

Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

Qiaoling Liu

Date

Improving Usefulness of Community Question Answering Services
Towards Better Searcher Satisfaction and Question Recommendation

By

Qiaoling Liu
Doctor of Philosophy

Computer Science and Informatics

Eugene Agichtein
Advisor

James Lu
Advisor

Jacob Eisenstein
Committee Member

Li Xiong
Committee Member

Accepted:

Dean of the Graduate School

Date

Improving Usefulness of Community Question Answering Services
Towards Better Searcher Satisfaction and Question Recommendation

By

Qiaoling Liu
M.S., Emory University, 2011

Advisor : Eugene Agichtein

Abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2014

Abstract

Improving Usefulness of Community Question Answering Services Towards Better Searcher Satisfaction and Question Recommendation

By Qiaoling Liu

Community-based Question Answering (CQA) sites, such as Yahoo! Answers and Quora, provide a promising way of finding and sharing information online. This thesis aims to improve the usefulness of CQA services towards better searcher satisfaction and question recommendation, by focusing on three important problems ignored in previous work:

(1) How to improve web searcher satisfaction using CQA services. This thesis proposes methods for a novel task of predicting web searcher satisfaction with existing answers in CQA, enabling better ranking of CQA pages for searchers. When searchers fail in web search, they may alternatively ask questions using CQA services. This thesis analyzes users' transition from searching to asking in terms of query and behavior characteristics, providing insights for predicting the transition.

(2) What contextual factors influence answerer behavior in CQA. This thesis analyzes the answerer behavior in a large scale CQA system, and explores when users tend to answer questions and how they tend to choose the questions to answer. Based on a user study, this thesis further explores how relevant web browsing context affects answerers' perceived ability, effort, and willingness to answer a question. The findings could inform the design of more intelligent question recommendation strategies in CQA systems.

(3) How to deploy question recommendation in real-time CQA systems. This thesis develops a scalable real-time CQA system with a mobile interface, which supports different question recommendation strategies. Based on two live user studies, this thesis further conducts both quantitative analysis of user behavior as well as qualitative analysis of user satisfaction with the system. The developed system and the reported analysis offer insights for designing real-time CQA systems and deploying question recommendation.

In summary, the work on predicting searcher satisfaction and understanding the transition from searching to asking would help improve searcher satisfaction using CQA systems, and the work on understanding answerer behavior and building a real-time CQA system would help improve question recommendation in CQA systems, making CQA services more useful.

Improving Usefulness of Community Question Answering Services
Towards Better Searcher Satisfaction and Question Recommendation

by

Qiaoling Liu
M.S., Emory University, 2011

Advisor : Eugene Agichtein

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2014

Acknowledgements

Firstly, I would like to thank my advisor Eugene Agichtein, who has been very supportive and helpful throughout my PhD study. For each of my projects, he provided insightful ideas and advices, from big picture to implementation details and from paper writing to presentation preparation. Meanwhile, I was given sufficient freedom and time to choose projects and methods. He always respected my interests and encouraged me to do my best. In particular, when I had a bad time after having a baby, he proactively contacted Professor Li Xiong (an experienced Mather) to chat with me and share feelings. Besides working in school, he also encouraged me to attend conferences and do internships. I have learned a lot from Eugene and from my PhD study. I am very grateful and proud to have him as my advisor.

I would also like to thank my other committee members, Professor James Lu, Professor Li Xiong, and Professor Jacob Eisenstein, for trying their best to make my defense happen given the super tight schedules at the end of the semester. Especially thank Jacob for driving to Emory for my defense from Georgia Tech right after his meeting, thank Li for the sweet chat during my life's breakdowns, and thank James for conscientiously chairing my defense. I also appreciate all their useful comments and suggestions about my thesis, which help me to make it better.

Another important part of my PhD study is working as a part-time student researcher with Yahoo!, which provides the opportunity for me to meet Yoelle Maarek, Dan Pelleg, Idan Szpektor, Gideon Dror, Evgeniy Gabrilovich, and Avihai Mejer. I want to thank them, especially Dan and Idan, for their precious time to have regular meetings and discussions with me, for their help with data extraction, experimentation and paper writing, as well as for the support, freedom, and understanding they provide. Thank Yahoo! Faculty Research Engagement Program for giving me this opportunity.

My last project was also collaborated with Tomasz Jurczyk and Professor Jinho Choi, who played important roles in making it accomplished. Tomasz worked very hard during the semesters and summer on the interface implementation and study design given his own heavy workload. Jinho provided a lot of helpful advices on the study design, paper writing, and long-term plan. Also thank the participants in the user studies for their effort, time, and feedback.

During my PhD study, I was very happy to have met my brilliant and cute lab mates, Qi Guo, Yu Wang, Dmitry Lagun, Denis Savenkov, Tomasz Jurczyk, Ablimit Aji, Haojian Jin, Alexander Kotov, Mikhail Ageev, Nikita Zhiltsov, Tianyong Hao, Julia Kiseleva, Akshatha Pai, and JongHo Shin. I want to thank them, especailly Qi, Yu, Dmitry, Denis, and Tomasz for their helpful discussion and chat where many great ideas came into being. I would also like to thank my friends in the MathCS department, in Atlanta, and in Seattle during my internships, who make my PhD life more colorful.

Last but not least, I would like to say thank you to my husband Cong Shi for his absolute love and support whenever I need it, to my little son Andrew Shi for coming to us and being so cute and sweet, to my parents and little sister for their forever love and care for me, and to my parents-in-law and sister-in-law for their great love and help. Without them, I cannot accomplish this thesis.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Contributions	8
1.3	Organization	11
2	Related Work	12
2.1	Improving Searcher Satisfaction using CQA Services	12
2.1.1	Question and Answer Retrieval in CQA	12
2.1.2	Searcher Satisfaction and Switching Behavior	13
2.1.3	Answer Quality and Asker Satisfaction	15
2.1.4	Query and Question Analysis	16
2.1.5	Improving Search Experience using CQA Data	18
2.2	Question Recommendation and Routing in CQA	19
2.3	Understanding User Behavior in CQA	24
2.4	Building Real-Time CQA Systems	25

2.5	Crowdsourcing and Social Networks	29
3	Predicting Web Searcher Satisfaction with Existing Answers	31
3.1	Problem and Approaches	33
3.2	Experimental Setup	41
3.3	Empirical Evaluation	47
3.4	Summary	54
4	Understanding When Searchers Become Askers	56
4.1	Dataset Preparation	60
4.2	Query and Behavior Analysis	62
4.2.1	Characteristics of Queries leading to Questions	62
4.2.2	Searcher Behavior Before Asking Questions	67
4.3	Queries vs. Questions	73
4.4	Question Analysis	78
4.5	Summary	82
5	Understanding Answerer Behavior for Better Question Recommendation	83
5.1	Modeling Answerer Behavior in CQA	84
5.1.1	Temporal Patterns in Answerer Behavior	85
5.1.2	Understanding How Answerers Choose Questions	92

5.2	Exploring Web Browsing Context for CQA	102
5.2.1	Study Design	104
5.2.2	Results	106
5.3	Summary	110
6	Building A Real-Time CQA System	112
6.1	System Overview	114
6.1.1	Front-end: mobile application	116
6.1.2	Back-end: server system	122
6.2	User studies	126
6.2.1	Statistics and survey responses	128
6.2.2	Question types and answer quality	130
6.2.3	Question recommendation strategies: PULL vs. PUSH	135
6.2.4	Recommendation from the main Page: question ranking	136
6.2.5	Recommendation via notification: user ranking	138
6.2.6	Tag ranking	141
6.2.7	Notification Settings	143
6.3	Discussion and Implications	144
6.4	Summary	147
7	Conclusions and Future Work	149

7.1	Summary of Thesis Work	149
7.2	Limitations and Future Work	157
	Bibliography	160

List of Figures

1.1	Interaction between users and CQA systems.	2
1.2	A subset of Google search results including resolved questions from Yahoo! Answers.	5
1.3	A resolved question on Yahoo! Answers.	6
3.1	The composite approach.	40
3.2	Distributions of the mean ratings of MTurk workers for query clarity, query-question match, and searcher satisfaction with answers.	45
3.3	Kendall's τ (a) and NDCG (b) relative improvements of the composite approach over Google's baseline on ranking answers for queries.	53
4.1	Example search (a) followed by a question posted by the same user on the Yahoo! Answers site with a satisfactory answer from the com- munity (b).	58
5.1	Basic question answering process in Yahoo! Answers.	86

5.2	Temporal patterns of answer activities in YA, showing the percentage of answers in the same hours aggregated by (a)months; (b)weeks; (c)days.	88
5.3	Example answering behavior for an active user over 1 day (a) and over a period of 2 hours (b).	89
5.4	The (a)Frequency and (b)Cumulative Distribution of the intervals between two successive answers for all active users.	91
5.5	The Cumulative Distribution (CDF) of user-based category coverage, which is the number of categories in which a user has posted answers across the entire dataset duration. The hollow circles represent user-based category coverage for top categories, and solid diamonds represent the leaf categories (a); The distribution of user entropy across all top (b) and leaf (c) categories: lower entropy indicates user activity focused on fewer categories.	94

5.6	(a)The Cumulative Distribution Function of session-based category coverage, which is the number of categories in which a user has posted answers in a single answer session. The hollow circles (solid diamonds) represent session-based category coverage for top (leaf) categories. (b)(c)The Probabilistic Distribution Function of session-based category change rate for leaf(b) and top(c) categories, which is the percentage of two successive answers in different categories posted by a user in a single answer session. Note that the session timeout threshold of 30m is used here.	96
5.7	The Probability Distribution Function(a) and Cumulative Distribution Function(b) of the positions in the list seen by a user, containing a question that was selected by the user to answer.	99
5.8	Illustration of the CONtextual QUEstion Recommender system (CON-QUER).	104
5.9	Correlation between answerer willingness to answer a question and answerer interest/ability/effort/context relevance/context helpfulness.	108
6.1	The main page of the mobile application.	117
6.2	Question thread and question post.	119
6.3	User profile and notification inbox.	121

List of Tables

2.1	Comparative statistics for Aardvark and RealQA (system built in this thesis).	27
3.1	Query clarity features (9 total).	36
3.2	Query-Question match features (23 total).	36
3.3	Answer quality features (37 total).	37
3.4	Regression results on searcher satisfaction.	49
3.5	Regression results on individual sub-tasks.	49
3.6	Sample (query, question, answer) tuples, with predictions and ground truth labels.	50
3.7	Mean Kendall's τ and NDCG results on ranking questions and answers for queries.	52
4.1	Statistics of words per query	63
4.2	Frequent words in SearchAsk queries and SearchOnly queries	65

4.3	Top frequent user action sequences in SearchAsk sessions and SearchOnly sessions (B: Begin a session, Q: Query, C_{qr} : Click on a Yahoo! Answers question result, C_{or} : Click on other result, A: Ask a question, E: End a session)	71
4.4	Statistics of length difference between a query and its associated question (number of words).	73
4.5	Overlap of content words (CW) between a query and its associated question.	75
4.6	Examples showing semantics difference between the query and the associated question.	76
4.7	Examples showing semantics difference between the query and question.	77
4.8	Categories with largest differences in assignment probability between questions coming from search and general questions	80
5.1	Dimension of the Yahoo! Answers dataset. The USER20 dataset focuses on answerers with at least 20 answers.	87
5.2	Answering session statistics for varying timeout values.	92
5.3	Features (50 total) used in the experiment	100
6.1	Functions in the systems used for the pilot and the main studies. . . .	115
6.2	Index structure for questions, users, and tags.	123

6.3	Statistics of the data collected from the user studies.	127
6.4	Survey responses. Ratings are scaled in $\{2, 1, 0, -1, -2\}$. Tuples in Question 5 represent percentages of notification, main page, and no preference. * indicates statistically significant difference according to the Mann-Whitney test at $p = 0.05$	131
6.5	Types of questions asked and their proportions.	132
6.6	Example questions and answers from the studies.	133
6.7	Statistics related to the answer quality.	134
6.8	Source of answers (considering only qualified users).	135
6.9	Question ranking in the main study.	137
6.10	Statistics of question recommendation notifications.	139
6.11	Comparing algorithms for ranking users to send recommendations for a newly posted question that is not annotated with any location. * indicates statistically significant difference according to the Mann-Whitney test at $p = 0.05$	141
6.12	Comparing algorithms for ranking users to send recommendations for a newly posted question that is annotated with a location.	141

6.13 Comparing algorithms for tag recommendation. * indicates statistically significant difference according to the Wilcoxon signed-rank test at $p = 0.05$	142
---	-----

Chapter 1

Introduction

1.1 Motivation

Community-based Question Answering (CQA) sites, such as Yahoo! Answers [114], Baidu Knows [7], Naver Knowledge-iN [80], and Quora [87], have gained substantial popularity over the recent years, providing an alternative way of online information seeking other than web search. Users resort to community help for a variety of reasons, from lack of proficiency in web search to seeking an answer accompanied by interaction with a real human. Although some of these CQA sites allow monetary payments in exchange for answering questions (e.g., JustAnswer [54] or Mahalo Answers [73]), answerers are usually attracted by social rewards and less tangible incentives, such as reputation or points [89, 78]. The CQA communities are mainly volunteer-driven, and their openness and accessibility appeal to millions of users; for example, the size of Yahoo! Answers surpassed 1 billion answers in 2010 [115], and Baidu Knows had over 300 million answered questions in 2014 [7].

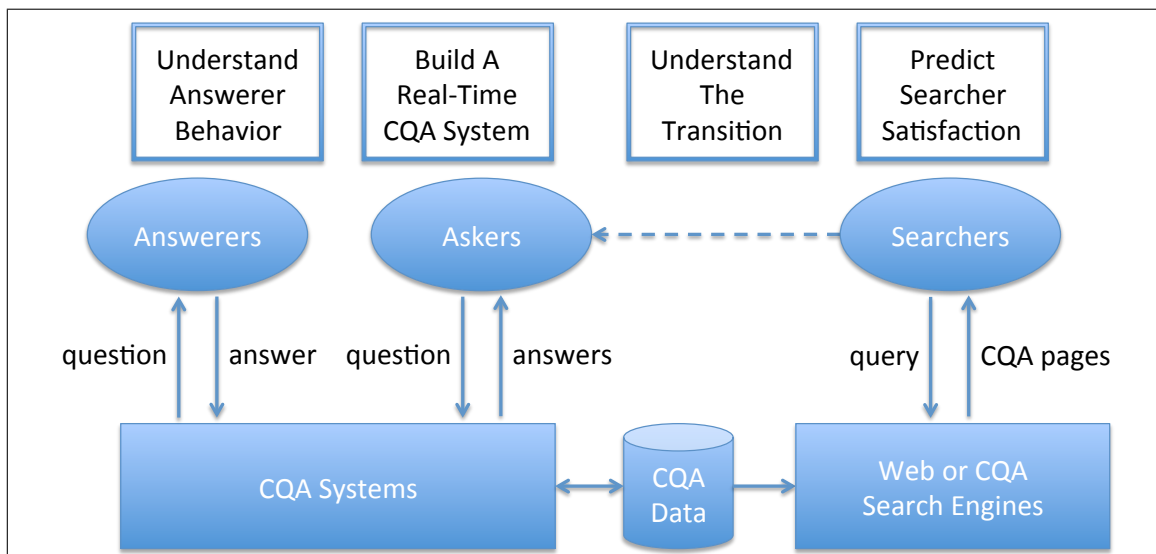


Figure 1.1: Interaction between users and CQA systems.

Users using CQA systems can play three different roles: askers, answerers, or searchers. Different roles interact with CQA systems differently as shown in Figure 1.1. Askers ask questions and get answers to their questions. Answerers view questions and provide answers to these questions. Searchers submit queries and get relevant CQA pages with relevant questions and answers as results.

A key to the success of CQA services is the quality and timeliness of the responses that users get. From the perspective of askers and answerers, askers care about the quality and timeliness of the answers to their questions, and answerers care about the quality of the questions they view and the effort to find appropriate questions to answer. One way to improve the experience of both askers and answerers is by question routing and recommendation, i.e., matching questions to potential answerers

who are most likely to provide a high-quality answer in a timely manner [25, 102, 31, 43]. Previous work has focused more on the matching algorithms, yet ignored two important problems in question recommendation.

First, the behavior of the answerers as well as the contextual factors that influence the answerer behavior and the quality and timeliness of the answers, are not well understood. Yet, understanding this is important for CQA systems to provide better question recommendation. To this end, this thesis analyzes the answerer behavior in a large scale CQA system, and explores when answerers tend to answer questions by analyzing both the overall and user-specific temporal activity patterns, and how they tend to choose questions to answer by analyzing the factors that may affect users' decisions of which questions to answer, including the question category, the question position in the list shown to users, and the surface patterns in the question text. Based on a controlled user study, this thesis further explores how relevant web browsing context could affect answerers' perceived ability, effort, and willingness to answer a question. The findings could inform the design of more intelligent question recommendation and routing strategies in CQA systems, especially when CQA moves towards the real-time setting.

Second, relatively few studies address how to deploy question routing and recommendation in real-time CQA systems and understanding user satisfaction and

preferences over different strategies. To this end, this thesis builds a real-time CQA system with a mobile interface, which supports different question recommendation strategies: users doing a pull of questions in the main page or being pushed questions via mobile notifications. The system is developed iteratively, incorporating insights from the analysis of two live user studies: a formative pilot study with the initial system design, and a more extensive study with the revised and improved interface and algorithms. Both quantitative analysis of user behavior in the system as well as qualitative analysis of user satisfaction with the system are conducted. The developed system and the reported analysis offer insights and implications for designing real-time CQA systems and deploying question recommendation in such systems.

As illustrated in Figure 1.1, besides askers, searchers could also benefit from CQA services as they might find existing answers satisfying their queries via search engines. In particular, today a large number of users, i.e., web searchers, already benefit from the public accessibility of CQA archives via all the major web search engines. Existing answers often satisfy information needs of users who submit queries to a web search engine, obtain results from a CQA site, such as the ones shown in Figure 1.2, select one of these results, and finally reach a resolved question page on the CQA site, as illustrated in Figure 1.3.

Understanding the search quality and searchers' satisfaction with the search re-

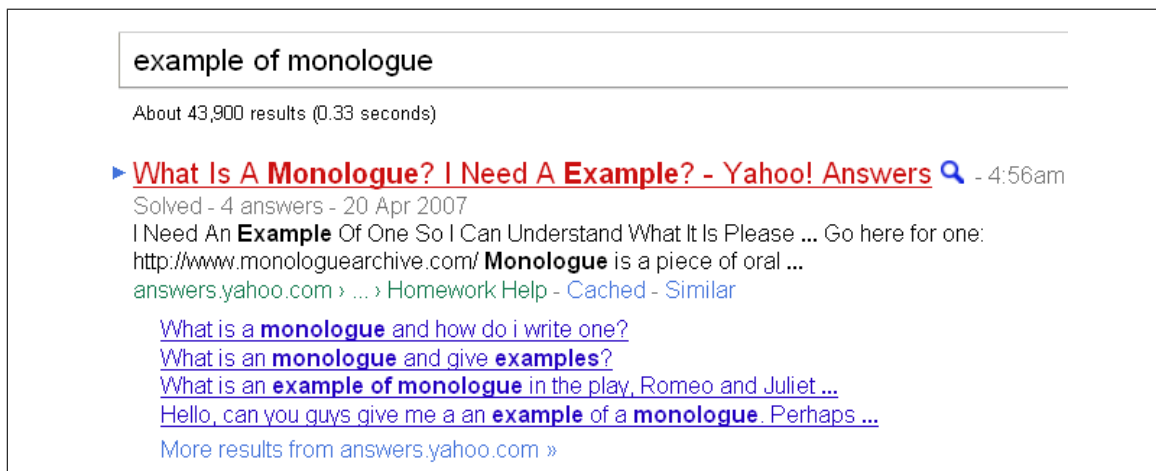



Figure 1.2: A subset of Google search results including resolved questions from Yahoo! Answers.

sults is important for improving search experience [45, 29, 40]. Yet, the understanding of searchers' satisfaction with the returned questions and answers in CQA pages has so far been under addressed. To this end, this thesis formulates a new problem of predicting the satisfaction of web searchers with answers in CQA. It analyzes a large number of web searches that result in a visit to a popular CQA site, and identifies unique characteristics of searcher satisfaction in this setting, namely, the effects of query clarity, query-to-question match, and answer quality. It then proposes and evaluates several approaches to predicting searcher satisfaction that exploit these characteristics. To the best of my knowledge, this is the first attempt to predict and validate the usefulness of CQA archives for external searchers, rather than for the original askers. The results suggest promising directions for improving and exploiting

Home > All Categories > Education & Reference > Homework Help > Resolved Question

Resolved Question
Show me another »




babyddk9...

What Is A Monologue? I Need A Example?

I Need An Example Of One So I Can Understand What It Is Please

4 years ago Report Abuse

Add to Answers Articles



missyvec...

Best Answer - Chosen by Asker

Go here for one: <http://www.monologuearchive.com/>

Monologue is a piece of oral or written literature (eg, a story, poem or part of a play) spoken by one person who exposes inner thoughts and provides insights into his or her character.

Good luck!

4 years ago Report Abuse

1
0

Asker's Rating: *****
That really help me Thanks =)

Action Bar: 2 ☆ Interesting! ✉ Email 💬 Comment (0) + Save

Figure 1.3: A resolved question on Yahoo! Answers.

CQA services in pursuit of satisfying even more web search queries.

For difficult queries, when searchers fail in web search, asking questions using CQA services might be an alternative solution [96]. To this end, this thesis performs a large-scale analysis of how searchers become askers. The logs of a major web search engine is studied to trace the transformation of a large number of failed searches into questions posted on a popular CQA site. Specifically, it analyzes the

characteristics of the queries, and of the patterns of search behavior that precede posting a question; the relationship between the content of the attempted queries and of the posted questions; and the subsequent actions the user performs on the CQA site. The work develops novel insights into searcher intent and behavior that lead to asking questions to the community, providing opportunities for having more frustrated web searchers benefit from CQA services.

In summary, this thesis aims to improve the usefulness of CQA services towards better searcher satisfaction and question recommendation, specifically by focusing on three important problems ignored in previous work: how to improve web searcher satisfaction using CQA services, what contextual factors influence answerer behavior in CQA, and how to deploy question recommendation in real-time CQA systems.

Figure 1.1 summarizes the thesis work around different user roles interacting with CQA systems. The work of predicting searcher satisfaction and understanding the transition from searching to asking is helpful for improving searcher satisfaction using CQA systems, while the work of understanding answerer behavior and building a real-time CQA system is helpful for improving question recommendation in CQA systems. All together, this helps make CQA services more useful with better experience of askers, answerers, and searchers.

1.2 Contributions

The main contributions of this thesis are summarized below:

- Methods to predicting web searcher satisfaction with existing answers in CQA archives [63] (Chapter 3).

This thesis formulates a new problem of predicting web searcher satisfaction with existing answers in CQA, and proposes regression models based on identifying three key aspects important to the problem, namely, query clarity, query-question match, and answer quality. It also demonstrates via experiments that incorporating the proposed predictor of searcher satisfaction leads to a significant improvement in ranking CQA results in web search over a state-of-the-art baseline. To the best of my knowledge, it is the first attempt to predict and validate the usefulness of CQA archives for external searchers, rather than for the original askers.

- A large-scale analysis of user transition from searching to asking in terms of query and behavior characteristics [64] (Chapter 4).

This thesis performs a large-scale study of the transformation of failed searches in a major web search engine into questions posted on a popular CQA site. It identifies the common characteristics of the queries, and of the patterns of

search behavior that precede posting a question, enabling search engines to better understand when searchers are unsatisfied by the returned results. It also investigates the relationship between the content of the attempted queries and of the posted questions, enabling better query intent understanding and automatic question formulation; Finally, it recognizes special characteristics of subsequent actions the searcher performs on the CQA site, allowing CQA sites to provide customized responses to users coming from search.

- A large-scale analysis of the temporal patterns of answerer activities and the contextual factors influencing answerer behavior [62] (Chapter 5).

On one hand, this thesis analyzes both the overall and user-specific temporal activity patterns, identifying not previously observed bursty patterns of activity in the individual answer sessions of many users, which results in a novel session-based analysis of the answerer activity. On the other hand, this thesis analyzes the factors that may affect users' decisions of which questions to answer and the effects of these factors, including the question category, the question position in the list shown to users, and the surface patterns in the question text. Realizing the importance of question category and question position in modeling answerer behavior informs intelligent question recommendation algorithms to make full use of category information and to be evaluated in a more realistic setting.

- A study showing that relevant web browsing context can have significant positive effects on the answerers' ability, effort and willingness to answer questions. [66] (Chapter 5).

This thesis performs a two-step user study in a lab setting: in Step 1 eliciting the baseline ratings independent of web browsing context, and in Step 2 quantifying the effects of web browsing context by measuring both absolute ratings and the change compared to Step 1. Then, this thesis observes significant positive effects of relevant web browsing context on the answerers' perceived effort, ability, and willingness to contribute answers. Moreover, it further identifies the characteristics of the questions or users for which the web browsing context is helpful. Exploring the effects of web browsing context on answerer behavior allows question recommendation algorithms to improve answerers' experience as well as to increase the likelihood and quality of their responses.

- A real-time CQA system that supports different question recommendation and user notification strategies [65] (Chapter 6).

This thesis builds a real-time CQA system with a mobile interface, which supports two question recommendation strategies: users doing a pull of questions in the main page or being pushed questions via mobile notifications. Then this thesis performs two live user studies with the system, and investigates users'

preference of different question recommendation strategies and the effectiveness of different ranking algorithms for each strategy, based on both quantitative analysis of user behavior as well as qualitative analysis of user satisfaction with the system. The developed system, and the reported findings and analysis, offer insights and implications for designing real-time CQA systems and deploying question recommendation in such systems.

1.3 Organization

The thesis is organized as follows: Chapter 2 overviews the related work. Then, Chapter 3 presents the methods for predicting searcher satisfaction in CQA and Chapter 4 shows the analysis on understanding how searchers become askers. In Chapter 5, the methods and results for analyzing answerer behavior in CQA are presented. Chapter 6 describes the real-time CQA system built and the user studies performed for evaluating the system. Finally, Chapter 7 summarizes the thesis work and discusses the implications, limitations, and future research directions.

Chapter 2

Related Work

This section discusses related work to the thesis, which can be organized into four parts. The first part is related work on improving searcher satisfaction using CQA services, which is reviewed in Section 2.1. The second part is related work on question recommendation and user behavior understanding in CQA, which is reviewed in Section 2.2 and Section 2.3 respectively. The third part is related work on building real-time CQA systems, which is reviewed in Section 2.4. The fourth part is related work on crowdsourcing and social networks, which is reviewed in Section 2.5.

2.1 Improving Searcher Satisfaction using CQA Services

2.1.1 Question and Answer Retrieval in CQA

Searching CQA archives with an input question, which allows efficiently reusing existing answers of similar questions to satisfy the new question, has been an active area

of research, and many retrieval models specialized to CQA content have been proposed. Example methods include applying statistical machine translation techniques [47, 113, 57, 123, 125, 124], incorporating question category information [19, 17, 18], using syntactic tree structure [105] or other question structure [27], computing question utility [97], enhancing term weighting [120, 75], learning latent topics [15, 50], and leveraging world knowledge of Wikipedia [127].

When relevant questions are found, however, their associated answers may suffer from low quality, which makes the results less useful for searchers. Therefore, some research effort has been done on combining both answer quality and relevance for retrieving answers [48, 9, 100, 106, 99].

This thesis focuses on estimating searcher satisfaction with a given retrieved question-answer pair for a query. During the process, some of these techniques are adapted and extended for matching the search query to the question and the answer content. Additionally, as will be shown, the estimated searcher satisfaction could be applied for effective re-ranking of the retrieved question-answer pairs for a query, resulting in a significant improvement over a state-of-the-art baseline.

2.1.2 Searcher Satisfaction and Switching Behavior

Significant research has been done on estimating searcher satisfaction or frustration in Web search [45, 29, 40, 2], which utilized query log information for the task, such

as relevance measures, as well as user behavior during the search session, including mouse clicks and time spent between user actions. What makes this task such a challenging problem is the large diversity in user goals [92], with a different definition of satisfaction for each, which requires developing unique satisfaction prediction models for the respective information needs.

This thesis focuses on satisfying the types of queries that are arguably difficult for a web search engine to satisfy and often require other people to answer [76, 77]. Specifically, some of these needs can be satisfied with existing answers from CQA archives. Hence, the goal of this thesis is to harness the unique structure of such archives for detecting web searcher satisfaction, which is not captured by standard query and session logs.

White and Dumais [109, 36] studied search engine switching behavior and developed models to predict the switching and its rationale. Although different types of searchers are focused on (they focused on searchers who turn to another search engine and issue more queries, while this thesis focuses on searchers who turn to CQA sites and post questions), both are interested in characterizing the types of queries and searcher behavior that lead to the switchings. The performed analysis shows both similar (e.g. longer sessions are more likely to involve a switch) and different characteristics (e.g. different last action before switching) compared to their study.

2.1.3 Answer Quality and Asker Satisfaction

The high variance in the perceived answer quality has been one of the main problems in CQA sites. Some studies attempted to assess the quality of answers and build a classifier to distinguish high-quality answers from the rest, based on non-textual features [48], or combination of textual, relationship, and usage features [3] or user expertise based features [100]. Another branch of work focused on predicting the best answer for a question [95] or rank all the answers for a question based on their quality [99, 106]. Other studies investigated via a controlled study what question askers can do to receive better answers from a CQA site [39].

Relatively few studies addressed the satisfaction of a user using CQA services. Most closely related to the thesis work, Agichtein et al. [68] attempted to predict whether the asker of a question will be satisfied with the received answers, and built models based on a variety of content, structure, and community-focused features for the task.

This thesis proposes to estimate the satisfaction of web searchers with the answers in CQA, as opposed to the original askers [68]. As answer quality is a key factor affecting searcher satisfaction, when building the feature set for estimating searcher satisfaction, the proposed method applies the features indicative of answer quality based on the analysis given in [3] as well as the top performing features for predicting

asker satisfaction as reported in [68].

2.1.4 Query and Question Analysis

Understanding query intent has been an active area of research because of its importance to Web search. Broder [12] firstly proposed to categorize query intent into navigational (to find a specific website), informational (to find information about a topic) and transactional (to perform a web-mediated activity). Rose and Levinson [92] further proposed a more refined hierarchy of search goals. Correspondingly, many query intent detection methods have been proposed [11]. Meanwhile, different dimensions for query intent have been characterized. For example, Calderón-Benavides et. al. [16] studied 9 dimensions of query intent along with their relationships and dependencies for better query intent identification. Donato et. al. [23] identified a novel subset of informational intent, namely research missions, which are cases when users' needs are too complex or too heterogeneous to be answered by a single Web page.

Besides query intent analysis, difficult queries [21, 20], long queries [8], and question-like queries [82] have also received special research attention. Cronen-Townsend et al. [21] found that query clarity scores, which computes the KL divergence between the query and collection language models, correlate with query performance using the TREC test sets. Carmel et al. [20] further showed that topic difficulty highly depends on the distances between the three components of a topic,

i.e., the query, the relevant documents, and the document collection. Bendersky and Croft [8] studied the long queries based on a large scale search log and showed the retrieval effectiveness of a query decreases as the query length increases. With an analysis of question formulation in Web search queries, Pang and Kumar [82] showed that users are becoming more likely to use question-queries, even over the long-standing search intents persisting over a 12-month period.

This thesis focuses on studying the types of queries that are arguably difficult for a web search engine to satisfy, often require human to answer [76, 77], and thus could be better handled by CQA sites. The observation that many searchers who were not satisfied with the search results finally posted a related question on a CQA site, inspires the analysis of how searchers become askers in this thesis.

On the CQA side, there is also research effort devoted to question analysis, e.g. distinguishing conversational and informational questions [38], identifying high quality questions [3], and investigating the effects of contexts in questions on answer quality [101]. This thesis uses their classification of contextual factors to analyze the semantic difference between the query and question posted by the same user for the same information need.

2.1.5 Improving Search Experience using CQA Data

Regarding taking the advantage of CQA sites to improve the experience of web searchers, some recent work has been proposed to automatically generate questions from queries [26, 121, 122]. A common idea of these methods is to use question templates for generating unseen questions. Also, the pairs of queries and their associated clicked questions from CQA sites played important roles, either in learning the templates or in building evaluation sets. Weber et al. [108] built a system to extract tips from Yahoo! Answers and serve them directly to “how-to” web queries instead of ranked web pages. Gao et al. [32] proposed to map keyword queries to questions on CQA sites, so as to find popular questions that capture the various information needs behind a query. Their method was based on the assumption that if a question represents an information need behind a query, the question must cover the query in terms of topic words, which are derived from noun phrases in questions and question category names. Si et al. [96] presented challenges and solutions to integrate CQA and web search.

This thesis focuses on different perspectives of using CQA sites and related data to help searchers address hard information needs, e.g., via improving query understanding and improving searcher satisfaction with CQA results, which is therefore orthogonal to the above work.

2.2 Question Recommendation and Routing in CQA

Question recommendation and routing, i.e., matching questions to potential answerers, has been proposed in CQA in order to improve the answer quality, reduce the waiting time to get an answer, and save the effort of answerers to find target questions. There are two ways for potential answerers to retrieve target questions: proactively do a pull of questions from CQA services (PULL), or let CQA systems push questions to them (PUSH). Accordingly, question recommendation can be done using two strategies.

The first question recommendation strategy is the PULL strategy, i.e., given a user, the system computes a ranked list of questions, which are recommended for the user to answer. Kabutoya et al. [55] applied a logistic regression model and computed six features using both collaborative filtering and content-based filtering schemes for this task. The results showed answer histories are more useful than question histories and content-based filtering more useful than collaborative filtering. The most two useful features were the probability of answering in the question category by the user, and the cosine similarity between the question and the user's answered questions. Dror et al. [25] applied a multi-channel recommendation model and computed various content signals (e.g., text and categories of questions and the associated answers) and social signals (e.g., user interactions with questions like ask-

ing, answering, and voting) for this task. The results showed category features were most important followed by textual features, and both were more important than social signals. Szpektor [102] pointed out that an effective personalized question recommendation should consider relevance, diversity, and freshness of the questions. A scalable solution was proposed based on representing questions and users as probability distribution trees and computing three feature vectors: latent topic vectors, lexical vectors, and category vectors. The results showed promoting diversity and freshness largely improved the number of answers than considering relevance alone.

The other question recommendation strategy is the PUSH strategy, i.e., given a newly posted question, the system computes a list of users, to whom the question will be sent in order to get answers. This process is also called question routing [31].

An attractive approach to improving the answer quality for CQA askers is to route a newly posted question to experts on the topic of the question. Therefore, identifying domain experts in CQA could help with the task, which has been actively studied. For example, Jurczyk et al. [53] formulated a graph structure for CQA systems and applied a web link analysis algorithm to discover authoritative users in topical categories. Bouguessa et al. [10] focused on automatically discriminating between authoritative and non-authoritative users by modeling users' authority scores as a mixture of gamma distributions for each topic. Si et al. [96] applied a weighted

and topic-sensitive link analysis algorithm based on user activity graph for finding domain experts. Beyond the CQA context, there has also been extensive work on expert finding in online forums such as [119, 129].

As a step forward, many methods have been proposed to identifying the best possible answerers for a particular question. Language modeling and topic modeling have been shown to be effective for capturing user interests based on user's answer history [67, 60, 34, 86, 61, 90, 112, 116, 117, 93, 104]. Besides user interests, researchers also found it important to consider user authority, answer quality, and related community signals like voting when ranking appropriate answerers [61, 59, 126, 128, 117, 49, 93, 104]. To further reduce the response time to new questions, methods that consider user availability and activity which would affect the likelihood for a user to answer a question have been proposed [61, 43, 59].

Brief descriptions about these question routing methods were provided below. A more detailed survey can be found in [31]. Liu et al. [67] casted the expert finding problem as an IR problem, by viewing a new question as a query to retrieve the user profiles as documents, and tested several language models for expert ranking. Guo et al. [34] developed a probabilistic generative model to obtain latent topics for questions and users, and incorporated both the topic-level and term-level information for routing new questions to potential answerers. Qu et al. [86] applied PLSA to

capture user interests in terms of topics based on their answering history. Liu et al. [61] integrated language model and Latent Dirichlet Allocation model for matching an answerer and a question, which also modeled user activity and authority as prior probability. Horowitz et al. [43] addressed the question routing problem in a real-time CQA system by considering user interest, connectedness, and availability. Li and King [59] proposed a language model based framework which considered answer quality and user availability when locating appropriate answerers. Li et al. [60] incorporated question category information into a category-sensitive language model for ranking users. Riahi et al. [90] tested four models for matching a new question to user profiles: language model, vector space model, Latent Dirichlet Allocation, Segmented Topic Model (STM), and found that STM achieved the best performance. Zhou et al. [128] casted the problem of question routing as a classification task, and developed both local and global features related to the question, user history, and question-user relationship. Xu et al. [112] proposed a dual role model based on PLSA and showed that the dual roles of users (as askers or answerers) have different influences on question recommendation. Zhou et al. [126] proposed a joint learning method considering both word mismatch and answer quality for question routing. Yan and Zhou [116] combined tensor model and topic model to the question routing task for capturing the semantic relations among asker, question, and answerer. Ji

and Wang [49] applied learning-to-rank methods to question routing which utilized the intrinsic relationships between the asker and answerers per question. Yang et al. [117] combined a probabilistic generative model and an extended PageRank algorithm to model user expertise and interests under different topics. Pedro and Karatzoglou [93] combined supervised ranking with topic modeling to jointly model both community feedback and text content topics for finding potential answerers. Tian et al. [104] utilized both topic modeling and collaborative voting to model user interests and expertise for predicting the best answerer for a new question.

Relatively few CQA systems support both the PUSH and PULL question recommendation strategies. To the best of my knowledge, the only work supporting both the two strategies in a single system is Aardvark [43]. In their study, more users answered via PUSH than PULL because users were willing to answer questions to help friends in their social networks. This thesis also studies users' preference of the PULL and PUSH strategies for question recommendation. However, the finding is different from the one found by Aardvark. One important reason might be the differences in terms of user connectivity level in the two systems. In RealQA, PULL is preferred, when a closely knit social network may not exist. Moreover, for the PULL strategy, this thesis considers a new factor for ranking questions, i.e., the number of existing answers of a question, which has not been investigated in previous work [102].

The analysis of answerer behavior performed in this thesis could further provide useful information for improving the question recommendation methods above. For example, the session-level patterns in the answerer behavior, and the importance of question category and ranking positions for choosing questions to answer is useful for improving the PULL question recommendation strategy. Moreover, the study of the effects of an answerer's Web browsing context on the answer contribution behavior is a first step towards saving answerers' effort to answer questions in the PUSH question routing process.

2.3 Understanding User Behavior in CQA

Understanding user behavior in CQA has been an active research area towards improving the effectiveness of CQA services. For example, Adamic et al. [1] analyzed the content properties and user interaction patterns across different Yahoo! Answers categories. Gyöngyi et al. [37] studied several aspects of user behavior in Yahoo! Answers, including users' activity levels, interests, and reputation. Guo et al. [35] studied the patterns of user contributions in knowledge sharing-oriented online social networks including a question answering social network. Nam et al. [78] investigated the motivation of top answerers in Naver Knowledge-iN, a popular Korean CQA system, including altruism, learning, competence and points. Aperjis et al. [6] studied

the speed-accuracy tradeoff faced by CQA askers, i.e., maximizing the information accuracy while minimizing the waiting time.

This thesis applies some previous analyzing methods to analyzing answerer behavior. More specifically, it applies the entropy measurement introduced by [1] when analyzing how focused an answerer is across categories. It also observes interesting aggregate temporal patterns of answer contributions which show some similarity with the patterns of user posting activities (including both question and answer posts) found in [35]. Yet, none of the above work has focused on comprehensively analyzing answerer behavior as in this thesis.

2.4 Building Real-Time CQA Systems

Real-time Question Answering

Real-time question answering (QA) systems have been designed to shorten the time for a question to be answered. These systems typically use some synchronous communication channels, e.g., instant message, SMS, mobile application notifications, emails, for asking and answering questions. In particular, these systems often use the PUSH strategy for question recommendation, i.e., pushing questions to potential answerers. For example, Aardvark automatically routes a question to people in the asker's extended social network who are most likely to answer [43]. The user rank-

ing algorithm considers user’s expertise in the question topic, connectedness to the asker, and availability to respond. IM-an-Expert provides an instant message service deployed in an organization to find an expert for a question and automatically create dialog sessions for the asker [110, 91]. The expert ranking algorithm is based on matching question text with user profiles using TF-IDF, an established method in information retrieval. Mimir, a market-based real-time QA system, broadcasts a question to all other users [44].

Some systems allow users to locate potential responders for new questions, but depend on the user to choose whom to ask. For example, the Quora service called “Online Now” enables an asker to find a list of experts who are currently online for his question, so that the user can choose whom to ask [88]. [79] proposed a system which can find a set of Twitter friends for a query based on availability, willingness, and knowledge. [74] presented methods to locate targeted strangers on Twitter for information solicitations.

The real-time CQA system built in this thesis, RealQA, supports both synchronous and asynchronous communication channels; it uses mobile application notifications as the synchronous channel, and the application main page as the asynchronous channel. It is most relevant to Aardvark. Table 2.1 compares the statistics reported in [43] and the statistics collected in the main study for RealQA, although

	Aardvark	RealQA
% of subjective questions	64.7%	71%
% of questions with local intent	10%	77%
% of questions answered	87.7%	89.3%
% of users received rec. questions	86.7%	66%
% of users clicked rec. questions	70%	74%
% of users answered rec. questions	38%	48%
% of users answered any questions	50.0%	74.3%

Table 2.1: Comparative statistics for Aardvark and RealQA (system built in this thesis).

the comparison is not quite fair due to different deployment settings and participation incentives. It is found that the nature of questions collected in RealQA is more subjective and local-intent, compared to that in Aardvark. Moreover, the proportion of users answered any questions (or recommended questions) is higher in RealQA. The main difference between the two systems is that Aardvark focuses more on social network while our system focuses more on location proximity when exploring recommendation and notification strategies.

Location-based Question Answering

Location-based QA systems have been designed to facilitate the information seeking and knowledge sharing about some geographic locations, e.g., LocalMind [69], LOCQL [70], and Naver KiN “Here” [83]. Typically, these systems allow users to post questions to users around a specific geographic location. A detailed classification of location-based CQA systems can be found in Table 1 in [83].

The system built in this thesis provides similar features, i.e., when posting a question, users can select “asking people around a specific location”. Then, the question recommendation algorithm uses this location information to recommend the question to users who are not only interested in answering but also close to that location.

Mobile-based Question Answering

Asking and answering questions from mobile devices have become increasingly popular [58, 84]. Most popular CQA services also provide mobile-friendly websites or mobile applications such as Yahoo! Answers [114], Quora [87], and Stackoverflow [98].

The portability of mobile devices also makes accurate location detection and real-time interaction with users easier. This motivates this thesis to support real-time and location-aware question answering services based on a mobile application. Recently, [85] studied the effect of mobile phone notifications on the daily lives of mobile users, and showed that an increasing number of notifications correlated with negative emotions. From the studies performed in this thesis, an observation found is that carefully sending notifications to users and let users have control resulted in better system performance and user satisfaction. [85] also found that silent notifications were not viewed slower than non-silent ones, which supports the design decision in this thesis to send silent notifications for recommendations.

2.5 Crowdsourcing and Social Networks

CQA systems, where users collaborate explicitly to find and share information, can be classified as one type of crowdsourcing systems on the Web [22]. Some crowdsourcing systems pay workers to perform tasks, and some ask for self-incentivised volunteers. With the increasing use of mobile devices, spatial crowdsourcing systems which focus on addressing location-based tasks have been proposed [56, 13, 4]. Similar to CQA systems, crowdsourcing systems also need to deal with problems like estimating user expertise levels and finding appropriate workers to assign a (location-based) task, although the settings and tasks could be diverse. Therefore, the general principles of solving these problems in crowdsourcing systems could be useful for designing question routing and recommendation in CQA systems, after being customized to the question answering tasks.

Social networking services like Facebook and Twitter are also examples of crowdsourcing systems [22], which can be used to address an information need by posting a question to one's online social network. To save users' effort, several methods have been proposed which allow users to locate potential responders from their network for new questions, e.g., [88, 79]. Towards a better understanding of different information seeking techniques, Morris et al. studied the types of questions people ask their social networks (e.g., Facebook, Twitter) and corresponding motivations

[76, 77]. Following the study, tradeoffs of social networks, search engines, and CQA were discussed: While search engines are good at finding more objective answers, both CQA and social networks enjoy great advantages of addressing subjective questions, e.g., those asking for opinions or recommendations, and CQA often becomes more attractive for highly personal topics (e.g., health, dating, religion, and finance) since anonymity is possible.

Chapter 3

Predicting Web Searcher Satisfaction with Existing Answers

Prior studies have mainly considered first-order effects of CQA, namely, the satisfaction of the question asker by the posted answers [68]. However, CQA has significant secondary benefits, i.e., a large number of web searchers could also benefit from CQA archives by finding existing answers that address their information needs, especially the difficult ones, via search engines. To understand how much this benefit is and how to maximize the benefit, it is necessary to understand what it means for a web searcher to be satisfied by an existing answer from a CQA archive, and to be able to predict this satisfaction. This chapter proposes and addresses this new problem of predicting the satisfaction of web searchers with existing CQA answers. The bulk of this work earlier appeared in SIGIR'11 [63].

One way to approach this problem would be to define features of queries, questions, and answers (and possibly pairs thereof), and solve it within the machine

learning paradigm. This can be done by constructing a labeled dataset of queries and answers, tagged by human judges based on how well they believe a query is satisfied by a given answer. This method is named as a direct approach.

Since queries are often quite short, and questions, which can be viewed as an intermediate link between queries and answers, are often not much longer, another way to approach the problem is through exogenous knowledge. To this end, this thesis identifies three key characteristics of searcher satisfaction, namely, query clarity, query-question match, and answer quality. It then collects separate human labels for each, and builds regression models for predicting these characteristics. Learning from these task-specific labels explicitly makes use of domain knowledge about the problem structure, which is not available in the above direct approach. It then uses the output of these individual regressors as features in a subsequent regression task, which aims to predict searcher satisfaction. This method is named as a composite approach. This approach also allows better understanding of how much the performance in the main prediction task can be improved by improving each of the individual regressors (by replacing the intermediate regression predictions with actual human labels). This additional interpretability of the model provides further insights into the problem.

The main contributions of this chapter include:

- Formulated a new problem, namely, predicting web searcher satisfaction with

existing answers in CQA. To the best of my knowledge, this is the first attempt to predict and validate the usefulness of CQA archives for external searchers, rather than for the original askers.

- Proposed two methods for solving this problem, a direct method and a composite method, based on identifying three key aspects important to the problem, query clarity, query-question match, and answer quality.
- Applied the methods to a standard ranking task in web search, where answers are treated as a semi-structured document collection, and showed significant improvement over a state-of-the-art baseline.

3.1 Problem and Approaches

This section first introduces the task of predicting searcher satisfaction by a CQA page, and then proposes approaches for representing and tackling this problem using regression algorithms.

Problem Description

The task of predicting searcher satisfaction by a CQA answer is defined as follows:

Given a search query S , a question Q , and an answer A originally posted in response to Q on a CQA site, predict whether A satisfies the query S .

Thus, instead of a web search satisfaction task that examines a $(query, webpage)$ pair, this thesis considers a different tuple $(query, question, answer)$, where the $(question, answer)$ pair has been extracted from the CQA page. The reason for using a more refined representation of $(question, answer)$ rather than a full web page is mostly for interpretability at a finer level.

To solve this prediction problem, this thesis proposes to break it down into three sub-tasks: *query clarity*, *query-question match* and *answer quality*. More specifically:

- The *query clarity task*, which should not be confused with traditional query difficulty in IR, consists of estimating whether the query may be viewed, and understood, as a question. This thesis hypothesizes that if a query is not understandable or ambiguous, a CQA site is unlikely to have an existing satisfactory answer for this query.
- The *query-question match task* consists of estimating whether the question is driven by the same or by a similar enough information need as the query. This is a prerequisite for the answer to have a chance to address the query. Furthermore, since most search result snippets will only show the question title, this match is a key driver for the searcher to select a specific CQA page: the question plays the role of a critical intermediary between the query and the answer.
- The *answer quality task* allows estimating the prior quality of the answer, with

respect to the original question, and thus relates to the previously studied asker satisfaction task [68]. In this thesis, answer quality characteristics are used not as the goal, but rather as additional input for the main task of predicting searcher satisfaction.

There are multiple advantages of breaking the main task into subtasks. First, it is helpful for better understanding and analyzing the problem structure, and devising more effective algorithms, as described next. Second, the resulting models become more interpretable and informative, by allowing the analysis of the performance for each subtask. Finally, answer quality and related prior information (taking advantage of meta-information in particular) may be computed offline within the CQA site using methods such as described in [3].

The searcher satisfaction task seems to be better modeled as a graded task, since it is easier for humans to judge satisfaction as a score within some range (see human annotation in Section 3.2). Therefore, this thesis treats the searcher satisfaction task as a regression problem. To this end, appropriate features need to be defined for learning the regressor. The following of this section will describe the features used to represent the information in the proposed task, and then formulate the direct and composite approaches.

- # of characters in the query.
- # of words in the query.
- # of clicks following the query.
- # of users who issued the query.
- # of questions clicked following the query.
- Overall click entropy of the query [103].
- User click entropy of the query [107].
- Query clarity score computed based on the language model built with approximately 3 million questions (using title, details and best answer) posted in 2009-2010 [21].
- WH-type of the query (whether it starts with ‘what’, ‘why’, ‘when’, ‘where’, ‘which’, ‘how’, ‘is’, ‘are’, ‘do’).

Table 3.1: Query clarity features (9 total).

- Match scores between the query and the question title/details/best-answer using the cosine/TFIDF/KL-divergence/BM25 retrieval models.
- The Jaccard/Dice/Tanimoto coefficient between the query and the question title.
- Ratio between the number of characters/words in the query and that in the question title/details.
- # of clicks on the question following this/any query.
- # of users who clicked the question following this/any query.

Table 3.2: Query-Question match features (23 total).

- # of characters/words in the answer.
- # of unique words in the answer.
- Ratio between the number of characters/words of the question (including title and details) and the answer.
- # of “thumbs up” minus “thumbs down”, divided by the total number of “thumbs” received by the answerer.
- # of “thumbs up” minus “thumbs down” received by the answerer.
- # of “thumbs up/down” received by the answerer.
- Match scores between the answer and the question title/details using cosine/TFIDF/KL-divergence/BM25 retrieval models.
- Percentage of users who voted this answer as the best.
- # of votes given by the voters for the answer.
- Best answer ratio for the answerer.
- Avg # of answers attracted by past questions of the asker.
- # of answers received by the asker in the past.
- Asker’s rating of the best answer to her previous question.
- Avg past rating by the asker.
- Time passed since the asker registered in Yahoo! Answers.
- # of previous questions resolved for the asker.
- Avg asker rating for best answers in the category.
- Avg voter rating for best answers in the category.
- Time of day when the question was posted.
- Avg # of answers per question in the category.
- Time passed since the answerer with most positive votes registered in Yahoo! Answers.
- Highest best answer ratio for any answerer of the question.
- Avg best answer ratio for all answerers of the question.
- Avg # of answers per hour in the category.
- Whether the best answer is chosen by the asker.
- Asker rating for choosing the answer as best answer.

Table 3.3: Answer quality features (37 total).

Features

By breaking down the main task into three subtasks, this thesis distinguishes between query clarity, query-question match, and answer quality features.

Since some of these subtasks were previously studied independently, such as [21, 103, 107] for query clarity, and [3, 68] for answer quality, this prior work can be leveraged in the construction of the feature set for the main task here. Each of the feature groups is described below, while the complete list of features is shown in Table 3.1, Table 3.2 and Table 3.3.

- *Query clarity features* include query length, click statistics for the query, and query entropy computed based on the click dataset. Besides, it also computes a query clarity score based on a language model of the CQA collection, as well as an indicator whether the query starts with a “WH” or other question word.

- *Query-question match features* include match scores computed by popular retrieval models such as cosine similarity, TFIDF, BM25, and KL-divergence language model. For measuring these scores, parts of a CQA page (the question title, question details and the best answer) are treated as separate documents, and match each such part against the query. Additional features include measures of the overlap between the query and question, such as Jaccard coefficient and length ratios, and co-occurrence statistics between the query and the given question from the click data.

- *Answer quality features* are of two types. The first type of features deals with the quality of the answer, and is mainly based on the analysis in [3]. The second type of features addresses the answer quality as predicting asker satisfaction, which maps to the third subtask. To this end, the top performing features for predicting asker satisfaction as reported in [68] are used.

Direct Approach: Logistic Regression

The first approach to estimating searcher satisfaction, which is named the direct approach, consists of simply training a regressor over all the features defined for a given *(query, question, answer)* tuple. The rationale here is to rely on the power of discriminative learning to optimally use all available features to predict the final target.

While many regression algorithms could be employed for this task, preliminary experiments were performed with a wide range of models, including Linear Regression, Gaussian Processes, Ridge Regression and Random Forests, which indicated Logistic Regression to be the most promising approach due to high variability and non-linear distribution of many of the input features.

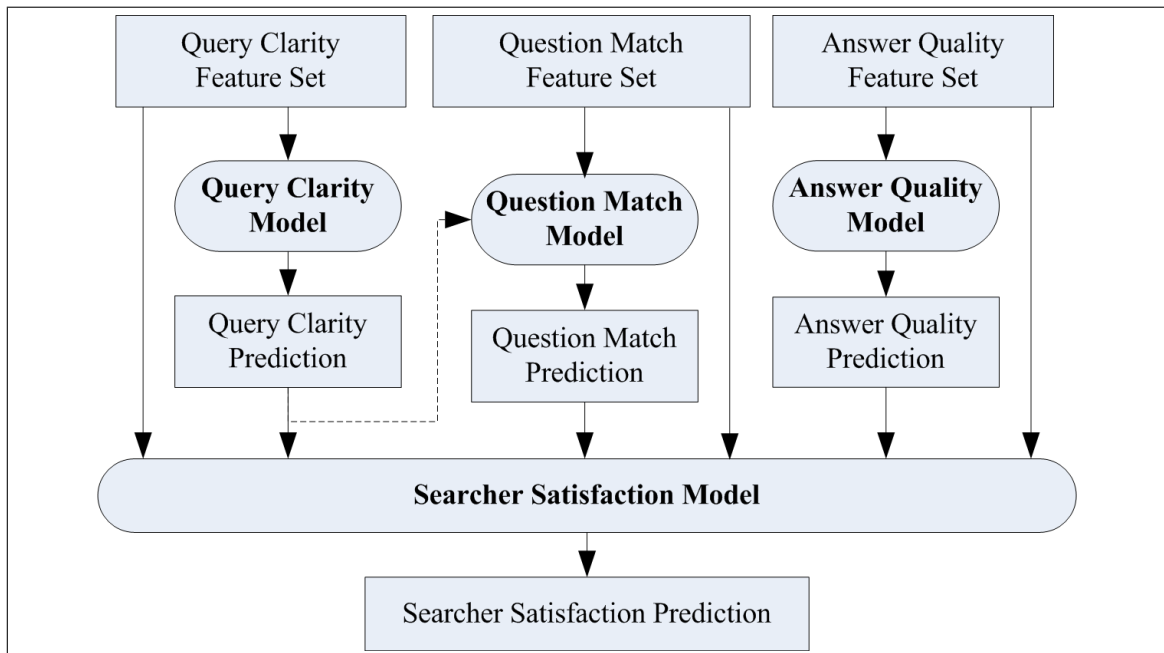


Figure 3.1: The composite approach.

Composite Approach

The second approach, called the composite approach, first trains a separate logistic regression model for each of the three subtasks defined above, and then combines their results for main task (predicting searcher satisfaction). Figure 3.1 depicts the high-level workflow of this approach. In this approach, each regressor is trained using a subset of features relevant for each subtask. Considering that query clarity may affect the question match, the query clarity prediction is also added as a feature to the query-question match regressor. Finally, the regression predictions for the three subtasks are provided as features for the final regressor to predict the overall searcher

satisfaction.

This composite approach presents several advantages over the direct approach. First, it is more flexible in terms of feature selection. The individual regressors could be trained either on the same feature set, i.e., the large feature vector used in the direct approach, or on different feature sets selected by suitable feature selection methods for each subtask. More importantly, the composite approach can take advantage of the advances made by the research community in each of the sub-tasks, to improve the prediction of overall searcher satisfaction.

3.2 Experimental Setup

This section first describes how a dataset is assembled from a sample of Google queries and a log of visits to Yahoo! Answers, and then describes the rating procedure to acquire the “ground truth” for searcher satisfaction, and the characteristics of the resulting data.

Dataset Preparation

To explore how searchers are satisfied with the existing answers in CQA sites, this thesis used a large sample of queries issued to Google’s search engine from Aug 24, 2010 to Aug 30, 2010 by users who selected as result (by clicking on it) a Yahoo! Answers link. This click dataset contains more than 37M clicks on 6M questions by

20M users following around 20M queries. By analyzing the distribution of this click data, 86% of the queries are issued by only one user; therefore, most of the queries are tail queries.

Since it is hard for human to label searcher satisfaction for such a big dataset, this thesis randomly sampled it to generate a smaller dataset consisting of 614 clicked questions following 457 queries issued by at least two users. These questions and corresponding answers may be biased to satisfy the searchers' information needs, as they are clicked from the search results. To correct this effect, this thesis further issued a random sample of 118 queries to Google's search engine with site restriction to Yahoo! Answers and crawled the top 20 results (all question pages due to the site restriction). Only questions posted in 2009-2010 are kept based on the available associated metadata. In total, the final dataset comprised of 1681 query-question pairs. The dataset with the associated human labels is publicly available through Yahoo's Webscope program (<http://webscope.sandbox.yahoo.com/catalog.php?datatype=l>, Dataset L16 - Yahoo! Answers Query to Questions).

Human Labeling

Amazon Mechanical Turk (MTurk) was used to collect human judgments on how an answer satisfies a search query. To better understand the effects of query clarity and query-question match on searcher satisfaction with answers, the MTurk workers were

also asked to label how clear the query is and how the question matches the query. The 3-scale rating method was used for all the rating tasks, {clear=1, medium=2, vague=3} for query clarity, {well matched=1, partially matched=2, not matched=3} for question match, and {highly satisfactory=1, somewhat satisfactory=2, not satisfactory=3} for searcher satisfaction. Each MTurk hit consists of 15 (*query, question, answer*) triples, and each hit is labeled by 5-7 workers.

To validate the labels of MTurk workers, six researchers were also asked to label the query clarity for all the queries. Then the agreement between the researchers and the MTurk workers was analyzed. This thesis first computed the average rating by researchers as well as by MTurk workers for each query, then used a threshold t to cast each average numerical rating nr into a binary rating br (if $nr \leq t$ then br =clear, else br =not clear), and finally computed the Fleiss's kappa coefficient[30] based on these binary ratings between the two sources. The highest kappa value 0.38 was achieved with a threshold of 1.3 (average agreement=0.70, average majority percentage=0.85). This analysis showed that the ratings from MTurk workers were reasonable.

For query-question match and searcher satisfaction, only ratings from the Mechanical Turk were collected. So this thesis used the same threshold strategy to cast each ordinal rating into a binary rating, then computed the Fleiss's kappa coefficient

for each MTurk HIT, and finally computed the average kappa value. The highest kappa value (0.34) was achieved with a threshold of 2 for query-question match (average agreement=0.85, average majority percentage=0.91), and the highest kappa value (0.25) was achieved with a threshold of 2 for searcher satisfaction (average agreement=0.76, average majority percentage=0.84).

From the above agreement analysis, we can see that although the kappa coefficient among MTurk workers is not high, possibly due to the careless rating of some MTurk workers, the average rating by all the MTurk workers shows a moderate agreement with researchers. Therefore, we use the average rating by MTurk workers as the ground truth in order to evaluate the prediction of query clarity, query-question match, and searcher satisfaction with answers.

Figure 3.2 shows the distributions of the resulting ground truth set. The x axis represents the mean over the ratings by all MTurk workers, with 1 standing for the highest score, e.g., clear/well-match/highly satisfactory for respectively query clarity/query-question match/searcher satisfaction with answers, and with 3 standing for the lowest score for each. The y axis represents the frequency count of ratings in each bucket of x. We can see that all the distributions are skewed, especially the query clarity one due to the bias of the click data. Distribution for question match and searcher satisfaction are more balanced after the search engine results

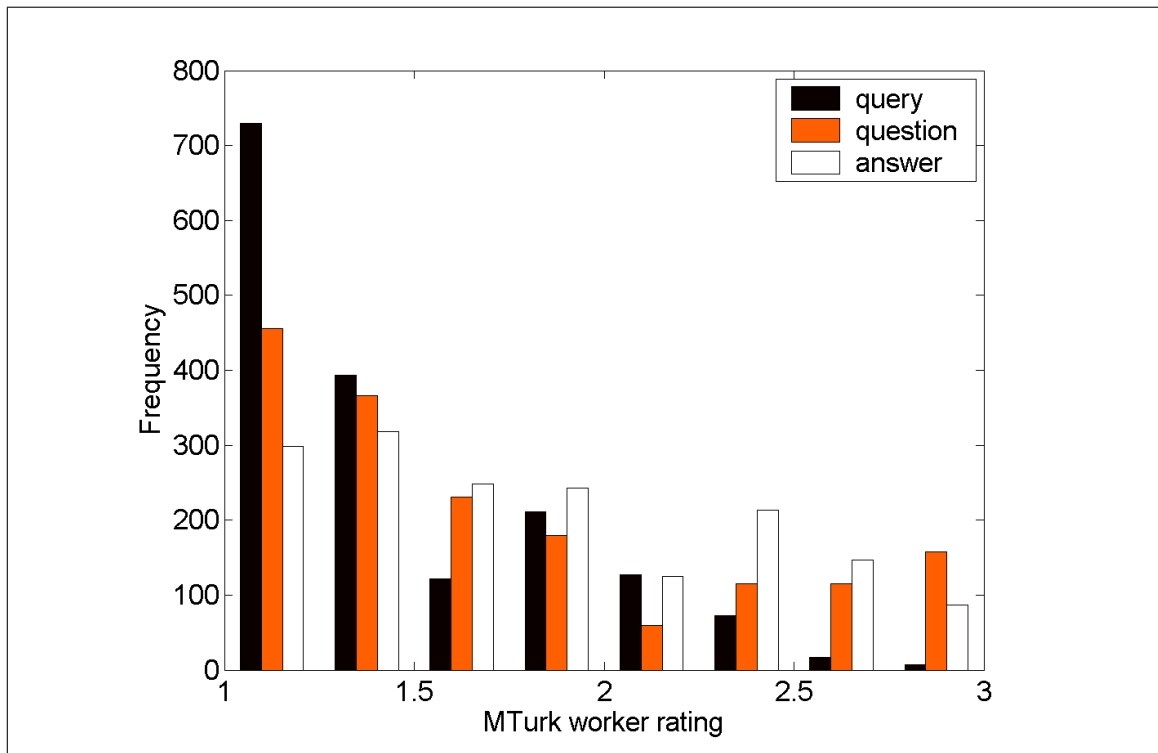


Figure 3.2: Distributions of the mean ratings of MTurk workers for query clarity, query-question match, and searcher satisfaction with answers.

were added. To better understand the relations between the three variables, the Pearson correlation between them were computed: the correlation between query clarity and searcher satisfaction is 0.1428, and the correlation between question match and searcher satisfaction is 0.6970.

Methods Compared

Four methods for estimating a searcher's satisfaction on Yahoo! Answers results are compared:

- **Google-derived baseline:** As described in Section 3.2, the top 20 results of Google were crawled by submitting our queries to Google search, with site restriction to the Yahoo! Answers site. As a result, a ranked list of question pages for each query was obtained from the search engine. The rank of each question page indicates how well this page satisfies the query. Since a search engine ranks results by maximizing searcher’s satisfaction with the overall result ordering, Google’s ranking of question pages was used as the baseline.

- **Direct approach:** This method implements the logistic regression approach described in Section 3.1.

- **Composite approach:** This method implements the composite approach described in Section 3.1.

- **Composite upper-bound:** This method trained the composite approach with the intermediate predictions for the query clarity and query-question match subtasks replaced with their human ratings. Since human judgments are expected to be more reliable than the automatic predictions, this method serves as an upper bound for the possible performance of the fully-automatic composite approach.

Evaluation Metrics and Setup

The main prediction task estimates searcher satisfaction for a query with one given answer, independently of other query-answer pairs. Hence the main evaluation of the

direct and composite approaches over all pairs is via root mean squared error (RMSE) and Pearson correlation between predictions and the human judgments. Both RMSE and Pearson correlation are standard performance estimators for regression problems.

When comparing the results of the proposed approaches to the Google-derived baseline described above, however, the above metrics are not applicable, since Google does not divulge an independent score for each query-answer pair. Therefore, two different metrics for ranking are used: (1) Kendall’s tau (τ) correlation that has often been used in IR [94] to compare two ranked lists, and (2) the popular NDCG metric often used in IR evaluation [46]. The MTurk ratings described above are used as ground truth to calculate these metrics.

For training and testing, stratified 10-fold cross-validation was used. This guarantees that the data distribution in each fold is similar to that of the entire dataset.

3.3 Empirical Evaluation

This section first presents the results for the main task of predicting searcher satisfaction, and compares the direct and the composite approaches. Then, the performance of the proposed methods was analyzed to identify key factors that affect the prediction accuracy. Finally, it presents the results of applying the proposed models to re-rank CQA answers in search engine results, showing significant improvements of

the ranking results over Google’s.

Direct vs. Composite Comparison

Table 3.4 shows the results on predicting searcher satisfaction using the proposed direct and composite approaches. The mean (\pm standard deviation) RMSE and Pearson correlation are reported over the ten cross-validation folds.

In the first two rows of Table 3.4, we see that the composite approach performs better than the direct approach on both correlation and RMSE metrics. This difference is statistically significant according to the Wilcoxon two-sided signed ranks test at $p = 0.01$ [111]. This observation is quite intuitive, since the composite approach takes advantage of additional knowledge, which is learned from the human labels for the query clarity and query-question match sub-tasks.

Now consider the last row in Table 3.4, which reports the upper bound performance of the composite approach. To estimate the upper bound, the individual regressors trained for the query clarity and query-question match sub-tasks are replaced with the actual (average) human scores for those tasks, and these scores are plugged as features into the composite regressor. Evidently, the performance of the composite method can be improved dramatically if its components, namely, the query clarity and the query-question match predictors, are improved. This flexibility of the composite approach constitutes a substantial advantage over the simpler direct

Method	Correlation	RMSE
Direct	0.608±0.042	0.222±0.009
Composite	0.618±0.054	0.217±0.011
Upper-bound	0.773±0.029	0.178±0.010

Table 3.4: Regression results on searcher satisfaction.

Task	Correlation	RMSE
query clarity	0.713±0.028	0.151±0.005
question match	0.702±0.043	0.218±0.014
answer quality	0.213±0.057	0.478±0.015

Table 3.5: Regression results on individual sub-tasks.

approach.

Analysis and Discussion

Table 3.5 details the performance of the individual regressors that were combined in the composite approach. Here the query-question match regressor also uses the query clarity prediction as a feature, as explained in Section 3.1. The answer quality regressor is trained using the asker satisfaction ground truth [68] (An asker is considered satisfied iff he selected the best answer and gave at least 3 “stars” for the quality). Again, the mean (\pm standard deviation) RMSE and Pearson correlation are reported over the ten folds.

By analyzing the predictions for searcher satisfaction by the composite regressor, we see that it can predict accurately both when the searcher is satisfied and not

No	Query, Question, Answer	Prediction (ground truth) of Query clarity, Query-question match, Searcher satisfaction
E1	Query: mexican halloween Question: Is dressing up like a mexican on cinco de mayo or halloween degrading to mexicans? Answer: Not really	1.74 (1.96) 2.04 (1.86) 2.70 (2.71)
E2	Query: how to stop loving someone you shouldn't Question: How do you stop loving someone you're really not suppose to love? Answer: ...Keep your mind occupy with work or school. whatever u can to keep your mind busy...	1.09 (1.14) 1.20 (1.0) 1.16 (1.14)
E3	Query: dtunes source Question: Ipod touch jailbreak dtunes and installous? Answer: Installous is currently incompatible with the safari download plugin, which is required for dTunes to work...	1.94 (2.03) 2.11 (2.0) 2.01 (1.0)
E4	Query: catsup vs ketchup Question: The condiment "KETCHUP" where did the name come from? Answer: The most popular theory is that the word ketchup was derived from "koe-chiap" or "ke-tsiap" in the Amoy dialect of China...	1.67 (1.29) 1.85 (2.43) 2.17 (1.14)
E5	Query: how much does it cost to send a letter to canada Question: How much does it cost to send a letter to canada? Answer: Go to your local post office and ask them.	1.22 (1.26) 1.24 (1.4) 1.89 (3.0)

Table 3.6: Sample (query, question, answer) tuples, with predictions and ground truth labels.

satisfied. A number of actual examples are shown in Table 3.6. In the first example (E1), the proposed method is able to correctly detect that the query is a little vague, the query-question match is low, and the answer is too simple to convince the searcher. On the other hand, in the second example (E2), the query is quite clear and matches the question well, and the answer provides helpful advice to the searcher. Again, the proposed method successfully predicts the overall searcher satisfaction with the answer, as well as individual sub-task scores (query clarity and query-question match).

To better understand the effectiveness of the proposed methods, error analysis was performed on the 30 cases where the difference between the prediction and the target was larger than 1. Two cases were found, E3 and E4 (Table 3.6), for which the system predicted lower searcher satisfaction than the ground truth. In the other cases, our system predicted higher than actual satisfaction—average prediction of 1.66 versus average ground truth of 2.65. The main reason for these large differences lies in the answer quality. In fact, more than half of the answers are not helpful at all (e.g., E5); other answers show negative opinions towards the askers, or contain ads. Thus, the error analysis confirms the importance of answer quality to searcher satisfaction, and also poses the challenge of more intelligent prediction of answer quality.

	Query-question match		Searcher satisfaction	
	τ	NDCG	τ	NDCG
Google	0.359	0.939	0.307	0.905
Direct	–	–	0.434(+41%)	0.928 (+2.5%)
Composite	0.301	0.919	0.437 (+42%)	0.928 (+2.5%)

Table 3.7: Mean Kendall’s τ and NDCG results on ranking questions and answers for queries.

Answer Ranking for Queries

One possible application of predicting searcher satisfaction is using this prediction for ranking CQA pages in Web search. To this end, Table 3.7 compares the quality of ranking produced by the proposed methods to that of Google’s ranking of results retrieved from the Yahoo! Answers site. Since the comparison is between the entire ranked lists of results returned by the proposed methods and by Google to the ranking induced by human (MTurk) labels, therefore, different metrics are used, namely, NDCG and Kendall’s τ . The prediction of searcher satisfaction by the proposed methods results in improvements over Google’s on both metrics. All improvements are statistically significant according to the Wilcoxon double-sided signed ranks test at $p = 0.01$ [111]. Interestingly, Google’s ranking of the questions (as opposed to answers) for a query is superior, which is to be expected due to additional information Google maybe used for ranking the questions (such as link structure and clicks), whereas this thesis focuses on predicting searcher satisfaction with the answers, where

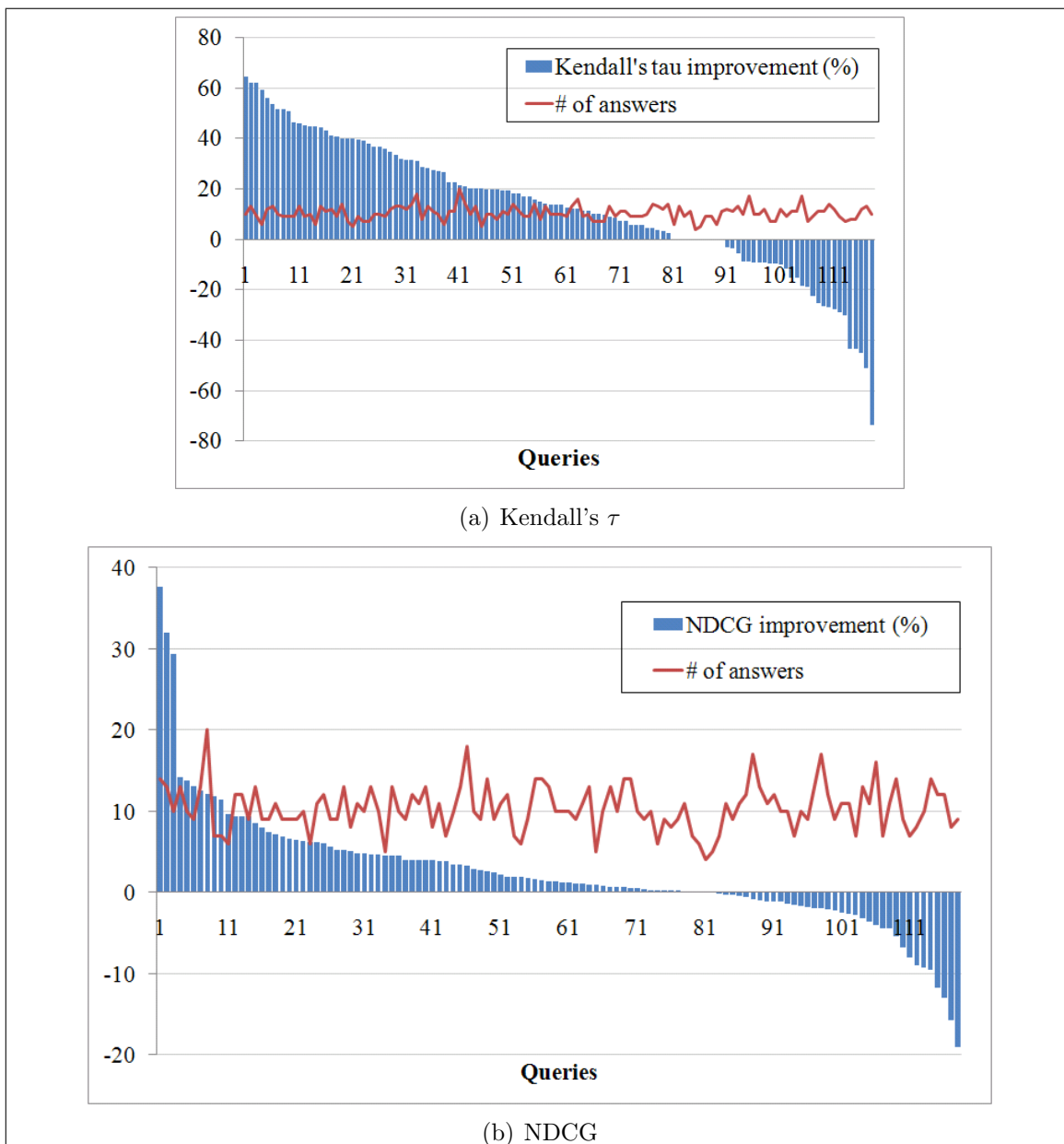


Figure 3.3: Kendall's τ (a) and NDCG (b) relative improvements of the composite approach over Google's baseline on ranking answers for queries.

the proposed methods perform better.

This thesis further analyzes these results by plotting the improvements over the Google baseline, for individual queries (Figure 3.3). Interestingly, it appears that the results do not depend on the number of answers to re-rank. In another experiment, it is found that the improvements are not correlated with query length. These results suggest that the proposed methods are robust for a wide range of queries, and are likely to remain stable for other conditions. In summary, the results show that the proposed satisfaction prediction allows the re-ranker to consistently outperform the state-of-the-art Google baseline, and could provide valuable input for other tasks.

3.4 Summary

This chapter formulated a novel task of predicting searcher satisfaction with answer pages from CQA sites. The task of predicting searcher satisfaction was decomposed into three sub-tasks, namely, predicting query clarity, query-question match, and answer quality. Two methods were formulated for solving the main prediction task. The direct method simply uses all the available features in a single regression model. The composite method first learns three separate regressors for each of the three sub-tasks, and then uses their predictions as features for solving the main task. Predictably, the performance of the composite method is statistically significantly

superior to that of the direct method. This can be explained due to its use of additional exogenous knowledge, which is learned from the human labels for each of the sub-tasks while training the three individual regressors. Furthermore, the composite approach is more flexible, and it can immediately benefit as the predictions in individual sub-tasks are improved. Indeed, when the predictions in each sub-task is replaced with actual human labels, the performance of the composite regressor is dramatically improved. An accurate predictor of searcher satisfaction can be used for improving ranking of CQA results in Web search. As demonstrated, the proposed methods produced better ranking of CQA answers than Google's search engine. To the best of my knowledge, this is the first attempt to predict and validate the usefulness of CQA archives for external searchers, rather than for the original askers. The results suggest promising directions for improving and exploiting CQA services in pursuit of satisfying even more Web search queries.

Chapter 4

Understanding When Searchers Become Askers

Chapter 3 observed that CQA pages could satisfy many information needs of web searchers and proposed the task and methods for predicting web searcher satisfaction with existing answers in CQA. This chapter focuses on analyzing the usefulness of CQA services for web searchers when they are not satisfied with the search results. The bulk of this work earlier appeared in SIGIR'12 [64].

While web search engines have significantly progressed in effectiveness and efficiency over the last decade, there still exist certain user needs that cannot be satisfied. This could be due to a number of reasons, such as the difficulty of expressing a complex need as a short search query, the lack of existing relevant content on the web, and for more “social” needs, for which the user prefers to interact with a real human. Actually, many such unsatisfied searches could be addressed by asking people via CQA services. For example, this thesis has observed that about 2% of web

search sessions performed by users who are also members of the Yahoo! Answers community, lead to a question posted to the community. Consider Figure 4.1(a), which depicts a sample search submitted to a major search engine. The searcher is not satisfied with the results, and eventually posts a related question on the Yahoo! Answers site, which is then answered to the searcher's satisfaction (Figure 4.1(b)).

To better understand the behavior of these users, as well as characterize the types of web searches that could be effectively handled by CQA sites, this thesis starts the analysis with web search sessions, traces the searcher through her visit to a CQA site, and analyzes the resulting questions posted for the community. More specifically, the study is organized around the following three research questions, each associated with a set of hypotheses:

Research Question 1: When do searchers turn to CQA for answers?

Hypothesis 1: Queries and information needs of search sessions that lead to posting questions are hypothesized to share common characteristics, and differ from general web searches in words and information needs (Section 4.2.1).

Hypothesis 2: Searchers who switch to CQA are hypothesized to exhibit common search behavior. For instance, they tend to click more on CQA results on the search result page, and their search sessions are longer, allowing to characterize different

how many bottles to buy for a newborn

5,120,000 results

WEB
IMAGES
VIDEO
SHOPPING
APPS
BLOGS
MORE ▾

Buy Similac® Bottles Sponsored Results

Similac Products are Available to Order at the Abbott® Store Online.
abbottstore.com/Similac

How many bottles should I buy my newborn when she arrives?
 [Oct 14, 2009] Best Answer: You should have 4-5 regular sized 8oz **bottles**, and maybe 3 of the 4oz. Congratulations! ;) I did it twice and worth it both times... ~ by Mrs T - Ellie's mummy! (7 comments)
answers.yahoo.com/question/index?qid=20091014135755AAxF19b - [Cached](#)
[More results from answers.yahoo.com »](#)

How many bottles should I buy for my newborn?
 [Mar 23, 2011] Best Answer: I would just **buy** the big **bottles**, because they will only use 3 ounces for a very short time. Some might use them longer than others, but you ... ~ by Moon M (12 comments)
answers.yahoo.com/question/index?qid=20110323104145AAD7Dwt - [Cached](#)

(a)

How many bottles should I purchase for my new baby? And what brand is best?

I am 9 mo. pregnant and still need to buy bottles. I will be trying to breast feed but I am unsure of how many bottles and what sizes I should buy. Is there anything else I will need for feeding and what brand do you recommend? Thanks!

Best Answer - Chosen by Voters

For The Brands you should ask friends or family members to what they thought was good...and if one brand was mentioned more...then I would get that.

For the size....they have different stages, the first 3 months will be smaller type bottles...and then they get a bit bigger.... just read the label and buy the 0-3 months size for now.

I personally had three.... I washed them Immediately after use... you could buy a little more if you feel like it.... I use the gerber brand, just because my baby did not want to

(b)

Figure 4.1: Example search (a) followed by a question posted by the same user on the Yahoo! Answers site with a satisfactory answer from the community (b).

types of users in the same spirit as [14] (Section 4.2.2).

Research Question 2: How do search queries relate to the associated questions posted on CQA sites?

Hypothesis 3: Queries and questions follow different word distributions. More specifically, words in queries are hypothesized to follow different distributions than those appearing in questions, and a clear vocabulary gap between these can be observed (Section 4.3).

Hypothesis 4: Questions are typically more specific than queries and include additional context (e.g., personal background) absent from the original queries. These differences are hypothesized to be reflected in the lexicographic differences between questions and queries, such as occurrence of personal pronouns or sentiment indicators (Section 4.3).

Research Question 3: How do searchers behave after transferring to the CQA site?

Hypothesis 5: The content and the topics of the questions posted after a search session are hypothesized to differ substantially from the general question distribution (Section 4.4).

Hypothesis 6: The question sessions after switching from searching are hypothesized to exhibit different characteristics than general question sessions and search sessions explored in depth in previous research work [52][23]. This thesis studies these different types of sessions in terms of duration and persistence for specific users and examines their behavior over time (Section 4.4).

The rest of this chapter is dedicated to answering the above questions and verifying the associated hypotheses.

4.1 Dataset Preparation

In order to understand how searchers become askers, this thesis collected a dataset that contains both the search session part and asking session part of each user who conducted a search session that resulted in posting a question. The dataset is derived from joining a sample of the query logs of the Yahoo! search engine and the Yahoo! Answers question logs, both for June 2011. To create this dataset, user actions were extracted from the query and question logs, e.g., posting queries and clicking on results from query logs, as well as posting questions and re-viewing them from the question logs. Then search sessions were constructed from these extracted actions, with a 30 minutes timeout as a session boundary. Question sessions have no temporal boundary, since every action in the session unambiguously refers to the question

posted by the asker.

After obtaining search sessions and question sessions, and mapping between some of them, two datasets were created. The first, termed *SearchAsk* dataset, contains search sessions that turned into question sessions. Only sessions that resulted in posting one and only one question were kept for simplicity of analysis later on. In addition, only sessions in which the posted question is “relevant” to a previously issued query were kept (if the query and the question share at least one non-stopword, they were considered relevant). The second dataset, termed *SearchOnly* dataset, consists of search sessions that did not turn into question sessions. In both datasets, only sessions for users that posted at least once in Yahoo! Answers were kept, since these users are aware of the site and know how to post a question there, thus removing the potential investment of effort for newcomers to join the site, and filtering out users that simply do not know where to ask questions.

Finally, among the 1,287,238 total sampled search sessions, 95.8% of them (1,233,279) are SearchOnly sessions, while SearchAsk sessions account for 1.65% (21231). Despite the sparsity of the SearchAsk sessions, understanding how searchers become askers in such sessions can still be helpful for improving the search experience of these users and perhaps more users.

A privacy-preserving subset of the data is publicly available through Yahoo’s

Webscope program (<http://webscope.sandbox.yahoo.com/catalog.php?datatype=l>, Dataset L21 - Yahoo! Answers Query To Questions).

4.2 Query and Behavior Analysis

4.2.1 Characteristics of Queries leading to Questions

The first interesting question is which queries are more likely to be unsuccessful for automated search, but instead are more amenable to be answered by a CQA site. To get such queries, this thesis examines each SearchAsk session, and extracts the queries that are issued before the question is posted, and are relevant to the question. Such queries are named *SearchAsk queries*. For comparison, it also extracts the queries in each SearchOnly session which are called *SearchOnly queries*. The following of the section explores how SearchAsk queries are different from SearchOnly queries in terms of length, words, frequency, and results.

Query Length Distribution Figure 4.2 compares the distribution of query length (in terms of number of words in the query) for the SearchAsk and SearchOnly queries. We can see that SearchAsk queries tend to be longer than SearchOnly queries, as 85% of the SearchOnly queries contain at most 5 words, while about 50% of the SearchAsk queries contain more than 5 words. Therefore, searchers issuing longer queries are more likely to turn to Yahoo! Answers to post a relevant question.

	SearchAsk queries	SearchOnly queries
Avg # of words	6.5	3.4
Avg # of stopwords	2.4	0.72
Avg % of stopwords	28%	11%
Avg word length	5.1	6.0

Table 4.1: Statistics of words per query

Table 4.1 compares the average word length per query and the average number of stopwords for the SearchAsk queries and SearchOnly queries. We can see that, on average, queries turning to questions tend to contain more words (but shorter words) than queries that do not turn to questions. The main reason could be that SearchAsk queries contain more stopwords (which are often short) than SearchOnly queries. Indeed, the percentage of stopwords in SearchAsk queries is over 2.5 times higher than in SearchOnly queries.

Query Words Distribution To better understand the difference between the content of SearchAsk queries and SearchOnly queries, this thesis compares their word distributions and shows the main difference in Table 4.2. We can see that SearchOnly queries are more likely to be navigational, e.g., to reach websites like Facebook or YouTube, or to find information related to the searcher’s common tasks such as looking up the weather, hunting for coupons, or finding a cooking recipe. In contrast, SearchAsk queries are more likely to start with question words (e.g., ‘how’, ‘what’), and tend to use more verbose natural language to express the needs of the

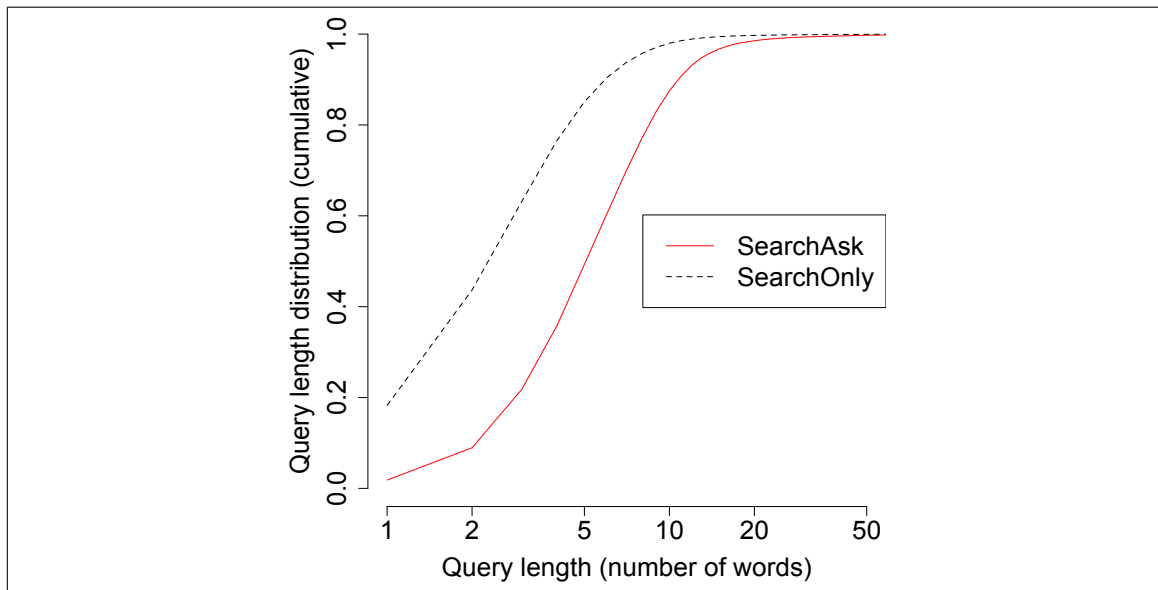


Figure 4.2: Distribution of query length

searchers (e.g., ‘want’, ‘to’, ‘know’) rather than using only keywords.

Query Frequency Distribution To verify the hypothesis from the above word distribution analysis that SearchAsk queries are more likely to be unique, this thesis computes the frequency of SearchAsk queries and SearchOnly queries in the 1-month query log. Note that the frequency of a query is computed as the number of search sessions containing the query. Figure 4.3 shows the results. We can see that over 90% of SearchAsk queries are tail (actually unique) queries, indicating the variety of the needs of searchers and the ways to express them. In contrast, SearchOnly queries contain more popular queries, e.g., around 20% of SearchOnly queries occur in more than 100 search sessions.

More likely in SearchAsk queries	
Words	to, a, be, i, how, do, my, can, what, on, in, the, for, have, get, with, you, if, yahoo, it
First words	how, what, can, be, why, i, do, my, where, yahoo, if, when, 0000, a, will, 00, best, who, which, should
Content words	yahoo, 00, use, 0, work, song, old, help, make, need, like, change, year, good, long, mail, answer, email, want, know
More likely in SearchOnly queries	
Words	facebook, youtube, google, lyric, craigslist, free, online, new, bank, game, map, ebay, county, porn, tube, coupon, recipe, home, city, park
First words	facebook, youtube, google, craigslist, ebay, the, you, gmail, casey, walmart, amazon, *rnr, justin, facebook.com, mapquest, netflix, face, fb, selena, home
Content words	facebook, youtube, google, craigslist, lyric, free, bank, map, ebay, online, county, porn, tube, coupon, recipe, anthony, weather, login, park, ca

Table 4.2: Frequent words in SearchAsk queries and SearchOnly queries

Query Results Distribution To better understand user needs behind SearchAsk queries, this thesis further examines the results returned in their search engine result pages (SERPs). A significant difference was found between SearchAsk and SearchOnly queries based on whether a SERP contains a Yahoo! Answers question page. As shown in Figure 4.4, a Yahoo! Answers question page occurs in the SERPs for half of the queries that eventually turn to questions, but for only 13% of SearchOnly queries. It is clear that SearchAsk queries are more likely to have a Yahoo! Answers question page in the SERP. This is not surprising. First, having a Yahoo! Answers question page in search results indicates that the query could be relevant to an existing Yahoo! Answers question. Therefore, answers from a hu-

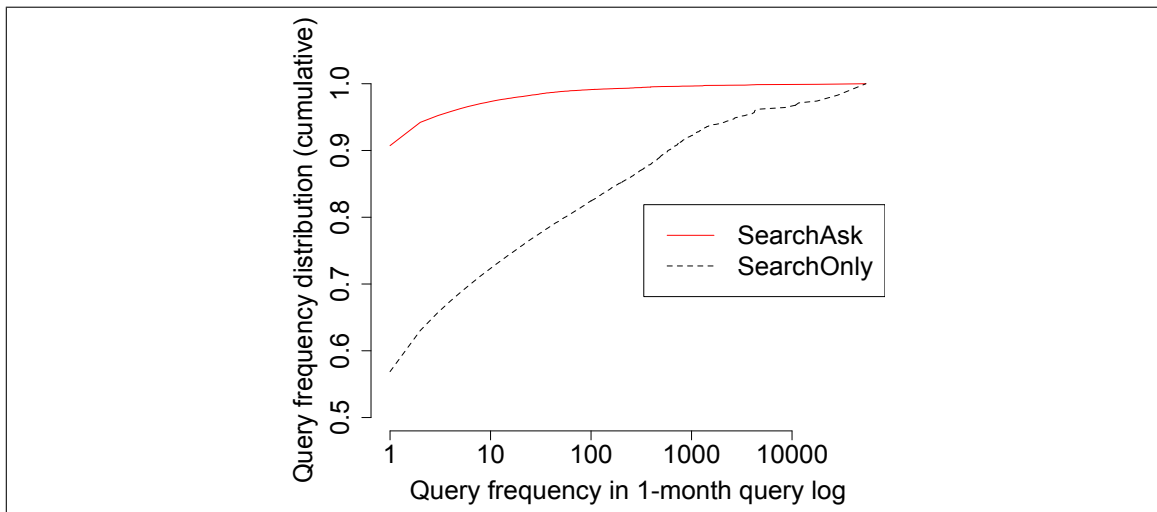


Figure 4.3: Distribution of query frequency

man might be more suitable to address the need behind the query, encouraging the searcher to post a question on Yahoo! Answers. Second, more impressions often leads to more clicks. After landing on the Yahoo! Answers site, the searcher might realize that a community might be able to answer her information need, and try posting a question.

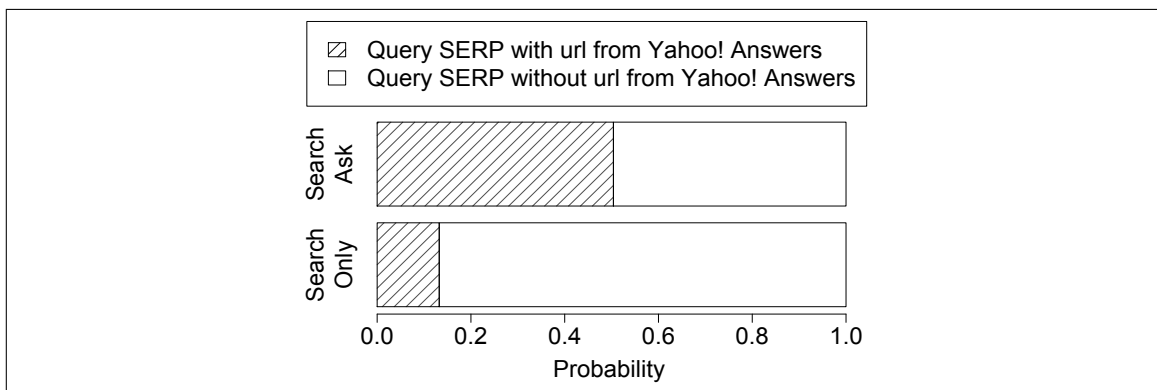


Figure 4.4: Distribution of query results

Summary of Query Characteristics As a summary of the above analysis, queries that are more likely to fail in search and lead to a question post on Yahoo! Answers tend to be longer, and use more verbose natural language to express the searchers' needs. The needs behind such queries tend to be more unique and complex than those associated with SearchOnly queries.

4.2.2 Searcher Behavior Before Asking Questions

To understand how searchers become askers, this thesis analyzes the searcher behavior in search sessions, with an associated question posted by the same user on Yahoo! Answers.

Last Action Before Question Asking First, this thesis examines what searchers do right before they start question asking, i.e., the last user action prior to a question being posted. This thesis found that the last search action before question asking is a click on Yahoo! Answers question result in 47.8% of the sessions, a click on other result in 31.2% of the sessions, and a query in 17.4% of the sessions. In about half of the sessions, the searcher posts a question right after clicking on a Yahoo! Answers question page from the search engine results. There may be several reasons for this. First, such a click indicates that the query is relevant to the clicked question, and therefore it probably carries an information need that would benefit

from a human response. Second, when the clicked Yahoo! Answers question page cannot satisfy the search need, it encourages the user to post a new question on Yahoo! Answers. Of course, it is also possible that a searcher had already decided to post a question when seeing the original SERP, and she then clicked on a Yahoo! Answers question result simply to navigate to the Yahoo! Answers site.

Distribution of Clicks To better understand the effects of clicking on a Yahoo! Answers question result on the transformation of searchers into askers, this thesis computes and compares the likelihood of such clicks in SearchAsk and SearchOnly sessions. Figure 4.5 shows the results. First, 21% of SearchOnly sessions and 81% of SearchAsk sessions contain a Yahoo! Answers question page in the SERPs. Next, after seeing a Yahoo! Answers question page in the SERPs, 81% of the searchers who turned to askers had clicked on a Yahoo! Answers question result while 19% of them hadn't; in contrast, 43% of the searchers in SearchOnly sessions seeing a Yahoo! Answers question result clicked on it while 57% of them didn't. Therefore, users in SearchAsk sessions are about twice as likely as in SearchOnly sessions to click on a Yahoo! Answers question page in the search results once seeing it. This indicates that searchers are more likely to post a question once clicking on a Yahoo! Answers question result.

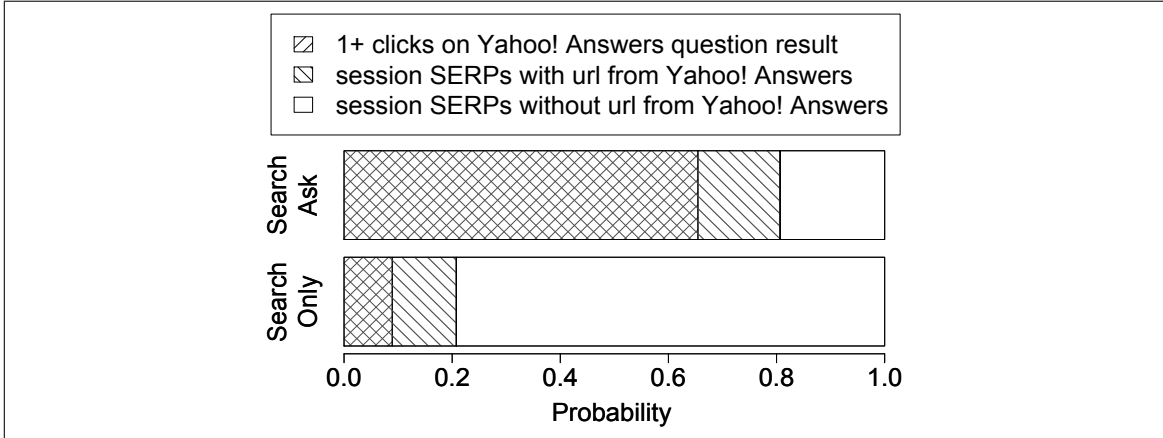


Figure 4.5: Distribution of clicks

Transitions between Actions To better understand searcher actions, this thesis further computes the probability of transitions between actions in SearchAsk and SearchOnly sessions respectively, and compares them in Figure 4.6. The transition probability between two actions a_i and a_j in SearchAsk (SearchOnly) sessions is computed using Maximum Likelihood estimation: $P(a_i, a_j) = N_{a_i, a_j} / N_{a_i}$, where N_{a_i, a_j} is the number of transitions from action a_i to action a_j in all SearchAsk (SearchOnly) sessions, and $N_{a_i} = \sum_{a_k} N_{a_i, a_k}$. SearchAsk transition probabilities are shown in red before the slash symbol, while SearchOnly transition probabilities are shown in black after the slash symbol. If we look at the transitions for SearchOnly sessions from the figure, we can see that after issuing a query, the searcher is very likely to click on other result, then with perhaps more queries and clicks on other result, and then ends the session. Clicking on a Yahoo! Answers question result is very unlikely.

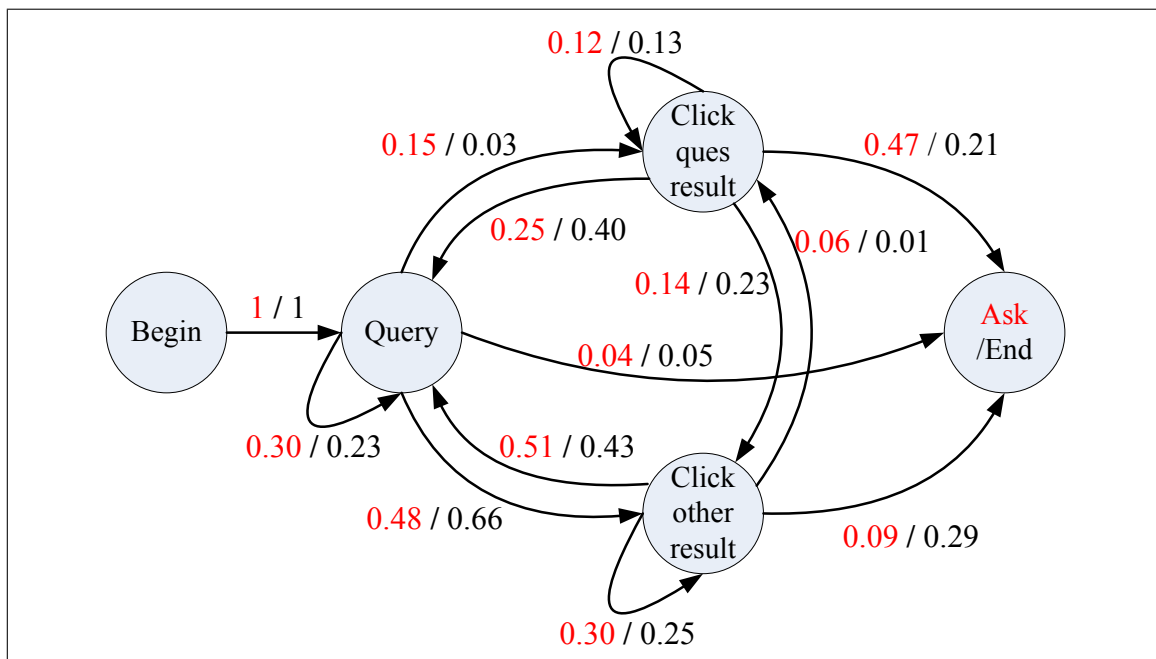


Figure 4.6: Transition probabilities for actions in SearchAsk (in red, before the slash symbol) and SearchOnly (in black, after the slash symbol) sessions. Note that two other actions (Pagination and Click interface) are ignored for simplicity.

However, in SearchAsk sessions, the searcher has a higher probability on clicking a Yahoo! Answers question result. After the click, the searcher in SearchAsk sessions would post a question on Yahoo! Answers for around half of the time.

Action Sequences Before Question Asking To better understand how searchers become askers, this thesis examines the user action sequences in SearchAsk sessions before the question post, and compares them with action sequences in SearchOnly sessions. Table 4.3 shows a sample of top frequent user action sequences. The top frequent path in SearchOnly sessions indicates navigational needs of the searchers,

SearchAsk sessions	Distribution
B Q C_{qr} A	10%
B Q C_{or} A	3.8%
B Q Q C_{qr} A	3.3%
B Q A	2.8%
B Q C_{or} Q C_{qr} A	2.0%
SearchOnly sessions	Distribution
B Q C_{or} E	30.2%
B Q C_{or} Q C_{or} E	7.1%
B Q E	6.1%
B Q C_{or} C_{or} E	3.6%
B Q Q C_{or} E	3.6%

Table 4.3: Top frequent user action sequences in SearchAsk sessions and SearchOnly sessions (B: Begin a session, Q: Query, C_{qr} : Click on a Yahoo! Answers question result, C_{or} : Click on other result, A: Ask a question, E: End a session)

i.e., they issue a query, click on a search result and leave the session. Such navigational cases account for 30% of total SearchOnly sessions. In contrast, the top frequent path in SearchAsk sessions indicates more “social” needs of the searchers, i.e., they issue a query, click on a search result of Yahoo! Answers question page, and then ask a question on Yahoo! Answers. Yet, the path distribution is more balanced for SearchAsk sessions. Moreover, clicks on Yahoo! Answers question results are common in the paths.

Session Size Distribution Finally, this thesis compares the distribution of session sizes for SearchAsk and SearchOnly sessions. Session size can be measured in several ways, e.g., by the number of (unique) queries issued by the searcher in the session,

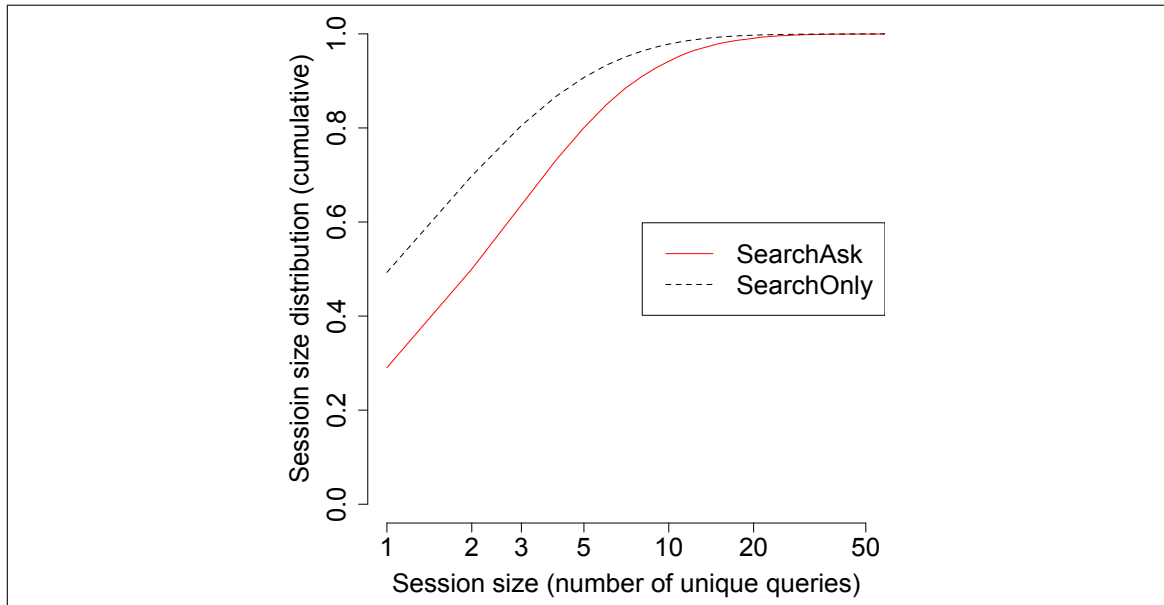


Figure 4.7: Distribution of session size

by the number of actions performed in the session, or by the duration that the session lasts. The first option is used in this analysis. The results are shown in Figure 4.7. While only one query is issued in the half of SearchOnly sessions, at least three different queries are issued in the half of SearchAsk sessions. The average session size is 2.5 for SearchOnly sessions and 3.8 for SearchAsk sessions. This shows that searchers tend to issue more queries in SearchAsk sessions, possibly because SearchOnly sessions contain more navigational needs, while SearchAsk sessions are associated with more difficult or complex needs, and thus require more effort in finding answers.

	Median	Avg	Max
question – query	42	66	1431
subject – query	3	4	27
content – query	31	55	1428

Table 4.4: Statistics of length difference between a query and its associated question (number of words).

4.3 Queries vs. Questions

After discovering the unique attributes of queries that lead to asking a question, this thesis next wants to understand better the process of turning a query into a question posted on Yahoo! Answers.

The most expected difference between queries and questions is their length. Table 4.4 shows these differences. From the table we can see that a question has 66 more words than its associated query on average. This indicates two things: first, as expected, questions are much more verbose, being natural language expressions, compared to the concise queries; second, since Yahoo! Answers questions are not limited in length, additional knowledge of the problem to be solved is added. Interestingly, the subject of the question is very close in length to the query, which shows that searchers still think in search-style writing for the subject. However, the content part of the question is significantly longer, and much more information is added in this question part.

This thesis next looks at word distribution differences, since they may point at

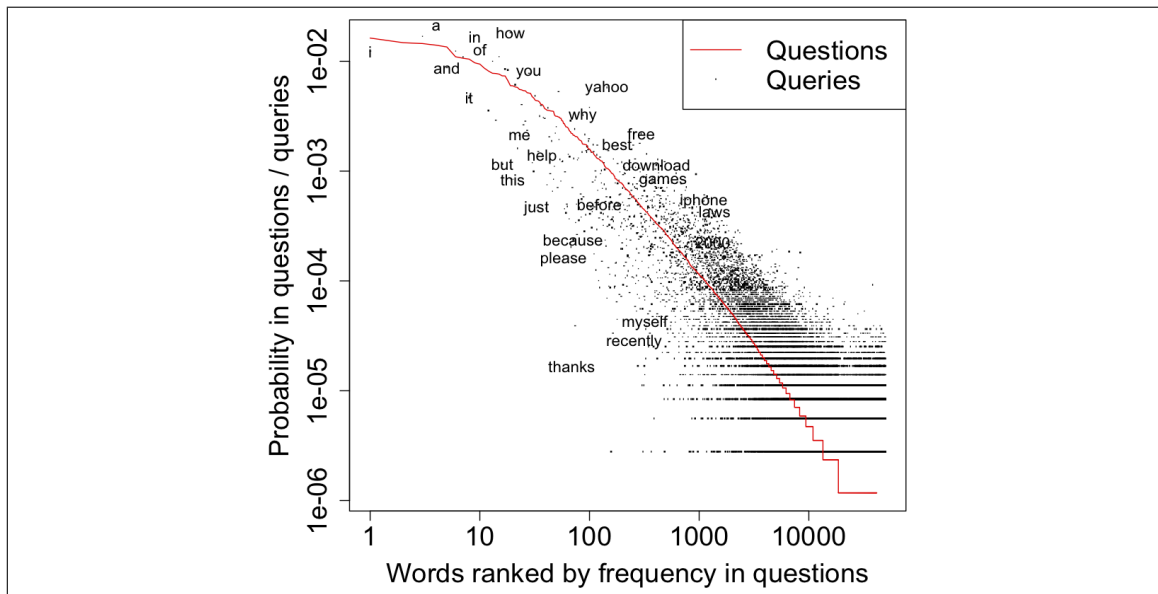


Figure 4.8: Word distributions over question words

the lexical gap between queries leading to questions and their associated posted questions. Figure 4.8 depicts the word occurrence distribution over word ranking by frequency for search-related questions. The most notable difference between the two distributions is that questions tend to be more personal and verbose, as captured by the abundant usage of the pronouns such as ‘*I*’, ‘*me*’, ‘*it*’ and ‘*this*’, connectives such as ‘*but*’, ‘*because*’, ‘*recently*’, and ‘*just*’, as well as sentiment indicators such as ‘*help*’, ‘*please*’, and ‘*thanks*’. Queries, on the other hand, tend to focus more on the things or actions that are searched for, with content words like ‘*best*’, ‘*free*’, ‘*download*’ and ‘*games*’ as well as question words like ‘*how*’, ‘*why*’ and ‘*what*’ occurring more frequently than in the associated questions corpus. Interestingly, one to four digit

	<i>?=question</i>	<i>?=subject</i>	<i>?=content</i>
$CW_? \supset CW_{query}$	31.4%	14.6%	14%
$CW_? = CW_{query}$	1.8%	6.2%	0.4%
$CW_? \subset CW_{query}$	0.7%	3.7%	17.1%
$CW_? \not\subset CW_{query}, CW_? \not\supset CW_{query}$	66.1%	75.5%	68.5%

Table 4.5: Overlap of content words (CW) between a query and its associated question.

figures, such as car model years, also appear more in question-related queries than in their associated questions, probably since they capture much of the essence of the target information need.

To further understand the semantic difference between composing a query and its related question, this thesis measured the distribution of query-question pairs in which the same words are used for both query and question, the pairs in which one is included in the other, and those pairs in which each contains words that do not occur in the other. Table 4.5 presents these statistics, while Table 4.6 and Table 4.7 provide examples of such pairs, annotated with the type of context added when switching from query to question, as been classified by [101], i.e., task, situation, attribute, limit, and thought.

Some interesting question composition patterns are evident from this analysis. First, in the majority of pairs (66%), both queries and questions contain unique words that do not occur in the other. This is somewhat surprising, since this thesis would expect more complete inclusion of the query terms in the question. However,

ID	Type of context added	Query	Question (Category, Subject, and Content)
1	N/A	what to serve with chicken salad	Food & Drink>Cooking & Recipes what can you serve with chicken salad?
2	thought	best nba players without a championship	Sports>Basketball Greatest NBA players to never win championship? Patricik ewing, reggie miller, charles barkley, karl malone? Who else?
3	task	pt cruiser ac fix	Cars & Transportation>Car Makes>Chrysler how much does it cost to fix an ac system in a pt cruiser?
4	task	solve $n^2 - 2n - 3 = 5000$	Education & Reference>Homework Help Algebra question, Need Help Pls!!!!? An owner of a key rings company found that the profit earned (in thousands of dollars) per day by selling n number of key rings is given by $n^2 - 2n - 3$, where n is the number of key rings in thousands. Find the number of key rings sold on a particular day when the total profit is \$5000. Thanx

Table 4.6: Examples showing semantics difference between the query and the associated question.

ID	Type of context added	Query	Question (Category, Subject, and Content)
5	limit	chocolate croissant menlo park	Dining Out>United States>San Jose Where can I get a good Chocolate Croissant near Menlo Park, CA? Something with thick, dark chocolate? And please, don't say La Boulanger.
6	situation, task	chicago fried chicken	Dining Out>United States>Chicago Where can I get really good fried chicken in the Lakeview area in Chicago? I really want fried chicken after watchin a special on TV. But I cant find any place near me that has decent priced chicken thats not fast food and is homemade and delicious. Any one know of a place?
7	situation, task	douglas az	Education & Reference>Higher Education (University +) Radiology schools in Arizona? Does any one know any schools in az that offer radiology degree programs, I moved to Douglas az and don't know any schools near to study radiology. If any one can help that would be great :)
8	attribute, situation, task	how many bottles to buy for a newborn	Pregnancy & Parenting>Newborn & Baby How many bottles should I purchase for my new baby? And what brand is best? I am 9 mo. pregnant and still need to buy bottles. I will be trying to breast feed but I am unsure of how many bottles and what sizes I should buy. Is there anything else I will need for feeding and what brand do you recommend? Thanks!

Table 4.7: Examples showing semantics difference between the query and question.

it seems that with the freedom of writing a free text question, searchers tend to rephrase some of the terms they used in their queries. For example, abbreviations and short terms are turned into their more complete forms, *e.g.* ‘AZ’ into ‘Arizona’ and ‘newborn’ into ‘new baby’ (see example 7 and 8 in Table 4.7). In addition, while 31% of the pairs do show complete inclusion of the query terms in the question, many times the query terms do not all appear in the question’s subject or content, but spread in both question parts. Table 4.6 and Table 4.7 show that most of the extensions of the query into a question include additional details that are related to the search task. Yet, many times details of the personal situation are added, such as the state of mind, *e.g.* “*after watching a special on TV*” (example 6 in Table 4.7).

One interesting future research is to automatically generate questions from queries [26, 121, 122]. However, adding context information to the question, such as the situation or limit is a difficult challenge. Still, expanding the query expression to an explicit question form may be possible for many cases, *e.g.* examples 1 and 3 in Table 4.6.

4.4 Question Analysis

As a final analysis, this thesis is interested in discovering the differences in lexicon, that is whether different words are used when composing a question by web searchers,

compared to typical askers.

As expected, this thesis finds that there is a large difference between the word distribution for the corpus of all questions posted in June 2011 and the distribution of the corpus of questions posted by searchers. In addition, the entropy of generating a word from the search-related question corpus is much lower, showing a more focused vocabulary. But what are the reasons for this large difference? It turned out to be mainly topical.

To measure this topical difference between the two types of questions, this thesis looked at the distribution of categories to which the questions in the two compared corpora are assigned. Table 4.8 shows the categories with largest differences in assignment probability, those that are preferred more in the general question corpus and in the search-related question corpus respectively. These lists show that searchers tend to ask informational questions [38] to get fact- or advice-oriented answers, such as how to fix the car or maintain one's garden, how to bake cookies, but also questions related to Yahoo products, such as Yahoo! Mail. On the other hand, regular askers are more likely to ask conversational questions [38] with a social flavor, such as discussions around music or sports events, politics and religions, and opinions on possible baby names.

After manually labeling 100 questions randomly sampled from the search-related

Categories more likely for general questions	Categories more likely for questions following search
Polls & Surveys (Entertainment & Music)	Maintenance & Repairs (Cars)
Singles & Dating	Law & Ethics
Religion & Spirituality	Dogs (Pets)
Politics	Pregnancy
Friends	Maintenance & Repairs (Home & Garden)
Mathematics	Renting & Real Estate
Diet & Fitness	Accounts & Passwords
Lesbian, Gay, Bisexual, and Transgendered	Other - Yahoo! Mail
Other - Beauty & Style	Military
Basketball	Problems with Service
Baby Names	Garden & Landscape
Adolescent	Cooking & Recipes

Table 4.8: Categories with largest differences in assignment probability between questions coming from search and general questions

question corpus, this thesis found none are conversational, showing a very different distribution compared to that 38% of Yahoo! Answers questions are conversational as reported in [38]. This thesis conjectures that this is because searchers usually turn to search engines to find information instead of starting conversations. Another kind of questions that are less likely searched first over the web are personal questions, in which the asker is interested in adding very personal details. These include topics such as diet and fitness advices, dating and style opinions. Finally, there are questions that are too complex, for which the asker knows the answer cannot be found on the web. A good example are Math questions, such as example 4 in Table 4.6.

To further investigate the differences between the two question types, this thesis removed the strong bias caused by the different category distributions within the

two corpora by sampling questions from the general question corpus based on the category distribution of the search-related question corpus. By comparing the word distribution between the sampled corpus and the search-related question corpus, this thesis found that hardly no topical differences remained. That is, the topical variation in the two corpora is more or less completely captured by the level of assigned categories, without more subtle topical differences evident. Still, there may be stylistic variations in question composition between searchers and typical askers. So this thesis examines the stylistic statistics for the general-sampled and search-related question corpora. The significant difference between the two corpora is the number of words per question: for the same topics, general questions contain 6% more words compared to search-related questions. Yet, interestingly, this attribute is due to more sentences that are written on average per general question, while if we look at the number of words per sentence, we see that surprisingly search-related questions have slightly more words in each sentence. This could be related to more information-focused nature of the questions posted after a search session, and suggests further investigation.

4.5 Summary

This chapter studies the unique properties of SearchAsk sessions: search sessions that turn into question composition. To the best of my knowledge, it is the first large-scale analysis of the user transition from searching to asking. What makes the work unique is the study of the explicit connection between the search query and the corresponding question from the same user for the same need. It provides insights into some specific needs that searchers try to express on search engines, yet are not satisfied by search results, and turn to human answerers instead. Various aspects of SearchAsk sessions are analyzed, including the differences between general search-engine queries and those belonging to a SearchAsk session, the transformation of a query into a natural language question and the question composition patterns, as well as other asking behavior of searchers, compared to general askers in a CQA service. The findings may contribute both to search-engine optimization, as well as to better user experience in CQA sites. Furthermore, as this chapter demonstrates, modeling the transformation of a query meant for an automated search engine into a fully specified question meant for human, provides a valuable tool for query intent and satisfaction analysis.

Chapter 5

Understanding Answerer Behavior for Better Question Recommendation

Chapter 3 and Chapter 4 explored how to improve web searcher satisfaction using CQA services. This chapter focuses on the problem of what contextual factors influence answerer behavior in CQA. In particular, it first explores the contextual factors that influence the answerer behavior in a large CQA system, and then studies the effects of the information context of the answerer at the time a question is received on the answerers' reported ability, effort, and willingness to answer questions. The goal is to inform the construction of question routing and recommendation strategies. The bulk of this work earlier appeared in ECIR'11 [62] and IIX'10 [66].

5.1 Modeling Answerer Behavior in CQA

Previous work has largely ignored a key problem in question recommendation, i.e., whether the potential answerer is likely to accept and answer the questions recommended to them in a timely manner. That is, even if the question is on a topic of past interest to the answerer, they may not have the opportunity or interest in answering the question at recommendation time.

To address this gap, this thesis explores the contextual factors that influence the answerer behavior in a large, popular CQA system, with the goal to inform the construction of real-time, online question routing and recommendation strategies that also take into account the behavior of real answerers. Specifically, the following two research questions are considered:

1. When do users tend to answer questions in a web-scale CQA system?
2. How do users tend to choose the questions to answer in CQA?

The overall approach is to analyze the answering behavior of a large group of Yahoo! Answers users, collected for more than 1 million questions over a period of one month in 2010. Specifically, for the first research question, this thesis analyzed both the overall and user-specific temporal activity patterns, identifying stable daily and weekly periodicities, as well as not previously observed bursty patterns of activity in

the individual answer sessions of many users. This observation is exploited to perform a novel session-based analysis of the answerer activity. For the second research question, this thesis analyzes the factors that may affect the users' decisions of which questions to answer. These factors include the question category (topic), the question position in the list shown to users, and the surface patterns in the question text. This thesis confirmed previous findings that users have favorite categories that attract most of their contributions, but interestingly the decisions for most users within a category are determined more by the rank position of the question in the list of available questions, than any other factors such as the text or the provenance of the question itself.

5.1.1 Temporal Patterns in Answerer Behavior

This section first describes the CQA dataset used for this analysis. It then shows the aggregate patterns of when answerers tend to answer questions, which confirms previous findings and validate our dataset construction. Then, it focuses on the novel contribution of this work, namely modeling the individual answerer activity within a single answer session. These analysis could help question recommender systems by suggesting when to begin recommending questions to a user in the first place, and how many questions to recommend to a user.

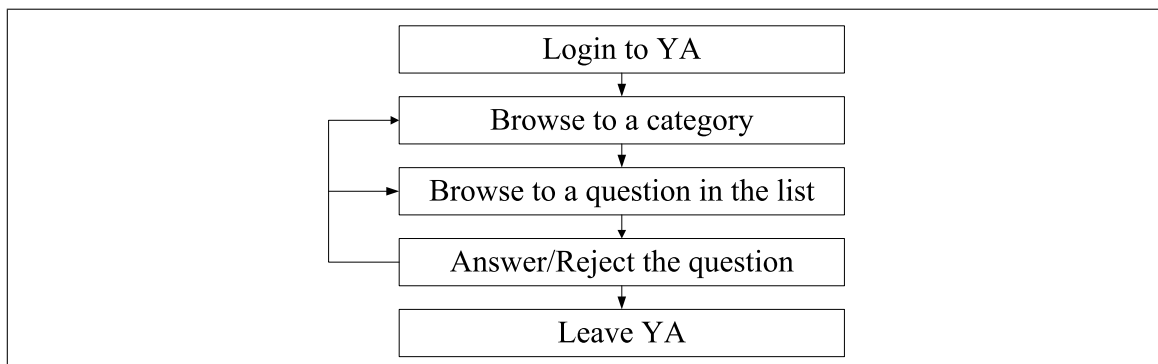


Figure 5.1: Basic question answering process in Yahoo! Answers.

The Yahoo! Answers Setting and Data

For this study the Yahoo! Answers (YA) website was chosen, as a large-scale, popular, and representative example of a CQA system. To clarify the terminology and the subsequent descriptions, the basic question answering model in YA is briefly summarized, which is representative of many other CQA sites. Figure 5.1 illustrates this process. After logging into the YA site, answerers can choose a category of interest to them to browse (including the root category “All categories”). Then they are shown a list of questions in that category among which they may answer some and skip others. This process is repeated when answerers browse to another category, until they eventually leave the site.

To construct the dataset, about 1M questions and 4.7M answers were crawled, covering the top 26 categories and 1659 leaf categories in YA, as of May 2010. Since inactive users reveal less information of their answering behavior, our analysis focuses

	Questions	Answers	Best Answers	Answerers	Askers	Users
ALL	1,056,945	4,734,589	1,056,945	433,902	466,775	726,825
USER20	933,746 (~88%)	3,319,985 (~70%)	751,633 (~71%)	45,543 (~10%)	419,395 (~90%)	437,493 (~60%)

Table 5.1: Dimension of the Yahoo! Answers dataset. The USER20 dataset focuses on answerers with at least 20 answers.

on active answerers who posted at least 20 answers during the period of time. This subset, called USER20, includes 45,543 answerers, accounting for about 10% of all answerers but 70% of all answers and best answers. Table 5.1 presents the statistics of the dataset in more detail.

Aggregate Temporal Pattern Analysis

First, this thesis analyzes the overall temporal patterns of answering activities in Yahoo! Answers with the strategy used in [35]. It bins all the answers by hours, aggregate answers in the same hours by months/weeks/days, and then normalizes the number of answers in each hour by the total number of answers. Based on the dataset, the answer activities in YA demonstrate strong monthly, weekly and daily patterns as shown in Figure 5.2. From Figure 5.2(a), we can see that the number of answers across the whole month is increasing, which indicates the growing popularity of YA. Figure 5.2(b) shows that the number of answers during the weekday is higher than that on the weekends, with Tuesdays and Wednesdays being the most active

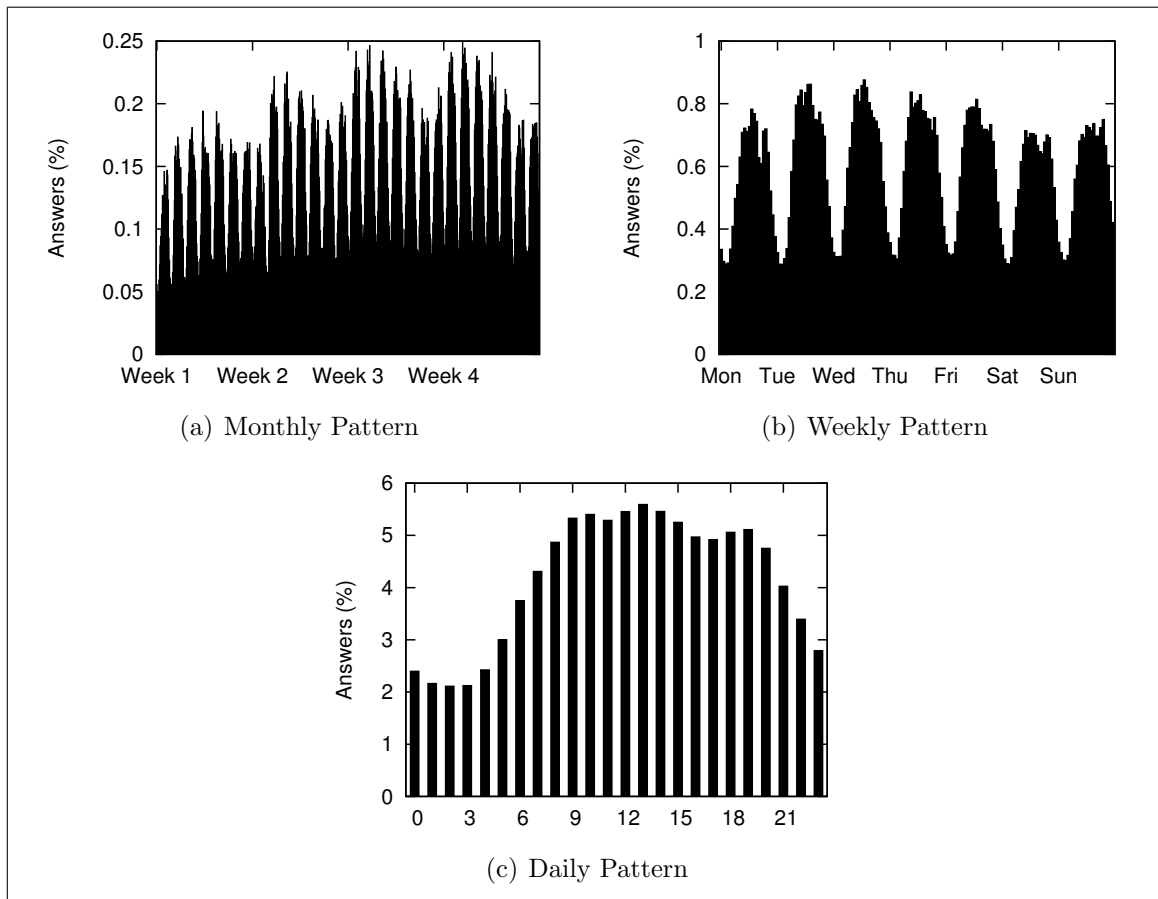


Figure 5.2: Temporal patterns of answer activities in YA, showing the percentage of answers in the same hours aggregated by (a)months; (b)weeks; (c)days.

weekdays. Based on Figure 5.2(c), there tend to be three peak times in a day for answering questions, 10:00, 13:00, and 19:00 (YA server time). The least active time for answering questions is 2:00-3:00 AM. These results are similar to those described in [35], but with a time shift in the daily pattern possibly due to a different time zone used in their study.

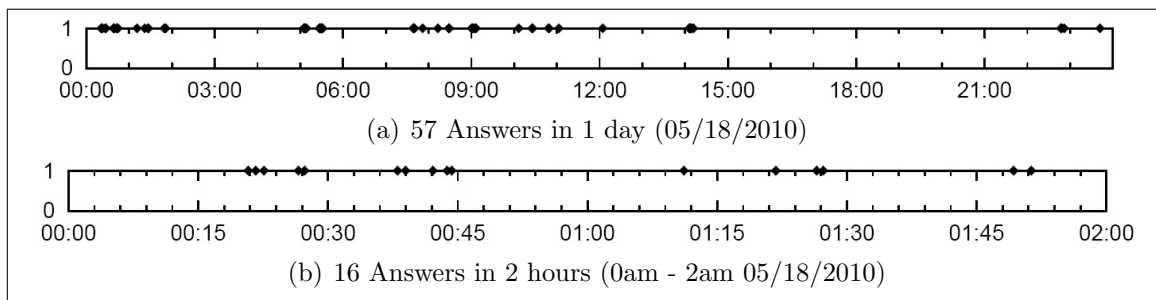


Figure 5.3: Example answering behavior for an active user over 1 day (a) and over a period of 2 hours (b).

Burstiness of Individual Answering Activities

This thesis now explores the temporal patterns of answering activity for individual users. It is found that users tend to post bursts of answers within short answering sessions, and then “disappear” for relatively long periods of inactivity. For example, Figure 5.3(a) illustrates the answer activities of an example user. The user answered 57 questions that day; however, the answering was not distributed uniformly, but was concentrated in relatively short bursts. To provide a better intuition, this thesis plots the user’s answering activities over a period of two hours, shown in Figure 5.3(b). We can see that some intervals between two successive answers are short (e.g. less than 3 minutes), but others are long (e.g. around 30 minutes), which presumably correspond to breaks between the answer sessions.

There may be two reasons for the long intervals between answers: it could be that it took the user a long time to provide the answer to a difficult question, or

that the user left Yahoo! Answers to do something else (a more likely scenario). Therefore, this thesis defines the continuous answer activities of a user as an answer session of the user. Understanding the number of questions that a user would answer continuously within a single answer session would be helpful for designing question recommender systems, e.g. how many questions to recommend to a user.

To detect answering session boundaries, this thesis adapts some of the methods proposed to determine Web search session boundaries (e.g., [33]). In the CQA setting, the time gap between the successive answers was chosen as the most intuitive metric. The distribution of intervals between two successive answers of a user is shown in Figure 5.4(a). As we can see, the frequency of intervals less than 8 hours long, forming a roughly power-law-like distribution. However, there are seven secondary peaks, corresponding to intervals of one to seven days. This thesis further zooms in to consider the intervals of one hour, shown in Figure 5.4(b), which shows that for over 70% of the cases, users post the next answer within 1 hour after the current answer.

Based on this observation, a timeout threshold is applied to detect the session boundaries. If the interval of two successive answers is larger than the threshold, they belong to different sessions, and to the same session otherwise. This thesis uses the methods in [42] to determine the optimal session timeout threshold, i.e., analyzing

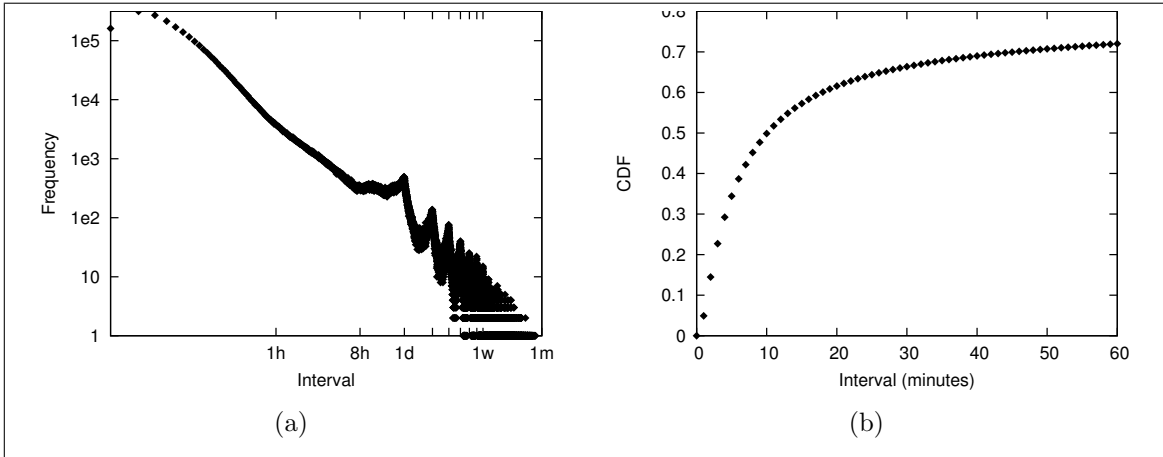


Figure 5.4: The (a)Frequency and (b)Cumulative Distribution of the intervals between two successive answers for all active users.

the effect of different session timeout thresholds on the proportions of sessions with different sizes. The proportion of 1-size sessions decreases quickly with the increase of session timeout threshold until 30 minutes, while the proportion of sessions with size 3-6 increases. After that, increasing timeout threshold has negligible impact on proportions of these sessions, especially when the session timeout threshold is larger than 40 minutes. Therefore, the session timeout threshold should be between 30 and 40 minutes.

Table 5.2 shows the session statistics computed based on the two thresholds for session detection. As we can see, the average session size is around 3 for both session threshold values. This means that users answer 3 questions in a session on average, providing guidance for designing real-time question recommender systems, e.g. three or more questions can be recommended to a user. To explore the average time that

Threshold	Session size	Session size (≥ 2)	Session duration	Answer time	Gap duration
30m	2.89±3.53	4.45±4.16	26.5m±27.7m	7.68m±6.70m	19.1h±33.8h
40m	3.13±3.86	4.69±4.48	32.2m±34.7m	8.73m±8.44m	20.6h±34.8h

Table 5.2: Answering session statistics for varying timeout values.

users spend on posting an answer, the average session duration is computed. A session duration is computed as the time between the posting of the first answer and the last answer in a session. For sessions with size $n \geq 2$ and duration d , the average answer time t can be computed as $t = \frac{d}{n-1}$. The results are shown in Table 5.2. As we can see, the average answer time appears to be about 8 minutes for both session threshold values (which also includes the time needed to choose the next question to answer).

In summary, the analysis above first focused on the answerer behavior in the aggregate (weekly and daily), and largely confirms previous findings, thus validating the data collection method. This thesis then considered session-based behavior of individual answerers, and identified a novel bursty behavior of the answerers.

5.1.2 Understanding How Answerers Choose Questions

Having analyzed when users would like to answer questions, this thesis now explores how they tend to choose the questions to answer. Based on the simplified answering process shown in Figure 5.1, it explores several factors that may affect the users'

decisions of which questions to answer, including question category (topic), the question’s rank in the list shown to users, question text, and the users’ previous answering history profile.

Question Category Effects

Browsing a category is a first step of an answering process in YA. Users can choose any category to browse, from top categories to leaf categories. If a category is not chosen explicitly, the root category “All categories” is used by default. To explore the effect of question category on users’ choices of which question to answer, this thesis first computes the category coverage of users. The category coverage of a user is the number of different categories in which the user has posted answers. The results shown in Figure 5.5(a) confirm that some users answer questions in more than one leaf categories within the same top category. Moreover, we can see that more than 90% of users post answers in less than 30 leaf categories (out of 1659 leaf categories present in our dataset).

Next, this thesis explores how focused the answers are across different categories, using the entropy measurement introduced by Adamic et al. [1]. The entropy of a user is defined as $H = -\sum_i p_i \log(p_i)$, where each i means a category covered by answers of the user and p_i means the percentage of answers of the user in that category. The results are shown in Figure 5.5(b) and 5.5(c). We can see that users

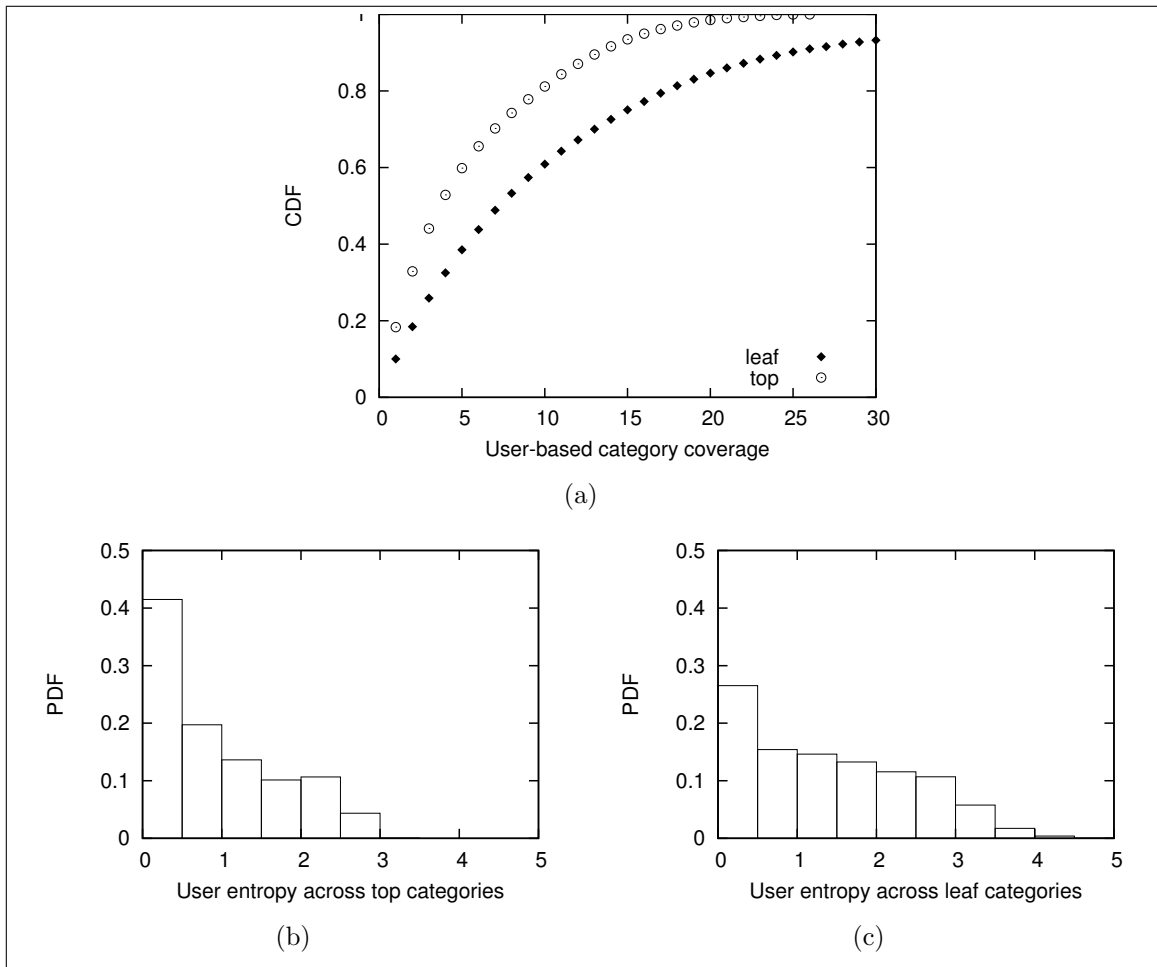


Figure 5.5: The Cumulative Distribution (CDF) of user-based category coverage, which is the number of categories in which a user has posted answers across the entire dataset duration. The hollow circles represent user-based category coverage for top categories, and solid diamonds represent the leaf categories (a); The distribution of user entropy across all top (b) and leaf (c) categories: lower entropy indicates user activity focused on fewer categories.

tend to be relatively focused to answer questions primarily on a handful of topics.

For real-time question recommender systems, it is also very important to know whether a user would like to answer questions in different categories within a single session. Therefore, this thesis also computes the session-based category coverage of users. The session-based category coverage of a user in a session is the number of different categories in which the user has posted answers in the session. The results are reported in Figure 5.6(a). As we can see, for around 70% of cases, the users post questions in just one leaf category in a single session.

To explore more about how users would change categories during his single answer session, this thesis also computes the change rate of categories, shown in Figure 5.6(b) and 5.6(c). We can see that in most cases they tend not to change throughout an answer session; however, in some cases they change at every chance. Understanding the user preference on category changes in a single session can be very helpful for improving the user experience in question recommender systems.

The above analysis shows that categories play an important role in deciding users' choices of which questions to answer, as they tend to be focused rather than diverse regarding the topics. In a single session, most users prefer to answer questions in only very few categories and to answer the next question in the same category.

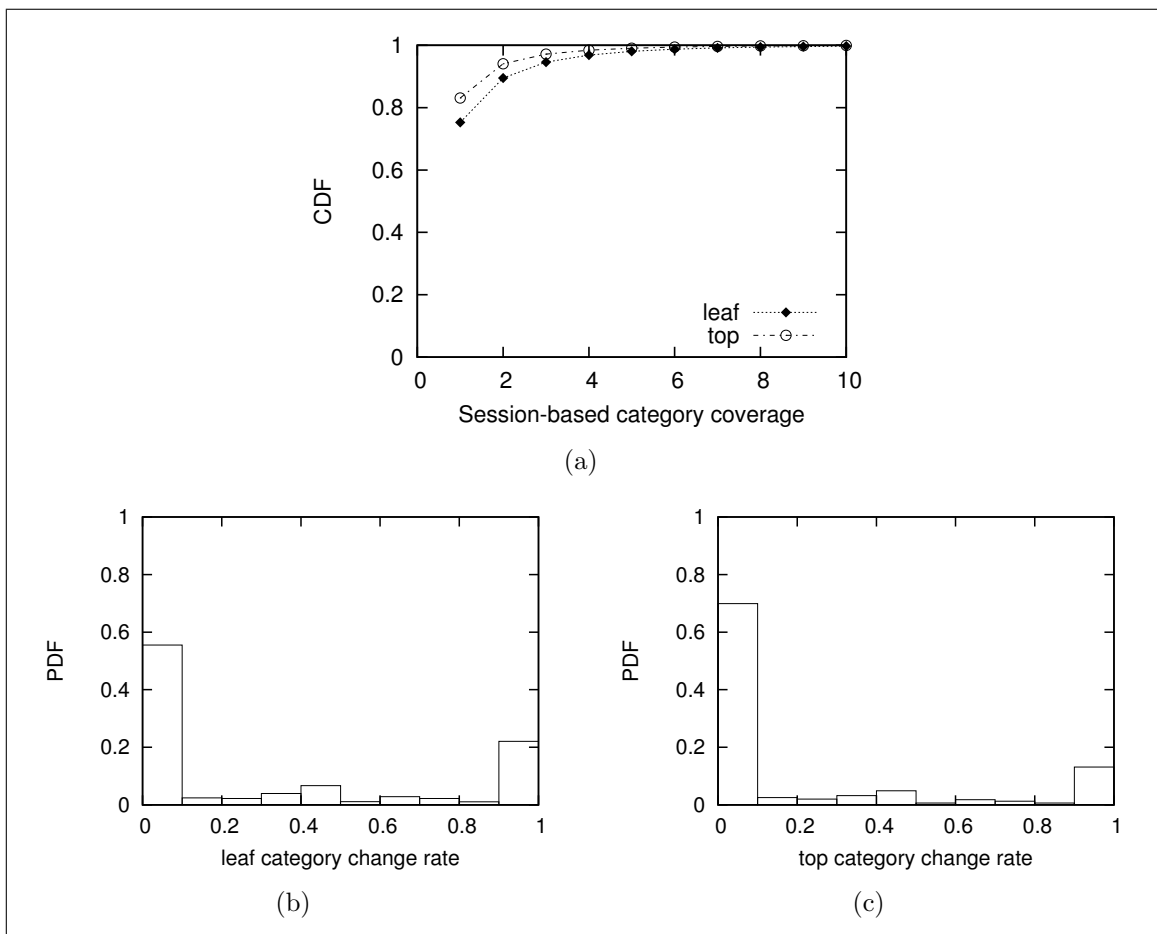


Figure 5.6: (a)The Cumulative Distribution Function of session-based category coverage, which is the number of categories in which a user has posted answers in a single answer session. The hollow circles (solid diamonds) represent session-based category coverage for top (leaf) categories. (b)(c)The Probabilistic Distribution Function of session-based category change rate for leaf(b) and top(c) categories, which is the percentage of two successive answers in different categories posted by a user in a single answer session. Note that the session timeout threshold of 30m is used here.

Question Rank Effects

According to the basic answering process shown in Figure 5.1, after choosing a category, the users will see a list of questions – by default, arranged in the order of most recent arrival. Then, the user will answer one or more questions in the list. This thesis posits that the users tend to examine the questions in order of listing and answer them in order of the examination. This examination hypothesis has extensive support from the web search result examination literature.

Therefore, this thesis proposes the following simple, yet surprisingly accurate model of answerer behavior that simply follows the order of the posted questions.

Ordered Question Examination Model (OQE): The Answerer repeatedly examines the questions in the order presented in the Category list (normally, in reverse order of arrival, most-recent first), and answers one of the top-K questions in the list - and then goes back and repeats.

To verify this OQE model, it is necessary to know the questions and their order that the answerers saw before choosing a question and posting an answer. However, it is difficult to recreate the exact list of questions that the answerers saw, so this thesis approximates the list based on the known characteristics of the YA site and the externally available data.

First, each answer is represented by a tuple $A(u_A, q_A, c_A, t_A)$, which means the

answer A is posted by user u_A for question q_A in category c_A at time t_A . Similarly, each question is represented by a tuple $Q(u_Q, c_Q, t_Q)$, which means the question Q is posted by user u_Q in category c_Q at time t_Q . Then, for each answer $A(u_A, q_A, c_A, t_A)$ of the user u_A , this thesis creates a ranked list of questions in the category c_A that are posted before the time t_A , ordered by their recentness, most recent first. More formally, the list with respect to $A(u_A, q_A, c_A, t_A)$ can be represented as

$$L_A = [Q_i(u_{Q_i}, c_{Q_i}, t_{Q_i}) \mid c_{Q_i} = c_A \wedge t_{Q_i} < t_A \wedge \forall j > i, t_{Q_j} < t_{Q_i}]$$

Note that in real scenarios, the answerers may browse any category from top to leaf. However, for simplicity, this thesis just assumes that answerers always browse to leaf categories before they answer questions. Also, it does not count in the user's time for submitting the answer A which will shift the estimated questions in L_A slightly, compared to the actual list. In addition, considering that YA shows 20 questions by default, and many answerers do not bother to click to the next page, this thesis focuses on the sublist $L_{A,20}$ containing the top 20 questions in the estimated list L_A . Then, based on the list $L_{A,20}$ for answer $A(u_A, q_A, c_A, t_A)$, the goal is to check whether the question q_A is in $L_{A,20}$. If yes, it is useful to know the rank of the hit, that is the i such that $Q_i = q_A$. This indicates that after the user u_A browsed to the category c_A , she chose to answer the question ranked at the i th ($1 \leq i \leq 20$) position

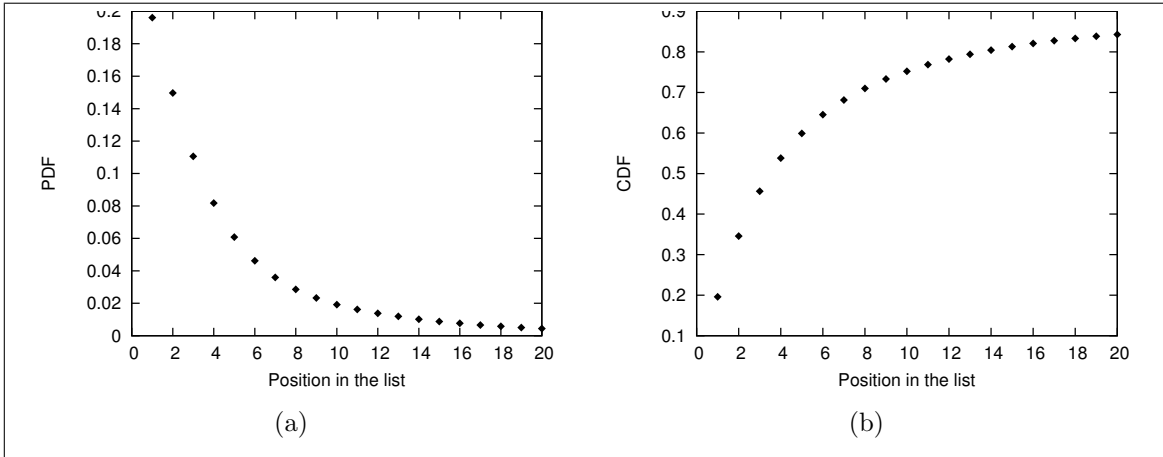


Figure 5.7: The Probability Distribution Function(a) and Cumulative Distribution Function(b) of the positions in the list seen by a user, containing a question that was selected by the user to answer.

in the list shown to her.

Figure 5.7 shows the distribution of the rank positions of the chosen questions. As we can see, the higher a question is ranked, the greater probability it is answered. In addition, while only top 20 questions in the list are considered, the OQE model achieves a recall of 0.84. This means that for 84% of the cases, users just choose questions from the first page they see to answer.

Question Text Effects

Next, this thesis tries to explore beyond the question rank, to understand how the question text affects users' choices of which questions to answer. So an experiment was performed that learns to find the target question q_A in the list of $L_{A,20}$ based on question text features. The learning-to-rank framework is used for this task. Given a

Position Information (4 total): <ul style="list-style-type: none"> * The position where the question is in the list $L_{A,20}$; * The delay of the question since it was posted until seen by the user. * The deviation of the above 2 feature values from the average values of the user.
Similarity (5 total): <ul style="list-style-type: none"> * The similarities between the question and user profile against the 4 fields and the whole profile.
Visual Quality (16 total): <ul style="list-style-type: none"> * The length of question subject/content. * The punctuation, capitalization and spacing density of question subject/content. * The deviation of the above 8 feature values from the average values of the user.
History (4 total): <ul style="list-style-type: none"> * The number of prior answers for the question seen by the user. * The number of prior questions asked by the question asker. * The deviation of the above 2 feature values from the average values of the user.
Keywords (21 total): <ul style="list-style-type: none"> * A vector of length 20 representing whether this question contains the 20 most frequent terms in popular questions (i.e. questions with more than 20 answers). * The number of 1s in the above vector.

Table 5.3: Features (50 total) used in the experiment

list of questions $L_{A,20}$ seen by a user, features representing the associated information (e.g. question text, user’s answering history) are derived to predict which question will be answered by the user.

Guided by reference [3], features are designed according to five layers: position information about questions, question-user similarities, visual quality of questions, popular keywords in questions, and history information about the questions. The complete list of features is shown in Table 5.3.

To make the experiments feasible, 1000 out of the 45,543 active users were ran-

domly selected to build the dataset for this experiment. For each user, the first half of her answers is used to build her user profile. Then, the next 1/4 of her answers is used as training data, and the last 1/4 of her answers as testing data for training and testing the ranker. The resulted dataset contains 15,226 answers and 304,434 questions for training, and 14,721 answers and 294,361 questions for testing.

Lucene (<http://lucene.apache.org/>) is used in the experiment to compute the similarity features between a user profile and the question. Each user profile is indexed as a document with four fields: the content of her answers; and the title, content, and category of the questions she answered. Then a question (including the title, content and category) is treated as 5 queries against the 4 different fields as well as the whole user profile. The 5 scores returned are used as the question-user similarity features. All the features are normalized by linear scaling to unit range. After computing all the features, this thesis then applies a learning-to-rank algorithm, SVM^{rank} [51], to rank the questions.

Since the goal is to find the target question q_A in the list of $L_{A,20}$, the results are evaluated by P@1. The baseline by only checking whether the question is ranked at position 1 achieves the P@1 of 0.24. This means around one fourth of cases, users answer questions ranked top 1 in the list they see. Although position features dominate the performance, including the additional text features provides a slight,

but statistically significant improvement of 4% ($p < 0.01$) over the simple position-only OQE model. Therefore, while the question text does affect answerers for choosing questions, the effect of the question text is not as large as that of category and rank.

5.2 Exploring Web Browsing Context for CQA

Question routing has been proposed as an attractive solution to improve CQA efficiency. Yet, how it would affect answerer behavior, especially the factors influencing the quality and timeliness of the answer contributions, are not well understood. Moreover, as CQA moves towards the real-time setting, it is also important to consider the interruptions and costly context switching such question routing and recommendation may cause for the answerers – who may be less willing to contribute answers when asked at an inopportune time. This could be an important factor for the adaption and success of question recommendation strategies in CQA systems.

As a first step towards modeling the answer contribution behavior in CQA, this thesis considers the effort (costs) associated with answering a question, from the answerer's perspective. Generally, it hypothesizes that the answerers' web browsing context at the time when a question is received affects the likelihood and quality of their responses. For example, if a user is busy paying bills online, she will likely not want to answer any questions at that time; in contrast, if she is browsing a

story about her favorite soccer team’s match, she may welcome a question about the team’s performance in that game. Further, it hypothesizes that relevant web browsing context may also help answerers to more easily and effectively respond to difficult questions that may otherwise go unanswered. Such effects will become increasingly important as CQA moves towards the real-time setting. While current CQA systems do not have access to the web browsing data, such information is already used by web search engines (via toolbars), and by contextual advertising systems (via ad content networks) – so in principle, access to web browsing data (e.g., via a CQA toolbar) could be acceptable to CQA users as well.

The envisioned CONtextual QUEstion Recommender system (CONQUER) is illustrated in Figure 5.8. An asker has a question “What is the difference between Phishing and Hacking”, and this question is routed by our system to the “best” user, who is currently in the most relevant information context to answer this question (in this case, browsing a Wikipedia page about Phishing). Other candidate answerers (indicated with dashed arrows) are otherwise engaged, and may not welcome the topic switch needed to answer the question.

As an initial evaluation of the CONQUER system, this thesis explores the effects of the web browsing context on the answerer ability, effort and willingness to answer questions. Specifically, the research questions are: 1) What are the effects of the

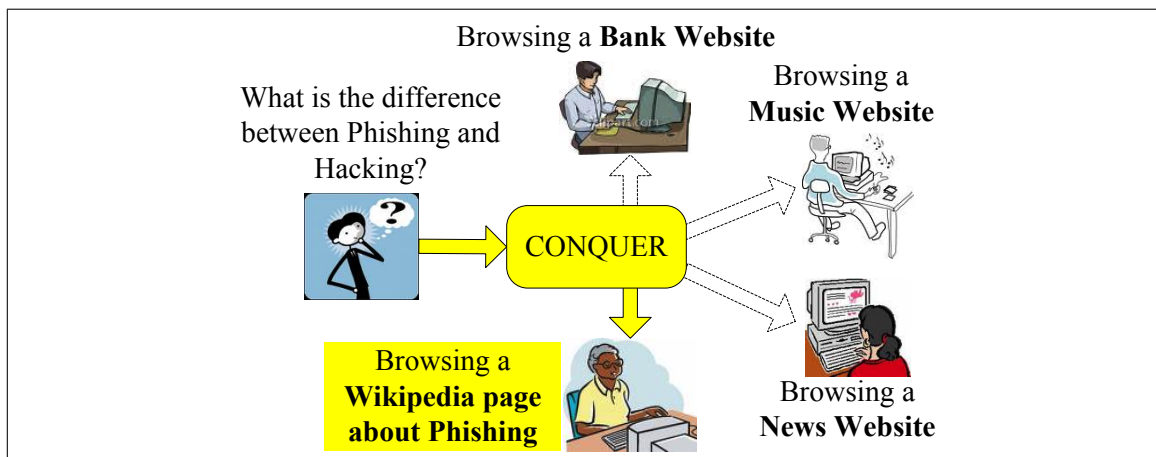


Figure 5.8: Illustration of the CONtExtual QUeStion REcommender system (CONQUER).

relevance of the web browsing context, on the answerers' perceived effort, ability, and willingness to contribute answers to questions? 2) What are the characteristics of the questions or users for which the web browsing context is helpful?

5.2.1 Study Design

To explore the two research questions outlined above, a two-step user study was performed in a lab setting. Step 1 was designed to elicit the baseline values of subjects' interest, ability, perceived effort, and willingness to answer questions independent of the web browsing context. Step 2 was designed to quantify the effect of web browsing context on subjects' ability, effort and willingness to answer questions, measuring both absolute values and relative to the values obtained in Step 1.

The questions for this study were drawn from the 40 most popular resolved ques-

tions in the “Computers & Internet” category on Yahoo! Answers (these questions are presumed to be of high quality since their popularity was rated by the Yahoo! Answers users). For half of these questions (randomly chosen), the most relevant Wikipedia page was selected by the researchers as the “ideal” web browsing context. In total, 17 Wikipedia pages were included (some pages were relevant to multiple questions), comprising the universe of the browsing contexts for this study. There were 10 participants in this study, all Computer Science graduate students. User ratings were collected with a Firefox browser extension.

Step 1: The subjects were shown 20 questions randomly drawn from our question set. For each question, the subjects were asked to rate their interest, ability, estimated effort, and willingness to answer the question. 1-5 scale was used for each rating after being explained to the subjects. For example, for the “interest” rating, 1 indicated “not at all interested” and 5 “extremely interested”. This step was used to establish the baseline values for each of the subjects.

Step 2: Immediately after step 1, the same subjects chose 10 most interesting topics from the 17 computer-related topics to browse. For each topic, the subjects were asked to browse the corresponding Wikipedia page as they normally would. After a one- minute delay, the subjects were then “recommended” four questions in the browser sidebar. At least one of the recommended questions was known to be

relevant to the page, and half of the recommended questions were those rated by the subject in Step 1. For each question, the subjects were asked to rate the page for relevance, helpfulness (for answering the question), and their interest, ability, estimated effort, and willingness to answer the question. The same 1-to-5 scale was used for each rating.

There is a limitation in the realism of the study: the subjects were prompted to choose a topic to browse (as opposed to browsing the web “naturally”). Also, the browsing context is restricted to Wikipedia pages. However, this study describes an interesting setting: the millions of Wikipedia pages comprise some of the most popular browsing destinations on the Web; Furthermore, the contextual effects on answerers’ willingness, effort, and ability to answer questions will generalize to other web browsing contexts as well.

5.2.2 Results

Before analyzing the results, consider the example question “What is the difference between Phishing and Hacking”. Three subjects rated this question in both Step 1 and Step 2 as they browsed the relevant Wikipedia page “<http://en.wikipedia.org/wiki/Phishing>”. Two of the three subjects rated their ability to answer higher in Step 2, while the corresponding ratings of the estimated effort had decreased; finally, for one of the subjects, the willingness to answer increased. Thus, for some users,

relevant browsing context can be helpful. However, it is found that the effects vary for different subjects and different questions. The rest of this section will analyze the results in more detail.

Agreement of Context Relevance Ratings

Recall, that the goal is to study the answerer behavior in both relevant and non-relevant web browsing contexts (measured by the relevance of the corresponding Wikipedia page to the question under consideration). So, this thesis first compared the page relevance as rated by the participants to the researchers' ratings used to construct the dataset. The average agreement between researcher and participant ratings was 0.95 (N=268, Kappa=0.72), thus validating that the relevant pages as chosen by researchers are indeed considered relevant by the participants.

Willingness Causes

Next, as the answerers' willingness to answer a question is an important factor to the effectiveness of CQA, this thesis analyzes its causes, specifically its overall connection with the answerers' interest, ability, and estimated effort to answer the question, as well as context relevance and context helpfulness. As shown in Figure 5.9, the strongest correlation is with the subjects' interest in the question ($r=0.59$), and their ability to answer the question well ($r=0.49$). Interestingly, there is also a weak overall

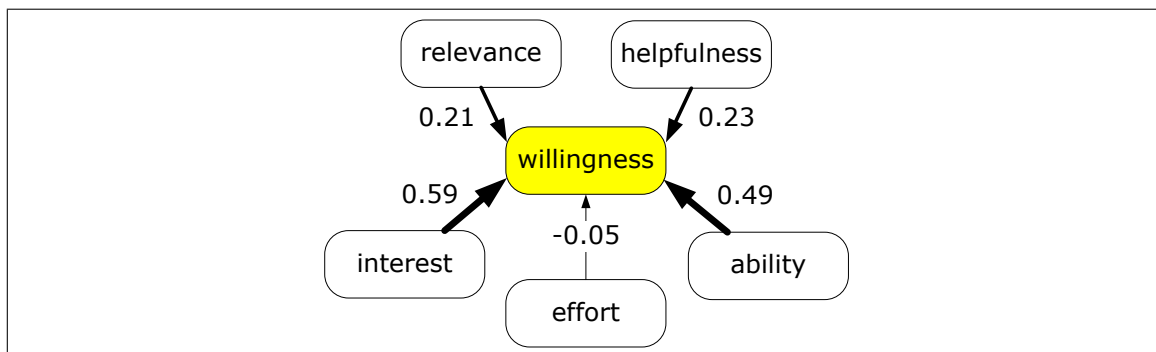


Figure 5.9: Correlation between answerer willingness to answer a question and answerer interest/ability/effort/context relevance/context helpfulness.

correlation with context relevance and helpfulness.

However, if the participants' ratings are examined individually, different patterns for answerer willingness are observed. For one subject, her willingness to answer a question highly correlated to her reported ability ($r=0.71$) and negatively correlated to her effort to answer the question ($r=-0.68$). In contrast, another subject's willingness to answer a question highly correlated to her effort to answer the question ($r=0.85$) – perhaps this subject enjoys challenging questions and opportunities to learn.

Effects of the Browsing Context

This thesis now explores how browsing context affects subject ability, effort, and willingness to answer the recommended questions. It examines the change of subject ratings from step 1 to step 2 for the same question, where in Step 1 a subject rated a question without any context, while in Step 2 the same questions was rated by

the same subject in relevant or non-relevant browsing context. When a question was recommended in a non-relevant browsing context, the subjects' reported ability to answer the same question was significantly decreased. One possible reason is that answering non-relevant questions would require a topic (context) change, and thus additional effort for the answerer. Interestingly, the subjects' estimates of effort and willingness to answer the same questions did not change. In contrast, when a question was recommended to subjects in a relevant browsing context, their estimates of the effort to answer the question were significantly reduced. This indicates that relevant browsing context does help reduce subjects' perceived effort to answer questions. Note that the p-values are computed using paired t-tests (one-tailed distribution).

To analyze which questions and subjects are affected by context, this thesis examines the effects of relevant browsing context for different question and subject groups, specifically the changes in the ratings and corresponding fraction of the subjects that are positively affected by relevant browsing contexts. These questions were first shown without context (Step 1), and then in relevant contexts (Step 2). When subjects had low prior ratings of ability, relevant contexts significantly increased the ratings of ability and significantly decreased the ratings of effort for 68% and 55% of the subjects, respectively. Interestingly, for more than 40% of the subjects with high prior ability, relevant context also significantly decreased their ratings of effort

and significantly increased their ratings of willingness to answer questions. In summary, for many subjects, especially those with low prior ratings of ability, relevant context increased their perceived ability and willingness to answer the question, and reduced their expected effort. However, some subjects reported that seeing a relevant Wikipedia page showed just how difficult it would be to answer some of these questions well, thus discouraging them from posting a potentially poor or even incorrect answer.

The benefits of relevant browsing context also varied according to the intrinsic question characteristics. For objective questions (manually rated by the researchers), relevant contexts significantly increased answerer willingness for 59% of the subjects, possibly due to the increase of their ability and the decrease of their effort to answer questions. In contrast, for subjective questions, while relevant contexts significantly decreased the estimated effort for 52% of the subjects, for most participants this did not increase their willingness to answer questions, which supports the intuition that subjective questions do not benefit from easy access to relevant information.

5.3 Summary

This chapter explored the contextual factors that influence the answerer behavior in a large, popular CQA system, with the goal to inform the construction of real-time,

online question routing and recommendation strategies. Specifically, it considered when users tend to answer questions in a large-scale CQA system, and how answerers tend to choose the questions to answer. The analysis could help develop more realistic evaluation methods for question recommendation, and provide valuable insights into answerer behavior.

A user study was further performed to better understand the role of web browsing context in answerer behavior in CQA. While many open issues remain, the results suggest that for some subjects, especially for those with lower prior ratings of ability, the browsing context can be very helpful. Some of the effects could be due to the inherent question characteristics.

Chapter 6

Building A Real-Time CQA System

Chapter 5 explored what factors influence answerer behavior in CQA, with the goal to inform the construction of question routing and recommendation strategies. This chapter focuses on the problem of how to deploy question recommendation in real-time CQA systems through building such a system.

More specifically, this thesis built a real-time CQA system called RealQA, utilizing a mobile application supporting instant notification and location detection, and a server backend handling recommendation and notification strategies. The main functions of the system include common CQA functions such as asking and answering questions, voting for questions and answers, retrieving a list of questions to answer. This thesis also investigates novel features specifically designed for real-time receiving recommendation notifications of newly posted questions, and receiving new answers for subscribed questions. User locations are recorded while the users interact with

the system, which are later used for real-time question recommendation.

The system was developed iteratively, incorporating insights from the analysis of two user studies - the pilot study and the main study, conducted with students at Emory University. The focus was on evaluating the overall functionality of the system during the pilot study, and on comparing different recommendation algorithms during the main study. Specifically, the following were investigated: 1) the types of questions asked and quality of answers; 2) user preference for question recommendation strategies: letting users pull a list of questions from the main page vs. pushing questions to the users via notifications; 3) the effectiveness of question-ranking and user-ranking algorithms for different recommendation strategies; 4) the effectiveness of question tag recommendation algorithms.

The main contributions of this chapter include:

- A novel, location-aware real-time CQA system named RealQA as a research platform (Section 6.1).
- Two user studies with two versions of the system, showing improvements with both qualitative survey analysis and quantitative behavior analysis (Section 6.2).
- Comparison of different algorithms for question ranking, user ranking, and tag ranking (Section 6.2).

- Insights, data, and shared code¹ for future studies (Section 6.3).

6.1 System Overview

The system consists of two parts: a front-end mobile application and a back-end server system. Users can post, upvote, and downvote questions and answers using the mobile app. From the main page of the mobile application, users can browse questions recommended by the server system. Users may also get questions recommended via mobile notifications. Moreover, users can subscribe to questions they want to get notifications about when new answers are posted for the questions.

The system was developed iteratively, incorporating insights from the analysis of two user studies. In both studies, users' interactions with the system and their survey responses were analyzed. From the pilot study, user feedback was received and the system was improved for both the front-end mobile application and the back-end server system based on the feedback. Feedback from users from the main study was also collected, which will be used for future improvement. Table 6.1 shows main functions and comparisons between systems used for the pilot and the main studies.

¹The data and code are available at <http://ir.mathcs.emory.edu/projects/realqa/>.

	Pilot Study	Main Study
<i>User related</i>		
· Getting notifications (e.g., question recommendations, answer updates)	2 recommendation algorithms with sound & vibration	3 recommendation algorithms silent notifications
· Updating user's tags of interest	Y	Y
· Turning off notifications	Y	Y
· Setting maximum # of recommendations per day	N	Y
· Managing user's notifications in the inbox	N	Y
· Participating in the weekly lottery	N	Y
<i>Question related</i>		
· Browsing recommended questions in the main page	1 assigned ranking algorithm via navigation menu	5 selectable algorithms via ranking options in main page
· Browsing all questions in the system	for asked/answered questions	for asked questions only
· Subscribing questions automatically	Y	Y
· Subscribing questions manually	Y	Y
· Browsing user's asked/answered/subscribed questions	Y	Y
· Asking questions (optionally, annotating locations)	Y	Y
· Answering questions	Y	Y
· Voting for questions and answers	Y	Y
· Getting recommended tags based on question text	N	Y
· Dismissing questions from the main page	N	Y

Table 6.1: Functions in the systems used for the pilot and the main studies.

6.1.1 Front-end: mobile application

The mobile application has been designed and built for smart phones using an Android OS (version 4.0 or higher, occupying 87.9% of the whole Android device market). The application is distributed through the official Android distribution platform, Google Play (<http://play.google.com>).

Registration, login, and navigation During the registration process, each user is asked to enter an username, a password, and tags of interests. After registration, the user is logged into the system, and directed to the main page. The *navigation drawer* is used to view different pages in the application, and can be prompted when the user presses the application icon in the top-left corner (Figure 6.1(a)).

Application Main page By default, the application main page shows a list of questions recommended to this user (Figure 6.1(b)). Each row consists of a question body, tags related to the question, the score of the question, the number of answers to the question, and the time since question is posted. The cross icon in the upper right corner of each question item can be used to dismiss this question. After dismissal, the user no longer sees this question from any of the recommendation question lists.

By clicking the choice box in the upper right corner, users can choose different question ranking algorithms (see the ‘server’ section for more details about the al-

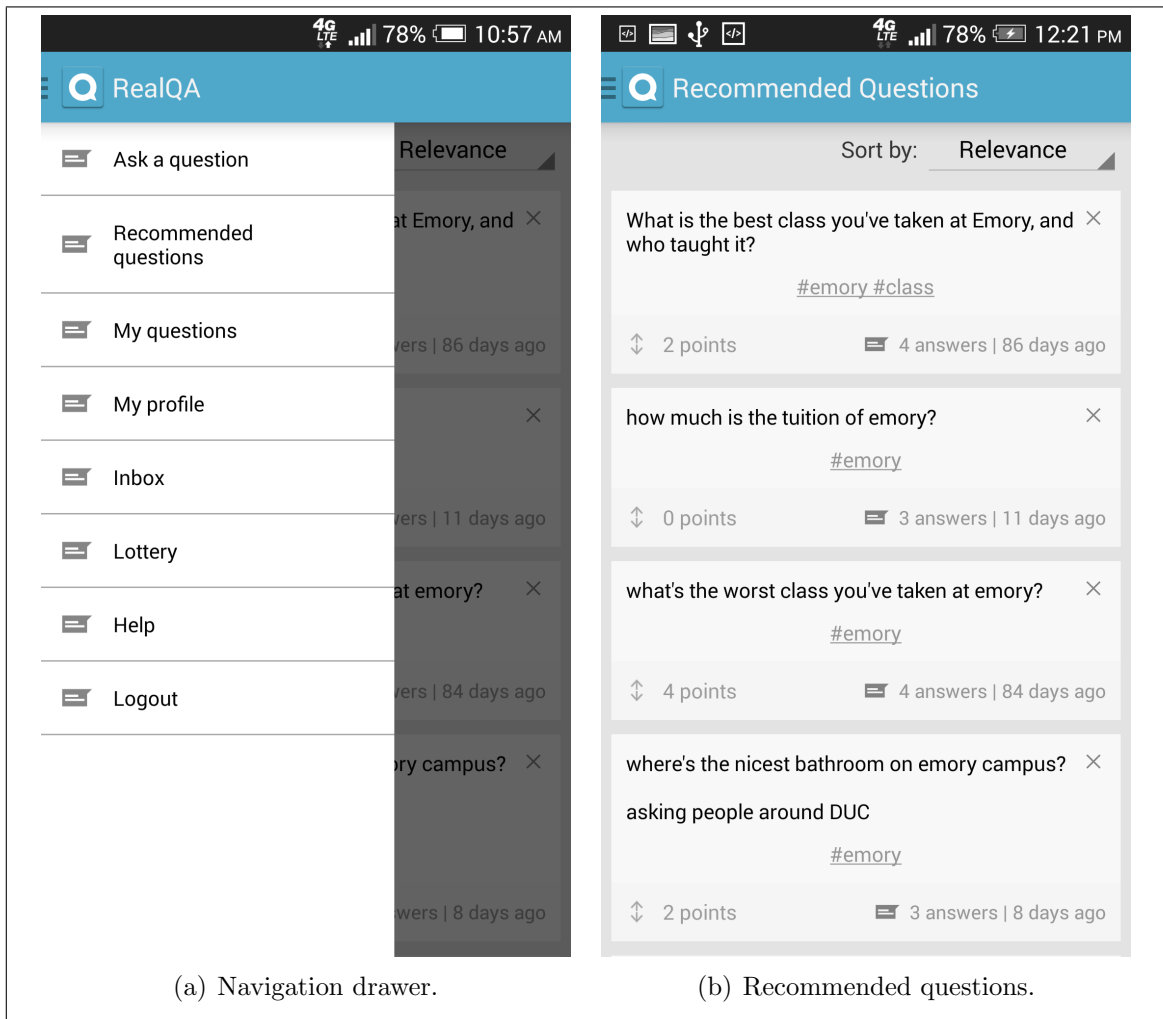


Figure 6.1: The main page of the mobile application.

gorithms). It also gives an option of retrieving all questions. Any click on either the question body or the number of answers displays the *question thread*, consisting of the question and all answers that have been posted for the question (Figure 6.2(a)). Additionally, two buttons are placed in the upper right corner; the left button lets the user subscribe to this question, and the right button lets the user post an answer to this question.

Posting a question A key aspect of the system is to post a new question in real-time (Figure 6.2(b)). When the user enters a question body, the system shows 3 recommended tags and an option to retrieve 10 recommended tags based on the question body (see the ‘server’ section for more details about the recommended tags).

Another important aspect is that the system allows the user to post a question to only people around a specific location. By default, each question is asked to people in all locations with no specific location annotated. If the user chooses “my current location”, the question is annotated with the user’s current location (using latitude and longitude). Other locations can be also chosen, including 21 popular locations on campus such as food courts or residence halls, in which case, the question is annotated with this particular location.

By clicking either the *Ask* button or the right upper corner icon, the question gets posted. The author is automatically subscribed to this question, and receives a

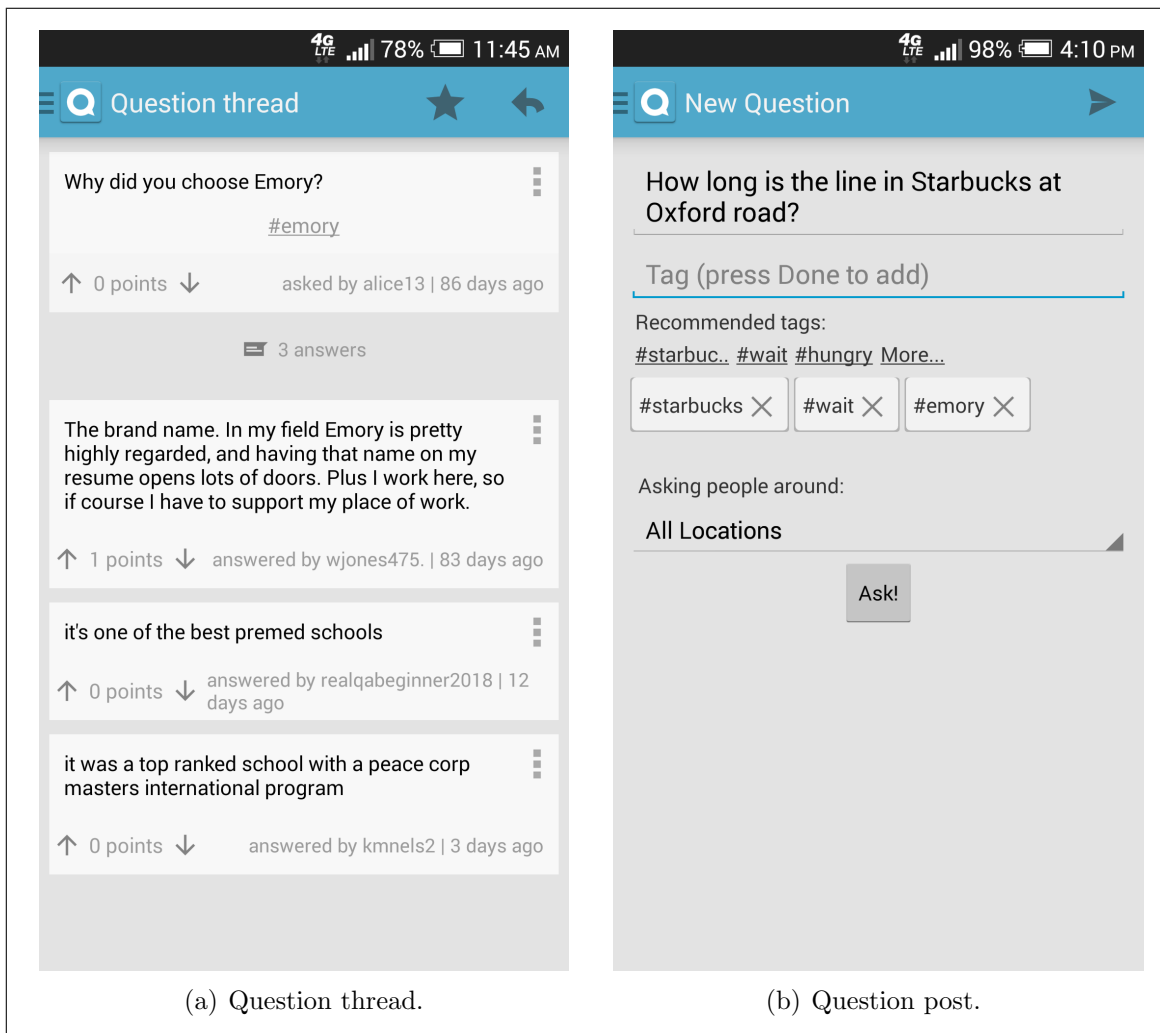


Figure 6.2: Question thread and question post.

notification whenever a new answer is posted for this question.

User profile and notification inbox Users can update their profiles from the *user profile* tab consisting of user statistics, notification settings, and tags settings (Figure 6.3(a)). From the notification settings, users can control whether to receive subscription notifications (when a subscribed question receives a new answer) and recommendation notifications (when a newly-posted question is recommended for the user to answer), as well as the number of maximum recommendations that users receive per day (the default is 5).

The *notification inbox* tab contains notifications that have been sent to the user (Figure 6.3(b)). The user can view the associated question by clicking on any notification. Notifications that have not been clicked are shown in bold.

Miscellaneous The *my questions* tab is similar to the *recommended questions* in Figure 6.1(b), except that the dismissal icon is not present and the right upper choice box include options: asked, answered, subscribed. The *lottery* tab provides information about the status of weekly lotteries. The *help* tab lets users see the terms of service, run the tutorial, or report bugs.

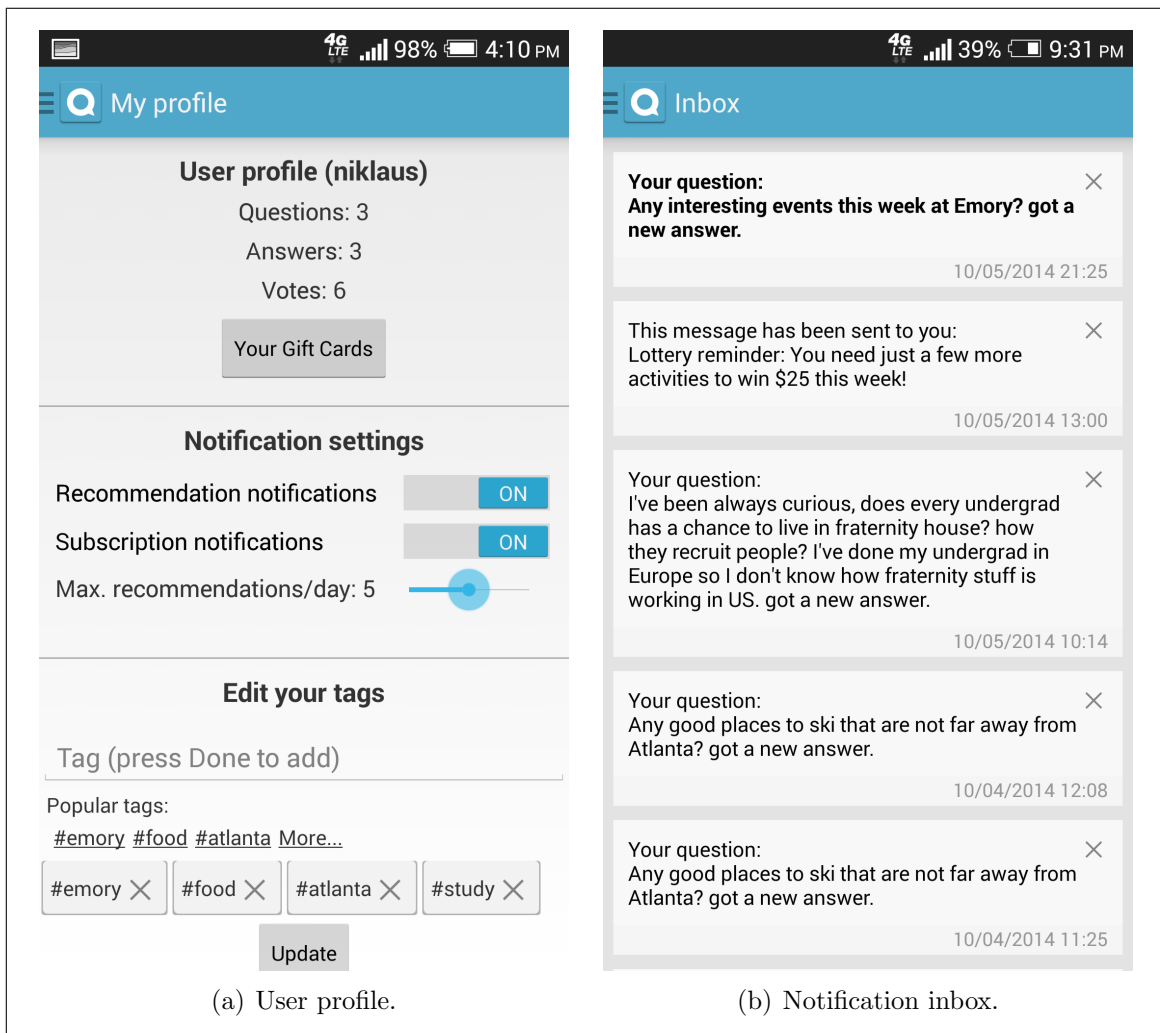


Figure 6.3: User profile and notification inbox.

6.1.2 Back-end: server system

For the implementation of the server system, this thesis adapted the Django REST framework [24] and extended the Django models from the Open Source Q&A system (OSQA [81]).

Search component

Haystack [41] and Elasticsearch [28] were adapted for implementing the search component: Haystack provides modular search for Django and Elasticsearch provides scalable real-time indexing and search. Adapting these tools allows the system to create indices for questions, users, and tags dynamically, as well as to search them in real time. Each document in the index contains two fields: a question text field and a tag field. Detailed content of each type of documents is provided in Table 6.2. Field level boosting was used to weigh the tag field twice more than the question text field. De-capitalization, stop word removal, and stemming were applied to the texts in documents and queries.

The system uses a vector space model for measuring the relevance between documents and search queries. All documents and queries are represented in a multi-dimensional vector space, where each dimension stands for a unique term, and its value stands for the TF-IDF score of the term. The relevance between a document

Document Type	Question Text Field and Tag Field
question	<ul style="list-style-type: none"> · Question body · Tags of this question
user	<ul style="list-style-type: none"> · Question bodies answered by this user. · Tags of questions answered by this user, and tags pre-entered in user's profile.
tag	<ul style="list-style-type: none"> · Question bodies with this tag. · The name of this tag.

Table 6.2: Index structure for questions, users, and tags.

and a query is measured by Lucene's practical scoring function [71]:

$$score(q, d) = \sum_{t \in q} (tf(t, d) \cdot idf(t)^2 \cdot norm(t, d)) \cdot queryNorm(q) \cdot coord(q, d)$$

$queryNorm(q)$ is the query normalization factor, $coord(q, d)$ is a score factor based on how many of the query terms are found in document d , $tf(t \in d)$ is the term frequency for term t in document d , $idf(t)$ is the inverse document frequency for term t , $norm(t, d)$ is the field length norm, combined with the index-time field-level boost.

Question Ranking Component

The system offers 5 different options for ranking questions when a pull (of questions) is requested by a user:

- **Relevance** - questions that are more relevant to the user's interests are ranked higher. To find relevant questions, search queries are extended with contents

in both question text field and tag field of user document, which provide information about the user's interests. This is the default option.

- **Freshness** - more recent questions are ranked higher.
- **Location** - if questions are annotated with locations, ones closer to the user's current location are ranked higher.
- **Popularity** - questions viewed by more unique users are ranked higher.
- **Answer count** - questions with a fewer number of answers are ranked higher.

Previous studies have shown that factors such as relevance and freshness are important for question recommendation [102]. During the pilot study, the system assigned a random ranking algorithm to each user; however, it lets the users choose the ranking algorithm during the main study, which is an improvement made from the analysis of the pilot study.

User ranking component

Given a newly posted question, the system randomly assigns one of the following algorithms for ranking users to send recommendation notifications for answering this question:

- **Matching questions** - This algorithm first finds the top 20 questions in the history that are similar to the new question using the search component, then ranks the answerers of these questions. The score of each answerer is measured by the sum of similarity scores between the new question and questions previously answered by the user [34].
- **Matching users** - This algorithm uses the new question as a query against the user documents, and returns the top ranked users using the search component.
- **Location proximity** - This algorithm computes the distance between user's and the question locations, and returns the users with closest distances. This algorithm is used for only questions that are annotated with a specific location.

The top 5 ranked users will be sent recommendation notifications for the question. The system during the pilot study always used the location proximity strategy for questions annotated with specific locations, and the matching users strategy for the rest. During the main study, random algorithms are assigned to new questions to evaluate the differences between these strategies.

Tag ranking component

When the user moves onto the tag input field after completing a new question, a list of recommended tags is displayed. The following algorithms are used for ranking

these tags:

- **Matching questions** - This algorithm is similar to the user ranking one, except the score of each tag is measured by the sum of similarity scores between the new question and questions annotated with this tag [118].
- **Matching tags** - This algorithm uses the new question as a query against the tag documents, and returns the top ranked tags using the search component.
- **Tag popularity** - This algorithm ranks the tags based on their use frequencies. A tag is ranked higher if more questions are annotated with this tag.

When a new question is entered, the user is randomly assigned one of the first two ranking algorithms, which returns the top 10 ranked tags. If the number of ranked tags is smaller than 10, then the third algorithm is used to fill in the rest.

6.2 User studies

To evaluate the system for real-time question answering, two user studies were conducted. During the pilot study, the focus was on evaluating the overall functionality of the system, while the focus was on comparing different recommendation algorithms during the main study. The key aspects of the studies are:

- Types of questions and quality of answers.

Statistics	Pilot	Main
# of registered users	27	35
# of qualified users	14	16
# of all questions	120	56
# of questions annotated with locations	5	6
% of answered questions	83%	89%
# of all answers	244	238
# of upvotes for questions	76	76
# of upvotes for answers	134	155
# of downvotes for questions	15	5
# of downvotes for answers	33	26
Avg. # of tags in user profiles	3.6	4.5
Avg. # of tags for a question	1.9	2.3

Table 6.3: Statistics of the data collected from the user studies.

- Preference of question recommendation strategies: letting users pull a list of questions in main page (PULL) vs. pushing questions to users via notifications (PUSH).
- Effectiveness of question ranking algorithms for PULL.
- Effectiveness of user ranking algorithms for PUSH.
- Effectiveness of tag recommendation algorithms.

Any student with an Android phone was eligible to participate in these studies. Each participant was asked to install and use the mobile application for a limited period of time. The runtime, minimum required activities, compensation, and initial database were different between two studies. The pilot study was run for a week in

July, 2014. In this study, participants received \$10 gift cards if they met the following requirements: 1) posted at least 5 questions, 10 answers, 10 votes; 2) performed 1 activity (ask, answer, or vote) per day for 5 days; 3) completed the survey. The initial database included 7 questions posted by the authors of this paper.

The main study was run for 3 weeks in Sep. 2014. In this study, participants received \$5 gift cards if they met the following requirements: 1) posted at least 1 question, 5 answers, 5 votes; 2) performed at least 1 activity per day for 3 days; 3) completed the survey. Moreover, users were encouraged to keep using the system for winning lotteries even after they received the gift cards; participants who completed the minimum required activities in a week were eligible to play the lottery held at the end of that week, and each winner received a \$25 gift card. The initial database included most questions, answers, tags, votes from the pilot study (some noisy data were discarded).

6.2.1 Statistics and survey responses

In both studies, users' interactions with the systems and their survey responses were analyzed. Table 6.3 shows statistics from the user studies. Qualified users are the ones who finished the minimum required activities. The percentage of qualified users is lower in the main study; one reason is that the pilot study was conducted during the summer when students had more free time, whereas the main study was conducted

during the regular school year. The difference in the number of all questions is most likely due to the different minimum required activities in each study. The percentages of answered questions are both high and comparable to the one provided by Aardvark (Table 2.1). The number of upvotes is much higher than the number of downvotes for both studies, similar as in Stack Overflow [5]. Users in the main study tended to utilize tags more, probably because they are generally younger than the ones in the pilot study.

Table 6.4 shows the survey questions and the responses from users collected during the studies. The average ratings for all questions increased from the pilot study to the main study. In fact, all the negative average ratings had turned to positive. This shows that the overall user satisfaction with the system in the main study had improved from the pilot study. From users' feedback from the pilot study, it is found that being able to ask a question to the local community was most important, and pushing the notifications about new answers to their questions was also important. Moreover, it is found that too many notifications, irrelevantly recommended questions, and the lack of options to sort recommended questions could annoy users. Based on the feedback, the system was improved for both the mobile application and the server system as described in Table 6.1 (e.g., sending silent notifications, allowing users to set the maximum number of recommendation notifications per day,

managing notification inbox, providing ranking options in the main page, sending answer updates to only askers, and using better recommendation algorithms).

From users' feedback from the main study, it is found that the most significant issue was having not enough participants. Some users mentioned about improving the quality of recommended questions, improving robustness of tag recommendation, as well as supporting discussion forums for answers.

6.2.2 Question types and answer quality

Types of questions To understand what types of questions had been asked, all questions were manually categorized, shown in Table 6.5, based on the question types discussed in [77, 43]. Many questions were with local intent, and meanwhile more subjective questions (recommendation, conversational, opinion) were found than objective ones (factual). The main topics of the questions were about food, study, and relaxing activities. Some example questions and answers from the study participants are shown in Table 6.6. Such characteristics of the questions were related to the deployment of the system within a campus setting. Compared to Aardvark, the percentages of subjective questions are similar; however, the percentage of questions with local intent is much higher in this study (see Table 2.1).

From the analysis of the question types, it is found that many questions were with local intent; however, only 5 and 6 questions were annotated with specific locations in

Survey questions	Pilot	Main
1. How satisfied are you with the answers?	0.71	1.13*
2. How would you rate the timeliness of receiving your answers?	0.57	0.94
3. How satisfied are you with the questions recommended via notification?	-0.86	0.75*
4. How satisfied are you with the questions recommended via main page?	-0.36	0.50*
5. Which do you prefer, question recommendation via notification or main page?	21%, 71%, 8%	25%, 56%, 19%
6. Which do you prefer for ranking recommended questions in main page?	-	relevant, fresh, popular
7. How satisfied are you with the tags recommended when asking a question?	-	0.75
8. How useful are the notifications about answer updates?	0.57	1.19
9. How useful are the notifications about question recommendations?	-0.43	0.81*
10. How useful is the “ask people around some place” feature?	0.57	0.63
11. What did you like (like best) about the system?
12. What did you dislike (dislike most) about the system?
13. Comments/Suggestions.

Table 6.4: Survey responses. Ratings are scaled in $\{2, 1, 0, -1, -2\}$. Tuples in Question 5 represent percentages of notification, main page, and no preference. * indicates statistically significant difference according to the Mann-Whitney test at $p = 0.05$.

Types	Pilot	Main
Local-intent	63.3%	76.8%
Time-sensitive	12.5%	14.3%
Recommendation	26.7%	50.0%
Conversational	48.3%	14.3%
Opinion	10.0%	7.1%
Factual	15.0%	28.6%
Food	10.8%	26.8%
Study	14.2%	14.3%
Entertainment	20.8%	19.6%
Annotated with locations	4.2%	10.7%

Table 6.5: Types of questions asked and their proportions.

the pilot and main studies, respectively. One reason is that such local intents are at the university or city level, while the locations provided by the system are at a finer granularity of the building or part-of-campus level. Therefore, users prefer to ask questions to people around all locations in the system, which is the default option. Future plan is to automatically detect the local intent based on the question text and tags, and dynamically suggest locations with various granularity in the future. As shown in Table 6.4 (Question 10), users found the function of “asking people around a specific location” somewhat useful in both studies ($0.57 \rightarrow 0.63$). This aspect can be improved if users have more freedom to select locations or areas.

Response latency For the analysis of the response latency, this thesis measured the duration between questions and their first answers being posted. The median of this duration is 3.2 hours and 36 minutes in the pilot and the main studies,

Q1: my mom is coming into town and I need to entertain her for the weekend. where do people bring their parents for fun things to do? #hungry #music #parents
A1: I would definitely take her to the botanic garden, the aquarium, and walking around emory. Make it a walking weekend with a focus on being active!
Q2: are there any upcoming food festivals in or around ATL? #food #atlanta #festival
A2: Taste of Atlanta! it's next month. great event.
Q3: Where can I get the most food at Emory (for a meal) for the best value? #food #emory #hungry
A3: the DUC is unlimited (all you can eat) for a meal swipe or dooley dollars. If you're tired of the DUC, try the Woodrec in Woodruff hall. you get up to 4 items for a meal swipe or dooley dollars. it's pretty filling, but not all you can eat.
Q4: is there free parking anywhere on campus? #broke
A4: the parking lots on campus open at different times. I know peavine opens at 4, but then fishbourne doesn't open till later 6. and don't risk parking overnight, I've seen people in the morning get tickets
Q5: What class do you recommend for any major to take? #emory #class #professors
A5: Astronomy with the physics department. No prior knowledge needed and lots of fun and pretty easy. Chaucer with Professor Morey was also excellent.

Table 6.6: Example questions and answers from the studies.

respectively. One reason for the faster latency in the main study can be found from the more active use of recommendation notifications (Table 6.8; 11% \rightarrow 24%). Mobile notifications are usually checked within minutes [85], which largely shortens the time for an user to see and answer a newly posted question. The proportion of questions being answered within the first 10 minutes is 34% in the main study here, which is lower than 57.2% as reported by Aardvark [43]. This number could be increased if the study had a larger user base. The improvement on timeliness of receiving answers in the main study over the pilot study is also supported by the survey responses shown in Table 6.4 (Question 2; 0.57 \rightarrow 0.94).

Quality of answers From the survey responses in Table 6.4 (Question 1), we see that users are mostly satisfied with the answer quality in both studies, and the satisfaction had increased in the main study (0.71 \rightarrow 1.13). Since the answer length is an important feature for predicting answer quality [3], this thesis measured the average answer length in each study and found that the average answer length in the main study was longer than the one in the pilot study (Table 6.7; 10.8 \rightarrow 12.4). Compared to Aardvark (median answer length: 22.2 as reported in [43]), the answer length tends to be shorter in the main study here (median answer length: 7). One reason is that the input device is limited to mobile devices in this study while users in Aardvark could also use desktop computers.

Measurements	Pilot	Main
Avg # of words in answers	10.8	12.4
Avg # of words in answers (score>0)	13.5	16.6
Avg time editing an answer	52s	40s
Avg time editing an answer (score>0)	66s	46s
% of answers with score>0	39%	31%

Table 6.7: Statistics related to the answer quality.

Another good indicator for the answer quality is the answer score, which is measured by subtracting the number of downvotes from the number of upvotes. It is found that answers with positive scores tended to be longer in both studies (13.5 vs. 10.8, 16.6 vs. 12.4), and users spent more time for editing them (66s vs. 52s, 46s vs. 40s). An interesting observation is that users in the main study tend to type faster,

Sources of answers	Pilot	Main
all questions	90 (44%)	25 (12%)
main page	80 (39%)	133 (61%)
recommendation notification	23 (11%)	51 (24%)
subscription notification	13 (6%)	7 (3%)
total	206 (100%)	216 (100%)

Table 6.8: Source of answers (considering only qualified users).

probably because they are generally younger than users in the pilot study.

6.2.3 Question recommendation strategies: PULL vs. PUSH

Which leads to more answers? Which is preferred by users? Table 6.8 shows which way the qualified users used the most to find questions they wanted. This thesis considers only behavior of qualified users here because behavior of unqualified users is too sparse and unreliable to be analyzed. For both studies, users answer more questions from the main page than from notifications (39% vs. 11%, 61% vs. 24%). Besides, we see the percentage of answers from recommendation notifications increased in the main study (11% \rightarrow 24%), implying the improvement of notification recommendation. On the other hand, the percentage of answers from subscription notification decreased (6% \rightarrow 3%). In addition, the percentage of answers from all questions list also decreased (44% \rightarrow 12%), as in the main study we removed the “all questions” choice from the navigation menu, but added it to the ranking options in the main page.

The preference result is also supported by the survey responses (Table 6.4, Question 5), i.e., the majority of users prefer main page rather than notification for question recommendation (71% vs. 21%, 56% vs. 25%). Meanwhile, more users express no preference between the two recommendation strategies in the main study (8% \rightarrow 19%), with the improvement made on these strategies.

The interpretation of these results highly depends on how PULL and PUSH are performed. Yet, the implication here is that it is important to allow users to pull questions, even in a real-time question answering system. This is also mentioned in Aardvark [43] that 16.9% of all users have proactively tried answering questions. However, in Aardvark more users answered via notification than via pulling questions. The explanation was that users were willing to answer questions to help their friends or connected people, but not everyone does so proactively. In the built system, the users were not as connected. This might be one important reason for the different preference of PULL and PUSH in this system compared to Aardvark.

6.2.4 Recommendation from the main Page: question ranking

Which ranking is viewed more and leads to more answers? How satisfied are users with main page recommendations? Which ranking is preferred by users? Table 6.9 shows the results from different ranking algorithms. Users

Ranking	Views	Answers	Answers per view
Relevance	336	87	0.259
Freshness	163	38	0.233
Popularity	21	4	0.190
Location	14	2	0.143
Answer count	8	2	0.250
all questions	114	25	0.219

Table 6.9: Question ranking in the main study.

mostly viewed questions by relevance, partially because it is the default option. When considering the likelihood of answering, ranking by relevance and freshness are among the best. This is consistent with the observation in [102] that freshness is an important factor besides relevance for question recommendation.

From the survey responses (Table 6.4; Question 4), users are somewhat satisfied with the main page recommendations in the main study (0.50), better than in the pilot study (-0.36). One important reason is that in the pilot study, each user was randomly assigned a single question ranking algorithm, whereas users chose the ranking algorithm by themselves in the main study, which was more preferred by the users. From survey responses in the main study (Table 6.4; Question 6), ranking questions by relevance (31%), freshness (31%), and popularity (31%) are among the best. An interesting observation is that although users claimed to be interested in popular questions, they were less likely to answer these questions.

6.2.5 Recommendation via notification: user ranking

How many recommendation notifications are sent? How many are clicked and answered? How satisfied are users with this? Table 6.10 shows the statistics about recommendation notifications. The average number of recommendations sent per question to qualified users in the main study is higher than the one in the pilot study (3.0 \rightarrow 4.3). This implies that a larger proportion of recommendations was sent to unqualified users during the pilot study. Looking at the recommendation notifications sent to qualified users, the likelihood of users clicking on them and the likelihood of users answering the recommended questions had increased (0.40 \rightarrow 0.52, 0.15 \rightarrow 0.28).² This shows that our system performed better in notification recommendations during the main study, thanks to several improvements (i.e., managing notification inbox, setting max recommendation notifications per day, and the recommendation algorithm). First, as notification inbox was not supported in the pilot study, users could only see the latest one, and miss some previous ones since their last check of mobile notifications. This would affect clicks and answers. Second, it is noticed that active users received more and more notifications per day in the pilot study. One reason is that the user ranking algorithm had a bias towards active

²Note that all the questions recommended to a user that got answered by him/her are counted here. A user might have not clicked on the recommendation notification but answered it from the main page.

	Pilot	Main
recommendations	356	239
avg rec. per question	3.0	4.3
clicks	142	124
click rate	0.40	0.52
answers	54	66
answer rate	0.15	0.28

Table 6.10: Statistics of question recommendation notifications.

users. Another reason is that the max recommendations per day was not limited in the pilot study in the first 3 days. To avoid the unbalance of user workloads getting worse, the limitation was set to 5 in the last 4 days. Users were not aware of this number. However, in the main study users were allowed to reset this number at any time of the study.

The improvement of the system regarding notification recommendations is also supported by survey responses. As shown in Table 6.4 (question 3), users' satisfaction with notification recommendations increased in the main study compared to the pilot study (-0.86 \rightarrow 0.75).

Which algorithm is better for user ranking? To evaluate the three algorithms used in the user ranking component, this thesis uses both click related metrics (considering users who clicked on a given question as the ground truth) and answer related metrics (considering users who answered a given question as the ground truth). Specifically, average precision, recall, and F1 scores are computed across all questions

using both click and answer based ground truths.

Table 6.11 and Table 6.12 show the results of comparing the three algorithms. When looking at questions that are not annotated with a specific location, the algorithm based on matching users performed better in both click and answer related metrics. First, this algorithm has a bias towards active users, because the document of an active user contains more answered questions and corresponding tags, which makes it more likely to match a new question. Meanwhile, active users are more likely to respond to recommended questions, e.g., by clicking on the question and answering it. The workload balance is handled using the maximum number of recommendations per day set by each user, therefore active users will not get over-annoyed. This is however different from the observation in [34] that matching questions is more effective than matching users for finding potential answerers for a question. First, questions in their setting, i.e., from Yahoo! Answers, is much more diverse in terms of topics than in the setting of this thesis, i.e., posted by university students, where the main topics are food, study, and entertainment. Therefore, active users are able to answer a larger percentage of questions. Second, the results in this thesis come from live user studies while their results are based on offline simulation experiments. Therefore, it is hard to predict how the results would change from offline simulation to online application.

Algorithm	Click based ground truth			Answer based ground truth		
	Prec@5	Rec@5	F1@5	Prec@5	Rec@5	F1@5
matching questions	0.49	0.30	0.35	0.20	0.36	0.24
matching users	0.60	0.37	0.44*	0.29	0.38	0.32

Table 6.11: Comparing algorithms for ranking users to send recommendations for a newly posted question that is not annotated with any location. * indicates statistically significant difference according to the Mann-Whitney test at $p = 0.05$.

Algorithm	Click based ground truth			Answer based ground truth		
	Prec@5	Rec@5	F1@5	Prec@5	Rec@5	F1@5
matching questions	0.53	0.27	0.35	0.15	0.44	0.21
location proximity	0.55	0.32	0.39	0.2	0.58	0.29

Table 6.12: Comparing algorithms for ranking users to send recommendations for a newly posted question that is annotated with a location.

When looking at questions that are annotated with a specific location, the algorithm based on location proximity showed better performance. This indicates that location proximity is an important factor other than relevance for ranking questions related to specific locations. Yet, more data is needed to do more meaningful analysis regarding the algorithms.

6.2.6 Tag ranking

Which algorithm is more effective for tag recommendation? To evaluate the algorithms used in the tag ranking component, this thesis uses the actual tags entered by the asker as the ground truth, and computes the average precision, recall and F1 scores across all the questions for both top 10 results and top 3 results of

Algorithm	Prec@3	Rec@3	F1@3	Prec@10	Rec@10	F1@10
matching questions	0.283	0.358	0.300	0.102	0.419	0.158
matching tags	0.220*	0.301*	0.240*	0.092*	0.401*	0.145*
tag popularity	0.289	0.376	0.309	0.111	0.453	0.173
combined live system	0.294*	0.400*	0.319*	0.116*	0.485*	0.180*

Table 6.13: Comparing algorithms for tag recommendation. * indicates statistically significant difference according to the Wilcoxon signed-rank test at $p = 0.05$.

each algorithm.

Table 6.13 shows the results of comparing the algorithms. For both top 3 and top 10 results, the live system combining the three algorithms achieved the best results on all metrics. The algorithm based on matching questions performed better than matching tags, as matching questions and then summing scores of questions for a tag would better filter out noisy tags in top results while matching tags directly are more likely to return noisy tags because their documents might well match the given question text. The proposed algorithm based on matching questions is similar to the SimilarityRank approach [118], with the distinction that this thesis sums the scores of similar questions for a tag as its score while their approach chose the max score of similar questions for a tag as its score. The precision and recall are comparable to the ones in [118].

As shown in Table 6.4 (question 7), users are somewhat satisfied with tag recommendations (0.75). This may be improved by applying more state-of-the-art tag recommendation algorithms [72], which will be the future work.

6.2.7 Notification Settings

How many users changed their default notification settings? How annoying are the notifications? In the pilot study, three users turned off recommendation notifications. In the main study, one user turned it off, and three users reset the number of max recommendation notifications per day (from 5 to 10, 3, 3 respectively). One user in the pilot study and no user in the main study turned off subscription notifications. This improvement is mainly because of being more careful about sending notifications in the main study, e.g., notifications being sent in silent mode, answerers not getting answer updates on their answered questions, at most 5 recommendation notifications per day and allowing users to reset it, and the improvement of the recommendation algorithms.

From survey responses, improved user experience about notifications was also observed. As shown in Table 6.4 (Question 9), in the pilot study users rated recommendation notifications somewhat annoying (-0.43), but in the main study users rated them useful (0.81). Meanwhile, in the main study users rated notifications about answer updates more useful (Question 8; 0.57 \rightarrow 1.19).

6.3 Discussion and Implications

By analyzing the types of questions posted in the studies, it is found a large proportion of questions asked are with local intent, subjective, and about food, study, and relaxing activities. Previous research on question recommendation focused more on matching questions and potential answerers based on interest, relevance and availability [31], while less effort was made on question recommendation considering types of questions, e.g., questions with local intent. The preliminary study results indicate that location proximity is an important factor for finding potential answerers to answer such questions besides relevance. More investigation on how to integrate location proximity with relevance and other factors for question recommendation in such cases is needed.

From the studies, it is learned that the majority of users prefer pulling questions to answer rather than being pushed questions to answer. However, there are some users who really like to receive notifications, for example, one user commented in survey that “I like the notifications. it’s a good way to get people to look at questions without having to browse the app.” Therefore, question recommendation systems would achieve better performance if such personal preference are detected and supported. One solution is to ask users during registration about their preferred settings of recommendation strategies, and then learn based on user behavior to

suggest change of the settings or even perform automatic change.

When users do a pull of questions to answer, it is found that overall ranking questions by relevance and freshness leads to higher answer rate, which is consistent with the observation in [102]. However, users express different preferences of relevance, freshness, and popularity for ranking the questions for this purpose. Therefore, question recommendation systems need to consider the different importance of factors for ranking questions for different users in the pulling mode. It is also found that the question dismissal function was not actively used (75% of the qualified users never dismissed any question from their pulled questions). A potential improvement is to consider users' feedback of dismissing a question in question ranking algorithms, and meanwhile inform users of the benefit.

For deciding which users to push a new question, it is found that ranking users by matching user profiles leads to better performance than by matching questions, which differs from the observation in [34], due to the differences in user base (university students vs Yahoo! Answers users) and in experiment type (live user studies vs offline simulation experiment). It is found that this better performing algorithm has a bias towards active users, who are more likely to answer questions. Therefore, question recommendation systems could rely on active users to contribute more answers, and will benefit from predicting active users before they actually behave actively. Yet,

control is needed to avoid over-annoying active users.

Regarding interface design for asking a question, it is learned from the studies that careful design of location annotation for questions is needed. There was a large gap between the number of questions with local intent and the number of questions that are annotated with a specific location by the asker, partially because the granularity of the predefined locations in the system is not so useful. Meanwhile, letting askers make manual annotation of locations is annoying, which takes extra effort of askers, especially to select diverse granularity of locations using a complex interface. Therefore, location-aware question answering systems might have user experience improved if they can automatically detect the local intent based the question text and tags, and dynamically suggest locations with various granularity to be annotated with the question.

The developed system was designed to be highly scalable. The Django REST framework [24] and Open Source QA models [81] used have proven their success to build scalable server systems. The recommendation algorithms are based on query evaluation using Elasticsearch [28], a scalable distributed real-time search software. The system could be used for multiple communities as well after extending the current limited location choices to cover more places using a map.

The presented studies have a few limitations. First, the data collected is from 27

users in 7 days and 35 users in 3 weeks in the pilot and the main studies, respectively. This limited user base and study period makes it hard to draw definitive conclusions from the data. Further, since users were mainly university students, observations in our studies may not generalize to other populations. Second, the two studies have similar but different settings, e.g., the pilot study was conducted in summer with more graduate students joined, while the main study was conducted in fall with more undergraduate students joined. Third, rewards were used as participation incentives. However, given that participants were rewarded for total activity, they could freely choose how to allocate their effort and time. Also, it is found that qualified users performed more activities than the required minimum, not just being perfunctory for rewards. Therefore, comparing recommendation strategies and ranking algorithms in this setting is still meaningful. Thus, the findings and results would still be likely to apply in more realistic settings.

6.4 Summary

This chapter presented RealQA, a real-time CQA system with a mobile front-end interface. The system provided two strategies for users to get recommended questions to answer: users doing a pull of questions in the main page or being pushed questions via mobile notifications. Two user studies were conducted to test the effectiveness

of the system. Based on user feedback and comments from the first study, both the front-end interface and back-end algorithms of the system were upgraded. Both users' self-reported satisfaction and behavior related metrics were improved in the main study compared to the pilot one. Different algorithms for question ranking, user ranking, and tag ranking were adapted and compared. The developed system, and the reported findings and analysis, offer insights and implications for designing real-time CQA systems, and provide a valuable platform for future research.

Chapter 7

Conclusions and Future Work

This chapter first summarizes the main findings and implications of the thesis work, and then discusses the limitations and future research directions.

7.1 Summary of Thesis Work

Community-based Question Answering (CQA) services allow users to find and share information by interacting with other users in the communities. Users using CQA systems can play three different roles: askers, answerers or searchers. One way to improve the experience of askers and answerers is by question routing and recommendation, while one way to improve the experience of searchers is by understanding searcher satisfaction. This thesis presented analysis and methods to improve the usefulness of CQA services towards better searcher satisfaction and question recommendation. Specifically, it contributes to the research on CQA by focusing on three important problems that have been under addressed in previous work:

(1) How to improve web searcher satisfaction using CQA services

By formulating a new problem of predicting web searcher satisfaction with existing answers in CQA, this thesis made the first attempt to predict and validate the usefulness of CQA archives for external searchers, rather than for the original askers. After identifying three key aspects for this task, namely, query clarity, query-question match, and answer quality, this thesis proposed a direct method that uses all the available features in a single regression model, and a composite method that learns three separate regressors for each of the three sub-tasks, and then uses their predictions as features for solving the main task.

Findings and Implications This thesis found that the composite method achieved better performance than the direct method, with the use of additional exogenous knowledge learned from the human labels for each of the sub-tasks while training the three sub-regressors. Furthermore, the composite approach is more flexible, and it can immediately benefit as the predictions in individual sub-tasks are improved. As demonstrated, the prediction of web searcher satisfaction by the proposed methods can be used for reranking CQA pages in web search, achieving a better ranking over a state-of-the-art baseline. Modeling the searcher satisfaction with answers in CQA has multiple benefits. First, CQA data provides a unique semi-structured source of

human answers for search engines, which are particularly useful for satisfying tail queries. Moreover, an accurate predictor of searcher satisfaction can be used for improved ranking of CQA results in web search. As demonstrated by the experiments, this can be achieved. In addition, if a search engine detects that a user is struggling with a search session, it could suggest posting a question on a CQA site, offering help with formulating a natural language question and choosing an appropriate category. The results suggest promising directions for improving and exploiting CQA services in pursuit of satisfying even more web search queries.

This thesis further considered the usefulness of CQA services for web searchers who are not satisfied with the search results, by analyzing the unique properties of search sessions that turn into question composition on CQA services. Various aspects of such SearchAsk sessions are analyzed, including the differences between general search-engine queries and those belonging to a SearchAsk session, the transformation of a query into a natural language question and the question composition patterns, as well as other asking behavior of searchers, compared to general askers in a CQA service.

Findings and Implications This thesis found that Queries which are more likely to fail in search and lead to a question post tend to be longer, and use more verbose natural language to express the searchers needs. The information needs behind

such queries tend to be more unique and complex than those associated with general web search queries. Searchers who switch to CQA exhibit common search behavior. For instance, they tend to click more on CQA results on the search result page, and their search sessions are longer. Words in queries follow different distributions than those appearing in questions, and a clear vocabulary gap between them can be observed. Questions are typically more specific than queries and include additional context (e.g., personal background) absent from the original queries. The content and the topics of the questions posted after a search session differ substantially from the general question distribution. The findings may contribute to both search-engine optimization and better user experience in CQA sites. For example, searchers are found not as patient as regular askers when waiting for answers to their questions. This finding may influence CQA sites to promote questions coming from searchers, if they want to retain their engagement. As another example, the analysis of the transitions between user actions in SearchAsk sessions, and especially the fact that question asking is typically preceded by viewing a CQA page, may help search engines. They might decide to detect such cases and explicitly promote the option of asking a question to the searcher, even before she resorts into doing it on her own. These provide opportunities for having more frustrated web searchers benefit from CQA services.

(2) What contextual factors influence answerer behavior in CQA

Through analyzing the answering behavior of a large number of users in a popular CQA system, this thesis explored when answerers tend to answer questions and how they tend to choose questions to answer. For the first research question, this thesis analyzed both the overall and user-specific temporal activity patterns. For the second research question, this thesis analyzed the factors that may affect users' decisions of which questions to answer, including the question category, the question position in the list shown to users, and the surface patterns in the question text.

Findings and Implications This thesis identified stable daily and weekly periodicities, as well as not previously observed bursty patterns of activity in the individual answer sessions of many users. This observation leads to a novel session-based analysis of the answerer activity. This thesis also confirmed previous findings that users have favorite categories that attract most of their contributions, but interestingly the decisions for most users within a category are determined more by the rank of the question in the list of available questions, than other factors such as the text or the provenance of the question itself. The results could help develop more accurate answerer behavior and prediction models; allow the development of more realistic evaluation methodology for question recommendation; and inform the design of bet-

ter question recommender strategies.

Through a controlled user study, this thesis further explored how relevant web browsing context affects answerers' perceived ability, effort, and willingness to answer a question, and the characteristics of the questions or users for which the web browsing context is helpful.

Findings and Implications The results showed that in many cases answerers' relevant browsing context increases their estimated ability and willingness while reduces their perceived effort to answer questions. Some intriguing cases were also found where a relevant page, shown in the right context, could discourage an answerer from posting a potentially poor or even incorrect answer. Moreover, the effect of relevant browsing context were found to vary for different question and subject groups. For some subjects, especially for those with lower prior ratings of ability, relevant browsing context can be very helpful. Meanwhile, objective questions are more likely to be answered in relevant web browsing context, based on users' perceived willingness. This study was a first step towards better understanding answerers' effort and potential willingness to answer a question, which is important for intelligent question recommendation in CQA systems, especially when CQA moves towards the real-time setting.

(3) How to deploy question recommendation in real-time CQA systems

This thesis developed a real-time CQA system with a mobile interface, which provides two strategies for users to get recommended questions to answer: users doing a pull of questions in the main page or being pushed questions via mobile notifications. Two live user studies were conducted: a formative pilot study with the initial system design, and a more extensive study with the revised interface and algorithms. Obvious improvements on both users' self-reported satisfaction and behavior-related metrics were observed by comparing the two studies.

Findings and Implications The results showed that the majority of users prefer pulling questions rather than being pushed questions to answer, though some users like to receive notifications more. Therefore, question recommendation would achieve better performance if such personal preferences are detected and supported. For the pulling strategy, ranking questions by relevance and freshness leads to higher answer rate. However, users may express different preferences of relevance, freshness, and popularity. Therefore, better question recommendation needs to consider the different importance of factors for ranking questions for different users in the pulling mode. For the pushing strategy, the better performing algorithm of ranking users for a new question has a bias towards active users who are more likely to answer

questions, suggesting that intelligent question routing could rely on active users to contribute more answers, and will benefit from predicting active users before they actually behave actively, with some control to avoid annoying them too much. The study results also suggest that location proximity is an important factor for finding potential answerers to answer questions with local intent. The developed system, and the reported findings and analysis, offer insights and implications for designing real-time CQA systems and deploying question recommendation in such systems, and provide a valuable platform for future research.

In summary, the work of predicting searcher satisfaction and understanding the transition from searching to asking is helpful for improving searcher satisfaction using CQA systems, while the work of understanding answerer behavior and building a real-time CQA system is helpful for improving question recommendation in CQA systems. All together, this helps make CQA services more useful with better experience of askers, answerers, and searchers.

From a broader perspective, this thesis also contributes to improving web search by proposing new data and methods for query intent analysis and query reformulation pattern analysis, and by utilizing the structure of community-generated web pages and available community signals. The real-time CQA system built in this thesis can also be applied to better evaluating question recommendation strategies and

automatic answer generation algorithms, via deploying the system in live user studies.

7.2 Limitations and Future Work

The results and findings in the thesis are promising. Yet, there are a few limitations in the proposed methods and analysis:

- The analysis and experiments in this thesis are mostly based on data from a large scale popular CQA system, i.e., Yahoo! Answers. Yet, each CQA system is a little different. For example, some CQA systems like StackOverflow uses tags to organize questions instead of using hierarchical categories as in Yahoo! Answers, and other CQA systems like JustAnswer let askers use money to attract answers instead of using less tangible incentives as in Yahoo! Answers. How the findings in this thesis will be generalized to other CQA systems need to be verified.
- The user studies performed for evaluating the built real-time CQA system have limited user base and study period, which makes it hard to draw definitive conclusions from the data. Also, since users were mainly university students, observations in the studies may not generalize to other population. In addition, the two studies have similar but different settings, e.g., the pilot study was conducted in summer with more graduate students joined, while the main study

was conducted in fall with more undergraduate students joined. Therefore, a larger user study with wider range of population is needed, in which A/B testing can be executed for more fair system comparisons.

Based on the thesis work, some interesting future research directions towards a better searcher satisfaction and question recommendation in CQA systems include:

- **Improving prediction of searcher satisfaction with CQA results:** As shown in Chapter 3, the improvement of addressing the sub-tasks, e.g., predicting query clarity, predicting query-question matching, and predicting answer quality, would help better address the main task of predicting searcher satisfaction. Therefore, applying more state-of-the-art methods for solving the sub-tasks into the system could be one direction. Another branch of potential work is to develop semi-supervised or unsupervised methods to predict the searcher satisfaction, e.g., by treating the queries leading to a transition from searching to asking as negative examples, as large numbers of human labels are hard to obtain.
- **Predicting transition from searching to asking:** A natural extension of the work on analyzing when searchers become askers is to predict such transition before users resort into doing it on their own, and help them with the transition, e.g., by suggesting which CQA service to use, what question and

associated category to submit. The findings from the analysis show some important features for this task, e.g., query length, query frequency, query words, query results, action transitions and action sequence in the search session.

- **Extending the real-time CQA system built:** The system could be extended to deploy more state-of-the-art question recommendation algorithms. Also, for questions with local intent, it would be helpful to investigate the integration of location proximity, relevance, and other factors like availability for routing such questions to potential answerers. In addition, more personalized question recommendation strategies can be considered, e.g., predicting which users prefer pulling strategy to pushing strategy, modeling how important users weigh factors like relevance, freshness and popularity for ranking questions when users do a pull.
- **Performing larger scale user studies:** To have more algorithms deployed and evaluated in the built real-time CQA system, a larger user base and a longer period of study time is needed. To expand the future studies to a larger user base, it would be helpful to build an iOS version of the mobile front-end and a web front-end for the system. To better attract users to use the system regularly in a longer period, it would be helpful to introduce more incentives such as reputation and points besides regular lotteries into the system.

Bibliography

- [1] Lada A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. Knowledge sharing and yahoo answers: everyone knows something. In *WWW*, pages 665–674, 2008.
- [2] Mikhail Ageev, Qi Guo, Dmitry Lagun, and Eugene Agichtein. Find it if you can: a game for modeling different types of web search success using interaction data. In *SIGIR*, pages 345–354, 2011.
- [3] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining, WSDM '08*, pages 183–194, New York, NY, USA, 2008. ACM.
- [4] Florian Alt, Alireza Sahami Shirazi, Albrecht Schmidt, Urs Kramer, and Zahid Nawaz. Location-based crowdsourcing: Extending crowdsourcing to the real world. In *Proceedings of the 6th Nordic Conference on Human-Computer In-*

- teraction: Extending Boundaries*, NordiCHI '10, pages 13–22, New York, NY, USA, 2010. ACM.
- [5] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. Discovering value from community activity on focused question answering sites: A case study of stack overflow. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 850–858, New York, NY, USA, 2012. ACM.
- [6] Christina Aperjis, Bernardo A. Huberman, and Fang Wu. Harvesting collective intelligence: Temporal behavior in yahoo answers. Jan 2010.
- [7] Baidu Knows. <http://zhidao.baidu.com/>. Accessed: 2014-10-28.
- [8] Michael Bendersky and W. Bruce Croft. Analysis of long queries in a large scale search log. In *WSCD*, 2009.
- [9] Jiang Bian, Yandong Liu, Eugene Agichtein, and Hongyuan Zha. Finding the right facts in the crowd: factoid question answering over social media. In *WWW*, pages 467–476, 2008.
- [10] Mohamed Bouguessa, Benoît Dumoulin, and Shengrui Wang. Identifying authoritative actors in question-answering forums: the case of yahoo! answers. In *KDD*, 2008.

- [11] David J. Brenes, Daniel Gayo-Avello, and Kilian Pérez-González. Survey and evaluation of query intent detection methods. In *Proc. of the 2009 workshop on Web Search Click Data*, pages 1–7.
- [12] Andrei Broder. A taxonomy of web search. *SIGIR Forum*, 36:3–10, September 2002.
- [13] M.F. Bulut, Y.S. Yilmaz, and M. Demirbas. Crowdsourcing location-based queries. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on*, pages 513–518, March 2011.
- [14] Georg Buscher, Ryen W. White, Susan Dumais, and Jeff Huang. Large-scale analysis of individual and task differences in search result page examination strategies. In *WSDM*, pages 373–382, 2012.
- [15] Li Cai, Guangyou Zhou, Kang Liu, and Jun Zhao. Learning the latent topics for question retrieval in community qa. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 273–281, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing.
- [16] Liliana Calderón-Benavides, Cristina González-Caro, and Ricardo Baeza-Yates. Towards a Deeper Understanding of the User’s Query Intent. In *Pro-*

- ceedings of the Query Representation and Understanding workshop at SIGIR 2010*, pages 21–24, 2010.
- [17] Xin Cao, Gao Cong, Bin Cui, and Christian S. Jensen. A generalized framework of exploring category information for question retrieval in community question answer archives. In *WWW*, pages 201–210, 2010.
- [18] Xin Cao, Gao Cong, Bin Cui, Christian S. Jensen, and Quan Yuan. Approaches to exploring category information for question retrieval in community question-answer archives. *ACM Trans. Inf. Syst.*, 30(2):7:1–7:38, May 2012.
- [19] Xin Cao, Gao Cong, Bin Cui, Christian Søndergaard Jensen, and Ce Zhang. The use of categorization information in language models for question retrieval. In *CIKM*, pages 265–274, 2009.
- [20] David Carmel, Elad Yom-Tov, Adam Darlow, and Dan Pelleg. What makes a query difficult? In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '06*, pages 390–397, New York, NY, USA, 2006. ACM.
- [21] Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. Predicting query performance. In *SIGIR*, pages 299–306, 2002.

- [22] Anhai Doan, Raghu Ramakrishnan, and Alon Y. Halevy. Crowdsourcing systems on the world-wide web. *Commun. ACM*, 54(4):86–96, April 2011.
- [23] Debora Donato, Francesco Bonchi, Tom Chi, and Yoelle Maarek. Do you want to take notes?: identifying research missions in Yahoo! search pad. In *WWW*, pages 321–330, 2010.
- [24] Django REST framework. <http://www.django-rest-framework.org/>. Accessed: 2014-10-05.
- [25] Gideon Dror, Yehuda Koren, Yoelle Maarek, and Idan Szpektor. I want to answer; who has a question?: Yahoo! answers recommender system. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11*, pages 1109–1117, New York, NY, USA, 2011. ACM.
- [26] Gideon Dror, Yoelle Maarek, Avihai Mejer, and Idan Szpektor. From query to question in one click: suggesting synthetic questions to searchers. In *WWW*, pages 391–402, 2013.
- [27] Huizhong Duan, Yunbo Cao, Chin-Yew Lin, and Yong Yu. Searching questions by identifying question topic and question focus. In *ACL*, pages 156–164, 2008.
- [28] Elasticsearch. <http://www.elasticsearch.org/>. Accessed: 2014-10-05.

- [29] Henry A. Feild, James Allan, and Rosie Jones. Predicting searcher frustration. In *SIGIR*, pages 34–41, 2010.
- [30] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378 – 382, 1971.
- [31] Bojan Furlan, Bosko Nikolic, and Veljko Milutinovic. A survey and evaluation of state-of-the-art intelligent question routing systems. *International Journal of Intelligent Systems*, 28(7):686–708, 2013.
- [32] Yunjun Gao, Lu Chen, Rui Li, and Gang Chen. Mapping queries to questions: towards understanding users’ information needs. In *SIGIR*, pages 977–980, 2013.
- [33] Daniel Gayo-Avello. A survey on session detection methods in query logs and a proposal for future evaluation. *Inf. Sci.*, 179(12):1822–1843, 2009.
- [34] Jinwen Guo, Shengliang Xu, Shenghua Bao, and Yong Yu. Tapping on the potential of q&a community by recommending answer providers. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM '08*, pages 921–930, New York, NY, USA, 2008. ACM.
- [35] Lei Guo, Enhua Tan, Songqing Chen, Xiaodong Zhang, and Yihong (Eric)

- Zhao. Analyzing patterns of user content generation in online social networks. In *KDD*, pages 369–378, 2009.
- [36] Qi Guo, Ryen W. White, Yunqiao Zhang, Blake Anderson, and Susan T. Dumais. Why searchers switch: understanding and predicting engine switching rationales. In *SIGIR*, pages 335–344, 2011.
- [37] Zoltán Gyöngyi, Georgia Koutrika, Jan Pedersen, and Hector Garcia-Molina. Questioning yahoo! answers. In *Proc. of the 1st Workshop on Question Answering on the Web*, 2008.
- [38] F. Maxwell Harper, Daniel Moy, and Joseph A. Konstan. Facts or friends?: distinguishing informational and conversational questions in social q&a sites. In *CHI*, pages 759–768, 2009.
- [39] F. Maxwell Harper, Daphne Raban, Sheizaf Rafaeli, and Joseph A. Konstan. Predictors of answer quality in online q&a sites. In *CHI*, pages 865–874, 2008.
- [40] Ahmed Hassan, Rosie Jones, and Kristina Lisa Klinkner. Beyond dcg: user behavior as a predictor of a successful search. In *WSDM*, pages 221–230, 2010.
- [41] Haystack. <http://haystacksearch.org/>. Accessed: 2014-10-05.
- [42] Daqing He and Ayse Göker. Detecting session boundaries from web user logs.

In *Proc. of the BCS-IRSG 22nd Annual Colloquium on Information Retrieval Research*, 2000.

- [43] Damon Horowitz and Sepandar D. Kamvar. The anatomy of a large-scale social search engine. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 431–440, New York, NY, USA, 2010. ACM.
- [44] Gary Hsieh and Scott Counts. Mimir: A market-based real-time question and answer service. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 769–778, New York, NY, USA, 2009. ACM.
- [45] Scott B. Huffman and Michael Hochster. How well does result relevance predict session satisfaction? In *SIGIR*, pages 567–574, 2007.
- [46] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20:422–446, October 2002.
- [47] Jiwoon Jeon, W. Bruce Croft, and Joon Ho Lee. Finding similar questions in large question and answer archives. In *CIKM*, pages 84–90, 2005.
- [48] Jiwoon Jeon, W. Bruce Croft, Joon Ho Lee, and Soyeon Park. A framework to predict the quality of answers with non-textual features. In *SIGIR*, 2006.

- [49] Zongcheng Ji and Bin Wang. Learning to rank for question routing in community question answering. In *Proceedings of the 22Nd ACM International Conference on Conference on Information and Knowledge Management, CIKM '13*, pages 2363–2368, New York, NY, USA, 2013. ACM.
- [50] Zongcheng Ji, Fei Xu, Bin Wang, and Ben He. Question-answer topic model for question retrieval in community question answering. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12*, pages 2471–2474, New York, NY, USA, 2012. ACM.
- [51] Thorsten Joachims. Training linear svms in linear time. In *KDD*, pages 217–226, 2006.
- [52] Rosie Jones and Kristina Lisa Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *CIKM*, 2008.
- [53] Pawel Jurczyk and Eugene Agichtein. Discovering authorities in question answer communities by using link analysis. In *CIKM*, pages 919–922, 2007.
- [54] JustAnswer. <http://www.justanswer.com/>. Accessed: 2014-10-28.
- [55] Yutaka Kabutoya, Tomoharu Iwata, Hisako Shiohara, and Ko Fujimura. Effective question recommendation based on multiple features for question answering communities. In *Proceedings of the Fourth International Conference on*

Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010, 2010.

- [56] Leyla Kazemi and Cyrus Shahabi. Geocrowd: Enabling query answering with spatial crowdsourcing. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems, SIGSPATIAL '12*, pages 189–198, New York, NY, USA, 2012. ACM.
- [57] Jung-Tae Lee, Sang-Bum Kim, Young-In Song, and Hae-Chang Rim. Bridging lexical gaps between queries and questions on large online q&a collections with compact translation models. In *EMNLP*, pages 410–418, 2008.
- [58] Uichin Lee, Hyanghong Kang, Eunhee Yi, Mun Yi, and Jussi Kantola. Understanding mobile q&a usage: An exploratory study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '12*, pages 3215–3224, New York, NY, USA, 2012. ACM.
- [59] Baichuan Li and Irwin King. Routing questions to appropriate answerers in community question answering services. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10*, pages 1585–1588, New York, NY, USA, 2010. ACM.
- [60] Baichuan Li, Irwin King, and Michael R. Lyu. Question routing in community

- question answering: Putting category in its place. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM '11*, pages 2041–2044, New York, NY, USA, 2011. ACM.
- [61] Mingrong Liu, Yicen Liu, and Qing Yang. Predicting best answerers for new questions in community question answering. In *Proceedings of the 11th International Conference on Web-age Information Management, WAIM'10*, pages 127–138, Berlin, Heidelberg, 2010. Springer-Verlag.
- [62] Qiaoling Liu and Eugene Agichtein. Modeling answerer behavior in collaborative question answering systems. In *ECIR*, pages 67–79, 2011.
- [63] Qiaoling Liu, Eugene Agichtein, Gideon Dror, Evgeniy Gabrilovich, Yoelle Maarek, Dan Pelleg, and Idan Szpektor. Predicting web searcher satisfaction with existing community-based answers. In *SIGIR*, pages 415–424, 2011.
- [64] Qiaoling Liu, Eugene Agichtein, Gideon Dror, Yoelle Maarek, and Idan Szpektor. When web search fails, searchers become askers: understanding the transition. In *SIGIR*, pages 801–810, 2012.
- [65] Qiaoling Liu, Tomasz Jurczyk, Jinho Choi, and Eugene Agichtein. Real-time community question answering: Exploring content recommendation and user notification strategies. In *Proceedings of the 20th International Conference on*

Intelligent User Interfaces, IUI '15, pages 50–61, New York, NY, USA, 2015. ACM.

- [66] Qiaoling Liu, Yandong Liu, and Eugene Agichtein. Exploring web browsing context for collaborative question answering. In *IiX*, pages 305–310, 2010.
- [67] Xiaoyong Liu, W. Bruce Croft, and Matthew Koll. Finding experts in community-based question-answering services. In *CIKM*, pages 315–316, 2005.
- [68] Yandong Liu, Jiang Bian, and Eugene Agichtein. Predicting information seeker satisfaction in community question answering. In *SIGIR*, pages 483–490, 2008.
- [69] Localmind. <http://www.localmind.com/>. Accessed: 2014-08-28.
- [70] LOCQL. <http://www.locql.com/>. Accessed: 2014-08-28.
- [71] Lucene’s practical scoring function. https://lucene.apache.org/core/4_0_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html. Accessed: 2014-10-05.
- [72] Zongyang Ma, Aixin Sun, Quan Yuan, and Gao Cong. Tagging your tweets: A probabilistic modeling of hashtag annotation in twitter. In *Proceedings of the 23rd ACM Conference on Information and Knowledge Management, CIKM '14*, 2014.

- [73] Mahalo Answers. <http://www.mahalo.com/answers/>. Accessed: 2011-1-19.
- [74] Jalal Mahmud, Michelle X. Zhou, Nimrod Megiddo, Jeffrey Nichols, and Clemens Drews. Recommending targeted strangers from whom to solicit information on social media. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces, IUI '13*, pages 37–48, New York, NY, USA, 2013. ACM.
- [75] Zhaoyan Ming, Tat-Seng Chua, and Gao Cong. Exploring domain-specific term weight in archived question search. In *CIKM*, pages 1605–1608, 2010.
- [76] Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich. A comparison of information seeking using search engines and social networks. In *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*, 2010.
- [77] Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich. What do people ask their social networks, and why?: A survey study of status message q&a behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '10*, pages 1739–1748, New York, NY, USA, 2010. ACM.
- [78] Kevin Kyung Nam, Mark S. Ackerman, and Lada A. Adamic. Questions in,

- knowledge in?: a study of naver's question answering community. In *CHI*, pages 779–788, 2009.
- [79] Arnab Nandi, Stelios Pappas, John C. Shafer, and Rakesh Agrawal. With a little help from my friends. In *ICDE*, pages 1288–1291, 2013.
- [80] Naver Knowledge-iN. <http://kin.naver.com/>. Accessed: 2014-10-28.
- [81] OSQA. <http://www.osqa.net/>. Accessed: 2014-08-28.
- [82] Bo Pang and Ravi Kumar. Search in the lost sense of "query": question formulation in web search queries and its temporal changes. In *Proc. of HLT-ACL 2011: Short Papers*, pages 135–140, 2011.
- [83] Sangkeun Park, Yongsung Kim, Uichin Lee, and Mark Ackerman. Understanding localness of knowledge sharing: A study of naver kin 'here'. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices & Services, MobileHCI '14*, pages 13–22, New York, NY, USA, 2014. ACM.
- [84] Dan Pelleg, Denis Savenkov, and Eugene Agichtein. Touch screens for touchy issues: Analysis of accessing sensitive information from mobile devices. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013.*, 2013.

- [85] Martin Pielot, Karen Church, and Rodrigo de Oliveira. An in-situ study of mobile phone notifications. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices & Services, MobileHCI '14*, pages 233–242, New York, NY, USA, 2014. ACM.
- [86] Mingcheng Qu, Guang Qiu, Xiaofei He, Cheng Zhang, Hao Wu, Jiajun Bu, and Chun Chen. Probabilistic question recommendation for question answering communities. In *WWW*, 2009.
- [87] Quora. <http://www.quora.com/>. Accessed: 2014-08-28.
- [88] Quora Online Now. <http://blog.quora.com/Getting-Answers-Faster>. Accessed: 2014-08-28.
- [89] Daphne Raban. Self-presentation and the value of information in Q&A web sites. *JASIST*, 60(12):2465–2473, 2009.
- [90] Fatemeh Riahi, Zainab Zolaktaf, Mahdi Shafiei, and Evangelos Milios. Finding expert users in community question answering. In *Proceedings of the 21st International Conference Companion on World Wide Web, WWW – CQA'12 Workshop*, pages 791–798, New York, NY, USA, 2012. ACM.
- [91] Matthew Richardson and Ryen W. White. Supporting synchronous social q&a throughout the question lifecycle. In *Proceedings of the 20th International*

- Conference on World Wide Web, WWW '11*, pages 755–764, New York, NY, USA, 2011. ACM.
- [92] Daniel E. Rose and Danny Levinson. Understanding user goals in web search. In *WWW*, pages 13–19, 2004.
- [93] Jose San Pedro and Alexandros Karatzoglou. Question recommendation for collaborative question answering systems with rankslda. In *Proceedings of the 8th ACM Conference on Recommender Systems, RecSys '14*, pages 193–200, New York, NY, USA, 2014. ACM.
- [94] Mark Sanderson and Ian Soboroff. Problems with kendall’s tau. In *SIGIR*, 2007.
- [95] Chirag Shah and Jeffrey Pomerantz. Evaluating and predicting answer quality in community QA. In *SIGIR*, 2010.
- [96] Xiance Si, Edward Y. Chang, Zoltán Gyöngyi, and Maosong Sun. Confucius and its intelligent disciples: Integrating social with search. *Proc. VLDB Endow.*, 3(1-2):1505–1516, September 2010.
- [97] Young-In Song, Chin-Yew Lin, Yunbo Cao, and Hae-Chang Rim. Question utility: A novel static ranking of question search. In *AAAI*, pages 1231–1236, 2008.

- [98] Stackoverflow. <http://stackoverflow.com/>. Accessed: 2014-08-28.
- [99] Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. Learning to rank answers on large online qa collections. In *ACL*, pages 719–727, 2008.
- [100] Maggy Anastasia Suryanto, Ee Peng Lim, Aixin Sun, and Roger H. L. Chiang. Quality-aware collaborative question answering: methods and evaluation. In *WSDM*, pages 142–151, 2009.
- [101] Saori Suzuki, Shin’ichi Nakayama, and Hideo Joho. Formulating effective questions for community-based question answering. In *SIGIR*, pages 1261–1262, 2011.
- [102] Idan Szpektor, Yoelle Maarek, and Dan Pelleg. When relevance is not enough: Promoting diversity and freshness in personalized question recommendation. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW ’13, pages 1249–1260, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.
- [103] Jaime Teevan, Susan T. Dumais, and Daniel J. Liebling. To personalize or not to personalize: modeling queries with variation in user intent. In *SIGIR*, 2008.
- [104] Yuan Tian, PavneetSingh Kochhar, Ee-Peng Lim, Feida Zhu, and David Lo. Predicting best answerers for new questions: An approach leveraging topic

- modeling and collaborative voting. In Akiyo Nadamoto, Adam Jatowt, Adam Wierzbicki, and JochenL. Leidner, editors, *Social Informatics*, volume 8359 of *Lecture Notes in Computer Science*, pages 55–68. Springer Berlin Heidelberg, 2014.
- [105] Kai Wang, Zhaoyan Ming, and Tat-Seng Chua. A syntactic tree matching approach to finding similar questions in community-based qa services. In *SIGIR*, pages 187–194, 2009.
- [106] X.J. Wang, X. Tu, D. Feng, and L. Zhang. Ranking community answers by modeling question-answer relationships via analogical reasoning. In *SIGIR*, 2009.
- [107] Yu Wang and Eugene Agichtein. Query ambiguity revisited: clickthrough measures for distinguishing informational and ambiguous queries. In *ACL*, 2010.
- [108] Ingmar Weber, Antti Ukkonen, and Aris Gionis. Answers, not links: Extracting tips from yahoo! answers to address how-to web queries. In *WSDM*, pages 613–622, 2012.
- [109] Ryen W. White and Susan T. Dumais. Characterizing and predicting search engine switching behavior. In *CIKM*, pages 87–96. ACM, 2009.

- [110] Ryen W. White, Matthew Richardson, and Yandong Liu. Effects of community size and contact rate in synchronous social q&a. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2837–2846, New York, NY, USA, 2011. ACM.
- [111] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83, 1945.
- [112] Fei Xu, Zongcheng Ji, and Bin Wang. Dual role model for question recommendation in community question answering. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 771–780, New York, NY, USA, 2012. ACM.
- [113] Xiaobing Xue, Jiwoon Jeon, and W. Bruce Croft. Retrieval models for question and answer archives. In *SIGIR*, pages 475–482, 2008.
- [114] Yahoo! Answers. <https://answers.yahoo.com/>. Accessed: 2014-08-28.
- [115] 1 Billion Answers Served! <http://yanswersblog.com/index.php/archives/2010/05/03/1-billion-answers-served/>. Accessed: 2011-01-19.
- [116] Zhenlei Yan and Jie Zhou. A new approach to answerer recommendation in community question answering services. In *Proceedings of the 34th European*

- Conference on Advances in Information Retrieval*, ECIR'12, pages 121–132, Berlin, Heidelberg, 2012. Springer-Verlag.
- [117] Liu Yang, Minghui Qiu, Swapna Gottipati, Feida Zhu, Jing Jiang, Huiping Sun, and Zhong Chen. Cqarank: Jointly model topics and expertise in community question answering. In *Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management*, CIKM '13, pages 99–108, New York, NY, USA, 2013. ACM.
- [118] E. Zangerle, W. Gassler, and G. Specht. Recommending #-tags in twitter. In *Proceedings of the Workshop on Semantic Adaptive Social Web (SASWeb 2011)*. *CEUR Workshop Proceedings*, volume 730, pages 67–78, 2011.
- [119] Jun Zhang, Mark S. Ackerman, and Lada Adamic. Expertise networks in online communities: structure and algorithms. In *WWW*, pages 221–230, 2007.
- [120] Weinan Zhang, Zhaoyan Ming, Yu Zhang, Liqiang Nie, Ting Liu, and Tat-Seng Chua. The use of dependency relation graph to enhance the term weighting in question retrieval. In *COLING*, pages 3105–3120, 2012.
- [121] Shiqi Zhao, Haifeng Wang, Chao Li, Ting Liu, and Yi Guan. Automatically generating questions from queries for community-based question answering. In *IJCNLP*, pages 929–937, 2011.

- [122] Zhicheng Zheng, Xiance Si, Edward Chang, and Xiaoyan Zhu. K2q: Generating natural language questions from keywords with user refinements. In *IJCNLP*, pages 947–955, 2011.
- [123] Guangyou Zhou, Li Cai, Jun Zhao, and Kang Liu. Phrase-based translation model for question retrieval in community question answer archives. In *ACL*, pages 653–662, 2011.
- [124] Guangyou Zhou, Fang Liu, Yang Liu, Shizhu He, and Jun Zhao. Statistical machine translation improves question retrieval in community question answering via matrix factorization. In *ACL*, pages 852–861, 2013.
- [125] Guangyou Zhou, Kang Liu, and Jun Zhao. Exploiting bilingual translation for question retrieval in community-based question answering. In *COLING*, pages 3153–3170, 2012.
- [126] Guangyou Zhou, Kang Liu, and Jun Zhao. Joint relevance and answer quality learning for question routing in community qa. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12*, pages 1492–1496, New York, NY, USA, 2012. ACM.
- [127] Guangyou Zhou, Yang Liu, Fang Liu, Daojian Zeng, and Jun Zhao. Improving

- question retrieval in community question answering using world knowledge. In *IJCAI*, 2013.
- [128] Tom Chao Zhou, Michael R. Lyu, and Irwin King. A classification-based approach to question routing in community question answering. In *Proceedings of the 21st International Conference Companion on World Wide Web, WWW – CQA’12 Workshop*, pages 783–790, New York, NY, USA, 2012. ACM.
- [129] Yanhong Zhou, Gao Cong, Bin Cui, Christian S. Jensen, and Junjie Yao. Routing questions to the right users in online communities. In *ICDE*, pages 700–711, 2009.