $G_k^q (\subseteq G^q)$ be the set of goals covered by the top k actions of the ranked list. In this work, we consider that goal g_i^q is covered when there exists at least one action relevant to g_i^q in the top k actions. Then the recall of goals I-rec at cutoff k is defined as:

$$\text{Irec@}k = \frac{|G_k^q|}{|G^q|} \ . \tag{3.11}$$

With D-nDCG and I-rec, D#-nDCG at cutoff k is computed as:

$$\text{D\#nDCG@}k = \gamma\text{Irec@}k + (1-\gamma)\text{DnDCG@}k \ , \tag{3.12}$$

where we let $\gamma = 0.5$, following the NTCIR INTENT [Sakai et al. 2013] and IMine [Liu et al. 2014, Yamamoto et al. 2016] tasks. We use D#-nDCG@8 as our primary metric since many of the conventional query suggestions provide eight suggestions to a query. The ranked list containing more relevant and diverse (in terms of q's goals) actions achieves higher D#-nDCG@8.

3.5 Experimental Results

3.5.1 Comparison with Baselines

Table 3.3 shows the results of D#-nDCG@k of the baseline and our methods described in Section 3.4.2 for two test collections. Here, we use $\lambda = 0.7$ and $\alpha = 0.6$ as parameters, which is the optimum D#-nDCG@8 for EnCollection, for evaluating JaCollection. We also use $\lambda = 0.4$ and $\alpha = 0.5$, which achieves the optimum D#-nDCG@8 for JaCollection, for evaluating EnCollection. From the table, we can see that both RelDivQA and RelOnlyQA outperform the baselines for both test collections. In addition, it can be seen that D#-nDCG of both QS1 and QS2 are quite low, compared with RelDivQA and RelOnlyQA. This result

Table 3.3: D#-nDCG@k for each test collection (highest values among methods are in bold).

		D#-nDCG@k			
		$k=1$	$k=3$	$k=5$	$k=8$
JaCollection	QS1	0.000	0.000	0.006	0.017
	QS2	0.000	0.000	0.000	0.006
	RelOnlyWeb	0.044	0.079	0.089	0.131
	RelDivWeb	0.044	0.079	0.089	0.131
	RelOnlyQA	**0.116**	**0.292**	0.368	**0.412**
	RelDivQA	**0.116**	**0.292**	**0.369**	**0.412**
EnCollection	QS1	0.051	0.068	0.087	0.106
	QS2	0.000	0.059	0.067	0.107
	RelOnlyWeb	0.081	0.126	0.157	0.238
	RelDivWeb	0.081	0.126	0.165	0.238
	RelOnlyQA	**0.133**	0.189	0.279	**0.390**
	RelDivQA	**0.133**	**0.267**	**0.304**	0.371

indicates that conventional query suggestions rarely provide alternative actions for a query, whereas cQA is an effective resource for mining alternative actions.

Also, one possible reason why the performance of **RelDivWeb** was less than **RelDivQA** is the most Web documents retrieved by the query is so relevant to the query that the alternative-actions rarely appeared in the documents. The two-sided Randomized Tukey's HSD test [Sakai 2014] revealed that, in terms of D#-nDCG@8, we observed the significant differences between all the pairs of the proposed methods and the baselines at the significant level $\alpha = 0.01$ for JaCollection. On the other hand, for EnCollection, we observed that the differences between the proposed methods and the query suggestions (QS1, QS2) were significant, while the differences between the proposed methods and the Web-based baselines (RelDivWeb, RelOnlyWeb) were not significant in terms of D#-nDCG@8.

In addition, having that our methods achieved the similar performance on different test collections, the experimental results suggest that our method is able to applicable to many cQA services. When we compare the results of RelDivQA and RelOnlyQA, we observe that the results of RelDivQA and RelOnlyQA are similar. The two-sided Randomized Tukey's HSD test revealed that the differences between RelDivQA and RelOnlyQA were not significant for all the metrics on both test collections. This implies that the combination of the relevance and diversity does not always help to improve the performance in our evaluation. One possible reason of this would be that the number of goals were small. As described in Section 3.4.3, each query has at most three goals, which is relatively small number compared with the existing test collection [Liu et al. 2014, Sakai et al. 2013]

Table 3.4: D#-nDCG@8 for different λ and α for both JACollection and EnCollection (highest value in bold).

	JaCollection											EnCollection									
α\λ	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**	α\λ	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
0.1	0.413	0.413	0.414	0.414	0.414	0.414	0.414	0.414	0.414	0.414	**0.1**	0.308	0.349	0.304	0.326	0.315	0.320	0.333	0.337	0.353	0.347
0.2	0.413	0.413	0.414	0.414	0.414	0.414	0.414	0.414	0.414	0.414	**0.2**	0.295	0.340	0.307	0.308	0.301	0.345	0.361	0.359	0.375	0.370
0.3	0.413	0.413	0.414	0.414	0.414	0.414	0.414	0.414	0.414	0.414	**0.3**	0.321	0.335	0.329	0.348	0.323	0.341	0.383	0.389	0.395	0.360
0.4	0.413	0.413	0.413	0.414	0.414	0.414	0.414	0.414	0.414	0.413	**0.4**	0.322	0.350	0.350	0.365	0.345	0.379	0.415	0.420	0.373	0.392
0.5	0.413	0.413	0.413	**0.414**	0.414	0.414	0.414	0.414	0.414	0.414	**0.5**	0.370	0.377	0.372	0.390	0.387	0.429	0.443	0.426	0.409	0.415
0.6	0.406	0.406	0.406	0.412	0.412	0.412	0.412	0.412	0.412	0.412	**0.6**	0.350	0.351	0.359	0.375	0.389	0.420	**0.448**	0.418	0.412	0.435
0.7	0.402	0.406	0.406	0.411	0.412	0.412	0.412	0.412	0.412	0.411	**0.7**	0.362	0.356	0.353	0.378	0.419	0.429	0.420	0.413	0.401	0.420
0.8	0.402	0.406	0.406	0.406	0.412	0.412	0.412	0.412	0.411	0.411	**0.8**	0.366	0.354	0.345	0.386	0.389	0.414	0.414	0.401	0.412	0.425
0.9	0.402	0.402	0.406	0.406	0.412	0.412	0.411	0.403	0.402	0.402	**0.9**	0.356	0.359	0.346	0.352	0.425	0.399	0.416	0.403	0.408	0.403
1.0	0.398	0.398	0.398	0.398	0.398	0.410	0.410	0.410	0.410	0.410	**1.0**	0.323	0.301	0.297	0.371	0.443	0.399	0.410	0.413	0.419	0.415

Table 3.5: D#-nDCG@8 of RelDivQA ($\lambda = 0.4$, $\alpha = 0.5$ for JaCollection, $\lambda = 0.7$, $\alpha = 0.6$ for EnCollection) for different domains.

	JaCollection	**EnCollection**
Health	0.566	0.479
Recreation	0.300	0.413
Education	0.235	0.414
ALL	0.414	0.448

3.5.2 Effect of Parameters

We evaluate D#-nDCG@8 obtained by our method by varying the parameters α and $\lambda = 0.1, 0.2, \ldots, 1.0$ to investigate how diversity and effectiveness affected the performance. Table 3.4 summarizes the results on each test collection. From the tables, we observe that the optimum parameters are $\alpha = 0.5, \lambda = 0.4$ for JaCollection (0.4142 for that condition and 0.4138 for the other cells denoting 0.414) and $\alpha = 0.6, \lambda = 0.7$ for EnCollection.

For JaCollection, as we can see from the table, we could not find the effect of changing the parameters. One possible reason of this is that our method generated the similar ranked list for different parameters. As described in Section 3.4.3, JaCollection contains less actions (3,899) than EnCollection does (8,125), which implies that the ranked lists for JaCollection generated by our method are similar to each other.

On the other hand, we found some trends in the results on EnCollection. The combination of alternativeness and effectiveness (e.g., $\alpha = 0.5$) improves D#-nDCG@8 as compared with the case where each was used alone (e.g., $\alpha = 0.1$ or 1.0). This result indicates that considering both alternativeness and effectiveness helps to find alternative actions for a query for EnCollection. Also, when we focus on $\alpha = 0.5$, D#-nDCG improves as we change λ from 1.0 to 0.7, which suggests that the alternativeness between two actions $\mathrm{alt}(a_i, a_j)$ enables a method to generate a diversified ranked list for EnCollection.

3.5.3 Effect of Domain

To investigate the effectiveness of our method with the different domains, Table 3.5 summarizes D#-nDCG@8 for three domains. Note that we used the optimum parameters for each test collection when computing D#-nDCG@8. From the table, we can observe that the results of the health domain achieved the best performance for both JaCollection and EnCollection. The possible explanation of this would be, in the health domain, people discuss about many possible solutions for solving their problems since they want to choose the effective and credible solution for their health. We thus could find many alternative actions from the cQA corpus.

Table 3.6: Example top 2 alternative actions retrieved by proposed and baseline methods for queries "chamomile tea" and "youth hostel". Relevant actions are in bold.

	query = "chamomile tea"		
	RelDivQA	**QS1**	**QS2**
1st.	**drink a cup of hot milk**	chamomile tea effect	chamomile tea effect
2nd.	put it on your eyes	how to make chamomile tea	camomile tea cough
	query = "youth hostel"		
	RelDivQA	**QS1**	**QS2**
1st.	**check out airbnb**	youth hostel dublin	youth hostel paris france
2nd.	stayed at a hostel	youth hostels in london	youth hostel association of india

3.5.4 Examples

Table 3.6 shows examples of the alternative actions retrieved by our method and baselines (QS1 and QS2). For example, for the query "chamomile tea," our method successfully ranked the alternative action "drink a cup of hot milk," which can achieve the goal behind the query "promote falling asleep" at the first rank, while the baselines QS1 and QS2 suggested the queries which specialize the input query (e.g., "chamomile tea effect"). Since the conventional query suggestions are not designed for providing alternative actions for a query, suggesting the alternative actions obtained by our method can complement the existing query suggestions and help a searcher make an improved decision on how to achieve his/her goal.

On the other hand, we can see that our method ranked the action "put it on your eyes," which seems a meaningless phrase, at the second rank of the query "chamomile tea." We found that many of the irrelevant actions retrieved by our method were such meaningless verbal phrases, which affected the performance of our method. We will discuss how to improve our method in Section 3.6.

3.6 Limitations

Our work has several limitations that we should acknowledge. First, we take a search result diversification approach to generating a ranked list of alternative actions. The problem of this approach is that the ranked list contains actions which achieve different goals and it would be difficult for a searcher to find the alternative actions which achieve *his/her* actual goal. Several possible solution to solve this problem would be clustering the alternative actions according to their goals or predicting the goal of the searcher by analyzing his/her behavior log.

Second, as shown in Table 3.3, D#-nDCG obtained by our method is relatively low, compared with the standard search result diversification problems [Liu et al. 2014, Sakai et al. 2013]. Currently our method just extract actions (verbal phrases) from the answers in the

cQA corpus and use them as the candidates for the ranked list. The problem is that these actions contain lots of irrelevant actions which cannot be alternative actions for a query. Some researchers proposed to use syntactic patterns to extract target entities from a text corpus [Li et al. 2010]. Applying such a method would enable us to extract the candidate alternative actions rather than actions from the cQA corpus.

Third, we have not enough considered about the correctness of an action which might be a critical problem especially for an action under the medical domain. In this study, we used the effectiveness of an action to use the ranking. However, a high effectiveness of an action does not guarantee that the action is correct since the effectiveness just measures a sort of popularity. In order to solve the problem of the correctness of the mined actions, we may verify the alternative-action with the domain specific knowledge-bases such as WebMD, PubMed as the effectiveness of actions under these knowledge-bases are guaranteed by the domain experts.

Lastly, we should acknowledge the belief of a searcher. People usually favor information that confirms their pre-existing beliefs and biases [Kahneman 2003]. Recently, White also revealed that the existence of the search biases in which users preferred affirmative information to their beliefs [White 2013]. His findings implies that, even if we successfully provide alternative actions to a searcher, he/she would not take them into consideration because he/she believes in the solution expressed by the query. Although it is challenging to change the belief of a searcher, many researchers attempted to support a searcher's credible or careful search, *e.g.*, by suggesting disputed sentences [Ennals et al. 2010, Yamamoto and Shimada 2016] or providing scores according to credibility criteria [Yamamoto and Tanaka 2011]. By applying such methodologies, we may raise the awareness of the searcher.

3.7 Summary

In this study, we addressed the alternative action mining problem. We defined the alternative mining problem as similar in the search result diversification. To our knowledge, our work is the first to study this problem. Also, we proposed leveraging a cQA corpus to address the alternative action mining problem. Our method iteratively computes two measures; (1) $\mathrm{alt}(\cdot,\cdot)$, which measures the alternativeness between two actions, and (2) $\mathrm{sim}_{\mathrm{goal}}(\cdot,\cdot)$, which measures how well two questions represent the same goal, by applying the SimRank algorithm to the question-action bipartite graph. Our method generates the diversified ranked list of alternative actions by applying the MMR algorithm according to the alternativeness between actions.

The experimental results indicated that, for Japanese test collection, our proposed method significantly outperformed two types of baselines, one used the conventional query suggestions and the other extracted alternative-actions from the Web documents, in terms of D#-nDCG@8. Also, for English test collection, our method significantly outperformed the baseline using the conventional query suggestions in terms of D#-nDCG@8.

We believe that our method can complement the conventional query suggestions and help a searcher make an improved decision on how to achieve his/her goal. We also found that the combination of the alternativeness and the effectiveness improved the performance for the English test collection, whereas we could not find this trend for the Japanese test collection.

As we discussed in Section 3.6, we have several limitations that affect the performance of our method. One possible direction would be improving the step for extracting candidate actions so that we can obtain more relevant alternative actions. Another interesting direction would be designing a new search interaction so that a search system can encourage a searcher to carefully compare the solutions suggested by the system and the one he/she initially comes up with.

4 Understanding how and when people search for the alternative information

4.1 Introduction

To obtain accurate information, previous studies have suggested that users should expose themselves to diverse information [Garrett and Resnick 2011, Liao and Fu 2014, Liao et al. 2015, Resnick et al. 2013] and carefully verify information by spending more time on a search task, checking the author information in a document during a search, *etc* [Yamamoto et al. 2018a,b]. However, users are likely to exhibit confirmation bias, where they seek information to confirm their beliefs via web search engines [White 2013, 2014, White and Hassan 2014, White and Horvitz 2015]. For example, the users who believe that green tea extract is helpful for losing weight might select only positive information about the green tea extract, even though researchers argue that it is ineffective and harmful [Chalasani et al. 2014, Jurgens et al. 2012]. Consequently, they are likely to refrain from diverse and careful searches for information.

Another line of research suggested that opinion and the quality of information affect people's confirmation bias and information behavior [Festinger 1964, Kaplan et al. 2016, Schwind et al. 2012]. For example, Schwind et al. reported that people's confirmation bias could be mitigated by presenting information that was inconsistent with their preference [Schwind et al. 2012]. Such researches imply that we would be able to design a search results ranking system that encourages people to search more carefully.

In this study, we focused on two *factors* of the document searchers encountered during their searches: (1) *opinion* that is either consistent or inconsistent with the searchers' beliefs prior to the search task, and (2) *credibility*, which represents the believability of the document. We examined the extent to which these factors encouraged or discouraged searchers from searching carefully. The aim of our study is to learn the characteristics of a ranking system that can encourage people to search more carefully by better understanding how these factors affect their search behaviors and *belief dynamics* (*i.e,* how their beliefs change after the search process).

More precisely, the purpose of the study is to answer the following research questions:

- **RQ1:**How do opinion and credibility encourage or discourage searchers to search for information carefully?
- **RQ2:**How do opinion and credibility affect searchers' decision outcomes in terms of belief dynamics?
- **RQ3:**Is there any relationship between search behaviors and belief dynamics?

The participants were required to use our search system to find the answer to four medical-related *yes-no* questions, such as *Are dairy products effective in improving hypertension?* Prior to the search task, they were asked to provide their beliefs in relation to the remedy, which was then defined as *prior belief*. We controlled the search results list by inserting a manually prepared document with a specific opinion and credibility level depending on the experiment condition and the participants' prior beliefs in the second rank of the list. The participants were also asked to provide their beliefs in relation to the remedy after the search task, which was then defined as *posterior belief*. By doing so, we were able to analyze the effects the opinion and credibility had on the search behaviors and belief dynamics.

The main findings of the study are summarized as follows: (1) the participants spent more effort searching by issuing more queries when they encountered documents that were inconsistent with their prior beliefs; (2) they tended to alter their beliefs when encountering an opinion that was inconsistent with their prior beliefs while they tended to retain their beliefs when encountering an opinion that was consistent with their prior beliefs. Also, such an effect depended on the level of credibility; and (3) the behavioral differences we found in (1) could not be observed according to the participants' belief dynamics as only the types of opinion in the clicked documents were highly different. These results imply that search engines could prevent users from polarization and thus, help them obtain accurate information, by presenting documents that are inconsistent with users' beliefs on the higher-rank of the results.

4.2 Concept

In this section, we define a *critical thinker* who is expected to carefully search for information and consider different points of view. Next, we introduce *opinion* and *credibility* as the document factors investigated in this study. Finally, we describe how we prepared the documents based on these factors.

4.2.1 Critical Thinker in Web Search

Ennis defined critical thinking as the process of careful thinking to determine what to believe or react to. Ideally, critical thinkers should seek alternatives, consider other points of view, and be aware of their own beliefs [Ennis 1987]. Consistent with this definition, Yamamoto et al. [Yamamoto et al. 2018a, Yamamoto and Yamamoto 2018, Yamamoto et al. 2018b], discussed the behaviors of critical thinkers in the context of web searching: They are expected to spend more time searching, issue more queries, browse more documents so as to make

comparisons, and increase the evidence to support their decisions. More precisely, in the present study, we assumed that, compared to the non-critical thinkers, critical thinkers: (1) issue more queries, (2) click on more documents, (3) check deeper ranked result, (4) spend more time on browsing a document, (5) consider more documents as evidence that support their decisions, and (6) consider more reliable documents as evidence.

The aim of our study is to understand the extent to which a document's opinion and credibility encourage or discourage the search behaviors described above (**RQ1**). We also investigate the effects document's opinion and credibility have on people's belief dynamics (**RQ2**) and its relationship with their behaviors (**RQ3**).

4.2.2 Factors Investigated in This Study

In this study, *opinion* and *credibility* are considered a document's major factors that affect users' search behaviors and belief dynamics. Opinion is the polarity of the document's content compared to the users' *prior beliefs*, beliefs held before performing a search. We investigated two types of opinions, *consistent*, where the polarity of the content supports the users' prior beliefs and *inconsistent*, where the polarity of the content opposes the users' prior beliefs. Opinion is considered because it impacts users' information selection. While belief-inconsistent opinions encourage users to select diverse information, belief-consistent opinions exacerbate their confirmation bias [Schwind et al. 2012]. Credibility is also taken into account in this study; it has been viewed as an important factor that influences the behaviors of system users [Moshagen et al. 2009]. In this study, we prepared *high* and *low* credibility documents separately. Previous studies proposed that people became less careful when information that seemed credible was presented and more careful when information that seemed less credible was presented [Hernandez and Preston 2012, Kammerer et al. 2013]. Therefore, we expected people to be less likely to search for information carefully when a high credibility document was presented and more likely to search for information carefully when a low credibility document was presented. Some studies have revealed that opinion and the quality of information are related [Festinger 1964, Kaplan et al. 2016]. Thus, we also expected different combinations of opinions and levels of credibility to have different impacts on users' search behaviors and decision outcomes.

4.2.3 Controlled Documents Preparation

To investigate the effects different opinions and credibility levels have on peoples' search behaviors and belief dynamics, we manually prepared documents based on these factors. We refer to them as *controlled documents* hereinafter.

As for opinion, to prepare documents that were consistent/in-consistent with the users' prior beliefs, we manually prepared the content that supports/opposes to each search task. For example, for a search task that required a participant to find an answer for *Are dairy products effective in improving hypertension?*, we collected two types of articles from the web; ones

(a) Low Credibility Controlled Document

(b) High Credibility Controlled Document

Figure 4.1: Screenshots of our controlled documents with (a) low credibility and (b) high credibility (translated from Japanese). With the exception of visual styles, both controlled documents show the same content in their articles.

that claimed that *dairy products are effective in improving hypertension*, and ones that claimed that *dairy products are NOT effective in improving hypertension*. The contents were selected from reliable sources that satisfied either of the following conditions: (1) the article explicitly refers to an academic paper or technical report presented at an academic conference, (2) the article is from a reliable medical knowledge base[1], and (3) the article in which author information is explicitly written. Table 4.1 shows the list of search tasks used in our study and the sources for preparing the controlled documents.

As for credibility, we developed the document layout in accordance with *surface credibility* [Fogg 2003]. Surface credibility is an evaluation criterion that the believability of information is judged based on appearance (as in, visual style). One typical example of surface credibility is that people are more likely to believe aesthetically appealing information [Sillence et al. 2006]. We selected the surface credibility because it is easy for even people without an educational background to understand and apply [Fogg 2003, Metzger et al. 2010, Robins et al. 2009] and many people apply surface credibility to evaluate web information [Metzger 2007]. As shown in Figure 4.1, the high credibility controlled document was designed to be clean, clear, and simple since the document with such components are likely be perceived as aesthetic [Karvonen 2000, Karvonen et al. 2000]. Meanwhile, the low credibility controlled

[1] https://www.cochranelibrary.com/

document was intentionally designed as less aesthetically appealing than the high credibility ones.

To avoid situations in which the differences in their content would affect the participants' search behaviors and decision outcomes, both the high and low credibility controlled documents presented the same content. Also, as a document's URL plays an important role in judging surface credibility [Fogg 2003], we prepared the domain names for each credibility controlled document [2]. As we will cover in Section 4.4.1, the majority of the participants rated our high credibility documents as more credible than our low credibility ones. We prepared 16 controlled documents (4 search tasks $\times$ 2 opinion types $\times$ 2 credibility levels) for the user study described in the next section.

4.3 Method

We introduce the experimental design in this section and describe the experiment procedure in detail. We also introduce the search system that was tailored for this study and how we manipulated the search results. Finally, we outline how the participants were recruited and filtered from the analysis.

4.3.1 Experimental Design

The experiment was conducted with two independent within-subject variables: *opinion* (consistent/inconsistent) and *credibility* (high/low). It generated four experiment conditions which are C_1 (inconsistent, high credibility); C_2 (inconsistent, low credibility); C_3 (consistent, high credibility); and C_4 (consistent, low credibility). Our experiment was repeated in that each participant was assigned to all four conditions. For each condition, the participants were assigned to one of the four medical *yes-no* questions adopted from previous literature [Yamamoto et al. 2018b], because medical information tends to have reliable evidence for both supporting and opposing opinions. Also, the degree of belief can be measured by *yes-no* questions [White 2013]. Table 4.1 details each question. The tasks were rotated according to the Graeco-Latin square design [Kelly 2009].

4.3.2 Procedure

First, participants were informed how their data would be used and asked for consent to use their data for research purposes. Only participants who agreed to these conditions could participate in the training task. In the training task, they were asked to find the answer for the question *Is Chinese medicine effective for atopic dermatitis?* using our search system to get familiar with our search system. Note that search results were not manipulated during the training task.

[2] High Credibility controlled document: http://www.med-japan.com/ ,
Low Credibility controlled document: http://www.kenko-blog-life.org/ (The word "kenko" means healthy in Japanese)

Table 4.1: Search tasks used in this study. Sources of controlled documents are also shown.

Search Task		Sources	
ID	**Questions**	**Supporting**	**Opposing**
1	Are dairy products effective in improving hypertension?	http://bit.ly/2Tr2mYG (refers to an academic paper)	http://bit.ly/2TmjsHo (refers to an academic paper)
2	Are isoflavones effective in relieving high cholesterol?	https://nkbp.jp/2CNTjM5 (refers to a technical report)	http://bit.ly/2Tshjdc (medical knowledge base)
3	Is ginseng effective in improving dementia?	http://bit.ly/2COPciQ (refers to a technical report)	http://bit.ly/2CPDV21 (medical knowledge base)
4	Is homeopathy effective in improving asthma?	http://bit.ly/2TrKGfA (explicit author information)	http://bit.ly/2Tq7Z9B (medical knowledge base)

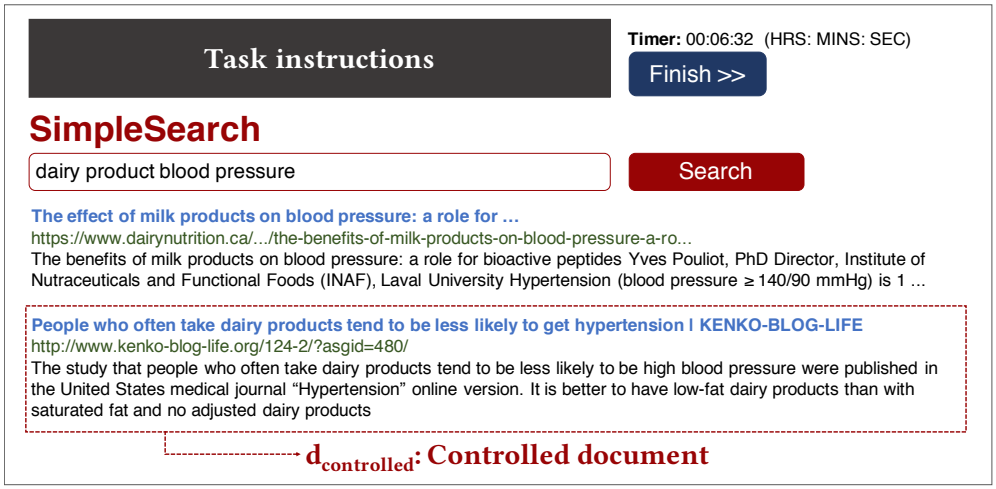

Figure 4.2: Screenshot of our search system. The controlled document was inserted at the second rank of search results.

After the training task, the participants were asked to complete four search tasks. Each search task was performed as follows. First, the participants were given the instructions that comprised the background of the symptom and a remedy for each search task. In order to avoid a situation in which the participants becoming biased because of the instructions, we tried to ensure the instructions were as neutral as possible. The following phrase is an example of the instructions (for task 1 in Table 4.1) that were presented to the participants:

Hypertension tends to be a disease of the heart and the vascular system. People with hypertension typically have a high mortality rate. Therefore, preventing and relieving hypertension is important. Suppose that you are currently having hypertension symptoms and the doctor tells you to lower your blood pressure. You have heard rumors that dairy products such as milk, yogurt, and cheese seem to be effective in improving hypertension. Therefore, in this task, please conduct a web search to examine whether dairy products are effective in improving hypertension.

After reading the instructions, the participants were asked to complete a pre-task questionnaire. First, the participants were asked to provide their beliefs in relation to the remedy by the question *Do you think dairy products are effective in improving hypertension?* with a four-point Likert scale (1: No; 2: Lean No; 3: Lean Yes; 4: Yes). This means that the participants had to decide their beliefs to be either of the two sides. A possible problem with using a four-point Likert scale will be discussed later in Section 4.5.2. We regarded the participants'

scores for this question to be their *prior beliefs*. Second, the participants were asked to rate their prior knowledge about the symptom and remedy on a four-point Likert scale (1: Not at all; 2: A little; 3: Good; 4: Excellent).

Subsequently, the participants performed the search task by using our search system. The details of our search system are explained in Section 4.3.3. During the search task, they were asked to use our system and were prohibited from using other commercial search engines. Also, they were allowed to issue any queries to the system and click on any documents returned by the system. The participants were not limited by time constraints. Once the participants felt satisfied, they could click the finish button at the top right of the search page.

At the end of each search task, the participants were asked to complete a post-task questionnaire. They were first asked to provide their beliefs about the remedy again on the same four-point Likert scale. We regarded their scores for this question as their *posterior beliefs*. Next, the participants were asked to select the documents that contain evidence to support their posterior beliefs. We provided a list of the documents they clicked on during the search task. They were asked to select one or more documents among the list.

After the four search tasks were finished, we asked the participants who clicked on the controlled document(s) during the four search tasks to evaluate the credibility of the controlled document(s) with the question *How likely were you to believe this document?* (1: Not at all likely to believe it; 2: Likely not to believe it; 3: Likely to believe it; 4: Very likely to believe it) for each controlled document they clicked, as we wanted to ensure that the majority of the participants perceived the high credibility controlled documents as credible and the low credibility ones as less credible. Thereafter, they had to complete an exit questionnaire that required them to provide demographic information (gender, education level, and search engine familiarity). Note that the participants had the option to refrain from answering the demographic questions on gender and educational background. Finally, the participants were asked if anything had bothered them during the search task. This questionnaire was prepared so that we could ascertain who had noticed that we manipulated the search results by inserting a controlled document in the search results.

4.3.3 Search System

Figure 4.2 shows the screenshot of our search system. In order to help the participants get familiar with our search system easily, we imitated the user interface of a commercial search engine with the task instructions above the search box. Figure 4.3 shows the overview of how the search results were generated. According to the query issued by the participant, the system first obtained 50 organic search results $d_1, d_2, \ldots, d_{50}$ via the Bing Web Search API[3]. The system then inserted the controlled document $d_{\text{controlled}}$ within the second rank of the search results and showed the top 50 search results, resulting in $d_1, d_{\text{controlled}}, d_2, \ldots, d_{49}$ being shown

[3] https://azure.microsoft.com/en-us/services/cognitive-services/

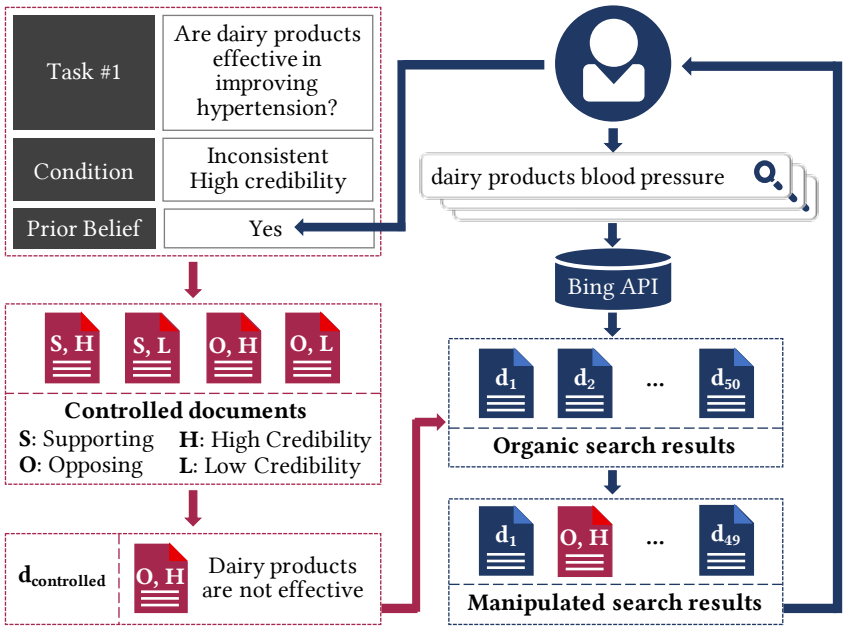

Figure 4.3: Search Result Manipulation Overview.

to the participant. Note that $d_{\text{controlled}}$ was selected based on the experiment condition and the participant's prior belief. For example, if the experiment condition was C_1 (inconsistent, high-credibility) and the participant's prior belief was either "Lean Yes" or "Yes", the system would present high-credibility $d_{\text{controlled}}$ containing information, which opposed the remedy (*i.e.,* there is *no* reliable evidence that the dairy products relieve the hypertension).

We expected that many of the participants would click on our controlled document during a search task as higher-ranked documents are likely to receive more attention [Guan and Cutrell 2007, Joachims et al. 2007]. We decided not to insert the $d_{\text{controlled}}$ in the first rank of the search results because we want to avoid situations where the participants doubted that the search results were not generated by a usual search engine or they doubted that the search results were manipulated based on their prior beliefs. Each search result was comprised of a title, a snippet, and a URL. For $d_{\text{controlled}}$, we used its title and the first 100 words of its body text as a title and snippet, respectively. The document was displayed in a separate tab when it was clicked.

Table 4.2: Demographics of participants.

Gender	*n*	Educational Background	*n*	Search Engine Familiarity	*n*
male	115	university educated	154	rarely use	1
female	141	not university educated	64	several times per week	22
N/A	4	N/A	42	once a day	45
				more than once a day	192

4.3.4 Participants

401 participants were recruited via Lancers.jp [4], a Japanese crowd-sourcing platform, from July 10 to August 3, 2018. The median and standard deviation of the task completion time were 22 minutes and 26 minutes, 17 seconds, respectively. Since we used a crowd-sourcing platform to conduct the experiment, controlling the quality is important. To do so, in the instruction page, we said that we might reject their submissions if we judged that they did not follow the task instruction, although we actually did not reject any submission of the 401 participants. Each participant received approximately $4 as compensation.

From the results of 401 participants, we removed 71 participants who clicked the controlled document less than two tasks. We further removed 24 participants who noticed the manipulation of the search results. Furthermore, we removed participants who used other search engines than ours, did not complete all the search tasks, or spent too much/too little time on the task. We collected the data of the search tasks in which a participant clicked on our controlled document and finally obtained 809 search tasks from 260 participants for analysis. The demographics of the 260 participants is shown in Table 4.2.

4.4 Results

The results are reported as follows. First, we give the participants' perception of the credibility of our controlled document. Second, we examine the extent to which the controlled documents affected the participants' search behavior (RQ1). Third, we explore the extent to which the controlled documents affected the participants' belief dynamics (RQ2). Finally, we discuss the relationship between search behaviors and belief dynamics (RQ3). For the analyses regarding RQ1–RQ3, as we only used the data of the search task in which the participants clicked on our controlled documents, the data became unbalanced. The distribution of 809 search tasks in terms of the experimental conditions are C_1: 210, C_2: 191, C_3: 213, and C_4: 195. Therefore, we employed a linear mixed-effects regression analysis which is applicable for the unbalanced data [Bates et al. 2015, Field et al. 2012].

[4] https://www.lancers.jp/

We constructed a linear mixed-effects regression model where *opinion*, *credibility* and their *interaction* were considered as fixed effects. In line with the recommendation of keeping the random effect structure maximal [Barr et al. 2011], the model initially included random intercepts and random slopes on participants and tasks. However, this model does not converge due to the data size. Therefore, we removed by-task random slopes because the tasks' variance is far smaller than participants' variance. Although removing the by-task random slopes decreases the generalizability of the model to the population of tasks, our model has high generalizability to the population of users [Barr et al. 2011, Winter 2013]. The final model we report contains random intercepts and slopes on participants and random intercepts on tasks. The formula for such a model [5] is $d \sim$ Opinion * Credibility + (1+ Opinion + Credibility | Participant) + (1 | Task), where d represents a dependent variable. We then performed the likelihood ratio test with the model and the *null model*, which only considers the random effects. The result is considered as significant if the model is statistically significant compared to the null model. The significance level in this study was set to 5%. Note that we applied log transformation to the temporal features, such as page dwell time.

4.4.1 Perception of Credibility

To ascertain whether the participants perceived our high (low) credibility documents as highly (low) credible, we calculated the average controlled documents' credibility score that had been obtained via the questionnaire for each credibility level separately. We found that high credibility controlled documents received higher average scores than low credibility controlled documents (high: $M = 2.89, SD = 0.80$; low: $M = 2.39, SD = 0.76$). The results showed that the majority of the participants agreed that the high credibility controlled documents were credible and low credibility ones were less credible. To validate the significance of the results, we conducted the linear mixed-effects regression analysis by constructing a model where the credibility score was regarded as a dependent variable and *credibility* was regarded as a fixed effect. Note that the random effect structure for this model was the same as we explained above. We found the main effect of the *credibility* on the credibility score ($\beta = 0.27, SE = 0.03, t = 9.49, p < 0.001$). This result implies that different credibility levels affect the controlled documents' credibility score assessed by the participants.

4.4.2 Data Annotation

For the later analysis, we collected the documents (excluding controlled documents) clicked on by the participants during the search task (which consisted of 1,103 documents) and hired three graduate students majoring computer science as assessors to annotate them. First, we asked the assessors to annotate the types of opinions on the document as either *supporting*, *opposing*, *supporting and opposing*, or *irrelevant* to the search task. The Fleiss' kappa

[5] The model is expressed in the syntax of lmer, a widely used mixed-effects fitting method contained in lme4 [Bates et al. 2015]

coefficient, which is used to measure the agreement among the assessors, for this annotation is 0.64, which indicates the substantial agreement among the assessors [Fleiss 1971].

In addition to the types of opinions, we also asked the same assessors to annotate the reliability of the clicked documents. As we described in Section 4.2.1, we expected that people are likely to consider *reliable* documents as evidence when they are careful in their search process. To measure the reliability of the documents, we followed the four criteria of the JAMA benchmark [Silberg et al. 1997]. The JAMA benchmark is the guideline that assesses the reliability of health care information on the web and has four criteria: (1) the document reveals its author (authorship), (2) updated date (currency), (3) references (attribution), (4) and sponsorship or ownership (disclosure). We asked the assessors to annotate whether the clicked document satisfies each of the four criteria. The Fleiss' kappa coefficients for the authorship, currency, attribution and disclosure were 0.72, 0.75, 0.58, and 0.68, respectively, all of which indicate the moderate to substantial agreement among the assessors. For later analyses, the annotation results were aggregated by majority vote, i.e., when at least two of the three assessors annotated the same label, it was considered as the gold standard label.

4.4.3 Search Behavior

As described in Section 4.2.1, our aim is to investigate to what extent a document's opinion and credibility encourage or discourage people to search for information carefully. The following search behaviors were analyzed for each participant during the search task:

- **# of Queries**: Number of queries issued in a search task.
- **# of Clicks**: Number of documents clicked on in a search task.
- **# of Consistent/Inconsistent Clicks**: Number of documents that were consistent/inconsistent with the participant's prior belief clicked on in a search task (excluding controlled documents).
- **Deepest Document Rank**: The lowest rank the participant clicked on.
- **Page Dwell Time**: Average time the participant spent on the documents.
- **SERP Dwell Time**: Average time the participant spent on the search results page.
- **Controlled Document Dwell Time**: Average time the participant spent on the controlled document.
- **Task Time Spent**: Amount of time the participant spent on the search task.

Note that the documents that contained both supporting and opposing opinions for the search task were not included in the analysis for both **# of Consistent Clicks** and **# of Inconsistent Clicks**.

We also analyzed the documents the participants selected as evidence that supported their posterior beliefs, which had been collected in the post-task questionnaire. We call these

Table 4.3: Mean (M) and standard deviation (SD) for each controlled document condition (C_1: inconsistent and high credibility, C_2: inconsistent and low credibility, C_3: consistent and high credibility, and C_4: consistent and low credibility) and statistical testing results of each search behavior of the model with different fixed effects (***: significance level at 0.001, **: 0.01, and *: 0.05). Coefficient (β), standard error (SE), t-statistic (t), and p-value (p) of each fixed effect are reported. The χ^2 and p-values of the model which is statistically significant compare to the null model are also reported.

	Condition								**Fixed Effects**													
	C_1		C_2		C_3		C_4		**Opinion**				**Credibility**				**Opinion × Credibility**					
	M	SD	M	SD	M	SD	M	SD	β	SE	t	p	β	SE	t	p	β	SE	t	p	χ^2	p
# of Queries	2.67	2.00	2.74	2.19	2.32	1.65	2.51	1.90	0.12	0.05	2.57	*	-0.07	0.05	-1.39	0.17	0.02	0.04	0.49	0.62	9.12	*
# of Clicks	6.11	3.52	6.32	3.51	5.68	2.82	6.02	3.32	0.14	0.08	1.87	0.06	-0.07	0.08	-0.81	0.42	0.03	0.07	0.46	0.65		
# of Consistent Clicks	1.50	1.75	1.24	1.70	1.38	1.82	1.34	1.91	0.01	0.06	0.11	0.92	0.07	0.06	1.17	0.24	0.05	0.06	0.83	0.40		
# of Inconsistent Clicks	2.21	2.47	2.31	2.46	2.03	2.18	2.45	2.49	0.02	0.06	0.28	0.78	-0.10	0.08	-1.20	0.23	0.08	0.06	1.34	0.18		
Deepest Document Rank	11.25	9.05	12.40	11.00	10.99	8.69	12.08	10.59	0.15	0.25	0.61	0.54	-0.45	0.27	-1.69	0.09	0.07	0.23	0.29	0.77		
Page Dwell Time (sec)	25.40	21.91	26.25	23.92	28.00	31.40	28.80	32.70	-0.02	0.02	-1.28	0.20	0.00	0.02	0.21	0.83	0.02	0.02	1.45	0.15		
SERP Dwell Time (sec)	38.54	29.30	41.08	35.63	41.90	45.10	41.28	36.06	0.00	0.01	-0.04	0.97	-0.01	0.02	-0.73	0.47	-0.01	0.01	-0.88	0.38		
Controlled Document Dwell Time (sec)	17.10	18.00	18.10	35.50	22.20	52.30	17.00	21.40	-0.04	0.02	-1.61	0.11	0.10	0.03	3.30	**	0.01	0.02	0.42	0.67	11.80	**
Task Time Spent (sec)	413.79	319.97	419.88	270.30	414.62	354.36	401.68	252.58	0.02	0.01	1.34	0.18	-0.01	0.01	-0.54	0.59	0.00	0.01	-0.03	0.97		
# of Evidence	2.61	1.70	2.60	1.63	2.29	1.28	2.43	1.50	0.10	0.04	2.41	*	-0.02	0.04	-0.58	0.56	0.02	0.04	0.52	0.60		
Authorship	0.17	0.50	0.18	0.40	0.16	0.37	0.22	0.47	-0.01	0.02	-0.41	0.68	-0.01	0.01	-0.99	0.32	0.01	0.01	0.71	0.48		
Attribution	0.29	0.56	0.25	0.49	0.30	0.57	0.26	0.56	-0.01	0.02	-0.33	0.74	0.02	0.02	1.22	0.22	0.00	0.02	0.10	0.92		
Currency	0.72	1.11	0.79	0.89	0.63	0.76	0.76	0.92	0.03	0.03	0.95	0.34	-0.05	0.03	-1.68	0.09	0.02	0.03	0.60	0.55		
Disclosure	0.78	1.06	0.84	0.98	0.66	0.76	0.72	0.94	0.06	0.03	1.92	0.06	-0.03	0.03	-0.97	0.33	0.00	0.03	0.01	0.99		

Table 4.4: Participants' behavior according to opinion types.

	Inconsistent		**Consistent**	
	M	*SD*	*M*	*SD*
# of Queries	2.70	2.09	2.41	1.77

documents *evidence documents* hereinafter. The following metrics regarding the evidence documents were examined:

- **# of Evidence**: Number of evidence documents selected by the participant.
- **JAMA Benchmark**: Four numbers of evidence documents that satisfied (1) **Authorship**, (2) **Attribution**, (3) **Currency**, and (4) **Disclosure**, respectively.

4.4.4 Document's Opinion and Credibility on Search Behavior (RQ1)

To answer this research question, we performed a linear mixed-effects regression analysis with each participants' search behavior. Table 4.3 presents the results of the analysis. The "Fixed Effects" column in Table 4.3 shows the coefficient, standard error, t-statistic, p-value of each fixed effect in the model. This column also reports the model's goodness of fit when the model is statistically significant compared to the null model.

We found a main effect of the *opinion* on the number of queries $(\beta = 0.12, p < 0.05)$. Table 4.4 shows the results of this behavior according to the types of opinion. The results in Table 4.4 revealed that the belief-inconsistent opinions encouraged the participants to issue more queries (inconsistent: $M = 2.70$, consistent: $M = 2.41$).

The results presented in Table 4.3 also revealed the main effect of *credibility* on the time spent on the controlled document $(\beta = 0.10\ , p < 0.01)$. Table 4.5 shows the results of this behavior according to the types of credibility. The results in Table 4.5 revealed that the participants preferred spending time on high credibility controlled documents (high: $M = 19.67$, low: $M = 17.55$).

As for the interaction between opinion and credibility, we observed no significant interaction effect between the two. From this study, we did not observe the data that suggests the effect the different combinations of opinions and credibility levels had on the participants' search behaviors.

In essence, we found that the belief-inconsistent opinions encouraged the participants to issue more queries. A possible explanation for this result could be the participants who encountered the belief-inconsistent documents tried to issue more queries to find more information. However, we need a further study to understand this effect since its difference seems not large.

4.4.5 Belief Dynamics (RQ2)

To answer this research question, our analysis was based on the answer obtained from the belief-related questionnaires that were asked before and after the search task. As for belief dynamics, in this study, we focused on the cases in which the participants changed (or did not change) their prior beliefs from one polar to the opposite polar after the search task. We refer to the cases in which the participants changed their prior beliefs from one polar to the opposite polar as they *altered* their beliefs, otherwise, *retained*.

We first performed a logistic mixed-effects regression analysis to investigate how opinion and credibility affected belief dynamics. Similar to Section 4.4.4, we constructed a logistic mixed-effects model where belief dynamics d was considered a binary dependent variable (1: alter, 0:retain). From the logistic mixed-effects regression analysis, we found the main effect of the document's *opinion* ($\beta = 0.71, SE = 0.10, z = 7.35, p < 0.001$) and *credibility* ($\beta = -0.21, SE = 0.09, z = -2.19, p < 0.05$) on the belief dynamics. This suggests that both opinion and credibility affected the participants' belief dynamics. Note that the model is statistically significant by the likelihood ratio test compared to the null model.

To further analyze the effects of opinion and credibility, Tables 4.6 (a)–(d) show the participants' beliefs after the search task (columns) compared to their beliefs prior to the search task (rows) with the different opinions and credibility conditions. Tables 4.6 (a) – (d), show that the participants were more likely to alter their beliefs from one polar to the opposite polar (①) when belief-inconsistent controlled documents were presented, especially from the *no* polar to *yes* polar, compared with the cases when belief-consistent controlled documents were presented (②). For example, when belief-inconsistent, high credibility controlled documents were presented, 71.43% (53.85% + 17.58%) of the participants who were "Lean No" prior to the search task altered their beliefs to either "Lean Yes" or "Yes" after the search task. On the contrary, as cells ③ and ④ show, the participants were more likely to retain their beliefs in the same polar when belief-consistent controlled documents were presented.

These results imply that the participants who encountered a belief-inconsistent controlled document during a search process were more likely to alter their beliefs (especially when their beliefs were either "No" or "Lean No" prior to the search task). On the other hand, the participants who encountered a belief-consistent controlled document were more likely to retain their beliefs.

As for credibility, we found that the participants were less likely to retain their beliefs when belief-consistent, low credibility controlled documents were presented, (④), compared with cases when belief-consistent and high credibility controlled documents were presented (③). This result implies that the effect the belief-consistent controlled document had depended on its credibility level. Participants who encountered a belief-consistent controlled document with higher credibility were more likely to retain their beliefs.

Table 4.5: Participants' behavior according to credibility level.

	High		Low	
	M	*SD*	*M*	*SD*
Controlled Document Dwell Time (sec)	19.67	39.26	17.55	29.20

Table 4.6: Percentage of belief dynamics for each controlled document condition. (For each prior belief, the polar that the majority of participants leaned towards in posterior belief is highlighted in gray). The presentation of this table is inspired by the paper [White 2013].

(a) C_1: Inconsistent, High Credibility (n = 210)

		Posterior Belief			
		No (n=21)	Lean-No (n=50)	Lean-Yes (n=104)	Yes (n=35)
Prior Belief	No (n =31)	16.13%	32.26%	25.81%	25.81%
	Lean No (n=91)	5.49%	23.08%	53.85%	17.58%
	Lean Yes (n=79)	13.92%	22.78%	50.63%	12.66%
	Yes (n=9)	-	11.11%	77.78%	11.11%

(b) C_2: Inconsistent, Low Credibility (n = 191)

		Posterior Belief			
		No (n=28)	Lean No (n=34)	Lean Yes (n=99)	Yes (n=30)
Prior Belief	No (n=31)	29.03%	19.35%	41.94%	9.68%
	Lean No (n=99)	10.10%	14.14%	56.57%	19.19%
	Lean Yes (n=56)	14.29%	25.00%	50.00%	10.71%
	Yes (n=5)	20.00%	-	40.00%	40.00%

(c) C_3: Consistent, High Credibility (n = 213)

		Posterior Belief			
		No (n=51)	Lean No (n=66)	Lean Yes (n=66)	Yes (n=30)
Prior Belief	No (n=47)	48.94%	31.91%	14.89%	4.26%
	Lean No (n=96)	25.00%	46.88%	27.08%	1.04%
	Lean Yes (n=65)	6.15%	9.23%	49.23%	35.38%
	Yes (n=5)	-	-	20.00%	80.00%

(d) C_4: Consistent, Low Credibility (n = 195)

		Posterior Belief			
		No (n=37)	Lean No (n=55)	Lean Yes (n=72)	Yes (n=31)
Prior Belief	No (n=37)	37.84%	37.84%	24.32%	-
	Lean No (n=99)	21.21%	34.34%	37.37%	7.07%
	Lean Yes (n=57)	3.51%	12.28%	45.61%	38.60%
	Yes (n=2)	-	-	-	100.00%

Table 4.7: Distribution of the clicked documents' opinion for each task (excluding controlled documents) where TaskID refers to task questions in Table 4.1.

TaskID	Supporting	Opposing	Supporting & Opposing
1	87	9	5
2	50	5	1
3	52	1	2
4	33	20	3

Finally, as we can see from Tables 4.6 (a) and (b), we observed that the participants were likely to lean their beliefs towards the *yes*-polar after the search task when belief-inconsistent documents were presented. The possible explanation for this result is that the majority of the organic search results leaned towards *yes*-polar. Table 4.7 shows the distribution of the opinions in the documents clicked on by the participants, from which we can observe that most search results leaned towards the supportive information about the search task. This result implies that the distributions of the opinions in the search results would have an impact on the searchers' decision outcomes, as also reported by [Pogacar et al. 2017, White 2014].

An interesting finding was that such an effect could not be observed for the participants who encountered the belief-consistent opinions. Even the distribution of the documents' opinion leaned towards *yes*; as shown in Table 4.7, many participants who believed *no* prior to the search task still retain their beliefs in *no* when the belief-consistent opinion was presented. This phenomenon might be caused by the effect of the participants' confirmation bias that was strengthened by belief-consistent opinions. Therefore, it was difficult to change their beliefs even when the majority of the search results contained the *yes*-polar information.

In summary, we found that both opinion and credibility affected the participants' belief dynamics. As for opinion, participants were likely to alter their beliefs from one polar to the opposite polar when a belief-inconsistent controlled document was presented. As for credibility, participants were likely to retain their beliefs when the belief-consistent, high credibility controlled document was presented.

4.4.6 Search Behaviors and Belief Dynamics (RQ3)

To explore the relationship between search behaviors and belief dynamics, we performed a logistic mixed-effects regression analysis. For each search behavior b_i investigated in Section 4.4.3, we constructed a logistic mixed-effects regression model M_i where the participant's belief dynamics d (0: retain, 1: alter) is considered as a dependent variable and b_i is considered as a fixed effect. We then evaluated the goodness of fit of M_i compared to the null model

Table 4.8: Mean (M) and standard deviation (SD) for each type of belief dynamics and statistical testing result of each search behavior with the belief dynamic as dependent variables (*: significance level at 0.05 ,***: significance level at 0.001). Coefficient (β), standard error (SE), and z-statistic (z) are also reported.

	Belief Dynamics				**Model**				
	Alter		**Retain**						
	M	SD	M	SD	χ^2	p	β	SE	z
# of Consistent Clicks	0.70	1.25	1.83	1.97	81.94	***	-0.53	0.07	-7.53
# of Inconsistent Clicks	3.10	2.60	1.65	2.06	67.54	***	0.31	0.04	7.37
Authorship	0.14	0.37	0.21	0.48	5.36	*	-0.42	0.20	-2.03
Attribution	0.23	0.51	0.30	0.57	6.14	*	-0.36	0.16	-2.20

by using the likelihood ratio test and reported the results in Table 4.8 when M_i is statistically significant.

We found the main effect of the *number of consistent clicks* (β = -0.53, $p < 0.001$), *number of inconsistent clicks* ($\beta = 0.31$, $p < 0.001$), *authorship* (β = -0.42, $p < 0.05$), and *attribution* (β = -0.36, $p < 0.05$) on the belief dynamics. Table 4.8 shows the results of these behaviors according to the types of belief dynamics. Table 4.8 revealed that the participants who altered their beliefs were less likely to click on the documents that were consistent with their prior beliefs (alter: $M = 0.70$, retain: $M = 1.83$) and were more likely to click on the documents that were inconsistent with their prior beliefs (alter: $M = 3.10$, retain: $M = 1.65$).

For the evidence related behaviors, Table 4.8 revealed that the participants who retained their beliefs were more likely to submit the evidence that satisfies authorship (alter: $M = 0.14$, retain: $M = 0.21$) and attribution (alter: $M = 0.23$, retain: $M = 0.30$). Although we could not provide a clear explanation for this result, this could be explained by the psychological phenomenon called *motivated reasoning* [Kaplan et al. 2016] in which people spend their efforts collecting information that confirms their prior beliefs to maintain their beliefs in the face of information that challenges their beliefs. However, these behavioral differences seem smaller than those in the number of clicked documents that were consistent/inconsistent with the participants' beliefs.

As Table 4.8 shows, while we found that the participants issued more queries when the belief-inconsistent documents were presented (Section 4.4.4), such a behavioral difference was not found in terms of the belief dynamics. Only the types of opinion of the clicked documents were highly different. This implies that the participants' belief dynamics and their search effort had few relationships.

4.5 Discussion

In this section, the implication of the results and limitations of the study are discussed.

4.5.1 Implications

Our findings imply that search engines could mitigate the confirmation bias by learning the users' prior beliefs through their search histories and presenting documents that are inconsistent with their prior beliefs at the higher rank in the search results, as they are likely to be influenced by the higher-ranking documents [Joachims et al. 2007]. By doing so, users can consider the different viewpoints and obtain accurate information. Recently, search engines have started summarizing the contradictory information about a query and presenting it as featured snippets to draw an attention from their users [6]. Presenting the belief-inconsistent information in the featured snippets might be helpful to encourage users to aware of their underlying confirmation bias.

As for credibility, we observed that the participants were likely to retain their beliefs when the highly credible and belief-consistent documents were presented (Section 4.4.5). This result implies that people might be more likely to believe the inaccurate content if it seems credible and confirms their prior beliefs. Recently, some major search engines encouraged content providers to follow design frameworks such as AMP, and search results that followed the framework would be quickly displayed to the users in a well-designed fashion[7,8]. One concern is that such search results would likely to be considered credible by users, since documents with high usability and visually aesthetic are often considered credible [Lindgaard et al. 2011, Zhang and Li 2004]. Therefore, both search engines and their users should be aware that such search results might strengthen users' confirmation bias.

4.5.2 Limitations

Our work has several limitations that should be acknowledged.

Effect of organic search results: It might be questioned whether the participants' behaviors and belief dynamics were affected by organic search results as well as by the controlled documents. Regarding the organic search results, we observed the results similar to White's [White 2014], where people were likely to lean their beliefs towards the dominant polar in the search results, as shown in Table 4.7 that the majority of search results were leaning towards *yes*. However, we also observed the effect controlled documents had on their belief-dynamics; the participants who believed *no* prior to the task still retained *no* as their posterior beliefs when the controlled document that was consistent with their prior beliefs was presented. This means that the participants' behaviors and belief dynamics were affected by

[6] https://www.blog.google/products/search/reintroduction-googles-featured-snippets/

[7] https://blog.google/products/search/search-results-are-officially-ampd/

[8] https://blogs.bing.com/Webmaster-Blog/September-2018/Introducing-Bing-AMP-viewer-and-Bing-AMP-cache/

Table 4.9: Distribution of prior knowledge for each task where TaskID refers to task questions in Table 4.1.

TaskID	not at all	a little	good	excellent
1	168	28	6	0
2	154	42	18	0
3	190	7	1	0
4	178	12	3	2

both the controlled documents and the organic search results. The current analyses reported in Section 4.4 cannot fully separate these two effects.

Noticing of the search result manipulation: We removed the participants who noticed our search result manipulation via the question in the exit questionnaire (Section 4.3.2). One limitation of such an approach is that we could not identify all the participants who noticed the manipulation. However, since we did not tell the participants the exact research questions investigated in the study, we still regarded the data of these participants as useful even if they felt strange about the presented search results.

Belief-related Scale: We used the 4-point Likert scale to measure the participants' prior and posterior beliefs, which means that we assumed they were able to decide either lean-yes or lean-no (Section 4.3.2). One problem with this assumption is that it would be difficult for participants to decide their beliefs (especially before starting the search task) when they had little knowledge about the topic; thus, their beliefs were weak. As shown in Table 4.9, most participants had little knowledge of the search tasks. However, we observed that the opinion affected the behaviors (Section 4.4.4) and belief dynamics (Section 4.4.5) of even such participants.

Generalizability: Our study cannot be generalized to other domains of search tasks. The question as to whether our findings are applicable to other domains remains open. For example, the political domain, where people are more likely to have strong prior beliefs and where complicated diverse opinions exist on the web. However, we think that the search tasks tackled in this study (*i.e.*, health-related search tasks that can be answered *yes* or *no*) are also an important potion of search tasks in our recent society.

4.6 Summary

In this study, we investigated the effects documents' opinion and credibility have on people's search behaviors and belief dynamics. The results were that the participants are likely to spend more effort in their search process by issuing more queries when the documents that were inconsistent with the participants' prior beliefs were presented. Furthermore, they tended

to alter their beliefs when an opinion that was inconsistent with their prior beliefs was presented, and they tended to retain their beliefs when an opinion that was consistent with their prior beliefs was presented. Also, the effect depended on the level of credibility. Finally, the participants' belief dynamics and search efforts had few relationships; only the types of opinion in the clicked documents were different. In the future, we aim to analyze the search tasks in which the participants did not click on the controlled documents because it would be interesting to observe whether they were affected by the title or the snippet of the controlled document even though they did not click on it. We also aim to develop a new re-ranking algorithm to encourage searchers to perform careful information searches.

5 Encourage the searchers to interact with the alternative information

5.1 Introduction

Major search engines, such as Google and Bing, have recently provided a set of questions and answers that are relevant to specific queries above the search results. For example, search engines return questions containing **Does green tea burn belly fat?** in response to the input query *"green tea weight loss"* and searchers can click on that question to reveal an answer. With this functionality, known as *People also ask*, searchers do not have to go through an excessive number of documents to fulfill their information needs [Chilton and Teevan 2011, Lin et al. 2003, Voorhees and Tice 1999]. Figure 5.1 shows an actual screenshot of *People also ask* for the query *"green tea weight loss"* generated by a commercial search engine.

Recently, various studies have revealed the effect of the search process on confirmation bias, where searchers seek information to confirm their beliefs. For example, White and his colleagues [White 2013, 2014, White and Hassan 2014, White and Horvitz 2015] revealed that search engines can be biased toward a particular remedy, which makes the searchers' decisions lean toward positive information related to the remedy. Furthermore, Pogacar et al. [Pogacar et al. 2017] reported that searchers' post-search decisions were likely to be inaccurate if the search results were biased toward inaccurate information. Suppanut et al. [Pothirattanachaikul et al. 2019] also revealed that searchers were likely to retain their beliefs when encountering a search result that provides information consistent with their beliefs.

While these studies revealed the effect of search results on confirmation bias, discovering the effect of a question answering system on confirmation bias is also important. Presenting *People also ask* without understanding its effect might strengthen searchers' confirmation bias. Questioners tend to ask questions to which a positive answer is considered as strong confirmation and a negative answer is considered as weak disconfirmation [Dardenne and Leyens 1995, Skov and Sherman 1986, Snyder and Swann 1978, Swann et al. 1982]. Moreover, questioners are likely to prefer a confirmation response over a disconfirmation response [Heritage 2013]. For example, searchers who believe that green tea extract is helpful for losing weight may click on the question **Does green tea burn belly fat?** that presents the answer that con-

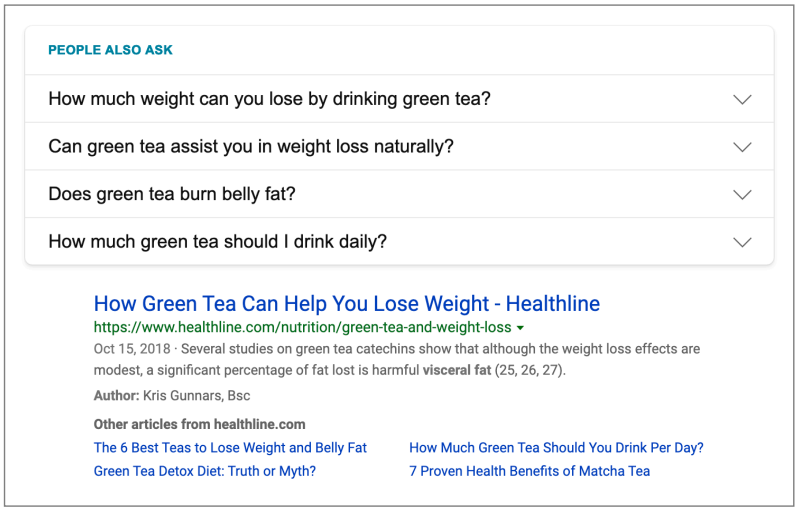

Figure 5.1: Search engine result page (SERP) where *People also ask* is displayed above organic search results.

firms the effectiveness of green tea extract on weight loss although researchers argue that it is ineffective and harmful [Chalasani et al. 2014, Jurgens et al. 2012].

To mitigate confirmation bias in Web searches, previous studies have suggested that people should be exposed to information that is not consistent with their beliefs [Pothirattanachaikul et al. 2019, Schwind et al. 2012] and should carefully verify information by spending more time on a search task, issuing more queries, *etc* [Edwards and Smith 1996, Yamamoto et al. 2018a,b]. In addition, Trope and Bassok suggested that, when testing the null hypothesis, people are likely to consider questions related to the *alternative* hypothesis when such questions are provided [Trope and Bassok 1982, 1983]. Such studies imply that *People also ask* can mitigate confirmation bias if it is carefully designed.

In this study, two experiments[1] were conducted to investigate the effect of *People also ask* on search behaviors and *beliefs* (*i.e.,* how people's beliefs change after the search process). In the first experiment, the effect of the *question*, *answer*, and the answer's *opinion* was investigated. The second experiment investigated the effect of the *alternative question*, i.e., a question related to the solution that can achieve the same goal as a query. The aim of this study is to learn the characteristics of a question answering system that mitigate confirmation bias. Specifically, the objective is to answer the following research questions.

- **RQ1:** How do the question and answer affect search behaviors and searchers' beliefs?

[1] The experiments were conducted in Japanese, and the information provided in the tables and figures is translated from Japanese to English if necessary.

- **RQ2:** How do the alternative questions affect search behaviors and searchers' beliefs?

For both experiments, participants were required to use a customized search system to find answer to three medically related *yes-no* questions (e.g., *Is acupuncture effective in relieving back pain?*). Both experiments were conducted as follows. Participants first assessed their beliefs in relation to the remedy (*acupuncture*). Such beliefs were then defined as a *prior belief*. Then, participants were led to a search system that was customized to display *People also ask* above fifty organic search results. Questions and answers were selected based on the experimental condition. After complete the search task, participants were then asked to assess their beliefs in relation to the remedy, which was then defined as *posterior belief*. By doing so, it was possible to analyze the effects of questions and answers on search behaviors and beliefs.

The main findings of this study can be summarized as follows. (1) Participants issued fewer queries and spent less time on SERPs when *People also ask* were presented. (2) Participants were less likely to interact with a SERP when they first encountered a belief-inconsistent answer. (3) We could not confirm the effect of *People also ask* on beliefs. The findings suggest that *People also ask* might not help to mitigate confirmation bias because, differing from encountering a belief-inconsistent document in search results, participants are likely to spend less effort on the search process (i.e., issue fewer queries) when they first encounter a belief-inconsistent answer. Moreover, an additional experiment is required to validate that participants who first encounter a belief-inconsistent answer are more likely to alter their beliefs as the number of such participants was inadequate.

5.2 Material

This section introduces the tasks which were used in both experiments and the method used to collect questions and answers.

5.2.1 Tasks

Four medical symptoms were adopted from a study by Pogacar et al. [Pogacar et al. 2017] as these symptoms have more than one possible remedy. For each symptom, two candidate remedies were selected from the Cochrane medical knowledge base as the *target* and *alternative* remedy. Medical experts know that the systematic reviews carried out by Cochrane meet the highest standard in evidence-based healthcare [Cipriani et al. 2011]. To form the tasks, we created questions related to the effectiveness of the target remedy for a given symptom. Table 5.1 lists the tasks, symptoms, target remedies, and alternative remedies. Note that in Table 1, Task 4 was used as the training task.

Table 5.1: Search tasks translated from Japanese. Candidate remedies and their Cochrane review URLs are shown.

Search Task			Remedies	
ID	**Symptom**	**Task**	**Target Remedy**	**Alternative Remedy**
1	back pain	Is acupuncture effective in relieving back pain?	acupuncture (https://bit.ly/2kbOlm0)	insoles (https://bit.ly/2m6YPDL)
2	eczema	Are moisturizers effective in relieving eczema?	moisturizers (https://bit.ly/2kqZyPN)	probiotics (https://bit.ly/2kBBCsN)
3	tooth decay	Are sealants effective in relieving tooth decay?	sealants (https://bit.ly/2kBRQlF)	chlorhexidine (https://bit.ly/2kpy9xH)
4	asthma	Is caffeine effective in relieving asthma?	caffeine (https://bit.ly/2kCcGBB)	homeopathy (https://bit.ly/2kBBWYx)

5.2.2 Questions and Answers

To collect the questions, a query in the format ”target remedy + symptom” (e.g., acupuncture back pain) was issued to Google and the top four *yes/no* questions were collected from *People also ask*. We used *yes/no* questions because we could pair the answers with any questions that were collected for the same task. Moreover, *yes/no* questions make us able to measure participants' degrees of beliefs easily [White 2013]. In addition, the query in the form of ”alternative remedy + symptom” was issued to Google in order to collect a single alternative-related *yes/no* question from *People also ask* that would be used in the second experiment.

To collect answers, workers from the Japanese crowdsourcing platform Lancers [2] were recruited. For each symptom and remedy in Table 5.1, the workers were asked to look for articles on the Internet that support/oppose the remedy. The workers were asked to collect articles that satisfied either of the following conditions: (1) the article explicitly refers to an academic paper or technical report presented at an academic conference; (2) the article is from a medical institute or government agency; and (3) the article explicitly reveals that the author is a medical expert. As Cochrane reviews are considered as the highest standard, it is very likely that their articles satisfy the established criteria resulting in a risk that workers would only submit articles from Cochrane. Thus, to obtain articles from diverse sources, workers were asked not to submit any articles from Cochrane. Finally, they were asked to submit the article and its URL that they think supports/opposes the remedy. Each worker received approximately $4 as compensation. For each remedy, we selected two supporting and opposing paragraphs from the submitted articles as the answers. Note that the selected text was easy even laypersons to understand. Table 5.2 lists the questions and example answers used in the experiments.

5.3 Method

This section first provides an overview of the method used in both experiments and then describes each experiment in detail. We also introduce the search system that was designed for this study and outline how participants were recruited.

5.3.1 Overview

First, participants were informed how their data would be used for research purposes and asked for consent. Only those who agreed to the conditions could participate in the training task (Table 5.1, Task 4: *Is caffeine effective for asthma?*). The participants completed this task using our search system. During training, *People also ask* containing four questions with supporting answers for caffeine was presented. After the training task, participants were asked to complete three search tasks as follows. First, for each search task, the participants were given instructions that comprised the background of the symptom and a target remedy. After

[2] https://www.lancers.jp/

Table 5.2: Questions and example answers that would be contained in People also ask (translated from Japanese). As the opinion was not manipulated in Experiment 2, the opposing answers were not required for the alternative question.

TaskID	Remedy	Question collected for People also ask	Example answer	
			Supporting	Opposing
1	**Acupuncture**	Does acupuncture relieve back pain?	Acupuncture is an effective treatment that relieves lower back pain for at least six months. It is said that the effectiveness of acupuncture is two times higher than the usual treatments.	There was no evidence that acupuncture was effective. Even if there is, it is considered as placebo.
		Does acupuncture help back muscle pain?		
		Is acupuncture good for upper back pain?		
		Does acupuncture help bad backs?		
	Insole	Do insoles help with back pain?	By using the insole, supporting the weight on the entire sole of the foot reduces the burden on the foot and thus reduces fatigue. It can be expected to treat a corn, knee pain, lower back pain, and bunion.	-
2	**Moisturizer**	Is Vaseline good for eczema?	Many doctors encourage that atopic dermatitis can be improved by keeping clothing and skin clean, taking care of moisturizing, not irritating the skin. By doing so, mild atopic dermatitis can be improved without using drug.	Applying a moisturizer directly to areas with eczema can worsen eczema. It is better to avoid applying the moisturizer to the areas with eczema because some moisturizers may not be applied over the ointment.
		Can I put lotion on my eczema?		
		Can a moisturizer cure eczema?		
		Can you use baby lotion to treat eczema?		
	Probiotics	Will probiotics help with psoriasis and eczema?	Probiotics as a therapeutic effect have been shown to improve skin rash especially in patients with food allergies and allergen positive cases.	-
3	**Sealants**	Are tooth sealants a good idea?	Sealant has a very high airtightness, and it fills the surface of the teeth and hardens, so there is no space for bacteria to propagate.	In the case of filling the back teeth with resin sealant, a heavy weight is applied to the teeth, so it is often seen that the teeth are decayed from the boundary between the sealant and the teeth.
		Are sealants for teeth necessary?		
		Does sealing teeth prevent cavities?		
		Do sealants help sensitive teeth?		
	Chlorhexidine	Does chlorhexidine kill MRSA?	Mouthwash is actually used in dental clinics. Reducing bacteria in the patient's mouth in advance leads to better treatment of cavity and periodontal disease	-

reading the instructions, the participants were asked to complete a pre-task questionnaire. They were asked to indicate their beliefs in relation to the target remedy by answering the question *Do you think acupuncture is effective in relieving back pain?* using a four-point Likert scale (1: No; 2: Lean no; 3: Lean yes; 4: Yes). The scale was structured such participants had to indicate whether they believed the remedy to be ineffective or effective. The scores for this question were considered as their *prior beliefs*. Then, participants were asked to rate their prior knowledge about the symptom and target remedy using a four-point Likert scale (1: Not at all; 2: A little; 3: Good; 4: Excellent).

Subsequently, participants performed the search task using our search system (Section 5.3.6). Note that they were not allowed to use other commercial search engines. During the search task, they were allowed to issue queries to the system and click on documents returned by the system without time constraints. Once they felt satisfied that they had completed the task, they could click the finish button at the top right of the search page (Figure 5.2).

At the end of each search task, participants were asked to complete a post-task questionnaire. They were first asked to provide their beliefs about the target remedy again using the same four-point Likert scale. Their scores for this question were regarded as their *posterior beliefs*. Next, they were asked to provide their satisfaction with the search system on a four-point Likert scale (1: Not satisfied; 2: Somewhat not satisfied; 3: Somewhat satisfied; 4: Satisfied). Finally, they were asked to select the documents that contain evidence that support their posterior beliefs. A list of the documents they clicked on during the search task was provided. They were asked to select one or more documents from the list.

After completing the three search tasks, participants had to complete an exit questionnaire that collected demographic information (gender, education level, and search engine familiarity). Note that participants did not have to answer the demographic questions on gender and educational background. Finally, they were asked if anything had bothered them during the search task to ascertain which participants had noticed the goal of the experiments.

5.3.2 Dependent Variables

For both experiments, two types of dependent variables were analyzed, i.e., search behaviors and beliefs. Regarding *search behaviors*, Yamamoto et al. [Yamamoto et al. 2018a, Yamamoto and Yamamoto 2018, Yamamoto et al. 2018b] discussed the behaviors of people who *search carefully*. Such people are expected to spend more time searching, issue more queries, browse more documents, check a deeper ranked result, spend more time on a search task, and increase the evidence to support their decisions. In this study, it is assumed that people who search carefully are likely to have the search behaviors mentioned above. The detail of each search behavior will be discussed in Section 5.4.1.

In this study *beliefs* were based on the answer obtained from the belief-related questionnaires that were asked before and after the search task. The focus was on cases in which participants changed (or did not change) their prior beliefs from one polarity to the opposite

polarity after the search task. We considered cases in which the participants changed their prior beliefs from one polar to the opposite polar as they *altered* their beliefs, and cases in which participants did not change prior beliefs from one polar to the opposite polar as *retained* beliefs.

5.3.3 Hypothesis

Experiment 1: Regarding **RQ1**, we investigated the effect of *People also ask*'s components including the *question*, *answer*, and answer's *opinion* that participants first encountered on search behaviors and beliefs. According to previous studies, people are likely to spend less effort on a SERP when presented by a question answering system [Belkin et al. 2000, Bernstein et al. 2012, Chilton and Teevan 2011, Lin et al. 2003]. We assumed that *People also ask* should have the same effect as other question answering systems. Therefore, we proposed the following hypothesis.

H1-1: Presenting *People also ask* discourages people from searching carefully compared to when it is not presented.

However, there are studies, which suggested that people are likely to search carefully and alter their beliefs after the search task when they encounter information that is inconsistent with their beliefs [Pothirattanachaikul et al. 2019, Schwind et al. 2012]. Thus, *People also ask* might help mitigate confirmation bias if information that is inconsistent with the searcher's beliefs is included. Based on these discussions, we proposed the following hypotheses.

H1-2: People who first encounter a belief-*inconsistent* answer are likely to search carefully compared to first encountering a belief-consistent answer.

H1-3: People who first encounter a belief-*consistent* answer are likely to retain their beliefs after the search compared to first encountering a belief-inconsistent answer.

Experiment 2: Regarding **RQ2**, we investigated the effect of the *alternative question*, i.e., the question contained in *People also ask* that is related to the solution that can achieve the same goal as a query on the search behaviors and beliefs. According to Sunstein, people should consider information that they might have not chosen in advance to mitigate confirmation bias and make more effective decisions [Sunstein 2017]. In this study, we consider such information as an alternative, i.e., a remedy that differs from the one explained in the search task but which allows people to improve the same symptom. An efficient practice to encourage people to consider an alternative is to present alternatives as questions as suggested by Trope and Bassok that people are likely to consider questions related to the alternative hypothesis when such questions are provided [Trope and Bassok 1982, 1983]. Therefore, we proposed the following hypotheses.

H2-1: Presenting an alternative question encourages people to search carefully.

H2-2: Presenting an alternative question encourages people to alter their beliefs related to the target remedy.

5.3.4 Experiment 1

Experimental design: Participants were presented with *People also ask* containing four questions, all of which were related to the target remedy. The answers were revealed only if the questions were clicked on. The experiment was manipulated with two conditions: *belief-consistent first* and *belief-inconsistent first*. For *belief-consistent first*, answers that were consistent with participants' beliefs were placed in the first and second ranked questions while inconsistent answers were placed in third and fourth ranked questions. For *belief-inconsistent first*, answers that were inconsistent with participants' beliefs were placed in the first and second ranked questions while consistent answers were placed in the third and fourth ranked questions. Therefore, we were able to balance the number of participants for each type of opinion as the higher ranked items were likely to receive more attention than lower ranked items [Guan and Cutrell 2007, Joachims et al. 2007]. In addition, we assigned the controlled condition where *People also ask* was not presented. Based on these settings, there are three experimental conditions: C_1 (controlled condition), C_2 (belief-consistent first), and C_3 (belief-inconsistent first). This experiment was repeated such that each participant was assigned all three conditions. For each condition, the participants were assigned one of the three tasks in Table 5.1. The tasks were rotated according to the Graeco-Latin square design [Kelly 2009].

Independent variables: In this experiment, *question*, *answer*, and *opinion* were major factors under investigation. *Question* is the binary variable indicating whether *People also ask* was presented. *Answer* is the variable indicating whether participants clicked on the question to see the answer. *Opinion* is the variable indicating whether participants first encounter a belief-consistent or a belief-inconsistent answer. The observations can be considered as hierarchical whereby factors are nested within each other such that the *answer* is nested within the *question*, which made it possible for participants to click on the question to see the answers only if the questions were presented. *Opinion* is nested within the *answer* such that participants were affected by the answer's opinion only if they clicked on the question.

Statistical analysis: To estimate the effect of the nested factors, orthogonal contrasts are required. Orthogonal contrasts are helpful in obtaining estimates of main and nested effects for mean comparisons between groups of data to obtain specific residuals [Nogueira 2004].

Apart from the *question* that was not nested under any factors, orthogonal contrasts were created separately for *answer* and *opinion*. For the analyses regarding RQ1, we conducted linear mixed-effects regression analysis, which is applicable for hierarchical and unbalanced data [Bates et al. 2015, Winter 2013].

We constructed a linear mixed-effects regression model where *question*, *answer*, and *opinion* were regarded as fixed effects. The model applied nested random effect structure,

i.e., **Question/Answer/Opinion** to simulate the hierarchical structure of the experiment [3]. Initially, we included random intercepts and random slopes; **1+Question/Answer/Opinion** on participants and tasks to maintain the maximal random effect structure to ensure that the Type I error is controlled [Barr et al. 2011]. However, we found that the model did not reach convergence due to the size of the observation data. Thus, we removed the by-participants and by-task random slopes from the model. The limitation of removing random slopes will be discussed in Section 5.5.2. The final model contains a nested random effect, random intercepts on participants, and random intercepts on tasks. The formula for such model is $d \sim$ Question + Answer + Opinion + (1|Question/Answer/Opinion) + (1 | Participant) + (1 | Task), where d represents a dependent variable. A likelihood ratio test was performed with the model and the *null model*, which considers only random effects. Likelihood ratio test is helpful to validate whether the fixed effects significantly improved the model as suggested by Field et al [Field et al. 2012]. The significance level in this study was set to 5%. Note that log transformation was applied to temporal features, such as page dwell time.

5.3.5 Experiment 2

Experimental design: In this experiment, the *rank* of the alternative question was manipulated based on whether the alternative question appeared on the *first* or *fourth* rank of *People also ask*. In addition, there was also controlled condition where all the questions presented in *People also ask* were related to the target remedy. All questions contained only the supporting answer toward both target and alternative remedies. Based on these settings, three experiment conditions were generated: C_1 (controlled condition), C_2 (alternative question ranked first), and C_3 (alternative question ranked fourth). Similar to Experiment 1, this experiment was repeated and tasks were rotated according to the Graeco-Latin square design.

Independent variables: In this experiment, *alternative question*, *answer*, and *alternative answer* were investigated. *Alternative question* indicated whether the alternative question is presented at rank 1, rank 4, or not presented (control). *Answer* indicates whether participants clicked on the question to see the answer. *Alternative answer* indicates whether participants clicked on alternative question once.

Statistical analysis: The observations are hierarchical such that participants could encounter an alternative answer only if the alternative question was presented and clicked on. Thus, orthogonal contrasts were created separately for the *alternative question* and *alternative answer*. Similar to the first experiment, a linear mixed-effects regression analysis was performed. A linear mixed-effects regression model was constructed, whereby *alternative question*, *answer*, and *alternative answer* were considered fixed effects. The model applied the following nested random effect structure: **Alternative question/Answer/Alternative answer**. The model also included random intercepts on participants and tasks. The final model

[3] The model is expressed in the syntax of lmer, a widely used mixed-effects fitting method contained in lme4 [Bates et al. 2015]

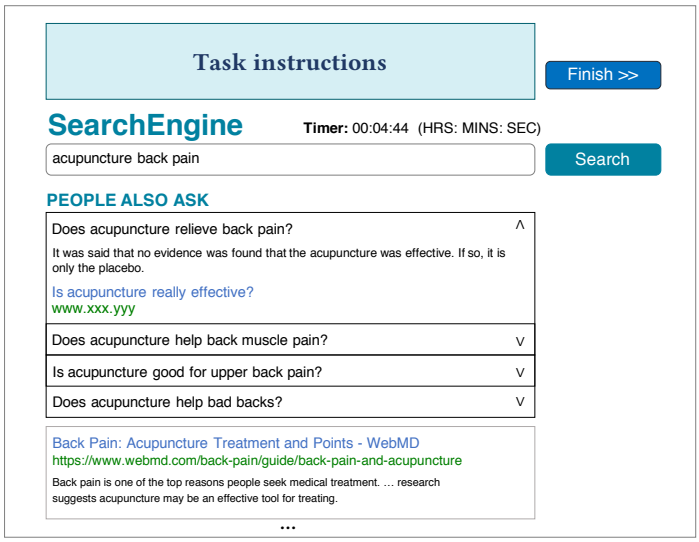

Figure 5.2: Screenshot of our search system (translated from Japanese). *People also ask* was presented above organic search results. The answer was revealed when the question was clicked on.

we reported contains nested random effects, random intercepts on participants, and random intercepts on tasks. The formula for such a model is $d \sim$ Alternative question + Answer + Alternative answer + (1|Alternative question/Answer/Alternative answer) + (1 | Participant) + (1 | Task).

5.3.6 Search System

Figure 5.2 shows a screenshot of our search system. To help the participants get familiar with our search system easily, we imitated the user interface of a commercial search engine with task instructions above the search box. The initial screen comprises the initial query "target remedy symptom" in the search box, *People also ask* located under the search box, and organic search results under *People also ask*. The questions and answers in *People also ask* were selected based on the experiment and condition. Each answer contained text, a title, a snippet, and a URL. Fifty organic search results were obtained from the Bing Web Search API[4] based on the initial query. Each search result comprised a title, a snippet, and a URL. For both the answer and organic search result, the Web page was displayed in a separate tab when either the title or URL was clicked. Note that the initial screen was the only screen that presented *People also ask*. The participants could freely issue any queries and received

[4] https://azure.microsoft.com/en-us/services/cognitive-services/

organic search results generated by Bing Web Search API, but *People also ask* would not appear afterward.

5.3.7 Participants

Participants were recruited via Lancers for both experiments. Since we used a crowdsourcing platform to conduct the experiment, controlling the quality is important. To control the quality, participants were informed (in the instruction page) that their submissions may be rejected if we judged they did not perform the task seriously. Each participant received approximately $4 as compensation.

Experiment 1: 276 participants were recruited from August 21 to August 22, 2019. Out of 276 participants, we removed 10 participants who did not perform the task seriously, and we removed 14 participants who accidentally closed the browser and attempted to resume the task. We then removed the data of four outliers who spent too little/too much time on the task. In total, 744 search tasks from 248 participants were obtained for analysis. The median task completion time was 17 minutes and 59 seconds, and the standard deviation was 15 minutes and 15 seconds. The demographic data for these 248 participants are shown in Table 5.3(a).

Experiment 2: 301 participants were recruited from September 27 to September 29, 2019. Out of 301 participants, we removed 28 participants who did not perform the task seriously, and we removed 15 participants who accidentally closed the browser and attempted to resume the task. We also removed a single outlier who spent too much time on the task. In total, 771 search tasks from 257 participants were obtained for analysis. The median of the task completion time were 17 minutes and 33 seconds, while the standard deviation was 17 minutes and 7 seconds. The demographic data for these 257 participants are shown in Table 5.3(b). Note that people who participated in Experiment 1 were not permitted to participate in Experiment 2.

5.4 Results

This section first discusses the effect of question and answer on participants' search behaviors and beliefs (RQ1). Then, the effect of an alternative question is discussed (RQ2).

5.4.1 Search Behavior

The following search behaviors were analyzed for each participant during the search task:

- **# of Queries**: Number of queries issued in a search task.
- **# of Clicks**: Number of documents clicked on in a search task (excluding clicks from the answers).
- **# of Evidence**: Number of evidence (excluding those from the answers).
- **Deepest Document Rank**: Lowest rank the participant clicked.

Table 5.3: Demographic data of participants.

Experiment 1 (a)					
Gender	*n*	Educational background	*n*	Search engine familiarity	*n*
male	151	university educated	160	rarely use	1
female	96	not university educated	68	several times per week	21
N/A	1	N/A	20	once a day	39
				more than once a day	187
Experiment 2 (b)					
Gender	*n*	Educational background	*n*	Search engine familiarity	*n*
male	127	university educated	159	rarely use	0
female	129	not university educated	69	several times per week	42
N/A	1	N/A	29	once a day	20
				more than once a day	195

- **Page Dwell Time**: Average time the participant spent on the documents.
- **SERP Dwell Time**: Average time the participant spent on the search results page.
- **Task Time Spent**: Amount of time the participant spent on the search task.

5.4.2 Effect of Question and Answer on Search Behavior and Beliefs (RQ1)

To verify hypotheses **H1-1** to **H1-3**, we performed a linear mixed-effects regression analysis with each participants' search behaviors obtained in Experiment 1. Table 5.4 presents the results of the analysis. The "Fixed Effects" column in Table 5.4 shows the coefficient, standard error, t-statistic, and p-value of each fixed effect.

Effect of question: We found a main effect of the *question* on the number of queries ($\beta = -0.73, p < 0.001$) and time spent on SERP ($\beta = 0.15, p < 0.01$). Table 5.5 shows the results of these behaviors according to whether questions were presented. As can be seen, the presence of questions discouraged participants from issuing queries (absent: $M = 2.37$, present: $M = 1.80$), and spending time on the SERP (absent: $M = 25.30$, present: $M = 24.99$).

Effect of answer: We found a main effect of the answer on the number of queries ($\beta = -0.27, p < 0.01$), number of document clicks ($\beta = -0.30, p < 0.05$), and time spent on search result page ($\beta = 0.20, p < 0.001$). Table 5.6 shows the results of these behaviors according to whether the participants encountered the answer. The results in Table 5.6 show that an answer encouraged participants to click on more documents (not encounter: $M = 4.67$,

Table 5.4: Mean (*M*) and standard deviation (*SD*) for each experimental condition in Experiment 1 (C_1: control, C_2: belief-consistent first, and C_3: belief-inconsistent first) and statistical testing results of each search behavior of the model with different fixed effects (***: significance level at 0.001, **: 0.01, and *: 0.05). Coefficient (β), standard error (*SE*), t-statistic (*t*), and p-value (*p*) of each fixed effect are reported. The χ^2 and *p*-values of the model which is statistically significant compared to the null model are also reported.

	Condition						**Fixed Effects**											
	C_1		C_2		C_3		**Question**				**Answer**				**Opinion**					
	M	*SD*	*M*	*SD*	*M*	*SD*	β	*SE*	*t*	*p*	β	*SE*	*t*	*p*	β	*SE*	*t*	*p*	χ^2	*p*
# of Queries	2.37	1.87	1.79	1.62	1.81	1.68	-0.73	0.11	-6.68	***	-0.27	0.09	-2.99	**	0.14	0.15	0.95	0.34	10.70	*
# of Clicks	4.47	3.19	4.70	2.88	4.73	3.19	0.03	0.15	0.22	0.83	-0.30	0.13	-2.30	*	0.61	0.21	2.94	**	9.94	*
Deepest Document Rank	11.82	11.96	11.89	10.92	12.06	11.06	0.58	0.73	0.79	0.43	0.59	0.59	1.00	0.32	3.00	0.98	3.05	**	10.20	*
Page Dwell Time (sec)	45.89	39.01	47.23	52.88	48.35	49.39	-0.01	0.05	-0.23	0.82	0.05	0.04	1.40	0.16	0.05	0.06	0.87	0.39		
SERP Dwell Time (sec)	25.31	22.61	25.93	22.80	24.06	20.76	0.15	0.05	3.14	**	0.20	0.04	3.14	***	-0.04	0.06	-0.59	0.56	10.10	*
Task Time Spent (sec)	400.05	335.43	386.50	320.26	390.87	324.35	0.01	0.03	0.36	0.72	0.05	0.03	1.74	0.08	0.11	0.04	2.40	*	8.27	*
# of Evidence	2.06	1.28	2.25	1.40	2.34	1.47	0.14	0.09	1.57	0.12	-0.13	0.07	-1.92	0.05	0.10	0.12	0.91	0.36		

Table 5.5: Participants' behavior according to whether questions were presented. (Experiment 1)

	Absent		**Present**	
	M	*SD*	*M*	*SD*
# of Queries	2.37	1.87	1.80	1.65
SERP Dwell Time (sec)	25.30	22.70	24.99	21.84

Table 5.6: Participants' behaviors according to whether they encountered answers. (Experiment 1)

	Not Encounter		**Encounter**	
	M	*SD*	*M*	*SD*
# of Queries	1.88	1.59	1.43	1.92
# of Clicks	4.67	3.09	4.94	2.79
SERP Dwell Time (sec)	22.29	19.66	38.60	26.80

encounter: $M = 4.94$), issue fewer queries (not encounter: $M = 1.88$, encounter: $M = 1.43$), and spend more time on the SERP (not encounter: $M = 22.29$, encounter: $M = 38.60$).

Effect of opinion: We found a main effect of the answer's *opinion* participants first encountered on document clicks ($\beta = 0.61, p < 0.01$), deepest document rank ($\beta = 3.00, p < 0.01$), and time spent on the search task ($\beta = 0.11, p < 0.05$). Table 5.7 shows the results of these behaviors according to the type of opinions encountered by the participants. The results demonstrate that belief-inconsistent answers discouraged participants from clicking on the documents (consistent: $M = 5.54$, inconsistent: $M = 4.40$), clicking on a deeper rank document (consistent: $M = 17.13$, inconsistent: $M = 10.70$), and spending time on the task (consistent: $M = 564.15$, inconsistent: $M = 417.39$).

Beliefs: For beliefs, we focused on whether participants changed their beliefs from one polar to the opposite polar after the search task. The cases in which participants changed their beliefs from one polar to the opposite polar after the search tasks were regarded as they *altered* their beliefs, otherwise *retained*. Tables 5.8 shows the participants' beliefs after the search task (columns) compared to their beliefs prior to the search task (rows). For example, when an answer that was consistent with the beliefs was first encountered, 84.61% (76.92% + 7.69%) of participants who were "Lean No" prior to the search task altered their beliefs to either "Lean Yes" or "Yes" after the search task. Tables 5.8 shows that most participants

Table 5.7: Participants' behaviors according to the answer's opinion they first encountered. (Experiment 1)

	Consistent		**Inconsistent**	
	M	*SD*	*M*	*SD*
# of Clicks	5.54	2.95	4.40	2.56
Deepest Document Rank	17.13	12.76	10.70	9.11
Task Time Spent (sec)	564.15	283.96	417.39	263.45

altered their beliefs toward yes polar after the search task. This may be explained by the fact that people who clicked on *People also ask* may also investigate organic search results, which are likely to lean toward *yes* polar or the remedy is effective [White 2013]. The results shown in Tables 5.8 demonstrate that participants who assessed their beliefs as *lean-yes* prior to the search task were less likely to retain their beliefs when they first encountered an answer that was inconsistent with their beliefs (②) compared to the case where the belief-consistent answer was first encountered (①). This result is consistent with the previous study which suggested that people are less likely to retain beliefs when they are exposed to a document that is inconsistent with their beliefs [Pothirattanachaikul et al. 2019, Schwind et al. 2012]. These results appear to support hypothesis **H1-3**, which indicates that people who encounter who first encounter with belief-*consistent* answer are likely to retain their beliefs compared to those who encounter a belief-inconsistent answer. To determine whether we can accept **H1-3**, we performed logistic mixed-effects regression analysis with question, answer, and opinion as independent variables. A logistic mixed-effects model was constructed using beliefs as a binary dependent variable *d* (1: alter, 0:retain). However, the model was not statistically significant compared to the null model ($\chi^2 = 7.27$, $p = 0.06$); thus, we cannot accept **H1-3**. This may be due to the size of participants who interacted with *People also ask*. Thus, further experiment might be required to increase the size of participants.

Summary: In summary, hypothesis **H1-1** was partially supported by the results because *question* discourages participants from issuing queries and spending time on SERP while *answer* encourages participants to issue fewer queries, spend more time on SERP, and click on more documents. For hypothesis **H1-2**, the results did not support the hypothesis because they suggested that participants who first encountered a *belief-inconsistent* answer are less likely to click on the documents, check a deeper rank document, and spend time with the search task. In this state, we cannot accept hypothesis **H1-3** because the logistic mixed-effects model was not statistically significant compare to the null model.

Table 5.8: Percentage of beliefs for Experiment 1 grouped by whether participants first encountered with belief-consistent answer. (for each prior belief, the polar that most participants leaned toward in posterior belief is highlighted in gray).

(a) Belief-consistent first (*n*=39)

		Posterior Belief			
		No (n=0)	Lean-No (n=2)	Lean-Yes (n=27)	Yes (n=10)
Prior Belief	No (n=2)	0.00%	0.00%	50.00%	50.00%
	Lean No (n=13)	0.00%	15.38%	76.92%	7.69%
	Lean Yes (n=20)	0.00%	0.00%	80.00%	20.00%
	Yes (n=4)	0.00%	0.00%	0.00%	100.00%

(b) Belief-inconsistent first (*n*=43)

		Posterior Belief			
		No (n=1)	Lean No (n=11)	Lean Yes (n=20)	Yes (n=11)
Prior Belief	No (n=5)	0.00%	20.00%	40.00%	40.00%
	Lean No (n=14)	0.00%	14.29%	71.43%	14.29%
	Lean Yes (n=23)	4.35%	34.78%	34.78%	26.09%
	Yes (n=1)	0.00%	0.00%	0.00%	100.00%

5.4.3 Effect of Alternative Question on Search Behaviors and Beliefs (RQ2)

To verify hypotheses **H2-1** and **H2-2**, linear mixed-effects regression analysis was performed with each participants' search behaviors obtained in Experiment 2. The results are presented in Table 5.9.

Effect of alternative question: We found a main effect of the *alternative question* on the number of queries ($\beta_a = -0.10, p < 0.01$) and number of evidence ($\beta_a = 0.09, p < 0.001$). Table 5.10 shows the results of this behavior according to whether the alternative question was presented. The results demonstrate that the presence of an alternative question encouraged participants to issue fewer queries (absent: $M = 2.37$, present: $M = 2.02$) and submit more evidence (absent: $M = 1.87$, present: $M = 2.15$).

Effect of answer: Similarly to Section 5.4.2, we found a main effect of answer on the number of queries ($\beta = -0.85, p < 0.001$), and time spent on the SERP ($\beta = 0.58, p < 0.001$). In addition, we found an effect of the answer on the number of evidence submitted by participants ($\beta = 0.28, p < 0.05$). This may due the to data size of this experiment which contained more search sessions than Experiment 1. The results shown in Table 5.11 demonstrate that answers encouraged participants to issue fewer queries (not encounter: $M = 2.26$, encounter: $M = 1.41$), submit more evidence (not encounter: $M = 2.00$, encounter:

Table 5.9: Mean (M) and standard deviation (SD) for each experimental condition in Experiment 2 (C_1: control, C_2: rank 1, and C_3: rank 4) and statistical testing results of each search behavior of the model with different fixed effects (***: significance level at 0.001, **: 0.01, and *: 0.05). Coefficient (β), standard error (SE), t-statistic (t), and p-value (p) of each fixed effect are reported. For the variable alternative question, β_a represents the effect of the alternative question while β_r represents the effect of its rank. The χ^2 and p-values of the model which is statistically significant compared to the null model are also reported.

	Condition						**Fixed Effects**															
	C_1		C_2		C_3		**Alternative question**								**Answer**				**Alternative answer**					
	M	SD	M	SD	M	SD	β_a	SE	t	p	β_r	SE	t	p	β	SE	t	p	β	SE	t	p	χ^2	p
# of Queries	2.37	2.12	1.94	2.09	2.10	2.13	-0.10	0.03	-3.05	**	-0.04	0.05	-0.74	0.46	-0.85	0.20	-4.27	***	-0.33	0.16	-2.04	*	15.9	**
# of Clicks	4.35	3.21	4.63	3.54	4.79	3.81	0.09	0.05	1.84	0.07	-0.11	0.09	-1.19	0.23	0.21	0.33	0.66	0.51	0.52	0.27	1.97	0.05		
Deepest Document Rank	9.69	10.47	10.57	10.34	10.36	10.30	0.26	0.18	1.41	0.16	0.12	0.32	0.39	0.70	-0.12	1.11	-0.11	0.92	-0.11	0.93	-0.12	0.90		
Page Dwell Time (sec)	56.72	69.98	50.93	51.78	53.27	55.48	-0.02	0.01	-1.56	0.12	-0.02	0.02	-0.82	0.41	-0.01	0.09	-0.14	0.89	-0.06	0.07	-0.88	0.38		
SERP Dwell Time (sec)	24.09	34.68	23.66	28.46	21.35	21.17	0.01	0.01	0.80	0.42	0.01	0.02	0.57	0.57	0.58	0.09	6.55	***	0.01	0.07	0.13	0.90	21.30	***
Task Time Spent (sec)	433.36	396.24	395.19	364.75	415.41	360.95	-0.02	0.01	-2.07	0.04	-0.02	0.02	-1.43	0.15	0.13	0.07	2.02	0.04	0.00	0.05	0.11	0.92		
# of Evidence	1.87	1.07	2.23	1.56	2.07	1.43	0.09	0.02	4.08	***	0.08	0.04	1.96	0.05	0.28	0.14	2.04	*	0.08	0.11	0.71	0.48	11.2	*

Table 5.10: Participants' behavior according to whether an alternative question was presented. (Experiment 2)

	Absent		**Present**	
	M	*SD*	*M*	*SD*
# of Queries	2.37	2.12	2.02	2.11
# of Evidence	1.87	1.07	2.15	1.50

$M = 2.36$) and spend more time on the SERP (not encounter: $M = 19.84$, encounter: $M = 42.00$).

Effect of alternative answer: We found a main effect of *alternative answer* on the number of queries ($\beta = -0.33, p < 0.05$). The results in Table 5.12 show that participants who encountered an alternative answer once were less likely to issue queries (not encounter: $M = 1.73$, encounter: $M = 0.53$).

Beliefs: We constructed a logistic mixed-effects model in which beliefs d was considered a binary dependent variable (1: alter, 0:retain). Similarly to Section 5.4.2, the model was not statistically significant compared to the null model ($\chi^2 = 8.11$, $p = 0.09$). Therefore, we cannot accept hypothesis **H2-2**. However, there is an interesting result to discussed. Although Tables 5.13 (a) and (b) indicate that most participants altered their beliefs toward *yes* polar after the search task, the participants were less likely to retain their beliefs in the same polar when the alternative question was presented at the fourth position in *People also ask* (④) compared to the first position (③). One possible explanation for this result is that highly salient verticals are more likely to be noticed at lower ranks [Arguello and Capra 2014]. The alternative question could be considered highly salient item because it satisfied the information need with the information that obviously differs from the other three questions. In addition, the result appears to support the suggestion that diverse information can mitigate the effect of confirmation bias [Sunstein 2017].

Summary: We found that the *alternative question* encouraged participants to issue fewer queries and submit more evidence. Similarly to the Section 5.4.2, we found that *answer* encourage participants to issue fewer queries, spend more time on SERP, and submit more evidence. In addition, we found that participants were less likely to issue queries when they encountered an *alternative answer*. Thus, the results partially support hypothesis **H2-1**. However, we cannot accept hypothesis **H2-2** because the model was not statistically significant compared to the null model.

Table 5.11: Participants' behaviors according to whether they encountered answers. (Experiment 2)

	Not Encounter		**Encounter**	
	M	*SD*	*M*	*SD*
# of Queries	2.26	2.13	1.41	1.96
SERP Dwell Time (sec)	19.84	24.21	42.00	43.00
# of Evidence	2.00	1.25	2.36	1.93

5.5 Discussion

The analysis revealed that *question* discouraged participants from issuing queries and spending time on the SERP, while the *answer* and its *opinion* encouraged participants to spend more time on SERP, click on more documents, check deeper rank documents, and spend more time on a task. We also found that the alternative question encouraged participants to submit more evidence and discouraged them from issuing queries. This section discusses the implications of the results and limitations.

5.5.1 Implications

Regarding **RQ1**, we found that the *question* discouraged participants from issuing queries and spending time on the SERP. This is consistent with previous studies, which suggested that people tend to spend less effort on a search task when presented by a question answering system [Belkin et al. 2000, Bernstein et al. 2012, Chilton and Teevan 2011, Lin et al. 2003]. We also found that the *answer* and *belief-consistent* answer encouraged participants to spend more time on SERP, click on more documents, check deeper rank documents, and spend more time on the task. However, the answer discouraged the participants from issuing queries. One possible explanation for these results is that participants who clicked on questions to check the answers performed the task seriously and behaved carefully even prior to participating to the experiment.

Although the previous study suggested that presenting information that is inconsistent with people's beliefs in organic search results is beneficial because it encourages people to perform a careful information search [Pothirattanachaikul et al. 2019], we did not observe such an effect in this study as belief-inconsistent answers discouraged participants from clicking on documents, checking the deeper rank documents, and spending time on the search tasks. This phenomenon may be explained in reference to a previous study, which suggested that questioners are likely to prefer confirmation responses over disconfirmations [Heritage 2013]. Therefore, participants may considered a belief-inconsistent answer as irrelevant information, which results in perceiving the SERP as being low quality. This implies that presenting belief-

Table 5.12: Participants' behaviors according to whether they encountered an alternative answer. (Experiment 2)

	Not Encounter		**Encounter**	
	M	*SD*	*M*	*SD*
# of Queries	1.73	2.17	0.53	0.68

inconsistent information as an answer in *People also ask* may not encourage people to conduct careful information search in contrast to presenting it within organic search results.

Regarding **RQ2**, although the effect of *People also ask* was consistent with findings from previous studies [Lin et al. 2003, Voorhees and Tice 1999] in which people spent less effort on the SERP when using a question answering system, participants appeared to become more careful with the search process as they submit more evidence when an alternative question was included in *People also ask*. However, further study may be required to investigate the polarity of the submitted evidence because participants may submit evidence that confirms their beliefs rather than evidence that is either inconsistent with their beliefs or related to the given alternative.

For both **RQ1** and **RQ2**, we could not confirm the effect of *question*, *answer*, *opinion*, and *alternative question* on beliefs. This may be due to the fact that the number of participants who clicked on *People also ask* was small. Therefore, participants' beliefs were not much affected by *People also ask* compared to organic search results. Although the results related to the effect of *opinion* and the rank of *alternative question* on beliefs appears to be promising, further experiment is required to validate the effect of these factors.

5.5.2 Limitations

This study has several limitations that should be acknowledged.

Number of participants who interact with People also ask: First, it might be questioned whether the findings are applicable to the general population of searchers because only a small number of participants clicked on *People also ask* ($\sim$20%). In addition, it might be questioned that search behaviors were actually affected by the organic search results or participants themselves behaved carefully prior to the task (especially the ones who clicked on the question to see the answer). The analysis presented in Section 5.4 cannot fully separate these effects and may thus not be generalizable to the general population of searchers.

Generalizability: Relative to search tasks, the results of this study cannot be generalized to other domains of search tasks. The question remains open whether the findings are applicable to other domains, e.g., the political domain, which is related to one's social identity. People may have a negative impression of an alternative political candidate from a party whose

Table 5.13: Percentage of beliefs for Experiment 2 grouped by whether the alternative question was placed at rank 1 or rank 4.

(a) Alternative question placed at rank 1 (n=257)

		Posterior Belief			
		No (n=7)	Lean-No (n=31)	Lean-Yes (n=139)	Yes (n=80)
Prior Belief	No (n=23)	13.04%	21.74%	30.43%	34.78%
	Lean No (n=58)	3.45%	17.24%	48.28%	31.03%
	Lean Yes (n=146)	0.68%	10.96%	63.01%	25.34%
	Yes (n=30)	3.33%	0.00%	40.00%	56.67%

(b) Alternative question placed at rank 4 (n=257)

		Posterior Belief			
		No (n=11)	Lean No (n=34)	Lean Yes (n=138)	Yes (n=74)
Prior Belief	No (n=25)	16.00%	12.00%	48.00%	24.00%
	Lean No (n=78)	3.85%	15.38%	62.82%	17.95%
	Lean Yes (n=134)	2.99%	12.69%	54.48%	29.85%
	Yes (n=20)	0.00%	10.00%	20.00%	70.00%

ideology opposes their beliefs. Nevertheless, search tasks tackled in this study (*i.e.*, health-related search tasks that can be answered with *yes* or *no*) are also important search tasks in contemporary society.

5.6 Summary

This study investigated the effects *People also ask* has on people's search behaviors and beliefs. The results demonstrate that the question discouraged participants from issuing queries and spending time on the SERP while the answer encouraged participants to issue fewer queries, spend more time on the SERP, and click more documents. However, we observed that participants who first encountered a belief-inconsistent answer were less likely to click on documents, check deeper rank documents, and spend time on the search task. In addition, we found that an alternative question encouraged participants to issue fewer queries and submit more evidence. In this study, we could not confirm an effect of *People also ask* on beliefs. These results imply that presenting *People also ask* is convenient for searchers because they tend to exert less effort on the search. Nonetheless, the presence of *People also ask* may not be able to mitigate confirmation bias. In future, we plan to conduct an experiment to validate the effect of *People also ask* on beliefs.

6 Conclusions

6.1 Summary

Search engines become a main information resource where searchers promptly obtain the information they need in a matter of seconds. To accurately obtain information, previous studies suggested that searchers should expose themselves to the information that are inconsistent with their beliefs and search carefully. Unfortunately, searchers are likely to exhibit confirmation bias; where they seek information to confirms their beliefs via Web search engines. Consequently, they refrain from diverse and careful searches for information and do not consider the information that are not consistent with their beliefs even such information could also satisfy their information need. Therefore, the goal of this thesis is to mitigate the confirmation bias in Web searches. Three research topics that we addressed in this thesis are summarized as follows:

- **Mining Alternative Information:** We define and tackle the *alternative action mining* problem. In particular, we attempt to develop a method for mining alternative actions for a given query. We define alternative actions as actions which share the same goal and define the alternative action mining problem as similar in the search result diversification. To tackle the problem, we propose leveraging a community Q&A (cQA) corpus for mining alternative actions. We propose a method to compute how well two actions can be alternative actions by using a question-answer structure in a cQA corpus. Our method builds a question-action bipartite graph and recursively computes how well two actions can be alternative actions. We conducted experiments to investigate the effectiveness of our method using two newly built test collections, each containing 50 queries. The experimental results indicated that our proposed method outperformed the query suggestion methods provided by the commercial search engines in terms of D#-nDCG.
- **Understanding how and when people search for the alternative information:** To obtain accurate information through web searches, people have to search for information carefully. This study investigates how the search behaviors and decision outcomes of searchers were affected by the documents they encountered during their search process. We focus on two document factors: (1) *opinion* (consistent and inconsistent) with the searchers' beliefs prior to the search task, and (2) *credibility* (high and low). We conducted a user study in which 260 participants were asked to perform health-related search tasks while controlling a search result with different opinions and credibility levels. The

results revealed that (i) the participants spent more effort searching by issuing more queries, when belief-inconsistent documents were presented; (ii) the documents' opinion and credibility affected their *belief dynamics*, (*i.e.,* how their beliefs changed after the search task); and (iii) their belief dynamics and search efforts had few relationships. These findings suggest that search engines could prevent users from polarization and thus, help them to obtain accurate information, by presenting documents that are inconsistent with users' beliefs on the higher-rank of the results.

- **Encourage the searchers to interact with the alternative information:** This study investigates the impact of collective questions and answers displayed on Search engine result pages (SERP), known as *People also ask* on searchers' behaviors and beliefs as suggested by Trope and Bassok [Trope and Bassok 1982, 1983] that people are likely to consider questions related to the alternative hypothesis when such questions are provided. Two experiments were conducted in which participants were asked to perform health-related search tasks. In both experiments, items in *People also ask* were manipulated. Experiment 1 focused on the effect of *question*, *answer*, and answer's *opinion*. Experiment 2 focused on the effect of the *alternative question*, i.e., a question related to the solution that can achieve the same goal as a query. The results revealed the following. (i) Participants issued fewer queries and spent less time on a SERP when *People also ask* were presented. (ii) Participants were less likely to interact with a SERP when they first encounter a belief-inconsistent answer. (iii) We could not confirm the effect of *People also ask* on beliefs at the current state. The findings suggest that *People also ask* might not help mitigate confirmation bias as participants are likely to spend less effort on the search process (i.e., issue fewer queries) when they first encounter a belief-inconsistent answer unlike when they encounter a belief-inconsistent document within the search results. An additional experiment is required to validate that participants who first encounter a belief-inconsistent answer are more likely to alter their beliefs as the number of such participants was inadequate.

Finally, the technical contributions of this thesis can be summarized as follows:

- We proposed the technique to extract alternative-actions which can achieve the same goal as an input query via the usage of the machine learning technique known as Conditional Random Field (CRF) and the novel ranking algorithm that was proposed based on the structure of community Q&A corpus.
- While previous study only focus on the effect of the belief-inconsistent information on the information selection, we uncovered the effect of *opinion* and *credibility* on the searchers' confirmation bias and search behaviors.

- We investigated the effect of *question-answering* system on search behaviors and confirmation bias. We found that such a system might not be appropriate to mitigate the confirmation bias as searchers tend to finished their search task with less efforts.

6.2 Future Directions

Several candidate topics need to be explored in future work. First, we should improve the step for extracting candidate actions so that we can obtain more relevant alternative actions. In this thesis, we relied on the machine learning technique called *conditional random field* (CRF) to extract the candidate actions from the community Q&A corpus. Unfortunately, CRF also extracted irrelevant actions (i.e., *think that*, *talk to you*) altogether with other candidate actions. In our opinion, it highly affected the quality of alternative-actions. Thus, this problem might be solved by the better natural language processing technique (i.e., BERT). Second, it is worth implementing a new re-ranking algorithm that encourages searchers to conduct careful information search. Nowadays, search engines tend to develop relevance-oriented functionalities to make the searchers satisfy their needs with less effort. With such functionality, searchers' confirmation bias tends to be exacerbated as they can satisfy their information need in an early manner. It would be better if search engines can make the searchers carefully conduct the search through the ranking system that diversifies the documents with different opinions. Finally, we are also interested in the comparison between *question suggestion* and *query suggestion*. As searchers might ignore the questions when they are contained in the vertical (in this study; *People also ask*), searchers might be more affected by the questions if they are directly suggested to the searchers. An eye-tracking study might be a candidate to observe the effects of *question suggestion* compare to the *query suggestion*.

Bibliography

L. M. Aiello, D. Donato, U. Ozertem, and F. Menczer. 2011. Behavior-driven clustering of queries into topics. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 1373–1382. ACM.

J. Arguello and R. Capra. 2014. The effects of vertical rank and border on aggregated search coherence and search behavior. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM'14, pp. 539–548.

D. J. Barr, R. Levy, C. Scheepers, and H. J.Tily. 2011. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3): 255–278.

D. Bates, M. Mächler, B. Bolker, and S. Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1): 1–48.

N. Belkin, A. Keller, D. Kelly, J. Perez-Carballo, C. Sikora, and Y. Sun. 2000. Support for question-answering in interactive information retrieval: Rutgers' trec-9 interactive track experience. In *The Text REtrieval Conference (TREC-9)*, pp. 352–821.

M. S. Bernstein, J. Teevan, S. Dumais, D. Liebling, and E. Horvitz. 2012. Direct answers for search queries in the long tail. In *Proceedings of the SIGCHI conference on human factors in computing systems*, CHI'12, pp. 237–246.

P. Boldi, F. Bonchi, C. Castillo, and S. Vigna. 2009. From "dango" to "japanese cakes": Query reformulation models and patterns. In *Proceedings of the IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies*, pp. 183–190.

C. Buchs, F. Butera, G. Mugny, and C. Darnon. 2004. Conflict elaboration and cognitive outcomes. *Theory Into Practice*, 43(1): 23–30.

J. Carbonell and J. Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proc. of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336.

N. P. Chalasani, P. H. Hayashi, H. L. Bonkovsky, V. J. Navarro, W. M. Lee, and R. J. Fontana. 2014. Acg clinical guideline: the diagnosis and management of idiosyncratic drug-induced liver injury. *The American Journal of Gastroenterology*, 109: 950–966.

L. B. Chilton and J. Teevan. 2011. Addressing people's information needs directly in a web search result page. In *Proceedings of the 20th international conference on World wide web*, WWW'11, pp. 27–36.

A. Cipriani, T. Furukawa, and C. Barbui. 2011. What is a cochrane review? *Epidemiology and psychiatric sciences*, 20(3): 231–233.

B. Dardenne and J.-P. Leyens. 1995. Confirmation bias as a social skill. *Personality and Social Psychology Bulletin*, 21(11): 1229–1239.

D. Donato, F. Bonchi, T. Chi, and Y. Maarek. 2010. Do you want to take notes?: identifying research missions in Yahoo! search pad. In *Proc. of the 19th International Conference on World Wide Web*,

pp. 321–330. ACM.

K. Edwards and E. E. Smith. 1996. A disconfirmation bias in the evaluation of arguments. *Journal of Personality and Social Psychology*, 71(1): 5.

R. Ennals, B. Trushkowsky, and J. M. Agosta. 2010. Highlighting disputed claims on the web. In *Proc. of the 19th International conference on World Wide Web*, pp. 341–350.

R. H. Ennis. 1987. *A taxonomy of critical thinking dispositions and abilities*. W H Freeman/Times Books/Henry Holt & Co.

L. Festinger. 1957. *A theory of cognitive dissonance*. Stanford University Press.

L. Festinger. 1964. *Conflict, decision, and dissonance*. Stanford University Press.

A. Field, J. Miles, and Z. Field. 2012. *Discovering statistics using R*. SAGE Learning.

J. L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5): 378.

J. L. Fleiss and J. Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3): 613–619.

B. J. Fogg. 2003. *Persuasive technology using computers to change what we think and do*. Morgan Kaufmann Publishers.

B. J. Fogg and H. Tseng. 1999. The elements of computer credibility. In *Proceedings of the 17th SIGCHI Conference on Human Factors in Computing Systems*, CHI'99, pp. 80–87.

K. R. Garrett and P. Resnick. 2011. Resisting political fragmentation on the internet. *Daedalus, Journal of the American Academy of Arts & Sciences*, 140: 108–120.

Z. Guan and E. Cutrell. 2007. An eye tracking study of the effect of target rank on web search. In *Proceedings of the 25th SIGCHI Conference on Human Factors in Computing Systems*, CHI'07, pp. 417–420.

F. M. Harper, D. Moy, and J. A. Konstan. 2009. Facts or friends?: distinguishing informational and conversational questions in social q&a sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 759–768.

A. Hassan and R. W. White. 2012. Task tours: helping users tackle complex search tasks. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pp. 1885–1889.

A. Hassan Awadallah, R. W. White, P. Pantel, S. T. Dumais, and Y.-M. Wang. 2014. Supporting complex search tasks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pp. 829–838.

J. Heritage. 2013. *Garfinkel and ethnomethodology*. John Wiley & Sons.

I. Hernandez and J. L. Preston. 2012. Disfluency disrupts the confirmation bias. *Journal of Experimental Social Psychology*, 49: 178–182.

I. L. Janis. 1982. *Groupthink*. Cengage Learning.

G. Jeh and J. Widom. 2002. SimRank: a measure of structural-context similarity. In *Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 538–543.

N. Jindal and B. Liu. 2006. Identifying comparative sentences in text documents. In *Proc. of the 29th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 244–251.

T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. 2007. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transaction on Information Systems (TOIS)*, 25(2).

E. Jonas, S. Schulz-Hardt, D. Frey, and N. Thelen. 2001. Confirmation bias in sequential information search after preliminary decisions: an expansion of dissonance theoretical research on selective exposure to information. *Journal of Personality and Social Psychology*, 80: 557–571.

R. Jones and K. L. Klinkner. 2008. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pp. 699–708.

T. M. Jurgens, A. M. Whelan, L. Killian, S. Doucette, S. Kirk, and E. Foy. 2012. Green tea for weight loss and weight maintenance in overweight or obese adults. *Cochrane Database of Systematic Reviews*, 12.

D. Kahneman. 2003. A perspective on judgment and choice: mapping bounded rationality. *American Psychologist*, 58(9): 697.

Y. Kammerer, I. Bråten, P. Gerjets, and H. I. Strømsø. 2013. The role of internet-specific epistemic beliefs in laypersons' source evaluations and decisions during web search on a medical issue. *Computers in Human Behavior*, 29(3): 1193–1203.

J. T. Kaplan, S. I. Gimbel, and S. Harris. 2016. Neural correlates of maintaining one's political beliefs in the face of counterevidence. *Scientific reports*, 6: 39589.

K. Karvonen. 2000. The beauty of simplicity. In *Proceedings on the 1st conference on Universal Usability*, CUU'00, pp. 85–90.

K. Karvonen, L. Cardholm, and S. Karlsson. 2000. Cultures of trust: a cross-cultural study on the formation of trust in an electronic environment. In *Proceedings of the 3rd Nordic Workshop on Security (NordSec'00)*, pp. 89–100.

D. Kelly. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval*, 3(1–2): 1–224.

S. Li, C.-Y. Lin, Y.-I. Song, and Z. Li. 2010. Comparable entity mining from comparative questions. In *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 650–658.

V. Q. Liao and W.-T. Fu. 2014. Can you hear me now?: mitigating the echo chamber effect by source position indicators. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW'14, pp. 184–196.

V. Q. Liao, W.-T. Fu, and S. S. Mamidi. 2015. It is all about perspective: an exploration of mitigating selective exposure with aspect indicators. In *Proceedings of the 33rd SIGCHI Conference on Human Factors in Computing Systems*, CHI'15, pp. 1439–1448.

J. Lin, D. Quan, V. Sinha, K. Bakshi, D. Huynh, B. Katz, and D. R. Karger. 2003. What makes a good answer? the role of context in question answering. In *Proceedings of the Ninth IFIP TC13 International Conference on Human-Computer Interaction*, pp. 25–32.

G. Lindgaard, C. Dudek, D. Sen, L. Sumegi, and P. Noonan. 2011. An exploration of relations between visual appeal, trustworthiness and perceived usability of homepages. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 18(1): 1:1–1:30.

Y. Liu, R. Song, M. Zhang, Z. Dou, T. Yamamoto, M. P. Kato, H. Ohshima, and K. Zhou. 2014. Overview of the ntcir-11 imine task. In *Proceeding of the 11th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*.

Y. Matias, D. Keysar, G. Chechik, Z. Bar-Yossef, and T. Shmiel, June 13 2017. Generating related questions for search queries. US Patent 9,679,027.

M. J. Metzger. 2007. Making sense of credibility on the web: models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology (JASIST)*, 58(13): 2078–2091.

M. J. Metzger, A. J. Flanagin, and R. B. Medders. 2010. Social and heuristic approaches to credibility evaluation online. *Journal of Communication*, 60(3): 413–439.

M. J. Metzger, A. J. Flanagin, A. Markov, R. Grossman, and M. Bulger. 2015. Believing the unbelievable: understanding young people's information literacy beliefs and practices in the united states. *Journal of Children and Media*, 9(3): 325–348.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pp. 3111–3119.

M. Moshagen, J. Musch, and A. S. Göritz. 2009. A blessing, not a curse: experimental evidence for beneficial effects of visual aesthetics on performance. *Ergonomics*, 52 10: 1311–20.

S. Nakamura, S. Konishi, A. Jatowt, H. Ohshima, H. Kondo, T. Tezuka, S. Oyama, and K. Tanaka. 2007. Trustworthiness analysis of web search results. In *Proc. of the 11th European Conference on Research and Advanced Technology for Digital Libraries*, pp. 38–49.

C. Nemeth and J. Rogers. 1996. Dissent and the search for information. *British Journal of Social Psychology*, 35: 67–76.

C. Nemeth, K. Mosier, and C. Chiles. 1992. When convergent thought improves performance: majority versus minority influence. *Personality and Social Psychology Bulletin*, 18(2): 139–144.

R. S. Nickerson. 1998. Confirmation bias: a ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2): 175–220.

M. C. S. Nogueira. 2004. Orthogonal contrasts: definitions and concepts. *Scientia Agricola*, 61(1): 118–124.

N. Okazaki, 2007. Crfsuite: a fast implementation of conditional random fields (crfs). http://www.chokkan.org/software/crfsuite/.

A. Omari, D. Carmel, O. Rokhlenko, and I. 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*, pp. 215–224.

F. A. Pogacar, A. Ghenai, M. D. Smucker, and C. L. Clarke. 2017. The positive and negative influence of search results on people's decisions about the efficacy of medical treatments. In *Proceedings of the 3rd ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR'17, pp. 209–216.

S. Pothirattanachaikul, T. Yamamoto, Y. Yamamoto, and M. Yoshikawa. 2019. Analyzing the effects of document's opinion and credibility on search behaviors and belief dynamics. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM'19, pp. 1653–1662.

P. Resnick, K. R. Garrett, T. Kriplean, S. A. Munson, and N. J. Stroud. 2013. Bursting your (filter) bubble: strategies for promoting diverse exposure. In *Proceedings of the 16th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW'13, pp. 95–100.

D. Robins, J. Holmes, and M. Stansbury. 2009. Consumer health information on the web: the relationship of visual design and perceptions of credibility. *Journal of the American Society for Information Science and Technology (JASIST)*, 61(1): 13–29.

T. Sakai. 2014. Metrics, statistics, tests. In *PROMISE Winter School 2013: Bridging between Information Retrieval and Databases*.

T. Sakai and R. Song. 2011. Evaluating diversified search results using per-intent graded relevance. In *Proc. of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1043–1052.

T. Sakai, Z. Dou, T. Yamamoto, Y. Liu, M. Zhang, R. Song, M. Kato, and M. Iwata. 2013. Overview of the NTCIR-10 INTENT-2 task. In *Proc. of the 10th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*.

C. Schwind and J. Buder. 2012. Reducing confirmation bias and evaluation bias: when are preference-inconsistent recommendations effective - and when not? *Computers in Human Behavior*, 28(6): 2280 – 2290.

C. Schwind, J. Buder, U. Cress, and F. W. Hesse. 2012. Preference-inconsistent recommendations: an effective approach for reducing confirmation bias and stimulating divergent thinking? *Computers & Education*, 58: 787–796.

W. M. Silberg, G. D. Lundberg, and R. A. Musacchio. 1997. Assessing, controlling, and assuring the quality of medical information on the internet: caveant lector et viewor - let the reader and viewer beware. *Journal of the American Medical Association*, 277(15): 1244–1245.

E. Sillence, P. Briggs, P. Harris, and L. Fishwick. 2006. A framework for understanding trust factors in web-based health advice. *International Journal of Human-Computer Studies*, 64(8): 697–713.

R. B. Skov and S. J. Sherman. 1986. Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. *Journal of Experimental Social Psychology*, 22(2): 93–121.

M. Snyder and W. B. Swann. 1978. Hypothesis-testing processes in social interaction. *Journal of Personality and Social Psychology*, 36(11): 1202.

C. R. Sunstein. 2017. *#Republic divided democracy in the age of social media*. Princeton Press.

W. B. Swann, T. Giuliano, and D. M. Wegner. 1982. Where leading questions can lead: The power of conjecture in social interaction. *Journal of Personality and Social Psychology*, 42(6): 1025.

Y. Trope and M. Bassok. 1982. Confirmatory and diagnosing strategies in social information gathering. *Journal of personality and social psychology*, 43(1): 22.

Y. Trope and M. Bassok. 1983. Information-gathering strategies in hypothesis-testing. *Journal of Experimental Social Psychology*, 19(6): 560–576.

K. Tsukuda, H. Ohshima, and K. Tanaka. 2014. Ranking of coordinate terms and hypernyms using a hypernym-hyponym dictionary. In *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies*, pp. 15–21.

E. M. Voorhees and D. M. Tice. 1999. The trec-8 question answering track evaluation. In *TREC*, volume 1999, p. 82. Citeseer.

T.-X. Wang, K.-Y. Tsai, and W.-H. Lu. 2014. Identifying real-life complex task names with task-intrinsic entities from microblogs. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 470–475.

I. Weber, A. Ukkonen, and A. Gionis. 2012. Answers, not links: extracting tips from Yahoo! answers to address how-to web queries. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, pp. 613–622.

R. W. White. 2013. Beliefs and biases in web search. In *Proceedings of the 36th SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'13, pp. 3–12.

R. W. White. 2014. Belief dynamics in web search. *Journal of the Association for Information Science and Technology*, 65(11): 2165–2178.

R. W. White and A. Hassan. 2014. Content bias in online health search. *ACM Transactions on the Web (TWEB)*, 8(4): 25:1–25:33.

R. W. White and E. Horvitz. 2015. Belief dynamics and biases in web search. *ACM Transactions on Information Systems (TOIS)*, 33(4): 18:1–18:46.

B. Winter. 2013. Linear models and linear mixed effects models in r with linguistic applications. *arXiv preprint arXiv:1308.5499*.

T. Yamamoto, S. Nakamura, and K. Tanaka. 2011. Extracting adjective facets from community q&a corpus. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pp. 2021–2024.

T. Yamamoto, T. Sakai, M. Iwata, C. Yu, J.-R. Wen, and K. Tanaka. 2012. The wisdom of advertisers: mining subgoals via query clustering. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pp. 505–514.

T. Yamamoto, Y. Liu, M. Zhang, Z. Dou, K. Zhou, I. Markov, M. P. Kato, H. Ohshima, and S. Fujita. 2016. Overview of the NTCIR-12 IMine-2 task. In *Proc. of the 12th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, pp. 94–123.

T. Yamamoto, Y. Yamamoto, and S. Fujita. 2018a. Exploring people's attitudes and behaviors toward careful information seeking in web search. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM'18)*, pp. 963–972.

Y. Yamamoto and S. Shimada. 2016. Can disputed topic suggestion enhance user consideration of information credibility in web search? In *Proc. of the 27th ACM Conference on Hypertext and Social Media*, pp. 169–177.

Y. Yamamoto and K. Tanaka. 2011. Enhancing credibility judgment of web search results. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1235–1244.

Y. Yamamoto and T. Yamamoto. 2018. Query priming for promoting critical thinking in web search. In *Proceedings of the 3rd ACM SIGIR Conference on Human Information Interaction & Retrieval*, CHIIR'18, pp. 12–21.

Y. Yamamoto, T. Yamamoto, H. Ohshima, and K. Hiroshi. 2018b. Web access literacy scale to evaluate how critically users can browse and search for web information. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci'18, pp. 97–106.

Z. Yang and E. Nyberg. 2015. Leveraging procedural knowledge for task-oriented search. In *Proc. of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 513–522.

E. Yilmaz, E. Kanoulas, M. Verma, B. Carterette, N. Craswell, and R. Mehrotra. Overview of the TREC 2015 Tasks track. In *Proceedings of the 24th Text REtrieval Conference*.

P. Zhang and N. Li. 2004. Love at first sight or sustained effect? the role of perceived affective quality on users' cognitive reactions to information technology. In *Proceedings of the 25th International Conference on Information Systems (ICIS'04)*, pp. 283–296.

Publications

Academic Journal

1. Suppanut Pothirattanachaikul, Takehiro Yamamoto, Sumio Fujita, Akira Tajima, Katsumi Tanaka, and Masatoshi Yoshikawa, Mining Alternative Action from Community Q&A Corpus, In *Journal of Information Processing*, Vol. 26, pp 427-438, 2018.

International Conference

1. Suppanut Pothirattanachaikul, Takehiro Yamamoto, Sumio Fujita, Akira Tajima, and Katsumi Tanaka, Mining Alternative Action from Community Q&A Corpus for Task-Oriented Web Search, In *Proceedings of the 2017 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2017)*, pp 607-614, 2017.
2. Suppanut Pothirattanachaikul, Takehiro Yamamoto, Yusuke Yamamoto, and Masatoshi Yoshikawa, Analyzing the Effects of Document's Opinion and Credibility on Search Behaviors and Belief Dynamics, In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM 2019)*, pp 1653–1662, 2019.
3. Suppanut Pothirattanachaikul, Takehiro Yamamoto, Yusuke Yamamoto, and Masatoshi Yoshikawa, Analyzing the Effects of "People also ask" on Search Behaviors and Beliefs, In *Proceedings of the 31st ACM International Conference on Hypertext and Social Media (HT 2020)*, pp 101–110, 2020.

Domestic Symposium

1. Suppanut Pothirattanachaikul, Takehiro Yamamoto, Sumio Fujita, Akira Tajima, Katsumi Tanaka, Finding and Suggesting Alternative Verbal Queries from Community Q&A Corpus, 第9回データ工学と情報マネジメントに関するフォーラム（DEIM 2017), 高山市, 2017年3月.
2. Suppanut Pothirattanachaikul, Takehiro Yamamoto, Yusuke Yamamoto, and Masatoshi Yoshikawa, 文書の意見と信憑性がユーザの検索行動および信念の変化に与える影響の分析, 第11回データ工学と情報マネジメントに関するフォーラム（DEIM 2019), 長崎市 , 2019年3月.

Workshop

1. Suppanut Pothirattanachaikul, Finding Alternative Goals from Q&A Corpus, 関西データベースワークショップ2016, 神戸市, 2016年9月.

2. Suppanut Pothirattanachaikul, Mining Alternative Action from Community Q&A Corpus for Task-Oriented Web Search, iDB特別セッション, Webとデータベースに関するフォーラム (WebDB Forum 2017), 東京都文京区, 2017年9月.

3. Suppanut Pothirattanachaikul, When do People Search for Information Carefully? The Effect of Inconsistency and Credibility, SIGIR Tokyo-Beijing Joint Workshop 2019, 北京, 2019年 1月.